Chapter

2

Organizing and Graphing Data

Are you one of those adults who do not drink caffeinated beverages at all? Or are you one of those who drink four or more cups or cans of caffeinated beverages a day? Or do you fall somewhere in the middle of these two extremes? According to a sample survey of adults, 22% said that they do not drink any caffeinated beverages at all, 25% said they drink four or more cups or cans a day. Of the remaining adults, 16% drink one cup or can a day, 21% drink two cups or cans a day, and 16% drink three cups or cans a day. (See Case Study 2–4.)

In addition to thousands of private organizations and individuals, a large number of U.S. government agencies (such as the Bureau of the Census, the Bureau of Labor Statistics, the National Agricultural Statistics Service, the National Center for Education Statistics, the National Center for Health Statistics, and the Bureau of Justice Statistics) conduct hundreds of surveys every year. The data collected from each of these surveys fill hundreds of thousands of pages. In their original form, these data sets may be so large that they do not make sense to most of us. Descriptive statistics, however, supplies the techniques that help condense large data sets by using tables, graphs, and summary measures. We see such tables, graphs, and summary measures in newspapers and magazines every day. At a glance, these tabular and graphical displays present information on every aspect of life. Consequently, descriptive statistics is of immense importance because it provides efficient and effective methods for summarizing and analyzing information.

This chapter explains how to organize and display data using tables and graphs. We will learn how to prepare frequency distribution tables for qualitative and quantitative data; how to construct bar graphs, pie charts, histograms, and polygons for such data; and how to prepare stem-and-leaf displays.

2.1 Raw Data

When data are collected, the information obtained from each member of a population or sample is recorded in the sequence in which it becomes available. This sequence of data recording is random and unranked. Such data, before they are grouped or ranked, are called **raw data**.

Definition

Raw Data Data recorded in the sequence in which they are collected and before they are processed or ranked are called *raw data*.

Suppose we collect information on the ages (in years) of 50 students selected from a university. The data values, in the order they are collected, are recorded in Table 2.1. For instance, the first student's age is 21, the second student's age is 19 (second number in the first row), and so forth. The data in Table 2.1 are quantitative raw data.

Table 2.1 Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

Suppose we ask the same 50 students about their student status. The responses of the students are recorded in Table 2.2. In this table, F, SO, J, and SE are the abbreviations for freshman, sophomore, junior, and senior, respectively. This is an example of qualitative (or categorical) raw data.

Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

The data presented in Tables 2.1 and 2.2 are also called **ungrouped data**. An ungrouped data set contains information on each member of a sample or population individually.

2.2 Organizing and Graphing Qualitative Data

This section discusses how to organize and display qualitative (or categorical) data. Data sets are organized into tables, and data are displayed using graphs.

2.2.1 Frequency Distributions

A sample of 100 students enrolled at a university were asked what they intended to do after graduation. Forty-four said they wanted to work for private companies/businesses, 16 said they wanted to work for the federal government, 23 wanted to work for state or local governments, and 17 intended to start their own businesses. Table 2.3 lists the types of employment and the number of students who intend to engage in each type of employment. In this table, the variable is the *type of employment*, which is a qualitative variable. The categories (representing the type of employment) listed in the first column are mutually exclusive. In other words, each of the 100 students belongs to one and only one of these categories. The number of students who belong to a certain category is called the *frequency* of that category. A **frequency distribution** exhibits how the frequencies are distributed over various categories. Table 2.3 is called a *frequency distribution table* or simply a *frequency table*.

Table 2.3 **Type of Employment Students Intend to Engage In**

	Type of Employment	Number of Students	
Variable ⟶			⟵ Frequency column
	Private companies/businesses	44	
Category ⟶	Federal government	16	⟵ Frequency
	State/local government	23	
	Own business	17	
		Sum = 100	

Definition

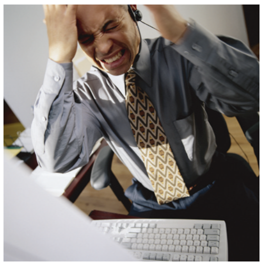
Frequency Distribution for Qualitative Data A *frequency distribution* for qualitative data lists all categories and the number of elements that belong to each of the categories.

Example 2–1 illustrates how a frequency distribution table is constructed for qualitative data.

■ EXAMPLE 2–1

Constructing a frequency distribution table for qualitative data.

A sample of 30 employees from large companies was selected, and these employees were asked how stressful their jobs were. The responses of these employees are recorded below where *very* represents very stressful, *somewhat* means somewhat stressful, and *none* stands for not stressful at all.



somewhat	none	somewhat	very	very	none
very	somewhat	somewhat	very	somewhat	somewhat
very	somewhat	none	very	none	somewhat
somewhat	very	somewhat	somewhat	very	none
somewhat	very	very	somewhat	none	somewhat

Construct a frequency distribution table for these data.

Solution Note that the variable in this example is *how stressful is an employee's job*. This variable is classified into three categories: very stressful, somewhat stressful, and not stressful at all. We record these categories in the first column of Table 2.4. Then we read each employee's response from the given data and mark a *tally*, denoted by the symbol |, in the second column of Table 2.4 next to the corresponding category. For example, the first employee's response is that his or her job is somewhat stressful. We show this in the frequency table by marking a tally in the second column next to the category *somewhat*. Note that the tallies are marked in blocks of five for counting convenience. Finally, we record the total of the tallies for each category in the third column of the table. This column is called the *column of frequencies* and is usually denoted by *f*. The sum of the entries in the frequency column gives the sample size or total frequency. In Table 2.4, this total is 30, which is the sample size.

1519T_c02 11/2/05 7:57 AM Page 29

Table 2.4 Frequency Distribution of Stress on Job

Stress on Job	Tally	Frequency (f)
Very	~~||||~~ ~~||||~~	10
Somewhat	~~||||~~ ~~||||~~ ||||	14
None	~~||||~~ |	6
		Sum = 30

2.2.2 Relative Frequency and Percentage Distributions

The **relative frequency** of a category is obtained by dividing the frequency of that category by the sum of all frequencies. Thus, the relative frequency shows what fractional part or proportion of the total frequency belongs to the corresponding category. A *relative frequency distribution* lists the relative frequencies for all categories.

Calculating Relative Frequency of a Category

$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

The **percentage** for a category is obtained by multiplying the relative frequency of that category by 100. A *percentage distribution* lists the percentages for all categories.

Calculating Percentage

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100$$

EXAMPLE 2–2

Determine the relative frequency and percentage distributions for the data of Table 2.4.

Constructing relative frequency and percentage distributions.

Solution The relative frequencies and percentages from Table 2.4 are calculated and listed in Table 2.5. Based on this table, we can state that .333 or 33.3% of the employees said that their jobs are very stressful. By adding the percentages for the first two categories, we can state that 80% of the employees said that their jobs are very or somewhat stressful. The other numbers in Table 2.5 can be interpreted the same way.

Notice that the sum of the relative frequencies is always 1.00 (or approximately 1.00 if the relative frequencies are rounded), and the sum of the percentages is always 100 (or approximately 100 if the percentages are rounded).

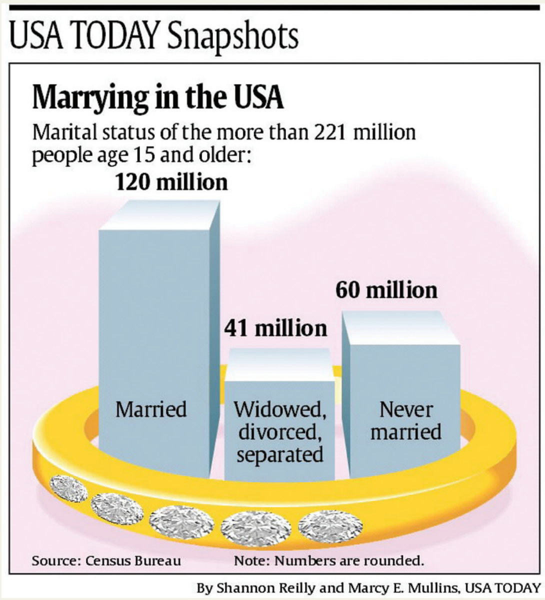
Table 2.5 Relative Frequency and Percentage Distributions of Stress on Job

Stress on Job	Relative Frequency	Percentage
Very	10/30 = .333	.333(100) = 33.3
Somewhat	14/30 = .467	.467(100) = 46.7
None	6/30 = .200	.200(100) = 20.0
	Sum = 1.000	Sum = 100

CASE STUDY

2–1

MARRYING IN THE USA



Source: *USA TODAY*, November 10, 2003. Copyright © 2003, *USA TODAY.* Chart reproduced with permission.

The above chart, reproduced from *USA TODAY,* shows a bar graph indicating the marital status of Americans aged 15 and older. According to the Census Bureau, there are around 221 million people in the United States who are 15 years of age or older. Of them, about 120 million are currently married, about 41 million are widowed/divorced/separated, and about 60 million have never married. Using these categories and numbers, we can write a frequency table, and then calculate the relative frequencies and percentages.

2.2.3 Graphical Presentation of Qualitative Data

All of us have heard the adage "a picture is worth a thousand words." A graphic display can reveal at a glance the main characteristics of a data set. The *bar graph* and the *pie chart* are two types of graphs used to display qualitative data.

Bar Graphs

To construct a **bar graph** (also called a *bar chart*), we mark the various categories on the horizontal axis as in Figure 2.1. Note that all categories are represented by intervals of the same width. We mark the frequencies on the vertical axis. Then we draw one bar for each category such that the height of the bar represents the frequency of the corresponding category. We leave a small gap between adjacent bars. Figure 2.1 gives the bar graph for the frequency distribution of Table 2.4.

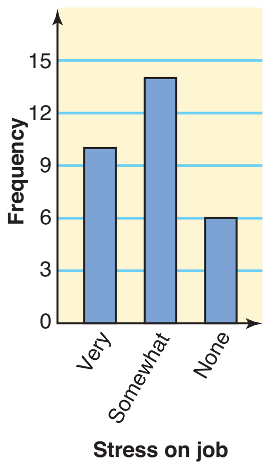


Figure 2.1 Bar graph for the frequency distribution of Table 2.4.

Definition

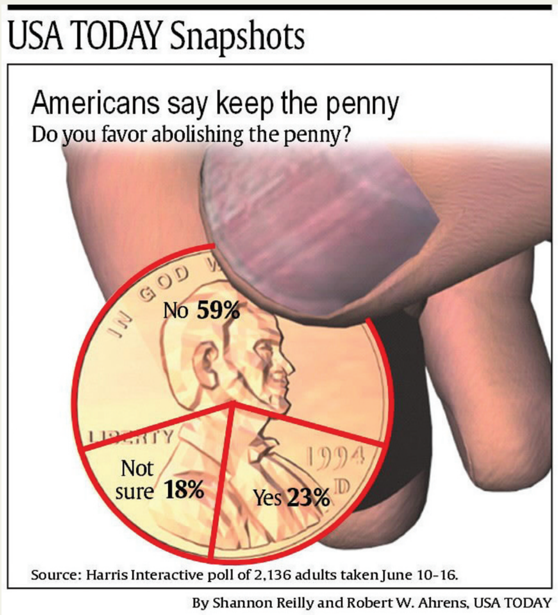
Bar Graph A graph made of bars whose heights represent the frequencies of respective categories is called a *bar graph.*

The bar graphs for relative frequency and percentage distributions can be drawn simply by marking the relative frequencies or percentages, instead of the class frequencies, on the vertical axis.

Sometimes a bar graph is constructed by marking the categories on the vertical axis and the frequencies on the horizontal axis.

2–2 CASE STUDY

AMERICANS SAY KEEP THE PENNY



The above pie chart shows the opinions of people on the issue of keeping or abolishing the penny. In a Harris Interactive poll of 2136 adults conducted in June 10–16, 2004, 59% of the respondents said that the penny should not be abolished, 23% said it should be abolished, and 18% were not sure. Using these categories and percentages, we can write a percentage distribution table.

Source: *USA TODAY*, August 5, 2004.

Pie Charts

A **pie chart** is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The whole pie (or circle) represents the total sample or population. Then we divide the pie into different portions that represent the different categories.

Definition

Pie Chart A circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories is called a *pie chart*.

As we know, a circle contains 360 degrees. To construct a pie chart, we multiply 360 by the relative frequency of each category to obtain the degree measure or size of the angle for the corresponding category. Table 2.6 shows the calculation of angle sizes for the various categories of Table 2.5.

Table 2.6 Calculating Angle Sizes for the Pie Chart

Stress on Job	Relative Frequency	Angle Size
Very	.333	360(.333) = 119.88
Somewhat	.467	360(.467) = 168.12
None	.200	360(.200) = 72.00
	Sum = 1.000	Sum = 360

Figure 2.2 shows the pie chart for the percentage distribution of Table 2.5, which uses the angle sizes calculated in Table 2.6.

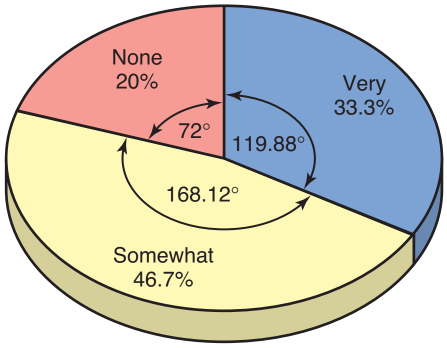


Figure 2.2 Pie chart for the percentage distribution of Table 2.5.

EXERCISES

CONCEPTS AND PROCEDURES

2.1 Why do we need to group data in the form of a frequency table? Explain briefly.

2.2 How are the relative frequencies and percentages of categories obtained from the frequencies of categories? Illustrate with the help of an example.

2.3 The following data give the results of a sample survey. The letters A, B, and C represent the three categories.

A	B	B	A	C	B	C	C	C	A
C	B	C	A	C	C	B	C	C	A
A	B	C	C	B	C	B	A	C	A

a. Prepare a frequency distribution table.
b. Calculate the relative frequencies and percentages for all categories.
c. What percentage of the elements in this sample belong to category B?
d. What percentage of the elements in this sample belong to category A or C?
e. Draw a bar graph for the frequency distribution.

2.4 The following data give the results of a sample survey. The letters Y, N, and D represent the three categories.

D	N	N	Y	Y	Y	N	Y	D	Y
Y	Y	Y	Y	N	Y	Y	N	N	Y
N	Y	Y	N	D	N	Y	Y	Y	Y
Y	Y	N	N	Y	Y	N	N	D	Y

a. Prepare a frequency distribution table.
b. Calculate the relative frequencies and percentages for all categories.
c. What percentage of the elements in this sample belong to category Y?
d. What percentage of the elements in this sample belong to category N or D?
e. Draw a pie chart for the percentage distribution.

APPLICATIONS

2.5 The data on the status of 50 students given in Table 2.2 of Section 2.1 are reproduced here.

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

a. Prepare a frequency distribution table.
b. Calculate the relative frequencies and percentages for all categories.
c. What percentage of these students are juniors or seniors?
d. Draw a bar graph for the frequency distribution.

2.6 Thirty adults were asked which of the following conveniences they would find most difficult to do without: television (T), refrigerator (R), air conditioning (A), public transportation (P), or microwave (M). Their responses are listed below.

R	A	R	P	P	T	R	M	P	A
A	R	R	T	P	P	T	R	A	A
R	P	A	T	R	P	R	A	P	R

a. Prepare a frequency distribution table.
b. Calculate the relative frequencies and percentages for all categories.
c. What percentage of these adults named refrigerator or air conditioning as the convenience that they would find most difficult to do without?
d. Draw a bar graph for the relative frequency distribution.

2.7 In a 2004 *USA TODAY* survey (*USA TODAY*, July 19, 2004), registered dietitians with the American Dietetic Association were asked, "What is the major reason people want to lose weight?" The responses were classified as *Health* (H), *Cosmetic* (C), and *Other* (O). Suppose a random sample of 20 dietitians is taken and these dietitians are asked the same question. Their responses are as follows.

H	H	C	H	O	C	C	H	C	O
O	H	C	H	H	C	H	H	O	H

a. Prepare a frequency distribution table.
b. Compute the relative frequencies and percentages for all categories.
c. What percentage of these dietitians gave *Health* as the major reason for people to lose weight?
d. Draw a pie chart for the percentage distribution.

2.8 The following data show the method of payment by 16 customers in a supermarket checkout line. Here, C refers to cash, CK to check, CC to credit card, D to debit card, and O stands for other.

C	CK	CK	C	CC	D	O	C
CK	CC	D	CC	C	CK	CK	CC

a. Construct a frequency distribution table.
b. Calculate the relative frequencies and percentages for all categories.
c. Draw a pie chart for the percentage distribution.

2.9 In the *MARS 2004 OTC/DTC* survey, U.S. adults were asked to rate their health. The table below summarizes their responses.

State of Health	Percentage of Responses
Excellent	17.0
Very good	36.2
Good	32.5
Fair	12.0
Poor	2.3

Source: USA TODAY, June 2, 2004.

Draw a pie chart for this percentage distribution.

2.10 In an exit poll taken during the 2004 presidential election, voters were asked to name the issue that most affected their vote for a candidate for presidency. The following table summarizes their responses.

Issue	Percentage of Responses
Moral values	22
Economy/jobs	20
Terrorism	19
Iraq	15
Health care	8
Taxes	5
Education	4

Source: United States General Exit Poll of 13,660 voters. *USA TODAY*, November 5, 2004.

As you will notice, these percentages add up to 93%. Assume that the remaining 7% of these voters named other issues and let us denote these issues as *Other.* Draw a bar graph to display these data.

2.3 Organizing and Graphing Quantitative Data

In the previous section we learned how to group and display qualitative data. This section explains how to group and display quantitative data.

2.3.1 Frequency Distributions

Table 2.7 gives the weekly earnings of 100 employees of a large company. The first column lists the *classes*, which represent the (quantitative) variable *weekly earnings*. For quantitative data, an interval that includes all the values that fall within two numbers, the lower and upper limits, is called a **class**. Note that the classes always represent a variable. As we can observe, the classes are nonoverlapping; that is, each value on earnings belongs to one and only one class. The second column in the table lists the number of employees who have earnings within each class. For example, nine employees of this company earn \$401 to \$600 per week. The numbers listed in the second column are called the **frequencies**, which give the number of values that belong to different classes. The frequencies are denoted by *f*.

Table 2.7 Weekly Earnings of 100 Employees of a Company

	Weekly Earnings (dollars)	Number of Employees f	
Variable →	← Variable	Frequency column ←	
	401 to 600	9	
	601 to 800	22	
Third class →	801 to 1000	39	← Frequency of the third class
	1001 to 1200	15	
	1201 to 1400	9	
Lower limit of the sixth class →	1401 to 1600	6	← Upper limit of the sixth class

For quantitative data, the frequency of a class represents the number of values in the data set that fall in that class. Table 2.7 contains six classes. Each class has a *lower limit* and an *upper limit*. The values 401, 601, 801, 1001, 1201, and 1401 give the lower limits, and the values 600, 800, 1000, 1200, 1400, and 1600 are the upper limits of the six classes, respectively. The data presented in Table 2.7 are an illustration of a **frequency distribution table** for quantitative data. Whereas the data that list individual values are called ungrouped data, the data presented in a frequency distribution table are called **grouped data**.

Definition

Frequency Distribution for Quantitative Data A *frequency distribution* for quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called *grouped data*.

To find the midpoint of the upper limit of the first class and the lower limit of the second class in Table 2.7, we divide the sum of these two limits by 2. Thus, this midpoint is

$$\frac{600 + 601}{2} = 600.5$$

The value 600.5 is called the *upper boundary* of the first class and the *lower boundary* of the second class. By using this technique, we can convert the class limits of Table 2.7 to **class boundaries**, which are also called *real class limits*. The second column of Table 2.8 lists the boundaries for Table 2.7.

1519T_c02 11/2/05 7:57 AM Page 35

EQA

Definition

Class Boundary The *class boundary* is given by the midpoint of the upper limit of one class and the lower limit of the next class.

The difference between the two boundaries of a class gives the **class width**. The class width is also called the **class size**.

Finding Class Width

$$\text{Class width} = \text{Upper boundary} - \text{Lower boundary}$$

Thus, in Table 2.8,

$$\text{Width of the first class} = 600.5 - 400.5 = 200$$

The class widths for the frequency distribution of Table 2.7 are listed in the third column of Table 2.8. Each class in Table 2.8 (and Table 2.7) has the same width of 200.

The **class midpoint** or **mark** is obtained by dividing the sum of the two limits (or the two boundaries) of a class by 2.

Calculating Class Midpoint or Mark

$$\text{Class midpoint or mark} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

Thus, the midpoint of the first class in Table 2.7 or Table 2.8 is calculated as follows:

$$\text{Midpoint of the first class} = \frac{401 + 600}{2} = 500.5$$

The class midpoints for the frequency distribution of Table 2.7 are listed in the fourth column of Table 2.8.

Table 2.8 Class Boundaries, Class Widths, and Class Midpoints for Table 2.7

Class Limits	Class Boundaries	Class Width	Class Midpoint
401 to 600	400.5 to less than 600.5	200	500.5
601 to 800	600.5 to less than 800.5	200	700.5
801 to 1000	800.5 to less than 1000.5	200	900.5
1001 to 1200	1000.5 to less than 1200.5	200	1100.5
1201 to 1400	1200.5 to less than 1400.5	200	1300.5
1401 to 1600	1400.5 to less than 1600.5	200	1500.5

Note that in Table 2.8, when we write classes using class boundaries, we write *to less than* to ensure that each value belongs to one and only one class. As we can see, the upper boundary of the preceding class and the lower boundary of the succeeding class are the same.

2.3.2 Constructing Frequency Distribution Tables

When constructing a frequency distribution table, we need to make the following three major decisions.

Number of Classes

Usually the number of classes for a frequency distribution table varies from 5 to 20, depending mainly on the number of observations in the data set.[1] It is preferable to have more classes as the size of a data set increases. The decision about the number of classes is arbitrarily made by the data organizer.

Class Width

Although it is not uncommon to have classes of different sizes, most of the time it is preferable to have the same width for all classes. To determine the class width when all classes are the same size, first find the difference between the largest and the smallest values in the data. Then, the approximate width of a class is obtained by dividing this difference by the number of desired classes.

Calculation of Class Width

$$\text{Approximate class width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}$$

Usually this approximate class width is rounded to a convenient number, which is then used as the class width. Note that rounding this number may slightly change the number of classes initially intended.

Lower Limit of the First Class or the Starting Point

Any convenient number that is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

Example 2–3 illustrates the procedure for constructing a frequency distribution table for quantitative data.

■ EXAMPLE 2–3

Constructing a frequency distribution table for quantitative data.

Table 2.9 (on next page) gives the total home runs hit by all players of each of the 30 Major League Baseball teams during the 2004 season. Construct a frequency distribution table.

Solution In these data, the minimum value is 135 and the maximum value is 242. Suppose we decide to group these data using five classes of equal width. Then,

$$\text{Approximate width of each class} = \frac{242 - 135}{5} = 21.4$$

Now we round this approximate width to a convenient number—say, 22. The lower limit of the first class can be taken as 135 or any number less than 135. Suppose we take 135 as the lower limit of the first class. Then our classes will be

135–156, 157–178, 179–200, 201–222, and 223–244

We record these five classes in the first column of Table 2.10 on page 37.

[1]One rule to help decide on the number of classes is Sturge's formula:

$$c = 1 + 3.3 \log n$$

where c is the number of classes and n is the number of observations in the data set. The value of log n can be obtained by entering the value of n on the calculator and pressing the *log* key.

Table 2.9 Home Runs Hit by Major League Baseball Teams During the 2004 Season

Team	Home Runs	Team	Home Runs
Arizona	135	Milwaukee	135
Atlanta	178	Minnesota	191
Baltimore	169	Montreal (now Washington)	151
Boston	222	New York Mets	185
Chicago Cubs	235	New York Yankees	242
Chicago White Sox	242	Oakland	189
Cincinnati	194	Philadelphia	215
Cleveland	184	Pittsburgh	142
Colorado	202	St. Louis	214
Detroit	201	San Diego	139
Florida	148	San Francisco	183
Houston	187	Seattle	136
Kansas City	150	Tampa Bay	145
Anaheim Angels[1]	162	Texas	227
Los Angeles Dodgers	203	Toronto	145

[1]In 2005, the Anaheim Angels changed their name to the Los Angeles Angels of Anaheim.

Now we read each value from the given data and mark a tally in the second column of Table 2.10 next to the corresponding class. The first value in our original data is 135, which belongs to the 135–156 class. To record it, we mark a tally in the second column next to the 135–156 class. We continue this process until all the data values have been read and entered in the tally column. Note that tallies are marked in blocks of fives for counting convenience. After the tally column is completed, we count the tally marks for each class and write those numbers in the third column. This gives the column of frequencies. These frequencies represent the number of teams that belong to each of the five different classes representing the total home runs. For example, 10 of the 30 Major League Baseball teams hit a total of 135–156 home runs during the 2004 season.

Table 2.10 Frequency Distribution for the Data of Table 2.9

Total Home Runs	Tally	f
135–156	~~||||~~ ~~||||~~	10
157–178	|||	3
179–200	~~||||~~ ||	7
201–222	~~||||~~ |	6
223–244	||||	4
		$\Sigma f = 30$

In Table 2.10, we can denote the frequencies of the five classes by f_1, f_2, f_3, f_4, and f_5, respectively. Therefore,

$$f_1 = \text{Frequency of the first class} = 10$$

Similarly,

$$f_2 = 3, \quad f_3 = 7, \quad f_4 = 6, \quad \text{and} \quad f_5 = 4$$

Using the Σ notation (see Section 1.8 of Chapter 1), we can denote the sum of the frequencies of all classes by Σf. Hence,

$$\Sigma f = f_1 + f_2 + f_3 + f_4 + f_5 = 10 + 3 + 7 + 6 + 4 = 30$$

The number of observations in a sample is usually denoted by n. Thus, for the sample data, Σf is equal to n. The number of observations in a population is denoted by N. Consequently, Σf is equal to N for population data. Because the data set on the total home runs by Major League Baseball teams in Table 2.10 is for all 30 teams, it represents the population. Therefore, in Table 2.10 we can denote the sum of frequencies by N instead of Σf. ■

Note that when we present the data in the form of a frequency distribution table, as in Table 2.10, we lose the information on individual observations. We cannot know the exact number of home runs hit by any particular Major League Baseball team from Table 2.10. All we know is that the home runs hit by 10 of these teams during the 2004 season are between 135–156, and so forth.

2.3.3 Relative Frequency and Percentage Distributions

Using Table 2.10, we can compute the relative frequency and percentage distributions the same way we did for qualitative data in Section 2.2.2. The relative frequencies and percentages for a quantitative data set are obtained as follows.

Calculating Relative Frequency and Percentage

$$\text{Relative frequency of a class} = \frac{\text{Frequency of that class}}{\text{Sum of all frequencies}} = \frac{f}{\Sigma f}$$

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100$$

Example 2–4 illustrates how to construct relative frequency and percentage distributions.

■ EXAMPLE 2–4

Constructing relative frequency and percentage distributions.

Calculate the relative frequencies and percentages for Table 2.10.

Solution The relative frequencies and percentages for the data in Table 2.10 are calculated and listed in the third and fourth columns, respectively, of Table 2.11 here. Note that the class boundaries are listed in the second column of Table 2.11.

Table 2.11 Relative Frequency and Percentage Distributions for Table 2.10

Total Home Runs	Class Boundaries	Relative Frequency	Percentage
135–156	134.5 to less than 156.5	.333	33.3
157–178	156.5 to less than 178.5	.100	10.0
179–200	178.5 to less than 200.5	.233	23.3
201–222	200.5 to less than 222.5	.200	20.0
223–244	222.5 to less than 244.5	.133	13.3
		Sum = .999	Sum = 99.9%

Using Table 2.11, we can make statements about the percentage of teams with home runs within a certain interval. For example, 33.3% of the Major League Baseball teams in this population hit total home runs between 135–156 during the 2004 season. By adding the percentages for the first two classes, we can state that about 43.3% of these teams hit home runs between 135–178 during the 2004 season. Similarly, by adding the percentages of the last two classes, we can state that about 33.3% of these teams hit home runs between 201–244 during the 2004 season. ■

2.3.4 Graphing Grouped Data

Grouped (quantitative) data can be displayed in a *histogram* or a *polygon*. This section describes how to construct such graphs. We can also draw a pie chart to display the percentage distribution for a quantitative data set. The procedure to construct a pie chart is similar to the one for qualitative data explained in Section 2.2.3; it will not be repeated in this section.

Histograms

A **histogram** can be drawn for a frequency distribution, a relative frequency distribution, or a percentage distribution. To draw a histogram, we first mark classes on the horizontal axis and frequencies (or relative frequencies or percentages) on the vertical axis. Next, we draw a bar for each class so that its height represents the frequency of that class. The bars in a histogram are drawn adjacent to each other with no gap between them. A histogram is called a **frequency histogram**, a **relative frequency histogram**, or a **percentage histogram** depending on whether frequencies, relative frequencies, or percentages are marked on the vertical axis.

Definition

Histogram A *histogram* is a graph in which classes are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on the vertical axis. The frequencies, relative frequencies, or percentages are represented by the heights of the bars. In a histogram, the bars are drawn adjacent to each other.

Figures 2.3 and 2.4 show the frequency and the relative frequency histograms, respectively, for the data of Tables 2.10 and 2.11 of Sections 2.3.2 and 2.3.3. The two histograms look alike because they represent the same data. A percentage histogram can be drawn for the percentage distribution of Table 2.11 by marking the percentages on the vertical axis.

The symbol –//– used in the horizontal axes of Figures 2.3 and 2.4 represents a break, called the **truncation**, in the horizontal axis. It indicates that the entire horizontal axis is not shown in these figures. Notice that the 0 to 134 portion of the horizontal axis has been omitted in each figure.

In Figures 2.3 and 2.4, we have used class limits to mark classes on the horizontal axis. However, we can show the intervals on the horizontal axis by using the class boundaries instead of the class limits.

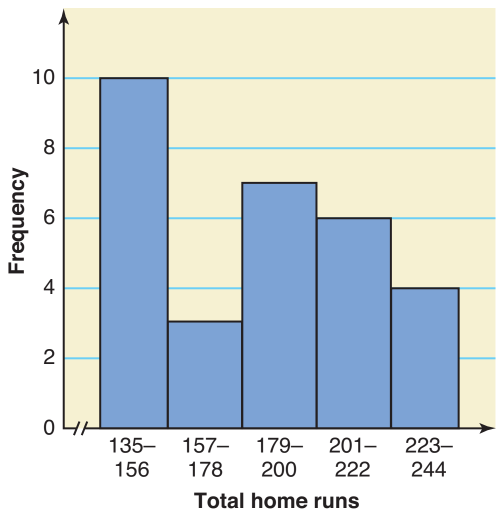
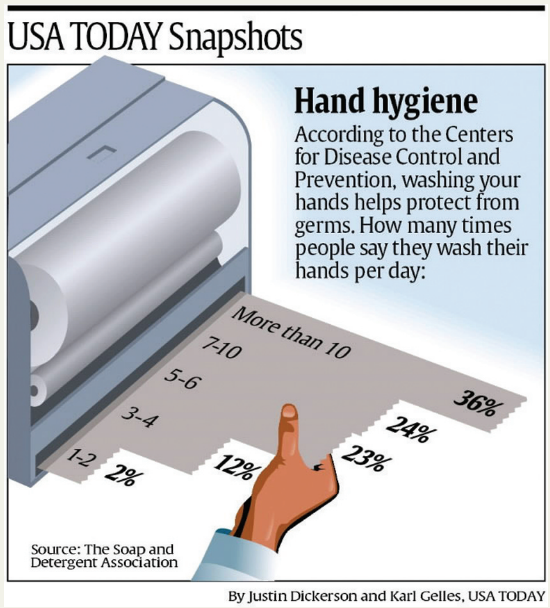


Figure 2.3 Frequency histogram for Table 2.10.

CASE STUDY

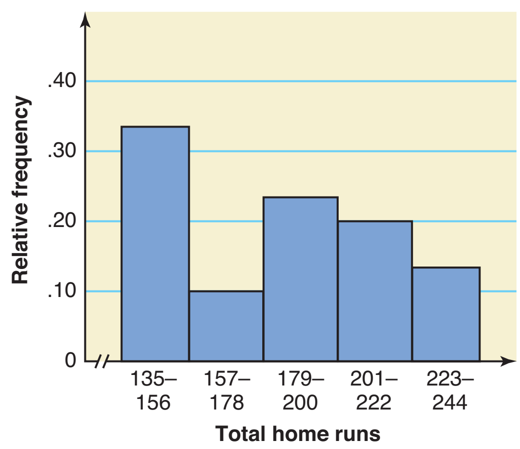
2–3 HAND HYGIENE



Source: *USA TODAY*, December 14, 2004. Copyright © 2004, *USA TODAY*. Chart reproduced with permission.

The above chart, reproduced from *USA TODAY*, gives the histogram for the percentage distribution of the number of times people say they wash their hands per day. As we can observe from the chart, 2% of people included in the survey said they wash their hands 1–2 times a day, 12% wash their hands 3–4 times a day, and so on. There are a couple of things you should note about this graph. First, the classes in the chart have different widths. For example, the first three classes (1–2, 3–4, and 5–6) have the same width, which is 2. However, the fourth class (7–10) has a width of 4. The fifth class (more than 10) is called an **open-ended class**. We know it has a lower limit of 11 but it has no upper limit. Second, the percentages for all classes in this chart add up to 97%. The chart does not say anything about the remaining 3% of the people. We can assume that these 3% of the people in the survey either did not wash their hands at all or they did not give an answer.

Figure 2.4 Relative frequency histogram for Table 2.11.



Polygons

A **polygon** is another device that can be used to present quantitative data in graphic form. To draw a **frequency polygon**, we first mark a dot above the midpoint of each class at a height equal to the frequency of that class. This is the same as marking the midpoint at the top of each bar in a histogram. Next we mark two more classes, one at each end, and mark their midpoints. Note that these two classes have zero frequencies. In the last step, we join the adjacent dots with straight lines. The resulting line graph is called a frequency polygon or simply a polygon.

A polygon with relative frequencies marked on the vertical axis is called a *relative frequency polygon*. Similarly, a polygon with percentages marked on the vertical axis is called a *percentage polygon*.

Definition

Polygon A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines is called a *polygon*.

Figure 2.5 shows the frequency polygon for the frequency distribution of Table 2.10.

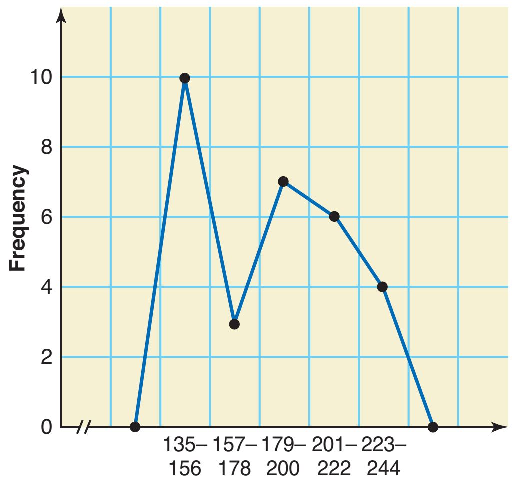


Figure 2.5 Frequency polygon for Table 2.10.

For a very large data set, as the number of classes is increased (and the width of classes is decreased), the frequency polygon eventually becomes a smooth curve. Such a curve is called a *frequency distribution curve* or simply a *frequency curve*. Figure 2.6 shows the frequency curve for a large data set with a large number of classes.

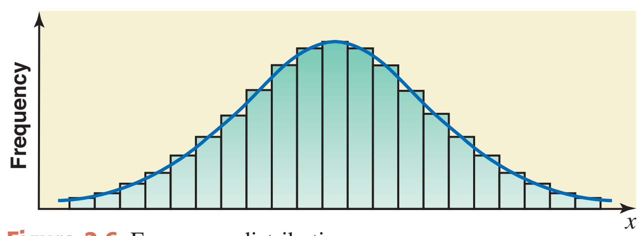


Figure 2.6 Frequency distribution curve.

2.3.5 More on Classes and Frequency Distributions

This section presents two alternative methods for writing classes to construct a frequency distribution for quantitative data.

Less Than Method for Writing Classes

The classes in the frequency distribution given in Table 2.10 for the data on home runs were written as 135–156, 157–178, and so on. Alternatively, we can write the classes in a frequency distribution table using the *less than* method. The technique for writing classes shown in Table 2.10 is more commonly used for data sets that do not contain fractional values. The *less than* method is more appropriate when a data set contains fractional values. Example 2–5 illustrates the *less than* method.

■ EXAMPLE 2–5

Constructing a frequency distribution using the less than method.

According to the American Petroleum Institute, the state taxes (in cents) per gallon of gasoline as of April 1, 2005 for all 50 states[2] are as follows.

18	8	18	21.5	18	22	25	23	14.5	7.5
16	25	19	18	20	24	16	20	25.2	23.5
23.5	19	20	18	17	27.75	25.4	23	18	14.5
17	31.9	26.6	21	26	16	24	31.1	30	16
22	20	20	24.5	20	17.5	28	20.5	32.9	14

Construct a frequency distribution table. Calculate the relative frequencies and percentages for all classes.

Solution The minimum value in this data set is 7.5 and the maximum value is 32.9. Suppose we decide to group these data using six classes of equal width. Then

$$\text{Approximate width of a class} = \frac{32.9 - 7.5}{6} = 4.23$$

We round this number to a more convenient number—say, 5. Then we take 5 as the width of each class. We can take the lower limit of the first class equal to 7.5 or any number lower than 7.5. If we start the first class at 5, the classes will be written as *5 to less than 10, 10 to less than 15,* and so on. The six classes, which cover all the data values, are recorded in the first column of Table 2.12. The second column lists the frequencies of these classes. A value in the data set that is 5 or larger but less than 10 belongs to the first class, a value that is 10 or larger

Table 2.12 Frequency, Relative Frequency, and Percentage Distributions of State Taxes on Gasoline

State Taxes Per Gallon of Gasoline (cents)	f	Relative Frequency	Percentage
5 to less than 10	2	.04	4
10 to less than 15	3	.06	6
15 to less than 20	15	.30	30
20 to less than 25	18	.36	36
25 to less than 30	8	.16	16
30 to less than 35	4	.08	8
	$\Sigma f = 50$	Sum = 1.00	Sum = 100%

[2]The data for the 50 states are entered (by row) in the following order: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming.

1519T_c02 11/2/05 7:57 AM Page 43

but less than 15 falls in the second class, and so on. The relative frequencies and percentages for classes are recorded in the third and fourth columns, respectively, of Table 2.12. Note that this table does not contain a column of tallies. ■

A histogram and a polygon for the data of Table 2.12 can be drawn in the same way as for the data of Tables 2.10 and 2.11.

Single-Valued Classes

If the observations in a data set assume only a few distinct (integer) values, it may be appropriate to prepare a frequency distribution table using *single-valued classes*—that is, classes that are made of single values and not of intervals. This technique is especially useful in cases of discrete data with only a few possible values. Example 2–6 exhibits such a situation.

■ EXAMPLE 2–6

Constructing a frequency distribution using single-valued classes.

The administration in a large city wanted to know the distribution of vehicles owned by households in that city. A sample of 40 randomly selected households from this city produced the following data on the number of vehicles owned.

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

Construct a frequency distribution table for these data using single-valued classes.

Solution The observations in this data set assume only six distinct values: 0, 1, 2, 3, 4, and 5. Each of these six values is used as a class in the frequency distribution in Table 2.13, and these six classes are listed in the first column of that table. To obtain the frequencies of these classes, the observations in the data that belong to each class are counted, and the results are recorded in the second column of Table 2.13. Thus, in these data, 2 households own no vehicle, 18 own one vehicle each, 11 own two vehicles each, and so on.

Table 2.13 Frequency Distribution of Vehicles Owned

Vehicles Owned	Number of Households (f)
0	2
1	18
2	11
3	4
4	3
5	2
	$\Sigma f = 40$

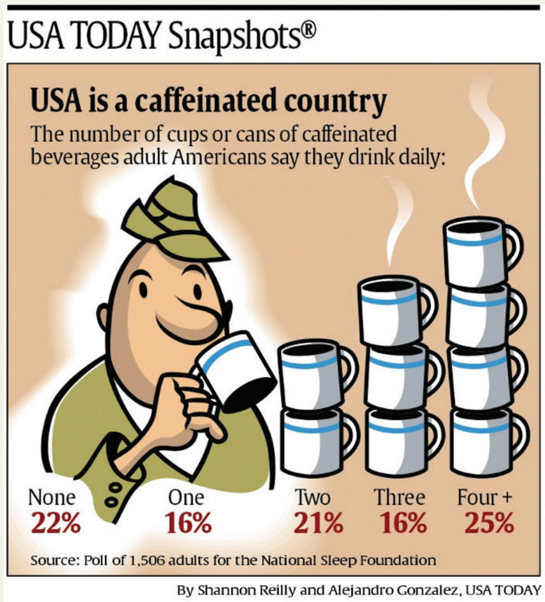
■

The data of Table 2.13 can also be displayed in a bar graph, as shown in Figure 2.7 on page 44. To construct a bar graph, we mark the classes, as intervals, on the horizontal axis with a little gap between consecutive intervals. The bars represent the frequencies of respective classes.

CASE STUDY

2–4

USA IS A CAFFEINATED COUNTRY



The above chart, reproduced from *USA TODAY*, shows the percentage distribution of the number of cups or cans of caffeinated beverages that adults in the United States drink per day. This distribution is based on a sample survey of 1506 adults. Here the classes are single-valued except the last one (Four Plus), which is an open-ended multiple-valued class.

Note that the chart does not show the typical bar graph that we learned to make in this chapter. In this chart, classes are not marked on the horizontal axis, percentages are not marked on the vertical axis, and the heights of bars do not show the percentages. Instead, the percentages are shown horizontally and bars show the number of cups or cans that belong to the corresponding percentage. Since 22% of the adults said that they do not drink any caffeinated beverage, there is no cup shown above *None.* For the category *One* cup with 16%, one cup is shown in the chart—the one being held by the person, and so on.

Source: USA TODAY, April 13, 2005. Copyright © 2005, *USA TODAY.* Chart reproduced with permission.

The frequencies of Table 2.13 can be converted to relative frequencies and percentages the same way as in Table 2.11. Then, a bar graph can be constructed to display the relative frequency or percentage distribution by marking the relative frequencies or percentages, respectively, on the vertical axis.

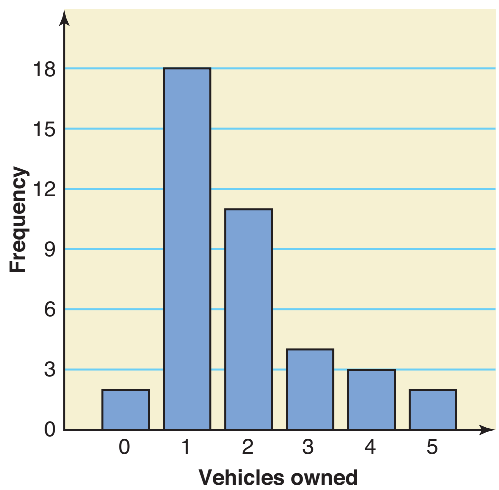


Figure 2.7 Bar graph for Table 2.13.

2.4 Shapes of Histograms

A histogram can assume any one of a large number of shapes. The most common of these shapes are

1. Symmetric
2. Skewed
3. Uniform or rectangular

A **symmetric histogram** is identical on both sides of its central point. The histograms shown in Figure 2.8 are symmetric around the dashed lines that represent their central points.

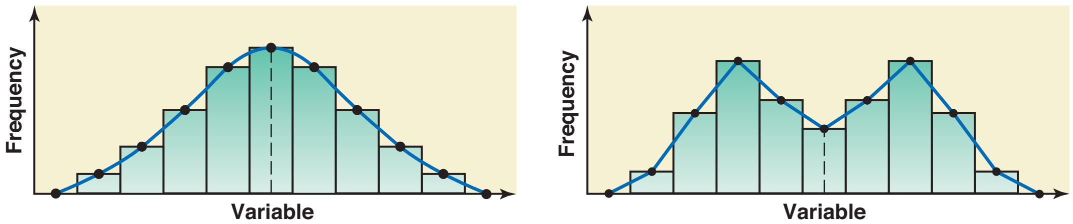


Figure 2.8 Symmetric histograms.

A **skewed histogram** is nonsymmetric. For a skewed histogram, the tail on one side is longer than the tail on the other side. A **skewed-to-the-right histogram** has a longer tail on the right side (see Figure 2.9*a*). A **skewed-to-the-left histogram** has a longer tail on the left side (see Figure 2.9*b*).

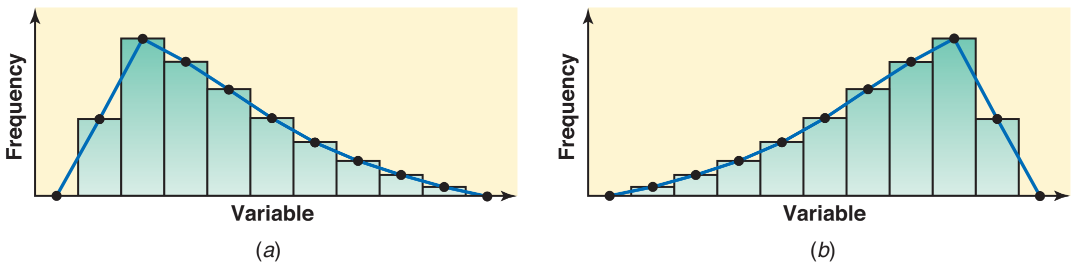


Figure 2.9 (*a*) A histogram skewed to the right. (*b*) A histogram skewed to the left.

A **uniform** or **rectangular histogram** has the same frequency for each class. Figure 2.10 is an illustration of such a case.

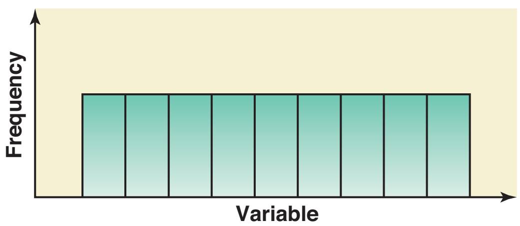


Figure 2.10 A histogram with uniform distribution.

Figures 2.11*a* and 2.11*b* display symmetric frequency curves. Figures 2.11*c* and 2.11*d* show frequency curves skewed to the right and to the left, respectively.

CASE STUDY 2–5

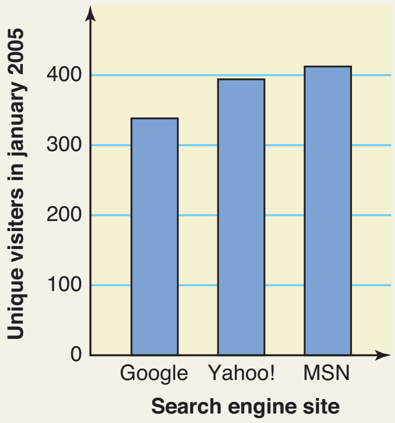
USING TRUNCATED AXES

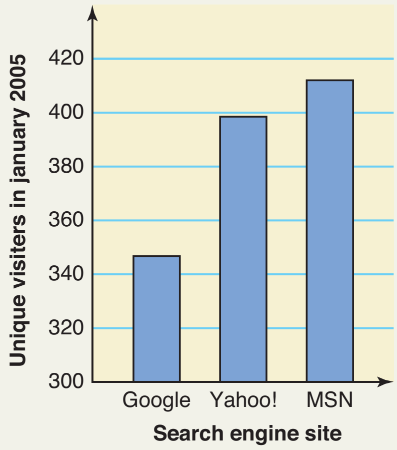
The following table gives the number of unique visitors (in millions) during January 2005 to the Web sites of three search engines—Google, Yahoo!, and MSN.

Search Engine Site	Unique Visitors During January 2005 (millions)
Google	344
Yahoo!	399
MSN	412

Source: Time, March 21, 2005.

The following two bar graphs are constructed for the data given in the above table. As you would observe, the figure on the left shows the complete vertical axis and the one on the right shows the truncated vertical axis. As is obvious, the two graphs give completely different impressions of the number of visitors to these three Web sites during the month of January 2005. The figure on the right exaggerates the differences in these numbers for the three sites. If you do not pay attention to the numbers on the vertical axis and to the truncation on the vertical axis, you may think that the number of visitors to these three sites in January 2005 varied a lot.





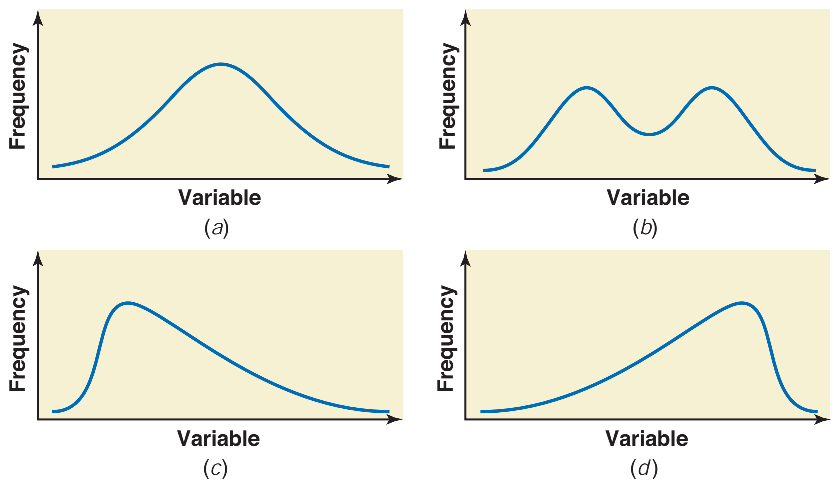


Figure 2.11 (*a*) and (*b*) Symmetric frequency curves. (*c*) Frequency curve skewed to the right. (*d*) Frequency curve skewed to the left.

◀ *Warning*

Describing data using graphs helps give us insights into the main characteristics of the data. But graphs, unfortunately, can also be used, intentionally or unintentionally, to distort the facts and deceive the reader. The following are two ways to manipulate graphs to convey a particular opinion or impression.

1. *Changing the scale* either on one or on both axes—that is, shortening or stretching one or both of the axes.
2. *Truncating the frequency axis*—that is, starting the frequency axis at a number greater than zero.

When interpreting a graph, we should be very cautious. We should observe carefully whether the frequency axis has been truncated or whether any axis has been unnecessarily shortened or stretched. Case Study 2–5 presents an example of this.

EXERCISES

CONCEPTS AND PROCEDURES

2.11 Briefly explain the three decisions that have to be made to group a data set in the form of a frequency distribution table.

2.12 How are the relative frequencies and percentages of classes obtained from the frequencies of classes? Illustrate with the help of an example.

2.13 Three methods—writing classes using limits, using the *less than* method, and grouping data using single-valued classes—were discussed to group quantitative data into classes. Explain these three methods and give one example of each.

APPLICATIONS

2.14 A sample of 80 adults was taken and these adults were asked about the number of credit cards they possess. The following table gives the frequency distribution of their responses.

Number of Credit Cards	Number of Adults
0 to 3	18
4 to 7	26
8 to 11	22
12 to 15	11
16 to 19	3

a. Find the class boundaries and class midpoints.
b. Do all classes have the same width? If so, what is this width?
c. Prepare the relative frequency and percentage distribution columns.
d. What percentage of these adults possess 8 or more credit cards?

2.15 The following table gives the frequency distribution of ages for all 50 employees of a company.

Age	Number of Employees
18 to 30	12
31 to 43	19
44 to 56	14
57 to 69	5

a. Find the class boundaries and class midpoints.
b. Do all classes have the same width? If yes, what is that width?
c. Prepare the relative frequency and percentage distribution columns.
d. What percentage of the employees of this company are age 43 or younger?

2.16 A data set on money spent on lottery tickets during the past year by 200 households has a lowest value of $1 and a highest value of $1167. Suppose we want to group these data into six classes of equal widths.

a. Assuming we take the lower limit of the first class as $1 and the width of each class equal to $200, write the class limits for all six classes.

b. What are the class boundaries and class midpoints?

2.17 A data set on monthly expenditures (rounded to the nearest dollar) incurred on fast food by a sample of 500 households has a minimum value of $3 and a maximum value of $147. Suppose we want to group these data into six classes of equal widths.

a. Assuming we take the lower limit of the first class as $1 and the upper limit of the sixth class as $150, write the class limits for all six classes.

b. Determine the class boundaries and class widths.

c. Find the class midpoints.

2.18 The following table lists the average attendance (rounded to the nearest hundred) per home game for the 2004 season for all 30 Major League Baseball teams. Note that the Montreal Expos are now the Washington Nationals.

Team	Average Attendance	Team	Average Attendance	Team	Average Attendance
A's	27,200	Tigers	24,000	Dodgers	43,100
Angels	41,700	Twins	23,600	Expos	9400
Blue Jays	23,500	White Sox	24,400	Giants	40,200
Devil Rays	16,100	Yankees	47,800	Marlins	22,100
Indians	22,400	Astros	38,100	Mets	29,000
Mariners	36,300	Braves	29,400	Padres	37,200
Orioles	34,300	Brewers	25,500	Phillies	40,600
Rangers	31,800	Cardinals	37,600	Pirates	21,100
Red Sox	35,000	Cubs	39,100	Reds	28,200
Royals	21,000	Diamondbacks	31,100	Rockies	29,600

Source: ESPN Sports Almanac 2005.

a. Construct a frequency distribution table. Take the classes as 9000–16,900, 17,000–24,900, 25,000–32,900, 33,000–40,900, and 41,000–48,900.

b. Calculate the relative frequencies and percentages for all classes.

c. Based on the frequency distribution, can you say whether the data are symmetric or skewed?

d. What percentage of these teams had average attendance of 25,000 or more?

2.19 Nixon Corporation manufactures computer monitors. The following data are the numbers of computer monitors produced at the company for a sample of 30 days.

24	32	27	23	33	33	29	25	23	28
21	26	31	22	27	33	27	23	28	29
31	35	34	22	26	28	23	35	31	27

a. Construct a frequency distribution table using the classes 21–23, 24–26, 27–29, 30–32, and 33–35.

b. Calculate the relative frequencies and percentages for all classes.

c. Construct a histogram and a polygon for the percentage distribution.

d. For what percentage of the days is the number of computer monitors produced in the interval 27–29?

2.20 The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45	52	48	41	56	46	44	42	48	53	51	53	51
48	46	43	52	50	54	47	44	47	50	49	52	

a. Make the frequency distribution table for these data.

b. Calculate the relative frequencies for all classes.

c. Construct a histogram for the relative frequency distribution.

d. Construct a polygon for the relative frequency distribution.

2.21 In its November 29, 2004 issue, *Business Week* magazine presented data on charitable contributions by S&P 500 companies during the 2003 fiscal year. The following table lists the cash contributions (in millions of dollars) of the top 15 companies from this list based on the cash gifts as percentage of their revenues.

Company	Cash Contributions During 2003 (millions of dollars)
Freeport-McMoRan	21.7
Corning	29.0
Avon Products	49.3
Newmont Mining	22.8
Computer Associates	15.3
General Mills	49.3
Fifth Third Bancorp	30.0
M & T Bank	13.7
Eli Lilly & Co.	51.1
Medtronic	31.0
Northern Trust	9.5
Janus Capital Group	3.3
Guidant	12.1
KeyCorp	18.6
Sallie Mae	14.1

Source: Company reports.

a. Construct a frequency distribution table. Take the classes as 2 to less than 12, 12 to less that 22, and so on.
b. Calculate the relative frequencies and percentages for all classes.

Exercises 2.22 through 2.26 are based on the following data.

The following table gives the crime rates per 100,000 people for five types of crimes for 26 states east of the Mississippi River. The rates are based on recent data from FBI *Uniform Crime Reports* that appeared in the *World Almanac and Book of Facts 2005*.

State	Murder Rate	Burglary Rate	Robbery Rate	Aggravated Assault Rate	Motor Vehicle Theft Rate
CT	2.3	493.8	117.3	170.4	334.4
ME	1.1	538.1	20.9	56.8	110.4
MA	2.7	517.2	111.5	342.5	413.6
NH	.9	379.4	32.4	92.9	152.5
RI	3.8	599.7	85.6	158.8	455.8
VT	2.1	565.9	12.5	71.7	124.7
NJ	3.9	511.0	161.9	193.0	416.0
NY	4.7	400.4	191.3	279.7	247.2
PA	5.1	450.8	139.1	227.5	266.0
IL	7.5	643.8	200.6	378.5	356.0
IN	5.9	691.7	107.4	214.1	329.4
MI	6.7	706.1	117.9	362.3	494.7
OH	4.6	868.2	156.5	148.2	374.5
WI	2.8	513.2	86.6	112.7	247.3
DE	3.2	663.3	142.9	408.5	378.6
FL	5.5	1060.5	194.9	529.4	529.6
GA	7.1	863.7	156.9	270.1	444.3

MD	9.4	728.5	245.8	489.5	623.3
NC	6.6	1196.3	146.7	290.5	298.9
SC	7.3	1065.1	140.6	626.5	410.7
VA	5.3	435.4	95.4	165.5	253.3
WV	3.2	537.1	36.5	176.4	216.3
AL	6.8	949.0	132.9	267.5	309.6
KY	4.5	680.6	74.8	173.1	213.8
MS	9.2	1030.5	116.9	178.0	331.6
TN	7.2	1056.5	162.4	507.8	457.8

2.22 **a.** Prepare a frequency distribution table for murder rates using five classes of equal widths.
b. Construct the relative frequency and percentage distribution columns.

2.23 **a.** Prepare a frequency distribution table for burglary rates using five classes of equal widths.
b. Construct the relative frequency and percentage distribution columns.

2.24 **a.** Prepare a frequency distribution table for robbery rates.
b. Construct the relative frequency and percentage distribution columns.
c. Draw a histogram and polygon for the relative frequency distribution.

2.25 **a.** Prepare a frequency distribution table for aggravated assault rates.
b. Calculate the relative frequencies and percentages for all classes.
c. Draw a histogram and a polygon for the percentage distribution.

2.26 **a.** Prepare a frequency distribution table for motor vehicle theft rates. Take 100 as the lower boundary of the first class and 100 as the width of each class.
b. Construct the relative frequency and percentage distribution columns.

2.27 The following table lists the earned run averages (ERAs) for the pitchers of all 16 National League Baseball teams for the 2004 season.

Team	ERA	Team	ERA
Arizona	4.98	Milwaukee	4.24
Atlanta	3.74	Montreal (now Washington)	4.33
Chicago	3.81	New York	4.09
Colorado	5.54	Philadelphia	4.45
Cincinnati	5.19	Pittsburgh	4.29
Florida	4.10	St. Louis	3.75
Houston	4.05	San Diego	4.03
Los Angeles	4.01	San Francisco	4.29

Source: The World Almanac and Book of Facts 2005.

a. Construct a frequency distribution table. Take 3.50 as the lower boundary of the first class and .50 as the width of each class.
b. Prepare the relative frequency and percentage distribution columns for the frequency table of part a.

2.28 The following data give the number of turnovers (fumbles and interceptions) by a college football team for each game in the past two seasons.

3	2	1	4	0	2	2	1	0	3	2	3
0	2	3	1	4	1	3	2	4	0	1	2

a. Prepare a frequency distribution table for these data using single-valued classes.
b. Calculate the relative frequencies and percentages for all classes.
c. In how many games did the team commit two or more turnovers?
d. Draw a bar graph for the frequency distribution of part a.

2.29 According to a survey by the U.S. Public Interest Research Group, about 79% of credit reports contain errors (*USA TODAY,* June 18, 2004). Suppose in a random sample of 25 credit reports, the number of errors found are as listed below.

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

a. Prepare a frequency distribution table for these data using single-valued classes.
b. Calculate the relative frequencies and percentages for all classes.
c. How many of these reports contained two or more errors?
d. Draw a bar graph for the frequency distribution of part a.

2.30 The following table gives the frequency distribution for the numbers of parking tickets received on the campus of a university during the past week for 200 students.

Number of Tickets	Number of Students
0	59
1	44
2	37
3	32
4	28

Draw two bar graphs for these data, the first without truncating the frequency axis and the second by truncating the frequency axis. In the second case, mark the frequencies on the vertical axis starting with 25. Briefly comment on the two bar graphs.

2.31 Eighty adults were asked to watch a 30-minute infomercial until the presentation ended or until boredom became intolerable. The following table lists the frequency distribution of the times that these adults were able to watch the infomercial.

Time (minutes)	Number of Adults
0 to less than 6	16
6 to less than 12	21
12 to less than 18	18
18 to less than 24	11
24 to less than 30	14

Draw two histograms for these data, the first without truncating the frequency axis. In the second case, mark the frequencies on the vertical axis starting with 10. Briefly comment on the two histograms.

2.5 Cumulative Frequency Distributions

Consider again Example 2–3 of Section 2.3.2 about the home runs hit by Major League Baseball teams. Suppose we want to know how many teams hit a total of 200 or fewer home runs during the 2004 season. Such a question can be answered using a **cumulative frequency distribution**. Each class in a cumulative frequency distribution table gives the total number of values that fall below a certain value. A cumulative frequency distribution is constructed for quantitative data only.

Definition

Cumulative Frequency Distribution A *cumulative frequency distribution* gives the total number of values that fall below the upper boundary of each class.

In a cumulative frequency distribution table, each class has the same lower limit but a different upper limit. Example 2–7 illustrates the procedure to prepare a cumulative frequency distribution.

EXAMPLE 2–7

Constructing a cumulative frequency distribution table.

Using the frequency distribution of Table 2.10, reproduced here, prepare a cumulative frequency distribution for the home runs hit by Major League Baseball teams during the 2004 season.

Total Home Runs	f
135–156	10
157–178	3
179–200	7
201–222	6
223–244	4

Solution Table 2.14 gives the cumulative frequency distribution for the home runs hit by Major League Baseball teams. As we can observe, 135 (which is the lower limit of the first class in Table 2.10) is taken as the lower limit of each class in Table 2.14. The upper limits of all classes in Table 2.14 are the same as those in Table 2.10. To obtain the cumulative frequency of a class, we add the frequency of that class in Table 2.10 to the frequencies of all preceding classes. The cumulative frequencies are recorded in the third column of Table 2.14. The second column of this table lists the class boundaries.

Table 2.14 Cumulative Frequency Distribution of Home Runs by Baseball Teams

Class Limits	Class Boundaries	Cumulative Frequency
135–156	134.5 to less than 156.5	10
135–178	134.5 to less than 178.5	$10 + 3 = 13$
135–200	134.5 to less than 200.5	$10 + 3 + 7 = 20$
135–222	134.5 to less than 222.5	$10 + 3 + 7 + 6 = 26$
135–244	134.5 to less than 244.5	$10 + 3 + 7 + 6 + 4 = 30$

From Table 2.14, we can determine the number of observations that fall below the upper limit or boundary of each class. For example, 20 Major League Baseball teams hit a total of 200 or fewer home runs.

The **cumulative relative frequencies** are obtained by dividing the cumulative frequencies by the total number of observations in the data set. The **cumulative percentages** are obtained by multiplying the cumulative relative frequencies by 100.

Calculating Cumulative Relative Frequency and Cumulative Percentage

$$\text{Cumulative relative frequency} = \frac{\text{Cumulative frequency of a class}}{\text{Total observations in the data set}}$$

$$\text{Cumulative percentage} = (\text{Cumulative relative frequency}) \cdot 100$$

Table 2.15 contains both the cumulative relative frequencies and the cumulative percentages for Table 2.14. We can observe, for example, that 66.7% of the Major League Baseball teams hit 200 or fewer home runs during the 2004 season.

Ogives

When plotted on a diagram, the cumulative frequencies give a curve that is called an **ogive** (pronounced *o-jive*). Figure 2.12 gives an ogive for the cumulative frequency distribution of Table 2.14.

Table 2.15 **Cumulative Relative Frequency and Cumulative Percentage Distributions for Home Runs Hit by Baseball Teams**

Class Limits	Cumulative Relative Frequency	Cumulative Percentage
135–156	10/30 = .333	33.3
135–178	13/30 = .433	43.3
135–200	20/30 = .667	66.7
135–222	26/30 = .867	86.7
135–244	30/30 = 1.000	100.0

To draw the ogive in Figure 2.12, the variable, which is total home runs, is marked on the horizontal axis and the cumulative frequencies on the vertical axis. Then the dots are marked above the upper boundaries of various classes at the heights equal to the corresponding cumulative frequencies. The ogive is obtained by joining consecutive points with straight lines. Note that the ogive starts at the lower boundary of the first class and ends at the upper boundary of the last class.

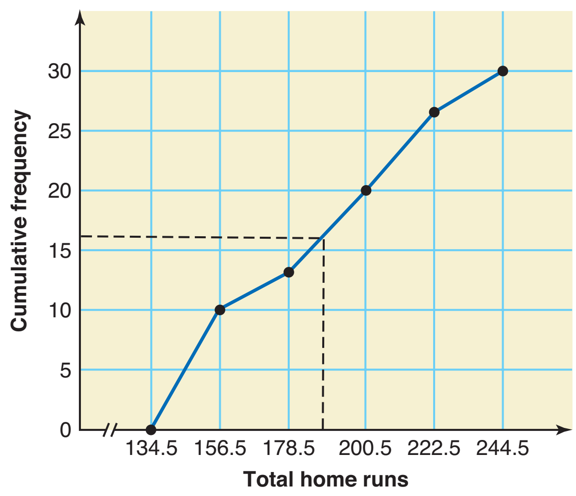


Figure 2.12 Ogive for the cumulative frequency distribution of Table 2.14.

Definition

Ogive An *ogive* is a curve drawn for the cumulative frequency distribution by joining with straight lines the dots marked above the upper boundaries of classes at heights equal to the cumulative frequencies of respective classes.

One advantage of an ogive is that it can be used to approximate the cumulative frequency for any interval. For example, we can use Figure 2.12 to find the number of Major League Baseball teams with 188 or fewer home runs. First, draw a vertical line from 188 on the horizontal axis up to the ogive. Then draw a horizontal line from the point where this line intersects the ogive to the vertical axis. This point gives the cumulative frequency of the class 135–188. In Figure 2.12, this cumulative frequency is (approximately) 16 as shown by the dashed line. Therefore, 16 baseball teams had 188 or fewer home runs during the 2004 season.

We can draw an ogive for cumulative relative frequency and cumulative percentage distributions the same way we did for the cumulative frequency distribution.

EXERCISES

CONCEPTS AND PROCEDURES

2.32 Briefly explain the concept of cumulative frequency distribution. How are the cumulative relative frequencies and cumulative percentages calculated?

2.33 Explain for what kind of frequency distribution an ogive is drawn. Can you think of any use for an ogive? Explain.

APPLICATIONS

2.34 The following table, reproduced from Exercise 2.14, gives the frequency distribution of the number of credit cards possessed by 80 adults.

Number of Credit Cards	Number of Adults
0 to 3	18
4 to 7	26
8 to 11	22
12 to 15	11
16 to 19	3

a. Prepare a cumulative frequency distribution.
b. Calculate the cumulative relative frequencies and cumulative percentages for all classes.
c. Find the percentage of these adults who possess 7 or fewer credit cards.
d. Draw an ogive for the cumulative percentage distribution.
e. Using the ogive, find the percentage of adults who possess 10 or fewer credit cards.

2.35 The following table, reproduced from Exercise 2.15, gives the frequency distribution of ages for all 50 employees of a company.

Age	Number of Employees
18 to 30	12
31 to 43	19
44 to 56	14
57 to 69	5

a. Prepare a cumulative frequency distribution table.
b. Calculate the cumulative relative frequencies and cumulative percentages for all classes.
c. What percentage of the employees of this company are 44 years of age or older?
d. Draw an ogive for the cumulative percentage distribution.
e. Using the ogive, find the percentage of employees who are age 40 or younger.

2.36 Using the frequency distribution table constructed in Exercise 2.18, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

2.37 Using the frequency distribution table constructed in Exercise 2.19, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

2.38 Using the frequency distribution table constructed in Exercise 2.20, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

2.39 Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the frequency distribution constructed in Exercise 2.23.

2.40 Using the frequency distribution table constructed for the data of Exercise 2.25, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions.

2.41 Refer to the frequency distribution table constructed in Exercise 2.26. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions by using that table.

2.42 Using the frequency distribution table constructed for the data of Exercise 2.21, prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions. Draw an ogive for the cumulative frequency distribution. Using the ogive, find the (approximate) number of companies in these 15 who made cash contributions of less than $35 million in 2003.

2.43 Refer to the frequency distribution table constructed in Exercise 2.27. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions. Draw an ogive for the cumulative frequency distribution. Using the ogive, find the (approximate) number of teams with an ERA of less than 4.20.

2.6 Stem-and-Leaf Displays

Another technique that is used to present quantitative data in condensed form is the **stem-and-leaf display**. An advantage of a stem-and-leaf display over a frequency distribution is that by preparing a stem-and-leaf display we do not lose information on individual observations. A stem-and-leaf display is constructed only for quantitative data.

Definition

Stem-and-Leaf Display In a *stem-and-leaf display* of quantitative data, each value is divided into two portions—a stem and a leaf. The leaves for each stem are shown separately in a display.

Example 2–8 describes the procedure for constructing a stem-and-leaf display.

■ EXAMPLE 2–8

Constructing a stem-and-leaf display for two-digit numbers.

The following are the scores of 30 college students on a statistics test.

75	52	80	96	65	79	71	87	93	95
69	72	81	61	76	86	79	68	50	92
83	84	77	64	71	87	72	92	57	98

Construct a stem-and-leaf display.

Solution To construct a stem-and-leaf display for these scores, we split each score into two parts. The first part contains the first digit, which is called the *stem*. The second part contains the second digit, which is called the *leaf*. Thus, for the score of the first student, which is 75, 7 is the stem and 5 is the leaf. For the score of the second student, which is 52, the stem is 5 and the leaf is 2. We observe from the data that the stems for all scores are 5, 6, 7, 8, and 9 because all the scores lie in the range 50 to 98. To create a stem-and-leaf display, we draw a vertical line and write the stems on the left side of it, arranged in increasing order, as shown in Figure 2.13.

Stems
↓
5 | 2 ⟵ Leaf for 52
6 |
7 | 5 ⟵ Leaf for 75
8 |
9 |

Figure 2.13 Stem-and-leaf display.

After we have listed the stems, we read the leaves for all scores and record them next to the corresponding stems on the right side of the vertical line. For example, for the first score we write the leaf 5 next to the stem 7; for the second score we write the leaf 2 next to the stem 5. The recording of these two scores in a stem-and-leaf display is shown in Figure 2.13.

Now, we read all the scores and write the leaves on the right side of the vertical line in the rows of corresponding stems. The complete stem-and-leaf display for scores is shown in Figure 2.14.

Figure 2.14 Stem-and-leaf display of test scores.

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

By looking at the stem-and-leaf display of Figure 2.14, we can observe how the data values are distributed. For example, the stem 7 has the highest frequency, followed by stems 8, 9, 6, and 5.

The leaves for each stem of the stem-and-leaf display of Figure 2.14 are *ranked* (in increasing order) and presented in Figure 2.15.

Figure 2.15 Ranked stem-and-leaf display of test scores.

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

■

As already mentioned, one advantage of a stem-and-leaf display is that we do not lose information on individual observations. We can rewrite the individual scores of the 30 college students from the stem-and-leaf display of Figure 2.14 or 2.15. By contrast, the information on individual observations is lost when data are grouped into a frequency table.

■ EXAMPLE 2–9

Constructing a stem-and-leaf display for three-and four-digit numbers.

The following data give the monthly rents paid by a sample of 30 households selected from a small city.

880	1081	721	1075	1023	775	1235	750	965	960
1210	985	1231	932	850	825	1000	915	1191	1035
1151	630	1175	952	1100	1140	750	1140	1370	1280

Construct a stem-and-leaf display for these data.

Solution Each of the values in the data set contains either three or four digits. We will take the first digit for three-digit numbers and the first two digits for four-digit numbers as stems. Then we will use the last two digits of each number as a leaf. Thus for the first value, which is 880, the stem is 8 and the leaf is 80. The stems for the entire data set are 6, 7, 8, 9, 10, 11, 12, and 13. They are recorded on the left side of the vertical line in Figure 2.16. The leaves for the numbers are recorded on the right side.

Figure 2.16 Stem-and-leaf display of rents.

6	30
7	75 50 21 50
8	80 25 50
9	32 52 15 60 85 65
10	23 81 35 75 00
11	91 51 40 75 40 00
12	10 31 35 80
13	70

■

Sometimes a data set may contain too many stems, with each stem containing only a few leaves. In such cases, we may want to condense the stem-and-leaf display by *grouping the stems*. Example 2–10 describes this procedure.

■ EXAMPLE 2–10

Preparing a grouped stem-and-leaf display.

The following stem-and-leaf display is prepared for the number of hours that 25 students spent working on computers during the past month.

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

Prepare a new stem-and-leaf display by grouping the stems.

Solution To condense the given stem-and-leaf display, we can combine the first three rows, the middle three rows, and the last three rows, thus getting the stems 0–2, 3–5, and 6–8. The leaves for each stem of a group are separated by an asterisk (*), as shown in Figure 2.17. Thus, the leaf 6 in the first row corresponds to stem 0; the leaves 1, 7, and 9 correspond to stem 1; and leaves 2 and 6 belong to stem 2.

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

Figure 2.17 Grouped stem-and-leaf display.

If a stem does not contain a leaf, this is indicated in the grouped stem-and-leaf display by two consecutive asterisks. For example, in the above stem-and-leaf display, there is no leaf for 7; that is, there in no number in the 70s. Hence, in Figure 2.17, we have two asterisks after the leaves for 6 and before the leaves for 8. ■

EXERCISES

■ CONCEPTS AND PROCEDURES

2.44 Briefly explain how to prepare a stem-and-leaf display for a data set. You may use an example to illustrate.

2.45 What advantage does preparing a stem-and-leaf display have over grouping a data set using a frequency distribution? Give one example.

2.46 Consider this stem-and-leaf display.

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

Write the data set that is represented by the display.

2.47 Consider this stem-and-leaf display.

2–3	18 45 56 * 29 67 83 97
4–5	04 27 33 71 * 23 37 51 63 81 92
6–8	22 36 47 55 78 89 * * 10 41

Write the data set that is represented by the display.

APPLICATIONS

2.48 The following data give the time (in minutes) that each of 20 students waited in line at their bookstore to pay for their textbooks in the beginning of Spring 2006 semester.

15	8	23	21	5	17	31	22	34	6
5	10	14	17	16	25	30	3	31	19

Construct a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order.

2.49 Following are the total yards gained rushing during the 2005 season by 14 running backs of 14 college football teams.

745	921	1133	1024	848	775	800
1009	1275	857	933	1145	967	995

Prepare a stem-and-leaf display. Arrange the leaves for each stem in increasing order.

2.50 Reconsider the data on the numbers of computer monitors produced at the Nixon Corporation for a sample of 30 days given in Exercise 2.19. Prepare a stem-and-leaf display for those data. Arrange the leaves for each stem in increasing order.

2.51 Reconsider the data on the numbers of computer keyboards assembled at the Twentieth Century Electronics Company given in Exercise 2.20. Prepare a stem-and-leaf display for those data. Arrange the leaves for each stem in increasing order.

2.52 Refer to Exercise 2.18. Rewrite those data by rounding each average attendance to the nearest thousand. For example, an attendance of 27,200 will be rounded to 27 thousand, and 41,700 will be rounded to 42 thousand. Prepare a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order.

2.53 These data give the times (in minutes) taken to commute from home to work for 20 workers.

10	50	65	33	48	5	11	23	39	26
26	32	17	7	15	19	29	43	21	22

Construct a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order. (*Note*: To prepare a stem-and-leaf display, each number in this data set can be written as a two-digit number. For example, 5 can be written as 05, for which the stem is 0 and the leaf is 5.)

2.54 The following data give the times served (in months) by 35 prison inmates who were released recently.

37	6	20	5	25	30	24	10	12	20
24	8	26	15	13	22	72	80	96	33
84	86	70	40	92	36	28	90	36	32
72	45	38	18	9					

a. Prepare a stem-and-leaf display for these data.
b. Condense the stem-and-leaf display by grouping the stems as 0–2, 3–5, and 6–9.

2.55 The following data give the money (in dollars) spent on textbooks by 35 students during the 2005–06 academic year.

565	528	270	220	245	368	210	265	350
345	530	705	490	158	320	505	457	487
617	721	635	438	475	702	538	720	460
540	390	560	570	706	430	268	638	

a. Prepare a stem-and-leaf display for these data using the last two digits as leaves.
b. Condense the stem-and-leaf display by grouping the stems as 1–3, 4–5, and 6–7.

2.7 DOTPLOTS

One of the simplest methods for graphing and understanding quantitative data is to create a dotplot. As with most graphs, statistical software should be used to make a dotplot for large data sets. However, Example 2–11 demonstrates how to create a dotplot by hand.

Dotplots can help us detect **outliers** (also called **extreme values**) in a data set. Outliers are the values that are extremely large or extremely small with respect to the rest of the data values.

Definition

Outliers or Extreme Values Values that are very small or very large relative to the majority of the values in a data set are called outliers or extreme values.

■ EXAMPLE 2–11

Creating a dotplot.

Table 2.16 lists the number of runs batted in (RBIs) during the 2004 Major League Baseball playoffs by members of the Boston Red Sox team with at least one at-bat. Create a dotplot for these data.

Table 2.16 Runs Batted In by Boston Red Sox Players in the 2004 Playoffs

Batter	RBIs	Batter	RBIs
D. Mientkiewicz	1	D. Mirabelli	0
D. Ortiz	19	J. Varitek	11
M. Ramirez	11	K. Millar	6
B. Mueller	3	G. Kapler	0
O. Cabrera	11	M. Bellhorn	8
J. Damon	9	P. Reese	0
T. Nixon	8	K. Youkilis	0

Source: http://www.sportsnetwork.com.

Solution Below we show how to make a dotplot for these data on RBIs.

Step 1. The minimum and maximum values in this data set are 0 and 19 RBIs, respectively. First we draw a horizontal line (let us call this the numbers line) with numbers that cover the given data as shown in Figure 2.18. Note that the numbers line in Figure 2.18 shows the values from 0 to 20.

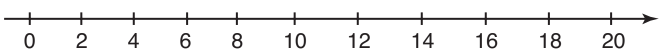


Figure 2.18 Numbers line.

Step 2. Place a dot above the value on the numbers line that represents each RBI listed in the table. For example, Doug Mientkiewicz had 1 RBI in the playoffs. Place a dot above 1 on the numbers line as shown in Figure 2.19. If there are two or more observations with the same value, we stack dots vertically above each other to represent those values. For example, as shown in Table 2.16, three players had 11 RBIs. We stack three dots (one for each player) above 11 on the numbers line as shown in Figure 2.19. Figure 2.19 gives the complete dotplot.

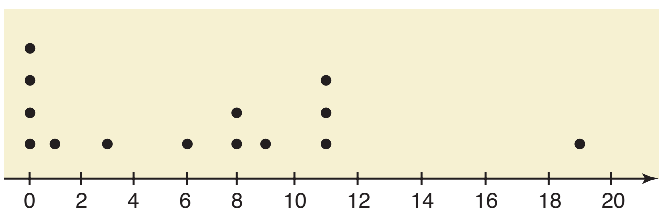


Figure 2.19 Dotplot for RBIs.

As we examine the dotplot of Figure 2.19, we notice that there are two clusters (groups) of data. Approximately half of the players had three or fewer RBIs and approximately half had between 6 and 11 RBIs. In addition, one player (David Ortiz) had 19 RBIs. Graphically, Ortiz's dot is far away from the general pattern of the data. When this occurs, we suspect that the data value could be an outlier. (In Chapter 3 in the box-and-whisker section we will learn a numerical method to determine if a data point should be classified as an outlier.) ■

Dotplots are also very useful for comparing two or more data sets. To do so, we create a dotplot for each data set with numbers lines for all data sets on the same scale. We place these data sets on top of each other, resulting in what are called **stacked dotplots**. Example 2–12 shows this procedure.

■ EXAMPLE 2–12

Comparing two data sets using dotplots.

Refer to Table 2.16 in Example 2–11 that gives the RBIs for players of the Red Sox baseball team during the 2004 playoffs. Table 2.17 below gives the RBIs for the players of the St. Louis Cardinals team during the 2004 playoffs. These two teams played in the 2004 World Series. Make dotplots for both sets of data and compare these two dotplots.

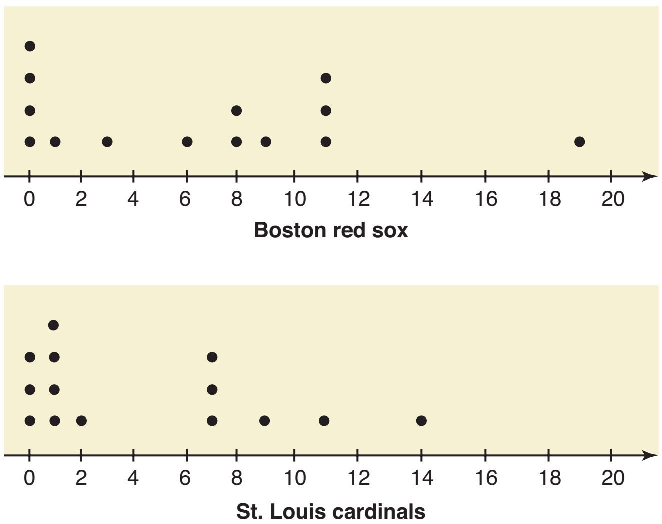
Table 2.17 **Runs Batted In by St. Louis Cardinals Players in the 2004 Playoffs**

Batter	RBIs	Batter	RBIs
A. Pujols	14	R. Sanders	1
L. Walker	11	M. Anderson	0
E. Renteria	7	S. Taguchi	1
R. Cedeno	1	S. Rolen	7
J. Edmonds	9	Y. Molina	0
T. Womack	2	J. Mabry	1
M. Matheny	7	H. Luna	0

Source: http://www.sportsnetwork.com.

Solution Figure 2.20 shows the dotplots for the RBIs for players of both these teams.

Figure 2.20 Dotplots of RBIs for Boston and St. Louis Baseball Teams.



Looking at these two dotplots, we can notice that each group has two clusters, and the clusters are in approximately the same areas. However, the St. Louis Cardinals had more players who had a lower number of RBIs (the first cluster) than the Boston Red Sox players. We also notice that among the groups of players with higher numbers of RBIs (the second cluster), more Red Sox players are distributed towards the higher end, while more Cardinals are distributed towards the lower end. ■

In practice, dotplots and other statistical graphs will be created using statistical software. The Technology Instruction section at the end of this chapter shows how we can do so.

EXERCISES

CONCEPTS AND PROCEDURES

2.56 Briefly explain how to prepare a dotplot for a data set. You may use an example to illustrate.

2.57 What is a stacked dotplot, and how is it used? Explain.

2.58 Create a dotplot for the following data set.

1	2	0	5	1	1	3	2	0	5
2	1	2	1	2	0	1	3	1	2

APPLICATIONS

2.59 Reconsider the data on the numbers of computer keyboards assembled at the Twentieth Century Electronics Company given in Exercise 2.20. Create a dotplot for those data.

2.60 Create a dotplot for the data on the number of turnovers (fumbles and interceptions) by a college football team for games in the past two seasons given in Exercise 2.28.

2.61 Reconsider the data on the numbers of errors found in 25 randomly selected credit reports given in Exercise 2.29. Create a dotplot for those data.

2.62 The following data give the number of times each of the 30 randomly selected account holders at a bank used that bank's ATM during a 60-day period.

3	2	3	2	2	5	0	4	1	3
2	3	3	5	9	0	3	2	2	15
1	3	2	7	9	3	0	4	2	2

Create a dotplot for these data and point out any clusters or outliers.

2.63 The following data give the number of times each of the 20 randomly selected male students from a state university ate at fast-food restaurants during a seven-day period.

5	8	10	3	5	5	10	7	2	1
10	4	5	0	10	1	2	8	3	5

Create a dotplot for these data and point out any clusters or outliers.

2.64 Reconsider Exercise 2.63. The following data give the number of times each of the 20 randomly selected female students from the same state university ate at fast-food restaurants during the same seven-day period.

0	0	4	2	4	10	2	5	0	5
6	1	1	4	6	2	4	5	6	0

a. Create a dotplot for these data.
b. Use the dotplots for male and female students to compare the two data sets.

2.65 The following table gives the number of stolen bases during the 2004 season by each Boston Red Sox player with 150 or more at-bats.

Player	Stolen Bases	Player	Stolen Bases
J. Damon	19	G. Kapler	5
D. Ortiz	0	P. Reese	6
M. Ramirez	2	O. Cabrera	4
M. Bellhorn	6	K. Youkilis	0
K. Millar	1	D. Mirabelli	0
J. Varitek	10	N. Garciaparra	2
B. Mueller	2	D. McCarty	1
D. Mientkiewicz	2		

Source: Sports Illustrated 2005 Almanac.

Create a dotplot for these data. Mention any clusters and/or outliers you observe.

USES AND MISUSES... BE SENSITIVE

As a budding statistician, the first task that you will perform is organizing your data. As this chapter showed, once organized, it is often convenient to display data in a graphical form. Though you were warned that truncating and changing the scale of axes might distort your data, the simple act of grouping your data can do the very same thing. This is especially important to remember when partitioning your data into classes.

Suppose that you were presented with a set of data on the ages of employees of a company. This phenomenon is known as sensitivity, and your goal is to present results that are not too sensitive to class boundaries and groupings. Let us look at an example to make the notion concrete. Suppose that you are given the following data representing ages of employees at Company X.

37	41	49	23
37	41	51	33
38	43	30	35
38	43	29	52
39	42	23	48
39	42	24	

The histogram in Figure 2.21 was made by partitioning the data by decade, thus getting the classes as 21–30, 31–40, and so on. Simple inspection tells us that most employees of the company are in their 30s and 40s, and that the ages are skewed (a little) to the younger side.

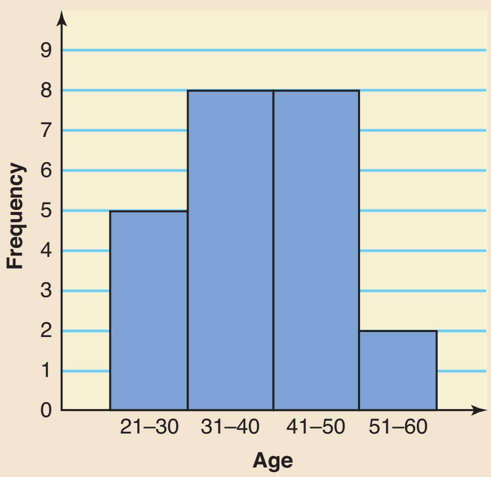


Figure 2.21 Histogram of the Ages of Company X's Workers.

But if we partition the data on the fifth year of the decade, thus getting the classes as 16–25, 26–35, and so on, we get a different picture as shown in Figure 2.22. Twelve of the 23 employees are between the ages of 36 and 45! Which picture is correct? The problem is that they both are.

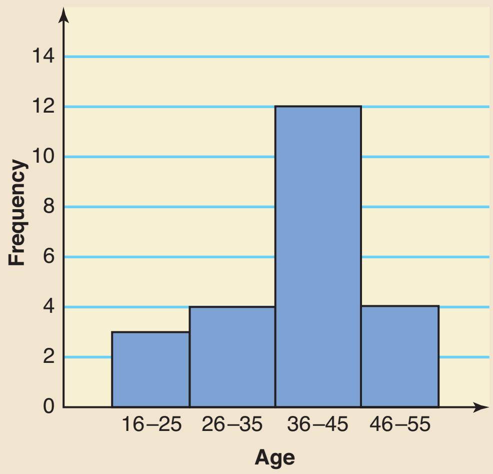


Figure 2.22 Histogram of the Ages of Company X's Workers.

The opposite of sensitivity is called *robustness*, and how do we present results in a robust way? When you are the statistician, be very careful when creating your classes. Follow the basic procedure in the text, rounding to a convenient class width when appropriate, but then look at your data and ask yourself how your data are distributed within the classes you have chosen. If the data seem to be skewed within a class or concentrated at the boundaries, think about how your interpretation would change if you were to shift your class boundaries a little. If your interpretation would change, you need to make a judgment about which set of classes is most appropriate. You might choose to tell your audience that the graph conceals something about the data. When you are reviewing another statistician's graphical results, think about what the choice of classes might be revealing or concealing.

Glossary

Bar graph A graph made of bars whose heights represent the frequencies of respective categories.

Class An interval that includes all the values in a (quantitative) data set that fall within two numbers, the lower and upper limits of the class.

Class boundary The midpoint of the upper limit of one class and the lower limit of the next class.

Class frequency The number of values in a data set that belong to a certain class.

Class midpoint or **mark** The class midpoint or mark is obtained by dividing the sum of the lower and upper limits (or boundaries) of a class by 2.

Class width or **size** The difference between the two boundaries of a class.

Cumulative frequency The frequency of a class that includes all values in a data set that fall below the upper boundary of that class.

Cumulative frequency distribution A table that lists the total number of values that fall below the upper boundary of each class.

Cumulative percentage The cumulative relative frequency multiplied by 100.

Cumulative relative frequency The cumulative frequency of a class divided by the total number of observations.

Frequency distribution A table that lists all the categories or classes and the number of values that belong to each of these categories or classes.

Grouped data A data set presented in the form of a frequency distribution.

Histogram A graph in which classes are marked on the horizontal axis and frequencies, relative frequencies, or percentages are marked on the vertical axis. The frequencies, relative frequencies, or percentages of various classes are represented by bars that are drawn adjacent to each other.

Ogive A curve drawn for a cumulative frequency distribution.

Outliers or **Extreme values** Values that are very small or very large relative to the majority of the values in a data set.

Percentage The percentage for a class or category is obtained by multiplying the relative frequency of that class or category by 100.

Pie chart A circle divided into portions that represent the relative frequencies or percentages of different categories or classes.

Polygon A graph formed by joining the midpoints of the tops of successive bars in a histogram by straight lines.

Raw data Data recorded in the sequence in which they are collected and before they are processed.

Relative frequency The frequency of a class or category divided by the sum of all frequencies.

Skewed-to-the-left histogram A histogram with a longer tail on the left side.

Skewed-to-the-right histogram A histogram with a longer tail on the right side.

Stem-and-leaf display A display of data in which each value is divided into two portions—a stem and a leaf.

Symmetric histogram A histogram that is identical on both sides of its central point.

Ungrouped data Data containing information on each member of a sample or population individually.

Uniform or **rectangular histogram** A histogram with the same frequency for all classes.

Supplementary Exercises

2.66 The following data give the political party of each of the first 30 U.S. presidents. In the data, D stands for Democrat, DR for Democratic Republican, F for Federalist, R for Republican, and W for Whig.

F	F	DR	DR	DR	DR	D	D	W	W
D	W	W	D	D	R	D	R	R	R
R	D	R	D	R	R	R	D	R	R

a. Prepare a frequency distribution table for these data.
b. Calculate the relative frequency and percentage distributions.
c. Draw a bar graph for the relative frequency distribution and a pie chart for the percentage distribution.
d. What percentage of these presidents were Whigs?

2.67 In a survey conducted by Harris Interactive for Tylenol PM, people were asked about how they cope with afternoon drowsiness (*USA TODAY*, October 19, 2004). Of the respondents, 35% said they drink a caffeinated beverage (C). Other responses were: taking a nap (N), going for a walk (W), or eating a sugary snack (S). Suppose that in a recent poll, 30 people were asked which one of the above choices they preferred when dealing with drowsiness. Their responses are given below.

C	C	N	C	W	N	S	C	N	C
S	S	W	C	C	N	S	N	C	C
N	C	C	W	W	C	W	N	C	S

a. Prepare a frequency distribution table for these data.
b. Calculate the relative frequencies and percentages for all classes.
c. Draw a bar graph for the frequency distribution and a pie chart for the percentage distribution.
d. What percentage of these respondents preferred to cope with afternoon drowsiness by taking a nap?

2.68 The following data give the numbers of television sets owned by 40 randomly selected households.

1	1	2	3	2	4	1	3	2	1
3	0	2	1	2	3	2	3	2	2
1	2	1	1	1	3	1	1	1	2
2	4	2	3	1	3	1	2	2	4

a. Prepare a frequency distribution table for these data using single-valued classes.
b. Compute the relative frequency and percentage distributions.
c. Draw a bar graph for the frequency distribution.
d. What percentage of the households own two or more television sets?

2.69 Twenty-four students from universities in Connecticut were asked to name the five current members of the U.S. House of Representatives from Connecticut. The number of correct names supplied by the students are given below.

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

a. Prepare a frequency distribution for these data using single-valued classes.
b. Compute the relative frequency and percentage distributions.
c. What percentage of the students in this sample named fewer than two of the representatives correctly?
d. Draw a bar graph for the relative frequency distribution.

2.70 The following data give the amounts spent on video rentals (in dollars) during 2005 by 30 households randomly selected from those who rented videos in 2005.

595	24	6	100	100	40	622	405	90
55	155	760	405	90	205	70	180	88
808	100	240	127	83	310	350	160	22
111	70	15						

a. Construct a frequency distribution table. Take \$1 as the lower limit of the first class and \$200 as the width of each class.
b. Calculate the relative frequencies and percentages for all classes.
c. What percentage of the households in this sample spent more than \$400 on video rentals in 2005?

2.71 The following data give the numbers of orders received for a sample of 30 hours at the Time-saver Mail Order Company.

34	44	31	52	41	47	38	35	32	39
28	24	46	41	49	53	57	33	27	37
30	27	45	38	34	46	36	30	47	50

a. Construct a frequency distribution table. Take 23 as the lower limit of the first class and 7 as the width of each class.
b. Calculate the relative frequencies and percentages for all classes.
c. For what percentage of the hours in this sample was the number of orders more than 36?

2.72 The following data give the amounts spent (in dollars) on refreshments by 30 spectators randomly selected from those who patronized the concession stands at a recent Major League Baseball game.

4.95	27.99	8.00	5.80	4.50	2.99	4.85	6.00
9.00	15.75	9.50	3.05	5.65	21.00	16.60	18.00
21.77	12.35	7.75	10.45	3.85	28.45	8.35	17.70
19.50	11.65	11.45	3.00	6.55	16.50		

a. Construct a frequency distribution table using the *less than* method to write classes. Take \$0 as the lower boundary of the first class and \$6 as the width of each class.
b. Calculate the relative frequencies and percentages for all classes.
c. Draw a histogram for the frequency distribution.

2.73 The following data give the repair costs (in dollars) for 30 cars randomly selected from a list of cars that were involved in collisions.

2300	750	2500	410	555	1576
2460	1795	2108	897	989	1866
2105	335	1344	1159	1236	1395
6108	4995	5891	2309	3950	3950
6655	4900	1320	2901	1925	6896

a. Construct a frequency distribution table. Take $1 as the lower limit of the first class and $1400 as the width of each class.
b. Compute the relative frequencies and percentages for all classes.
c. Draw a histogram and a polygon for the relative frequency distribution.
d. What are the class boundaries and the width of the fourth class?

2.74 Refer to Exercise 2.70. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions by using the frequency distribution table of that exercise.

2.75 Refer to Exercise 2.71. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the frequency distribution table constructed for the data of that exercise.

2.76 Refer to Exercise 2.72. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the frequency distribution table constructed for the data of that exercise.

2.77 Construct the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions by using the frequency distribution table constructed for the data of Exercise 2.73.

2.78 Refer to Exercise 2.70. Prepare a stem-and-leaf display for the data of that exercise.

2.79 Construct a stem-and-leaf display for the data given in Exercise 2.71.

2.80 The following table gives the revenues (in millions of dollars) for the seven National Hockey League teams with the largest revenues during the 2003–04 season (*Forbes*, November 29, 2004).

Team	Revenue (millions of dollars)
New York Rangers	118
Toronto Maple Leafs	117
Philadelphia Flyers	106
Dallas Stars	103
Detroit Red Wings	97
Colorado Avalanche	99
Boston Bruins	95

Draw two bar graphs for these data, the first without truncating the axis on which revenues are marked and the second by truncating this axis. In the second graph, mark the revenues on the vertical axis starting with $90 million. Briefly comment on the two bar graphs.

2.81 The following table lists the average price per gallon for unleaded regular gasoline in the United States from 1997 to 2004. Note that the average price for 2004 is for the months from January to June only.

Year	Average Price Per Gallon (dollars)
1997	1.234
1998	1.059
1999	1.165
2000	1.510
2001	1.461
2002	1.358
2003	1.591
2004	1.819

Source: Energy Information Administration.

Draw two bar graphs for these data, the first without truncating the axis on which the gasoline prices are marked and the second by truncating this axis. In the second graph, mark the prices on the vertical axis, starting with $1.00. Briefly comment on the two bar graphs.

2.82 Reconsider the data on the times (in minutes) taken to commute from home to work for 20 workers given in Exercises 2.53. Create a dotplot for those data.

2.83 Reconsider the data on the numbers of orders received for a sample of 30 hours at the Timesaver Mail Order Company given in Exercise 2.71. Create a dotplot for those data.

2.84 Twenty-four students from a university in Oregon were asked to name the five current members of the U.S. House of Representatives from their state. The following data give the numbers of correct names given by these students.

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

Create a dotplot for these data.

2.85 The following data give the numbers of visitors during visiting hours on a given evening for each of the 20 randomly selected patients at a hospital.

3	0	1	4	2	0	4	1	1	3
4	2	0	2	2	2	1	1	3	0

Create a dotplot for these data.

Advanced Exercises

2.86 The following frequency distribution table gives the age distribution of drivers who were at fault in auto accidents that occurred during a one-week period in a city.

Age	f
18 to less than 20	7
20 to less than 25	12
25 to less than 30	18
30 to less than 40	14
40 to less than 50	15
50 to less than 60	16
60 and over	35

a. Draw a relative frequency histogram for this table.
b. In what way(s) is this histogram misleading?
c. How can you change the frequency distribution so that the resulting histogram gives a clearer picture?

2.87 Refer to the data presented in Exercise 2.86. Note that there were 50% more accidents in the *25 to less than 30* age group than in the *20 to less than 25* age group. Does this suggest that the older group of drivers in this city is more accident-prone than the younger group? What other explanation might account for the difference in accident rates?

2.88 Suppose a data set contains the ages of 135 autoworkers ranging from 20 to 53.
a. Using Sturge's formula given in footnote 1 on page 36, find an appropriate number of classes for a frequency distribution for this data set.
b. Find an appropriate class width based on the number of classes in part a.

2.89 Statisticians often need to know the shape of a population to make inferences. Suppose that you are asked to specify the shape of the population of weights of all college students.
a. Sketch a graph of what you think the weights of all college students would look like.
b. The following data give the weights (in pounds) of a random sample of 44 college students. (Here, F and M indicate female and male.)

123 F	195 M	138 M	115 F	179 M	119 F	148 F	147 F
180 M	146 F	179 M	189 M	175 M	108 F	193 M	114 F
179 M	147 M	108 F	128 F	164 F	174 M	128 F	159 M
193 M	204 M	125 F	133 F	115 F	168 M	123 F	183 M
116 F	182 M	174 M	102 F	123 F	99 F	161 M	162 M
155 F	202 M	110 F	132 M				

i. Construct a stem-and-leaf display for these data.
ii. Can you explain why these data appear the way they do?

c. Now sketch a new graph of what you think the weights of all college students look like. Is this similar to your sketch in part a?

2.90 Consider the two histograms given in Figure 2.23 that are drawn for the same data set. In this data set, none of the values are integers.

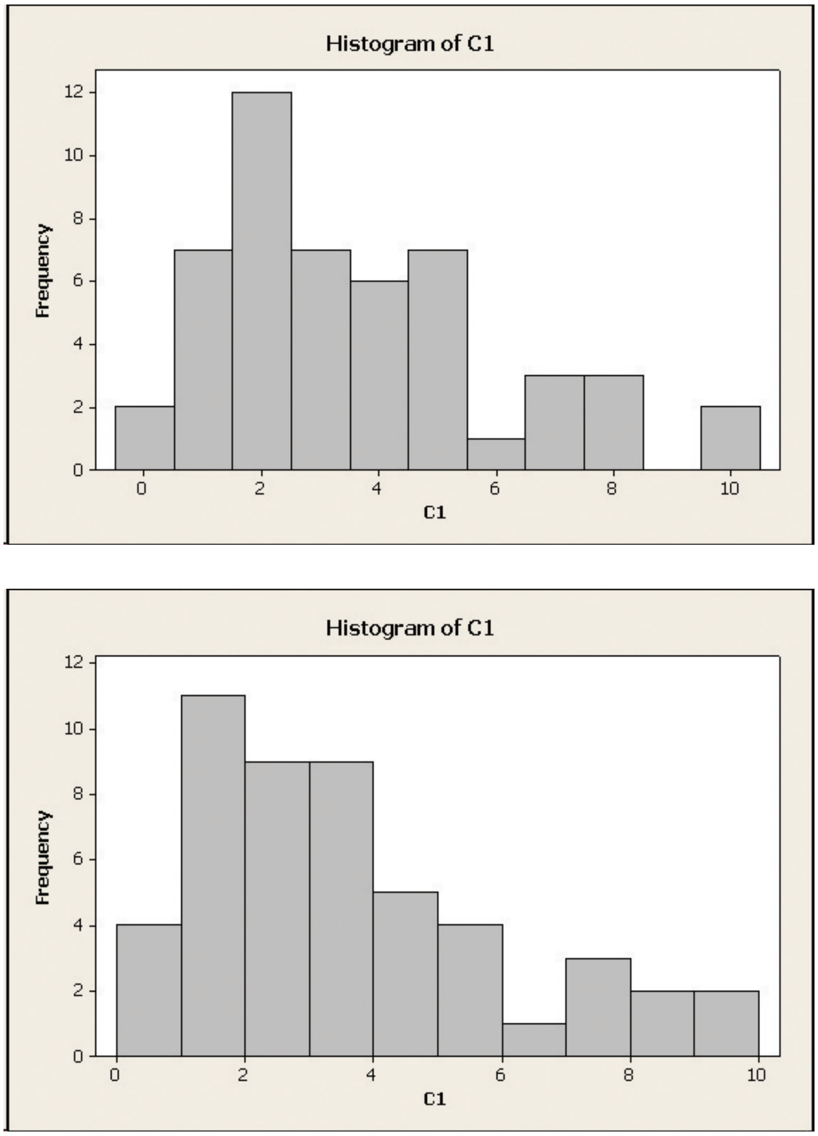


Figure 2.23 Two Histograms for the Same Data.

a. What are the endpoints and widths of classes in each of the two histograms?
b. In the first histogram, of the observations that fall in the interval that is centered at 8, how many are actually between the left endpoint of that interval and 8? Note that you have to consider both histograms to answer this question.
c. Observe the leftmost bars in both histograms. Why is the leftmost bar in the first histogram misleading?

2.91 Refer to the data on weights of 44 college students given in Exercise 2.88. Create a dotplot of all 44 weights. Then create stacked dotplots for the weights of male and female students. Describe the similarities and differences in the distributions of weights of male and female students. Using all three dotplots, explain why you cannot distinguish the lightest males from the heaviest females when you consider only the dotplot of all 44 weights.

2.92 The pie chart in Figure 2.24 shows the percentage distribution of ages (i.e., the percentages of all prostate cancer patients falling in various age groups) for men who were recently diagnosed with prostate cancer.

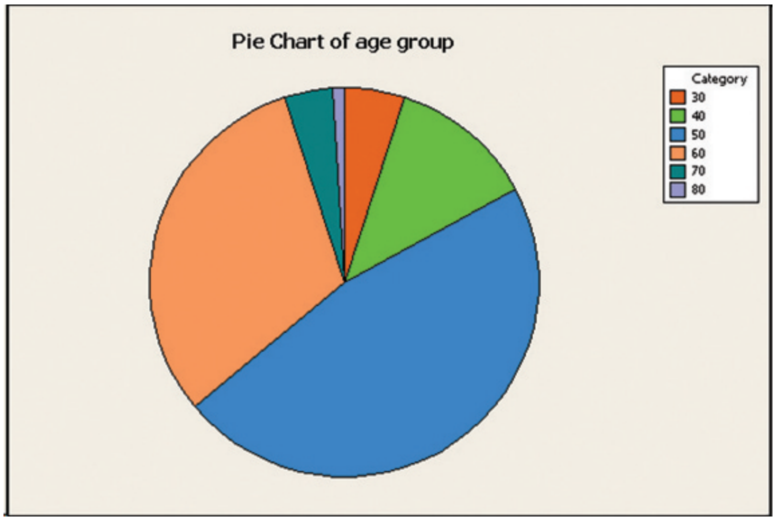


Figure 2.24 Pie Chart of Age Groups.

a. Are more or fewer than 50% of these patients in their 50s? How can you tell?
b. Are more or fewer than 75% of these patients in their 50s and 60s? How can you tell?
c. A reporter looks at this pie chart and says, "Look at all these 50-year-old men who are getting prostate cancer. This is a major concern for a man once he turns 50." Explain why the reporter cannot necessarily conclude from this pie chart that there are a lot of 50-year-old men with prostate cancer. Can you think of any other way to present these cancer cases (both graph and variable) to determine if the reporter's claim is valid?

Self-Review Test

1. Briefly explain the difference between ungrouped and grouped data and give one example of each type.

2. The following table gives the frequency distribution of times (to the nearest hour) that 90 fans spent waiting in line to buy tickets to a rock concert.

Waiting Time (hours)	Frequency
0 to 6	5
7 to 13	27
14 to 20	30
21 to 27	20
28 to 34	8

Circle the correct answer in each of the following statements, which are based on this table.
a. The number of classes in the table is 5, 30, 90
b. The class width is 6, 7, 34
c. The midpoint of the third class is 16.5, 17, 17.5
d. The lower boundary of the second class is 6.5, 7, 7.5
e. The upper limit of the second class is 12.5, 13, 13.5
f. The sample size is 5, 90, 11
g. The relative frequency of the second class is .22, .41, .30

3. Briefly explain and illustrate with the help of graphs a symmetric histogram, a histogram skewed to the right, and a histogram skewed to the left.

4. Twenty elementary school children were asked if they live with both parents (B), father only (F), mother only (M), or someone else (S). The responses of the children follow.

M	B	B	M	F	S	B	M	F	M
B	F	B	M	M	B	B	F	B	M

a. Construct a frequency distribution table.
b. Write the relative frequencies and percentages for all categories.
c. What percentage of the children in this sample live with their mothers only?
d. Draw a bar graph for the frequency distribution and a pie chart for the percentages.

5. A large Midwestern city has been chronically plagued by false fire alarms. The following data set gives the number of false alarms set off each week for a 24-week period in this city.

10	4	8	7	3	7	10	2	6	12	11	8
1	6	5	13	9	7	5	1	14	5	15	3

a. Construct a frequency distribution table. Take 1 as the lower limit of the first class and 3 as the width of each class.
b. Calculate the relative frequencies and percentages for all classes.
c. What percentage of these weeks had 9 or fewer false alarms?
d. Draw the frequency histogram and polygon.

6. Refer to the frequency distribution prepared in Problem 5. Prepare the cumulative percentage distribution using that table. Draw an ogive for the cumulative percentage distribution.

7. Construct a stem-and-leaf display for the following data, which give the times (in minutes) 24 customers spent waiting to speak to a customer service representative when they called about problems with their Internet service provider.

12	15	7	29	32	16	10	14	17	8	19	21
4	14	22	25	18	6	22	16	13	16	12	20

8. Consider this stem-and-leaf display:

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

Write the data set that was used to construct this display.

9. Make a dotplot for the data given in Problem 5.

Mini-Projects

■ MINI-PROJECT 2–1

Using the data you gathered for the mini-project in Chapter 1, prepare a summary of that data set that includes the following.

a. Prepare an appropriate type of frequency distribution table for one of the quantitative variables and then compute relative frequencies and cumulative relative frequencies.
b. Create a histogram, a stem-and-leaf display, and a dotplot of the data. Comment on any symmetry or skewness, and on the presence of clusters and any potential outliers.
c. Make stacked dotplots of the same variable (as in parts a and b) based on the values of one of your categorical variables. For example, if your quantitative variable is GPAs of students, your categorical variable could be gender. Comment on the similarities and differences between the distributions for the different values of your categorical variable.

■ MINI-PROJECT 2–2

Watch four hours of each of two types of television shows (comedy, news, drama, sports, and so forth) and record the duration of each commercial to the nearest second. This can include commercials that are shown between the shows. List these durations of various commercials and the type of the show. Using this information, write a brief report that covers the following.

a. Prepare an appropriate type of frequency distribution table for the quantitative variable and then compute relative frequencies and cumulative relative frequencies.

b. Create a histogram, a stem-and-leaf plot, and a dotplot for these data. Comment on any symmetry or skewness, and on the presence of clusters and any potential outliers.

c. Make stacked dotplots of the same variable for each of the two types of television shows. Comment on the similarities and differences between the distributions for the different types of shows.

DECIDE FOR YOURSELF

Deciding About Statistical Properties

Look around you. Graphs are everywhere. Business reports, newspapers, magazines, and so forth are all loaded with graphs. Unfortunately, some people feel that the primary purpose of graphs is to provide a break from the humdrum text. Executive summaries will often contain graphs so that CEOs and executive vice presidents need only to glance at these graphs to assume that they understand everything without reading more than a paragraph or so of the report. In reality, the usefulness of graphs is somewhere between the fluff of the popular press and the quick answer of the boardroom.

Here you are asked to interpret some graphs, primarily by using them to compare distributions of a variable. As we will discuss in Chapter 3, some of our concerns have to do with the location of the center of a distribution and the variability or spread of a distribution. We can use graphs to compare the centers and variability of two or more distributions.

In practice, the graphs are made using statistical software, so it is important to recognize that computer software is programmed to use the same format for each graph of a specific type, unless you tell the software to do differently. For example, consider the two histograms in Figures 2.25 and 2.26 that are drawn for two different data sets.

1. Examine the two graphs of Figures 2.25 and 2.26.

2. Explain what is meant by the statement "the shapes of the two distributions are the same."

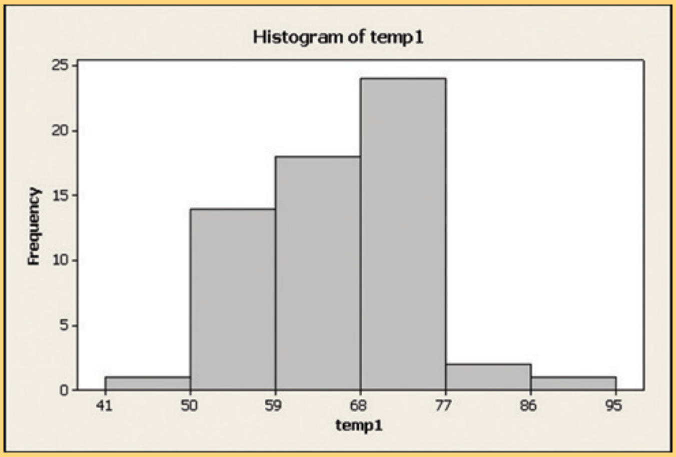


Figure 2.25 Histogram of Data Temp 1.

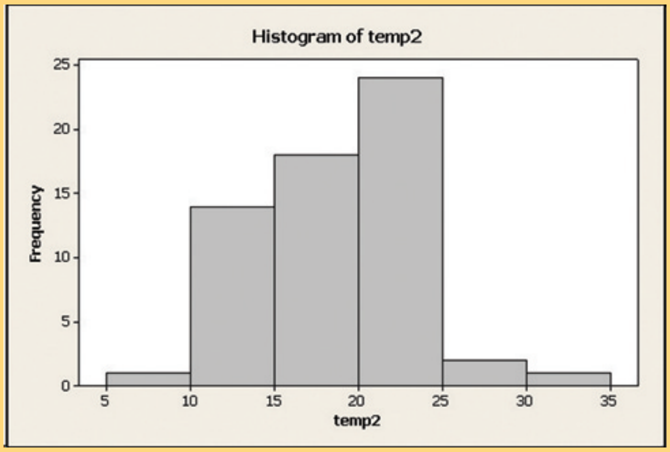


Figure 2.26 Histogram of Data Temp 2.

3. Does the fact that the shapes of the two distributions are the same imply that the centers of the two distributions are the same? Why or why not? Explain.

4. Does the fact that the shapes of the two distributions are the same imply that the spreads of the two distributions are the same? Why or why not? Explain.

5. It turns out that the same variable was represented in the two graphs, but with different units of measurement. Can you figure out the units?

Another situation that is important to compare is when two graphs cover a similar range but have different shapes, such as the histograms in Figures 2.27 and 2.28.

1. Examine the two histograms of Figures 2.27 and 2.28.

2. These two distributions have the same center, but do not have the same spread. Decide which distribution has the larger spread and explain the reasoning behind your decision.

Answer all the above questions again after reading Chapter 3.

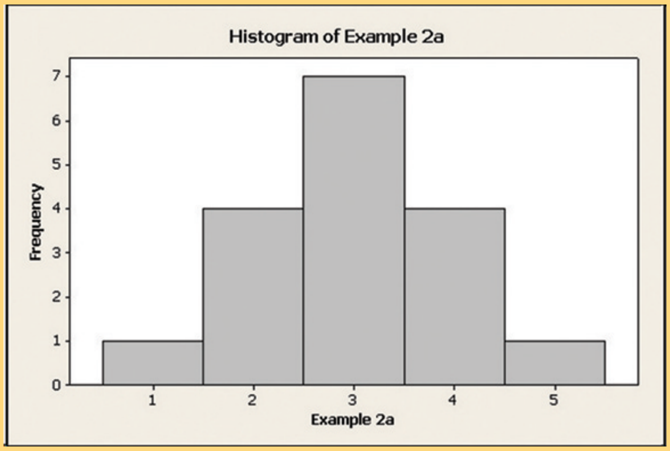


Figure 2.27 Histogram of Example 2a.

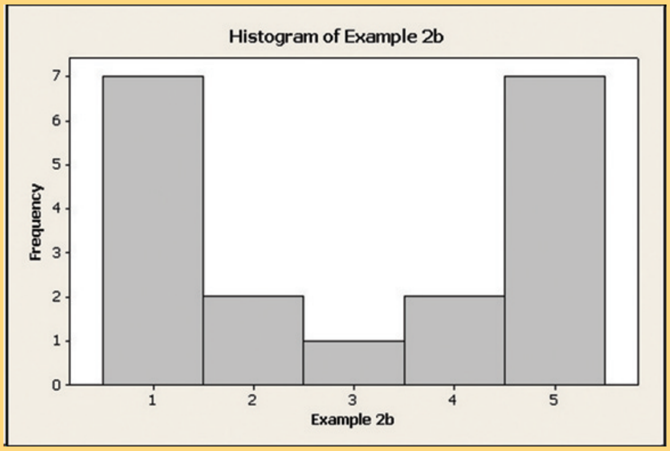
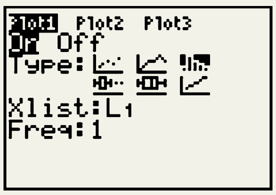


Figure 2.28 Histogram of Example 2b.

TECHNOLOGY INSTRUCTION — Organizing Data

TI-84



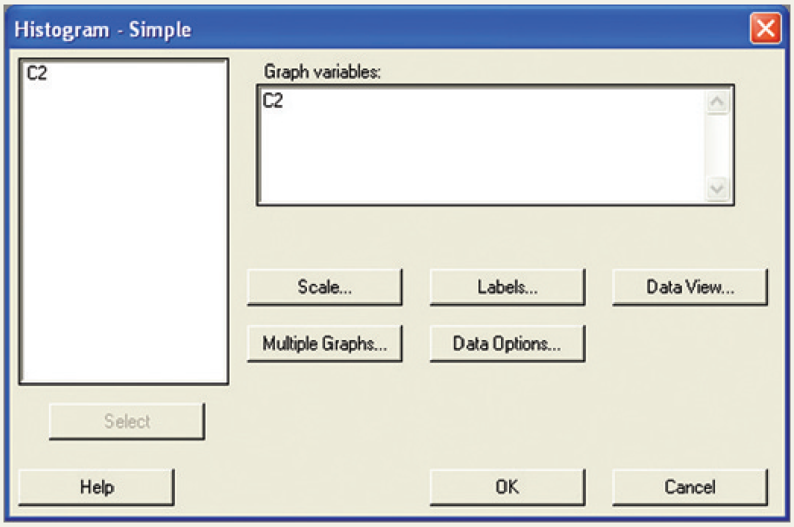
Screen 2.1

1. To create a frequency histogram for a list of data, press **STAT PLOT**, which is accessed by pressing **2nd > Y=**. The **Y=** key is located at the top left on the calculator.
2. Make sure that only one plot is turned on. If more than one plot is turned on, turn off the unwanted plots by using the following steps. Press the number corresponding to the plot you wish to turn off. A screen similar to **Screen 2.1** will appear. Use the arrow keys to move the cursor to the **Off** button, then press **Enter**. Now use the arrow keys to move to the row with **Plot1**, **Plot2**, and **Plot3**. If there is another plot that you need to turn off, select that plot by moving the cursor to it, press **Enter**, and repeat the previous procedure. If there is no plot that you need to turn off, move the cursor to the plot you want to use and press **Enter**.
3. In the **Type** rows, use the right arrow to move to the third column in the first row, which looks like a histogram, and press **Enter**. Move to **Xlist** to enter the name of the list where the data are located. Press **2nd>Stat**, then use the up and down arrows to move through the list names until you find the list you want to use. Press **Enter**. Leave the **Freq** setting at 1.
4. To see the graph, select **ZOOM > 9**, where **ZOOM** is the third key in the top row. This sets the window settings to display your graph.
5. If you would like to change the class width, select **WINDOW**, which is the second key in the top row. Change the value of **Xscl** to the desired width, then press **GRAPH**, which is the fifth key in the top row.
6. If you would like to see the interval endpoints and the number of observations in each class (which is given by the height of the corresponding bar), press **TRACE**, then use the left and right arrows to move between bars. When you are done, press **CLEAR**.

MINITAB

1. To create a bar graph for categorical data entered in column C1, select **Graph>Bar Chart**. In the dialog box obtained, select **Counts of unique values** and **Simple** and click **OK**. In the new dialog box you obtain, type **C1** in the box below **Categorical Variables** and click **OK**. This will produce a bar graph for the data.
If you have categorical data in a frequency table with categories entered in column C1 and frequencies in column C2, select **Graph>Bar Chart**. In the dialog box obtained, select **Values from a table** from choices below **Bars represent**, select **Simple** from graphs, and click **OK**. In the new dialog box you obtain, type **C2** in the box below **Graph variables** and **C1** in the box below **Categorical variable**. Click **OK**. This will produce a bar graph for the data.

2. To create a pie chart for categorical data, you can use raw data or a frequency table. After you enter data in the MINITAB worksheet, select **Graph>Pie Chart**. In the dialog box obtained, select either **Chart raw data** or **Chart values from a table**, then fill in the required column names, and click **OK**. This will produce a pie chart for the data.

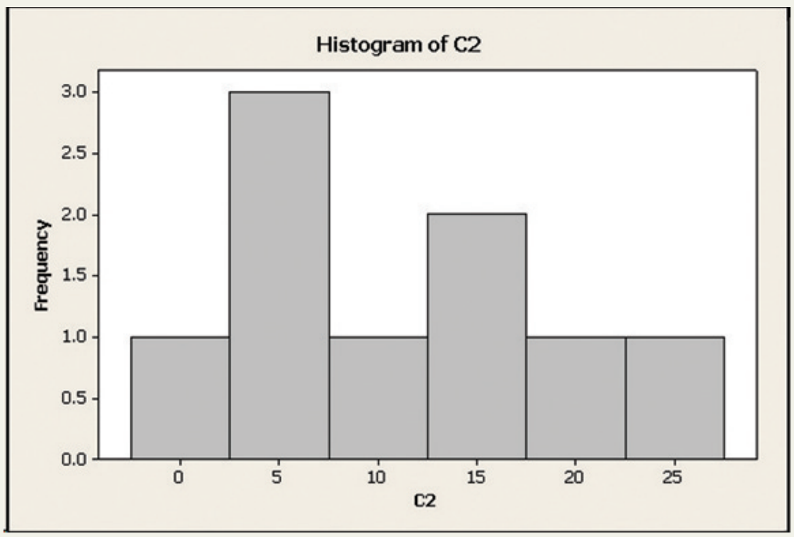


Screen 2.2

3. To create a frequency histogram for a quantitative data set entered in column C2, select **Graph>Histogram**, then select **Simple**, and click **OK**. In the dialog box you obtain, enter the name of your column, e.g., **C2** (see **Screen 2.2**) in the box below **Graph variables** and click **OK** to create the histogram. MINITAB will produce **Screen 2.3** with histogram.

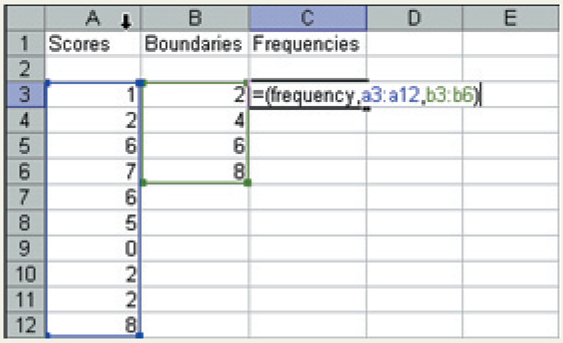
4. To create a stem-and-leaf display for a quantitative data set entered in column C1, select **Graph>Stem-and-Leaf**, enter the name of the column that contains data in the box below **Graph variables** and click **OK** to create the stem-and-leaf display.

5. To create a dotplot for a quantitative data set entered in column C1, select **Graph>Dotplot**, then select the appropriate dotplot graph from the dialog box and click **OK**. In the new dialog box you obtain, enter the name of your column, e.g., **C1** in the box below **Graph variables**, and click **OK** to create the dotplot.



Screen 2.3

Excel



Screen 2.4

1. To create a frequency distribution for a range of numerical data in Excel, decide how many categories you will have. Choose class boundaries between the categories so that you have one fewer boundary than classes. Type the class boundaries into Excel.

2. Select where you want the class frequencies to appear, and select a range of one more cell than the number of boundaries you have.

3. Type **=frequency**(.

4. Select the range of cells of numerical data, and then type a comma.

5. Select the range of class boundaries, and then type a right parenthesis. (See Screen 2.4.)

6. Press **Control-Shift-Enter**. The frequencies should appear.

7. If you would prefer relative frequencies, replace Steps 5 and 6 by the following:
   - **a.** Select the range of class boundaries, then type **)/count**(.
   - **b.** Select the range of cells of numerical data, then type a right parenthesis.
   - **c.** Press **Control-Shift-Enter**. The relative frequencies should appear.

8. To plot frequencies as bar charts, pie charts, and so on for 1, select **Insert>Chart** and follow the instructions in the Chart Wizard.

TECHNOLOGY ASSIGNMENTS

TA2.1 Construct a bar graph and a pie chart for the frequency distribution prepared in Exercise 2.5.

TA2.2 Construct a bar graph and a pie chart for the frequency distribution prepared in Exercise 2.6.

TA2.3 Refer to Data Set V that accompanies this text (see Preface and Appendix B) on the times taken to run the Manchester Road Race for a sample of 500 participants. From that data set, select the 6th value and then select every 10th value after that (i.e., select the 6th, 16th, 26th, 36th . . . values). This subsample will give you 50 measurements. (Such a sample selected from a population is called a *systematic random sample*.) Construct a histogram for these data. Let the software you use decide on classes and class limits.

TA2.4 Refer to Data Set I that accompanies this text on the prices of various products in different cities across the country. Select a subsample of 60 from the column that contains information on telephone charges and then construct a histogram for these data.

TA2.5 Construct a histogram for the data from Exercise 2.20 on the numbers of computer keyboards assembled. Use the classes given in that exercise. Use the midpoints to mark the horizontal axis in the histogram.

TA2.6 Prepare a stem-and-leaf display for the data given in Exercise 2.48.

TA2.7 Prepare a stem-and-leaf display for the data of Exercise 2.53.

TA2.8 Prepare a bar graph for the frequency distribution obtained in Exercise 2.28.

TA2.9 Prepare a bar graph for the frequency distribution obtained in Exercise 2.29.

TA2.10 Make a pie chart for the frequency distribution obtained in Exercise 2.19.

TA2.11 Make a pie chart for the frequency distribution obtained in Exercise 2.29.

TA2.12 Make a dotplot for the data of Exercise 2.64.

TA2.13 Make a dotplot for the data of Exercise 2.65.