CHAPTER 2

Describing Data Sets

Numbers constitute the only universal language.

Nathaniel West

People who don't count won't count.

Anatole France

In this chapter we learn methods for presenting and describing sets of data. We introduce different types of tables and graphs, which enable us to easily see key features of a data set.

CONTENTS

2.1 INTRODUCTION

It is very important that the numerical findings of any study be presented clearly and concisely and in a manner that enables one to quickly obtain a feel for the essential characteristics of the data. This is particularly needed when the set of data is large, as is frequently the case in surveys or controlled experiments. Indeed, an effective presentation of the data often quickly reveals important features such as their range, degree of symmetry, how concentrated or spread out they are, where they are concentrated, and so on. In this chapter we will be concerned with techniques, both tabular and graphic, for presenting data sets.

Frequency tables and frequency graphs are presented in Sec. 2.2. These include a variety of tables and graphs—line graphs, bar graphs, and polygon graphs—that are useful for describing data sets having a relatively small number of distinct values. As the number of distinct values becomes too large for these forms to be effective, it is useful to break up the data into disjoint classes and consider the number of data values that fall in each class. This is done in Sec. 2.3, where we study the histogram, a bar graph that results from graphing class frequencies. A variation of the histogram, called a stem-and-leaf plot, which uses the actual data values to represent the size of a class, is studied in Sec. 2.4. In Sec. 2.5 we consider the situation where the data consist of paired

Introductory Statistics. DOI:10.1016/B978-0-12-804317-2.00002-3

Table 2.1 A Frequency Table of Sick Leave Data

Value	Frequency	Value	Frequency
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

values, such as the population and the crime rate of various cities, and introduce the scatter diagram as an effective way of presenting such data. Some historical comments are presented in Sec. 2.6.

2.2 FREQUENCY TABLES AND GRAPHS

The following data represent the number of days of sick leave taken by each of 50 workers of a given company over the last 6 weeks:

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

Since this data set contains only a relatively small number of distinct, or different, values, it is convenient to represent it in a *frequency table*, which presents each distinct value along with its frequency of occurrence. Table 2.1 is a frequency table of the preceding data. In Table 2.1 the frequency column represents the number of occurrences of each distinct value in the data set. Note that the sum of all the frequencies is 50, the total number of data observations.

Example 2.1

Use Table 2.1 to answer the following questions:

(a) How many workers had at least 1 day of sick leave?
(b) How many workers had between 3 and 5 days of sick leave?
(c) How many workers had more than 5 days of sick leave?

Solutions

(a) Since 12 of the 50 workers had no days of sick leave, the answer is $50 - 12 = 38$.
(b) The answer is the sum of the frequencies for values 3, 4, and 5; that is, $4 + 5 + 8 = 17$.
(c) The answer is the sum of the frequencies for the values 6, 7, 8, and 9. Therefore, the answer is $0 + 5 + 2 + 1 = 8$.

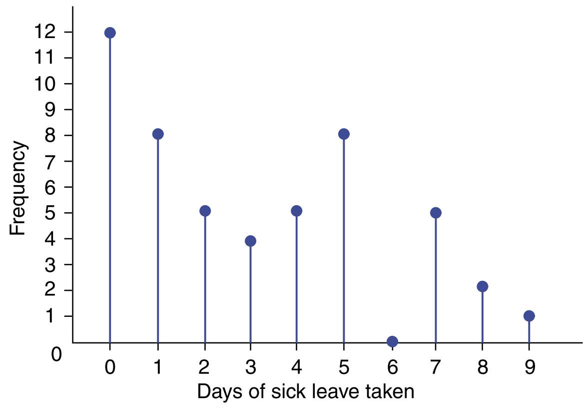


FIGURE 2.1 A line graph.

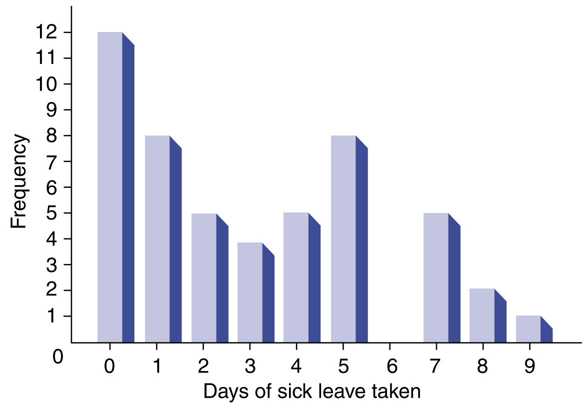


FIGURE 2.2 A bar graph.

2.2.1 Line Graphs, Bar Graphs, and Frequency Polygons

Data from a frequency table can be graphically pictured by a *line graph*, which plots the successive values on the horizontal axis and indicates the corresponding frequency by the height of a vertical line. A line graph for the data of Table 2.1 is shown in Fig. 2.1.

Sometimes the frequencies are represented not by lines but rather by bars having some thickness. These graphs, called *bar graphs*, are often utilized. Figure 2.2 presents a bar graph for the data of Table 2.1.

Another type of graph used to represent a frequency table is the *frequency polygon*, which plots the frequencies of the different data values and then connects the plotted points with straight lines. Figure 2.3 presents the frequency polygon of the data of Table 2.1.

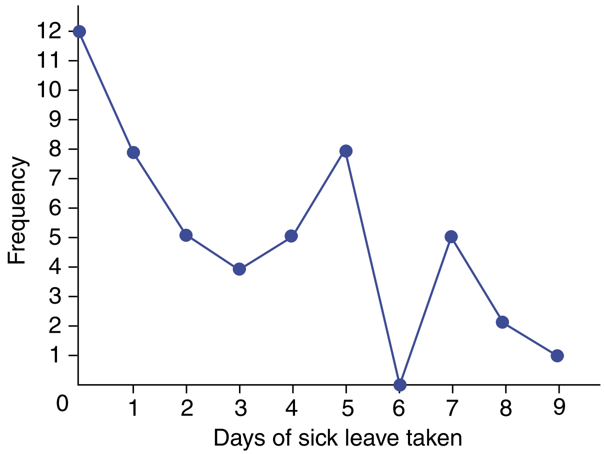


FIGURE 2.3 A frequency polygon.

Table 2.2 Frequency Table of a Symmetric Data Set

Value	Frequency	Value	Frequency
0	1	4	2
2	2	6	1
3	3	0	0

A set of data is said to be *symmetric* about the value x_0 if the frequencies of the values $x_0 - c$ and $x_0 + c$ are the same for all c. That is, for every constant c, there are just as many data points that are c less than x_0 as there are that are c greater than x_0. The data set presented in Table 2.2, a frequency table, is symmetric about the value $x_0 = 3$.

Data that are "close to" being symmetric are said to be *approximately symmetric*. The easiest way to determine whether a data set is approximately symmetric is to represent it graphically. Figure 2.4 presents three bar graphs: one of a symmetric data set, one of an approximately symmetric data set, and one of a data set that exhibits no symmetry.

2.2.2 Relative Frequency Graphs

It is sometimes convenient to consider and plot the *relative* rather than the absolute frequencies of the data values. If f represents the frequency of occurrence of some data value x, then the *relative frequency* f/n can be plotted versus x, where n represents the total number of observations in the data set. For the data of Table 2.1, $n = 50$ and so the relative frequencies are as given in Table 2.3. Note that whereas the sum of the frequency column should be the total number of observations in the data set, the sum of the relative frequency column should be 1.

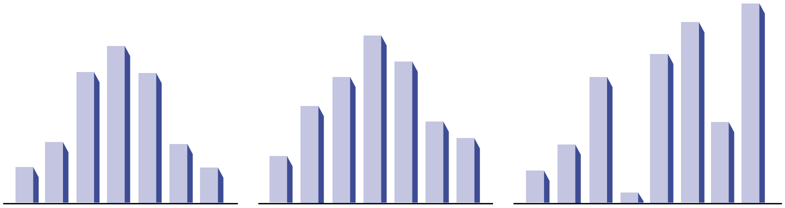


FIGURE 2.4 Bar graphs and symmetry.

Table 2.3 Relative Frequencies, $n = 50$, of Sick Leave Data

Value x	Frequency f	Relative frequency f/n
0	12	$\frac{12}{50} = 0.24$
1	8	$\frac{8}{50} = 0.16$
2	5	$\frac{5}{50} = 0.10$
3	4	$\frac{4}{50} = 0.08$
4	5	$\frac{5}{50} = 0.10$
5	8	$\frac{8}{50} = 0.16$
6	0	$\frac{0}{50} = 0.00$
7	5	$\frac{5}{50} = 0.10$
8	2	$\frac{2}{50} = 0.04$
9	1	$\frac{1}{50} = 0.02$

A polygon plot of these relative frequencies is presented in Fig. 2.5. A plot of the relative frequencies looks exactly like a plot of the absolute frequencies, except that the labels on the vertical axis are the old labels divided by the total number of observations in the data set.

To Construct a Relative Frequency Table from a Data Set

Arrange the data set in increasing order of values. Determine the distinct values and how often they occur. List these distinct values alongside their frequencies f and their relative frequencies f/n, where n is the total number of observations in the data set.

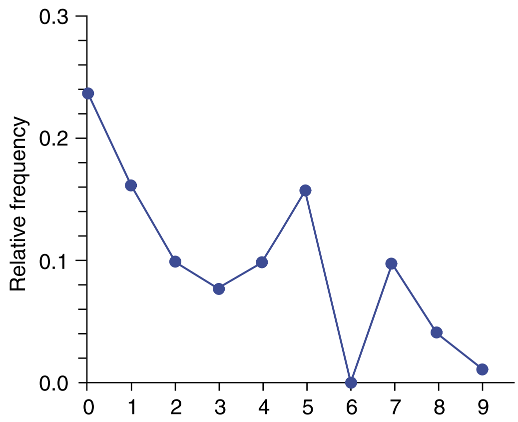


FIGURE 2.5 A relative frequency polygon.

Example 2.2

The Masters Golf Tournament is played each year at the Augusta National Golf Club in Augusta, Georgia. To discover what type of score it takes to win this tournament, we have gathered all the winning scores from 1980 to 2016.

The Masters Golf Tournament Winners

Year	Winner	Score	Year	Winner	Score
1980	Severiano Ballesteros	275	1999	J.M. Olazabal	280
1981	Tom Watson	280	2000	Vijay Singh	278
1982	Craig Stadler	284	2001	Tiger Woods	272
1983	Severiano Ballesteros	280	2002	Tiger Woods	276
1984	Ben Crenshaw	277	2003	Mike Weir	281
1985	Bernhard Langer	282	2004	Phil Mickelson	279
1986	Jack Nicklaus	279	2005	Tiger Woods	276
1987	Larry Mize	285	2006	Phil Mickelson	281
1988	Sandy Lyle	281	2007	Zach Johnson	289
1989	Nick Faldo	283	2008	Trevor Immelman	280
1990	Nick Faldo	278	2009	Angel Cabrera	276
1991	Ian Woosnam	277	2010	Phil Mickelson	272
1992	Fred Couples	275	2011	Charl Schwartzel	274
1993	Bernhard Langer	277	2012	Bubba Watson	278
1994	J.M. Olazabal	279	2013	Adam Scott	279
1995	Ben Crenshaw	274	2014	Bubba Watson	280
1996	Nick Faldo	276	2015	Jordan Spieth	270
1997	Tiger Woods	270	2016	Danny Willett	283
1998	Mark O'Meara	279			

(a) Arrange the data set of winning scores in a relative frequency table.
(b) Plot these data in a relative frequency bar graph.

Solution

(a) The 37 winning scores range from a low of 270 to a high of 289. This is the relative frequency table:

Winning score	Frequency f	Relative frequency $f/37$
270	2	0.054
271	0	0.000
272	2	0.054
274	2	0.054
275	2	0.054
276	4	0.108
277	3	0.081
278	3	0.081
279	5	0.135
280	5	0.135
281	3	0.081
282	1	0.027
283	2	0.054
284	1	0.027
285	1	0.027
289	1	0.027

(b) The following is a relative frequency bar graph of the preceding data.

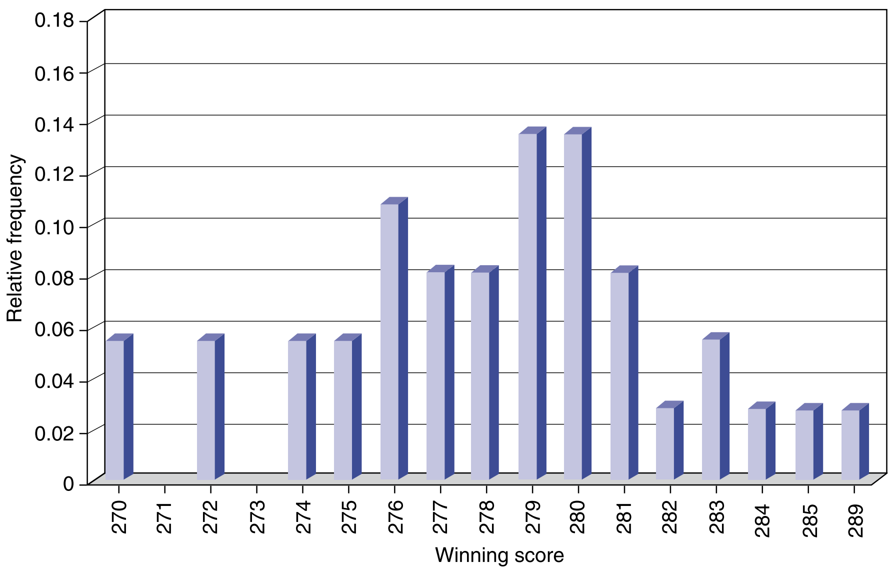


Table 2.4 Murder Weapons

Type of weapon	Percentage of murders caused by this weapon
Handgun	52
Knife	18
Shotgun	7
Rifle	4
Personal weapon	6
Other	13

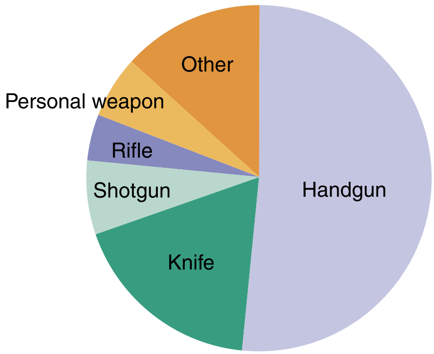


FIGURE 2.6 A pie chart.

2.2.3 Pie Charts

A *pie chart* is often used to plot relative frequencies when the data are nonnumeric. A circle is constructed and then is sliced up into distinct sectors, one for each different data value. The area of each sector, which is meant to represent the relative frequency of the value that the sector represents, is determined as follows. If the relative frequency of the data value is f/n, then the area of the sector is the fraction f/n of the total area of the circle. For instance, the data in Table 2.4 give the relative frequencies of types of weapons used in murders in a large midwestern city in 1985. These data are represented in a pie chart in Fig. 2.6.

If a data value has relative frequency f/n, then its sector can be obtained by setting the angle at which the lines of the sector meet equal to 360 f/n degrees. For instance, in Fig. 2.6, the angle of the lines forming the knife sector is $360(0.18) = 64.8°$.

PROBLEMS

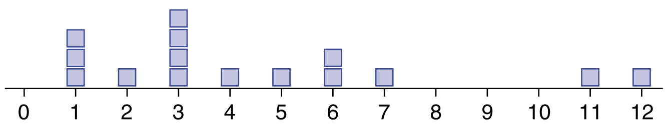
1. The following data represent the sizes of 30 families that reside in a small town in Guatemala:

   5, 13, 9, 12, 7, 4, 8, 6, 6, 10, 7, 11, 10, 8, 15,
   8, 6, 9, 12, 10, 7, 11, 10, 8, 12, 9, 7, 10, 7, 8

   (a) Construct a frequency table for these data.
   (b) Using a line graph, plot the data.
   (c) Plot the data as a frequency polygon.
2. The following frequency table relates the weekly sales of bicycles at a given store over a 42-week period.

Value	0	1	2	3	4	5	6	7
Frequency	3	6	7	10	8	5	2	1

   (a) In how many weeks were at least 2 bikes sold?
   (b) In how many weeks were at least 5 bikes sold?
   (c) In how many weeks were an even number of bikes sold?
3. Fifteen fourth-graders were asked how many blocks they lived from school. The results are displayed in the following graph.



   (a) What is the maximum number of blocks any student lives from school?
   (b) What is the minimum number of blocks?
   (c) How many students live less than 5 blocks from school?
   (d) How many students live more than 4 blocks from school?
4. Label each of the following data sets as symmetric, approximately symmetric, or not at all symmetric.

   A: 6, 0, 2, 1, 8, 3, 5
   B: 4, 0, 4, 0, 2, 1, 3, 2
   C: 1, 1, 0, 1, 0, 3, 3, 2, 2, 2
   D: 9, 9, 1, 2, 3, 9, 8, 4, 5

5. The following table lists all the values but only some of the frequencies for a symmetric data set. Fill in the missing numbers.

Value	Frequency
10	8
20	
30	7
40	
50	3
60	

6. The following are the scores of 32 students who took a statistics test:

   55, 70, 80, 75, 90, 80, 60, 100, 95, 70, 75, 85, 80, 80, 70, 95,
   100, 80, 85, 70, 85, 90, 80, 75, 85, 70, 90, 60, 80, 70, 85, 80

   Represent this data set in a frequency table, and then draw a bar graph.
7. Draw a relative frequency table for the data of Prob. 1. Plot these relative frequencies in a line graph.
8. The following data represent the time to tumor progression, measured in months, for 65 patients having a particular type of brain tumor called *glioblastoma*:

   6, 5, 37, 10, 22, 9, 2, 16, 3, 3, 11, 9, 5, 14, 11, 3, 1, 4, 6, 2, 7,
   3, 7, 5, 4, 8, 2, 7, 13, 16, 15, 9, 4, 4, 2, 3, 9, 5, 11, 3, 7, 5, 9,
   3, 8, 9, 4, 10, 3, 2, 7, 6, 9, 3, 5, 4, 6, 4, 14, 3, 12, 6, 8, 12, 7

   (a) Make up a relative frequency table for this data set.
   (b) Plot the relative frequencies in a frequency polygon.
   (c) Is this data set approximately symmetric?
9. The following relative frequency table is obtained from a data set of the number of emergency appendectomies performed each month at a certain hospital.

Value	0	1	2	3	4	5	6	7
Relative frequency	0.05	0.08	0.12	0.14	0.16	0.20	0.15	0.10

   (a) What proportion of months has fewer than 2 emergency appendectomies?
   (b) What proportion of months has more than 5?
   (c) Is this data set symmetric?
10. Relative frequency tables and plots are particularly useful when we want to compare different sets of data. The following two data sets relate the number of months from diagnosis to death of AIDS patients for samples of male and female AIDS sufferers in the early years of the epidemic.

Males	15	13	16	10	8	20	14	19	9	12	16	18	20	12	14	14
Females	8	12	10	8	14	12	13	11	9	8	9	10	14	9	10	

   Plot these two data sets together in a relative frequency polygon. Use a different color for each set. What conclusion can you draw about which data set tends to have larger values?

Average Number of Days with Precipitation of 0.01 Inch or More

State	City	Length of record (yr.)	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual
AL	Mobile	46	11	10	11	7	8	11	16	14	10	6	8	10	123
AK	Juneau	43	18	17	18	17	17	16	17	18	20	24	19	21	220
AZ	Phoenix	48	4	4	4	2	1	1	4	5	3	3	3	4	36
AR	Little Rock	45	9	9	10	10	10	8	8	7	7	7	8	9	103
CA	Los Angeles	52	6	6	6	3	1	1	1	0	1	2	4	5	36
	Sacramento	48	10	9	9	5	3	1	0	0	1	3	7	9	58
	San Diego	47	7	6	7	5	2	1	0	1	1	3	5	6	43
	San Francisco	60	11	10	10	6	3	1	0	0	1	4	7	10	62
CO	Denver	53	6	6	9	9	11	9	9	9	6	5	5	5	89
CT	Hartford	33	11	10	11	11	12	11	10	10	9	8	11	12	127
DE	Wilmington	40	11	10	11	11	11	10	9	9	8	8	10	10	117
DC	Washington	46	10	9	11	10	11	10	10	9	8	7	8	9	111
FL	Jacksonville	46	8	8	8	6	8	12	15	14	13	9	6	8	116
	Miami	45	6	6	6	6	10	15	16	17	17	14	9	7	129
GA	Atlanta	53	11	10	11	9	9	10	12	9	8	6	8	10	115
HI	Honolulu	38	10	9	9	9	7	6	8	6	7	9	9	10	100
ID	Boise	48	12	10	10	8	8	6	2	3	4	6	10	11	91
IL	Chicago	29	11	10	12	12	11	10	10	9	10	9	10	12	127
	Peoria	48	9	8	11	12	11	10	9	8	9	8	9	10	114
IN	Indianapolis	48	12	10	13	12	12	10	9	9	8	8	10	12	125
IA	Des Moines	48	7	7	10	11	11	11	9	9	9	8	7	8	107
KS	Wichita	34	6	5	8	8	11	9	7	8	8	6	5	6	86
KY	Louisville	40	11	11	13	12	12	10	11	8	8	8	10	11	125
LA	New Orleans	39	10	9	9	7	8	11	15	13	10	6	7	10	114
ME	Portland	47	11	10	11	12	13	11	10	9	8	9	12	12	128
MD	Baltimore	37	10	9	11	11	11	9	9	10	7	7	9	9	113
MA	Boston	36	12	10	12	11	12	11	9	10	9	9	11	12	126

(*Continued*)

(*Continued*)

State	City	Length of record (yr.)	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual
MI	Detroit	29	13	11	13	12	11	11	9	9	10	9	12	14	135
	Sault Ste. Marie	46	19	15	13	11	11	12	10	11	13	13	17	20	165
MN	Duluth	46	12	10	11	10	12	13	11	11	12	10	11	12	134
	Minneapolis-St. Paul	49	9	7	10	10	11	12	10	10	10	8	8	9	115
MS	Jackson	24	11	9	10	8	10	8	10	10	8	6	8	10	109
MO	Kansas City	15	7	7	11	11	11	11	7	9	8	8	8	8	107
	St. Louis	30	8	8	11	11	11	10	8	8	8	8	10	9	111
MT	Great Falls	50	9	8	9	9	12	12	7	8	7	6	7	8	101
NE	Omaha	51	6	7	9	10	12	11	9	9	9	7	5	6	98
NV	Reno	45	6	6	6	4	4	3	2	2	2	3	5	6	51
NH	Concord	46	11	10	11	12	12	11	10	10	9	9	11	11	125
NJ	Atlantic City	44	11	10	11	11	10	9	9	9	8	7	9	10	112
NM	Albuquerque	48	4	4	5	3	4	4	9	9	6	5	3	4	61
NY	Albany	41	12	10	12	12	13	11	10	10	10	9	12	12	134
	Buffalo	44	20	17	16	14	12	10	10	11	11	12	16	20	169
	New York	118	11	10	11	11	11	10	10	10	8	8	9	10	121
NC	Charlotte	48	10	10	11	9	10	10	11	9	7	7	8	10	111
	Raleigh	43	10	10	10	9	10	9	11	10	8	7	8	9	111
ND	Bismarck	48	8	7	8	8	10	12	9	9	7	6	6	8	97
OH	Cincinnati	40	12	11	13	13	11	11	10	9	8	8	11	12	129
	Cleveland	46	16	14	15	14	13	11	10	10	10	11	14	16	156
	Columbus	48	13	12	14	13	13	11	11	9	8	9	11	13	137
OK	Oklahoma City	48	5	6	7	8	10	9	6	6	7	6	5	5	82
OR	Portland	47	18	16	17	14	12	9	4	5	8	13	18	19	152
PA	Philadelphia	47	11	9	11	11	11	10	9	9	8	8	9	10	117
	Pittsburgh	35	16	14	16	14	12	12	11	10	9	11	13	17	154
RI	Providence	34	11	10	12	11	11	11	9	10	8	8	11	12	124

(*Continued*)

State	City	Length of record (yr.)	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual
SC	Columbia	40	10	10	11	8	9	9	12	11	8	6	7	9	109
SD	Sioux Falls	42	6	6	9	9	10	11	9	9	8	6	6	6	97
TN	Memphis	37	10	9	11	10	9	8	9	8	7	6	9	10	106
	Nashville	46	11	11	12	11	11	9	10	9	8	7	10	11	119
TX	Dallas-Fort Worth	34	7	7	7	8	9	6	5	5	7	6	6	6	78
	El Paso	48	4	3	2	2	2	4	8	8	5	4	3	4	48
	Houston	18	10	8	9	7	9	9	9	10	10	8	9	9	106
UT	Salt Lake City	59	10	9	10	9	8	5	5	6	5	6	8	9	91
VT	Burlington	44	14	12	13	12	14	13	12	12	12	12	14	15	154
VA	Norfolk	39	10	10	11	10	10	9	11	10	8	8	8	9	114
	Richmond	50	10	9	11	9	11	9	11	10	8	7	8	9	113
WA	Seattle	43	19	16	17	14	10	9	5	6	9	13	18	20	156
	Spokane	40	14	12	11	9	9	8	4	5	6	8	12	15	113
WV	Charleston	40	16	14	15	14	13	11	13	11	9	10	12	14	151
WI	Milwaukee	47	11	10	12	12	12	11	10	9	9	9	10	11	125
WY	Cheyenne	52	6	6	9	10	12	11	11	10	7	6	6	5	99
PR	San Juan	32	16	13	12	13	17	16	19	18	17	17	18	19	195

Source: U.S. National Oceanic and Atmospheric Administration, *Comparative Climatic Data*.

11. Using the data of Example 2.2, determine the proportion of winning scores in the Masters Golf Tournament that is
    (a) Below 280
    (b) 282 or higher
    (c) Between 278 and 284 inclusive

The table on the following three pages gives the average number of days in each month that various cities have at least 0.01 inch of precipitation. Problems 12 through 14 refer to it.

12. Construct a relative frequency table for the average number of rainy days in January for the different cities. Then plot the data in a relative frequency polygon.
13. Using only the data relating to the first 12 cities listed, construct a frequency table for the average number of rainy days in either November or December.
14. Using only the data relating to the first 24 cities, construct relative frequency tables for the month of June and separately for the month of December. Then plot these two sets of data together in a relative frequency polygon.
15. The following table gives the number of deaths on British roads in 1987 for individuals in various classifications.

Classification	**Number of deaths**
Pedestrians	1699
Bicyclists	280
Motorcyclists	650
Automobile drivers	1327

Express this data set in a pie chart.

16. The following data, taken from *The New York Times*, represent the percentage of items, by total weight, in the garbage of New York City. Represent them in a pie chart.

Organic material (food, yard waste, lumber, etc.)	37.3
Paper	30.8
Bulk (furniture, refrigerators, etc.)	10.9
Plastic	8.5
Glass	5
Metal	4
Inorganic	2.2
Aluminum	0.9
Hazardous waste	0.4

17. The following give the winning scores of the Masters Golf tournament from 2005 through 2009. Use them in conjunction with data given in Example 2.2 to obtain a relative frequency table of all winning scores from 1990 to 2009. Also, use the data given in Example 2.2 to obtain a relative fre-

quency table of all winning scores from 1970 to 1989. Do winning scores appear to have changed much over the past 20 years?

Year	Winner	Score
2005	Tiger Woods	276
2006	Phil Mickelson	281
2007	Zach Johnson	289
2008	Trevor Immelman	280
2009	Angel Cabrera	276

2.3 GROUPED DATA AND HISTOGRAMS

As seen in Sec. 2.2, using a line or a bar graph to plot the frequencies of data values is often an effective way of portraying a data set. However, for some data sets the number of distinct values is too large to utilize this approach. Instead, in such cases, we divide the values into groupings, or *class intervals*, and then plot the number of data values falling in each class interval. The number of class intervals chosen should be a trade-off between (1) choosing too few classes at a cost of losing too much information about the actual data values in a class and (2) choosing too many classes, which will result in the frequencies of each class being too small for a pattern to be discernible. Although 5 to 10 class intervals are typical, the appropriate number is a subjective choice, and of course you can try different numbers of class intervals to see which of the resulting charts appears to be most revealing about the data. It is common, although not essential, to choose class intervals of equal length.

The endpoints of a class interval are called the *class boundaries*. We will adopt the *left-end inclusion convention*, which stipulates that a class interval contains its left-end but not its right-end boundary point. Thus, for instance, the class interval 20–30 contains all values that are both greater than *or equal to* 20 and less than 30.

The data in Table 2.5 represent the blood cholesterol levels of 40 first-year students at a particular college. As a prelude to determining class size frequencies, it is useful to rearrange the data in increasing order. This gives the 40 values of Table 2.6.

Table 2.5 Blood Cholesterol Levels

213	174	193	196	220	183	194	200
192	200	200	199	178	183	188	193
187	181	193	205	196	211	202	213
216	206	195	191	171	194	184	191
221	212	221	204	204	191	183	227

Table 2.6 Blood Cholesterol Levels in Increasing Order

171,	174,	178,	181,	183,	183,	183,	184,	187,	188,	191,	191,	191,	192,	193,	193,	193,	194,	194,	195,
196,	196,	199,	200,	200,	200,	202,	204,	204,	205,	206,	211,	212,	213,	213,	216,	220,	221,	221,	227

Table 2.7 Frequency Table of Blood Cholesterol Levels

Class intervals	Frequency	Relative frequency
170–180	3	$\frac{3}{40} = 0.075$
180–190	7	$\frac{7}{40} = 0.175$
190–200	13	$\frac{13}{40} = 0.325$
200–210	8	$\frac{8}{40} = 0.20$
210–220	5	$\frac{5}{40} = 0.125$
220–230	4	$\frac{4}{40} = 0.10$

Since the data range from a minimum value of 171 to a maximum of 227, the left-end boundary of the first class interval must be less than or equal to 171, and the right-end boundary of the final class interval must be greater than 227. One choice would be to have the first class interval be 170 to 180. This will result in six class intervals. A frequency table giving the frequency (as well as the relative frequency) of data values falling in each class interval is seen in Table 2.7.

Note: Because of the left-end inclusion convention, the values of 200 were placed in the class interval of 200 to 210, not in the interval of 190 to 200.

A bar graph plot of the data, with the bars placed adjacent to each other, is called a *histogram*. The vertical axis of a histogram can represent either the class frequency or the relative class frequency. In the former case, the histogram is called a *frequency histogram* and in the latter a *relative frequency histogram*. Figure 2.7 presents a frequency histogram of the data of Table 2.7.

It is important to recognize that a class frequency table or a histogram based on that table does not contain all the information in the original data set. These two representations note only the *number* of data values in each class and not the actual data values themselves. Thus, whereas such tables and charts are useful for illustrating data, the original raw data set should *always* be saved.

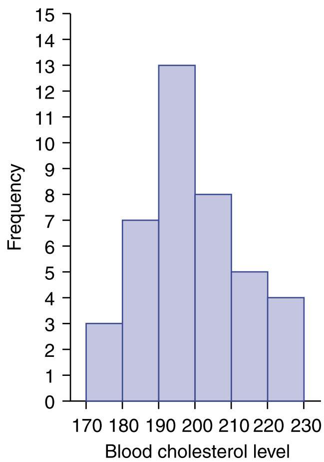


FIGURE 2.7 Frequency histogram for the data of Table 2.7.

To Construct a Histogram from a Data Set

1. Arrange the data in increasing order.
2. Choose class intervals so that all data points are covered.
3. Construct a frequency table.
4. Draw adjacent bars having heights determined by the frequencies in step 3.

The importance of a histogram is that it enables us to organize and present data graphically so as to draw attention to certain important features of the data. For instance, a histogram can often indicate

1. How symmetric the data are
2. How spread out the data are
3. Whether there are intervals having high levels of data concentration
4. Whether there are gaps in the data
5. Whether some data values are far apart from others

For instance, the histogram presented in Fig. 2.7 indicates that the frequencies of the successive classes first increase and then decrease, reaching a maximum in the class having limits of 190 to 200. The histograms of Fig. 2.8 give valuable information about the data sets they represent. The data set whose histogram is on the left side of Fig. 2.8(a) is symmetric, whereas the one on the right side is not. The data set represented on the left side of Fig. 2.8(b) is fairly evenly spread out, whereas the one for the right side is more concentrated. The data set represented by the left side of Fig. 2.8(c) has a gap, whereas the one represented on the right side has certain values far apart from the rest.

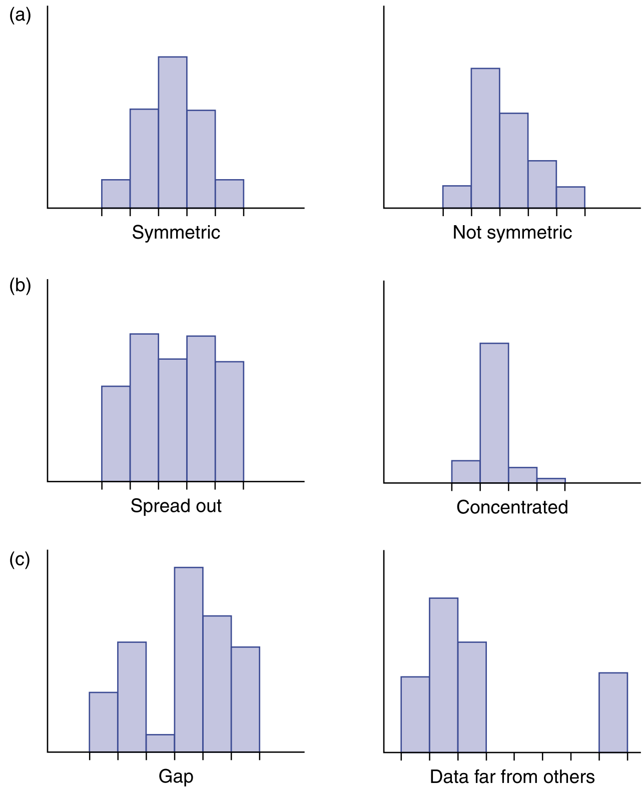


FIGURE 2.8 Characteristics of data detected by histograms. (a) symmetry, (b) degree of spread and where values are concentrated, and (c) gaps in data and data far from others.

■ Example 2.3

Table 2.8 gives the birth rates (per 1000 population) in each of the 50 states of the United States. Plot these data in a histogram.

Solution

Since the data range from a low value of 12.4 to a high of 21.9, let us use class intervals of length 1.5, starting at the value 12. With these class intervals, we obtain the following frequency table.

Class intervals	Frequency	Class intervals	Frequency
12.0–13.5	2	18.0–19.5	2
13.5–15.0	15	19.5–21.0	0
15.0–16.5	22	21.0–22.5	2
16.5–18.0	7		

Table 2.8 Birth Rates per 1000 Population

State	Rate	State	Rate	State	Rate
Alabama	14.2	Louisiana	15.7	Ohio	14.9
Alaska	21.9	Maine	13.8	Oklahoma	14.4
Arizona	19.0	Maryland	14.4	Oregon	15.5
Arkansas	14.5	Massachusetts	16.3	Pennsylvania	14.1
California	19.2	Michigan	15.4	Rhode Island	15.3
Colorado	15.9	Minnesota	15.3	South Carolina	15.7
Connecticut	14.7	Mississippi	16.1	South Dakota	15.4
Delaware	17.1	Missouri	15.5	Tennessee	15.5
Florida	15.2	Montana	14.1	Texas	17.7
Georgia	17.1	Nebraska	15.1	Utah	21.2
Hawaii	17.6	Nevada	16.5	Vermont	14.0
Idaho	15.2	New Hampshire	16.2	Virginia	15.3
Illinois	16.0	New Jersey	15.1	Washington	15.4
Indiana	14.8	New Mexico	17.9	West Virginia	12.4
Iowa	13.1	New York	16.2	Wisconsin	14.8
Kansas	14.2	North Carolina	15.6	Wyoming	13.7
Kentucky	14.1	North Dakota	16.5		

Source: Department of Health and Human Services.

A histogram plot of these data is presented in Fig. 2.9.

A histogram is, in essence, a bar chart that graphs the frequencies or relative frequencies of data falling into different class intervals. These class frequencies can also be represented graphically by a frequency (or relative frequency) polygon. Each class interval is represented by a value, usually taken to be the midpoint of that interval. A plot is made of these values versus the frequencies of the class intervals they represent. These plotted points are then connected by straight lines to yield the frequency polygon. Such graphs are particularly useful for comparing data sets, since the different frequency polygons can be plotted on the same chart. ■

■ Example 2.4

The data of Table 2.9 represent class frequencies for the systolic blood pressure of two groups of male industrial workers: those aged 30 to 40 and those aged 50 to 60.

It is difficult to directly compare the blood pressures for the two age groups since the total number of workers in each group is different. To remove this

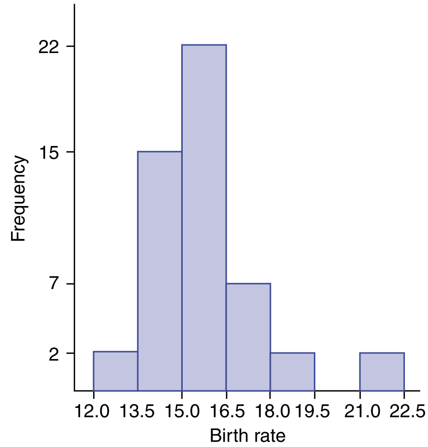


FIGURE 2.9 A histogram for birth rates in the 50 states.

Table 2.9 Class Frequencies of Systolic Blood Pressure of Two Groups of Male Workers

Blood pressure	Number of workers	
	Aged 30–40	Aged 50–60
Less than 90	3	1
90–100	17	2
100–110	118	23
110–120	460	57
120–130	768	122
130–140	675	149
140–150	312	167
150–160	120	73
160–170	45	62
170–180	18	35
180–190	3	20
190–200	1	9
200–210		3
210–220		5
220–230		2
230–240		1
Total	**2540**	**731**

Table 2.10 Relative Class Frequencies of Blood Pressures

Blood pressure	Percentage of workers	
	Aged 30–40	Aged 50–60
Less than 90	0.12	0.14
90–100	0.67	0.27
100–110	4.65	3.15
110–120	18.11	7.80
120–130	30.24	16.69
130–140	26.57	20.38
140–150	12.28	22.84
150–160	4.72	9.99
160–170	1.77	8.48
170–180	0.71	4.79
180–190	0.12	2.74
190–200	0.04	1.23
200–210		0.41
210–220		0.68
220–230		0.27
230–240		0.14
Total	**100.00**	**100.00**

difficulty, we can compute and graph the *relative* frequencies of each of the classes. That is, we divide all the frequencies relating to workers aged 30 to

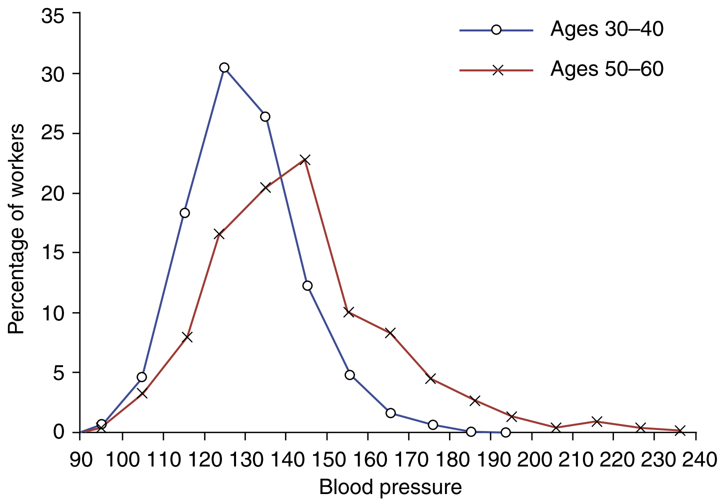


FIGURE 2.10 **Relative frequency polygons for the data of Table 2.10.**

39 by 2540 (the number of such workers) and all the frequencies relating to workers aged 50 to 59 by 731. This results in Table 2.10.

Figure 2.10 graphs the relative frequency polygons for both age groups. Having both frequency polygons on the same graph makes it easy to compare the two data sets. For instance, it appears that the blood pressures of the older group are more spread out among larger values than are those of the younger group. ■

PROBLEMS

1. The following data set represents the scores on intelligence quotient (IQ) examinations of 40 sixth-grade students at a particular school:

   114, 122, 103, 118, 99, 105, 134, 125, 117, 106, 109, 104, 111, 127,
   133, 111, 117, 103, 120, 98, 100, 130, 141, 119, 128, 106, 109, 115,
   113, 121, 100, 130, 125, 117, 119, 113, 104, 108, 110, 102

   (a) Present this data set in a frequency histogram.
   (b) Which class interval contains the greatest number of data values?
   (c) Is there a roughly equal number of data in each class interval?
   (d) Does the histogram appear to be approximately symmetric? If so, about which interval is it approximately symmetric?

2. The following data represent the daily high temperature (in degrees Celsius) on July 4 in San Francisco over a sequence of 30 years:

   22.8, 26.2, 31.7, 31.1, 26.9, 28.0, 29.4, 28.8, 26.7, 27.4, 28.2,

30.3, 29.5, 28.9, 27.5, 28.3, 24.1, 25.3, 28.5, 27.7, 24.4,
29.2, 30.3, 33.7, 27.5, 29.3, 30.2, 28.5, 32.2, 33.7

(a) Present this data set in a frequency histogram.
(b) What would you say is a “typical” July 4 temperature in San Francisco?
(c) What other conclusions can be drawn from the histogram?

3. The following data (in thousands of dollars) represent the net annual income for a sample of taxpayers:

47, 55, 18, 24, 27, 41, 50, 38, 33, 29, 15, 77, 64, 22, 19, 35, 39, 41,
67, 55, 121, 77, 80, 34, 41, 48, 60, 30, 22, 28, 84, 55, 26, 105, 62,
30, 17, 23, 31, 28, 56, 64, 88, 104, 115, 39, 25, 18, 21, 30, 57, 40,
38, 29, 19, 46, 40, 49, 72, 70, 37, 39, 18, 22, 29, 52, 94, 86, 23, 36

(a) Graph this data set in a frequency histogram having 5 class intervals.
(b) Graph this data set in a frequency histogram having 10 class intervals.
(c) Which histogram do you think is more informative? Why?

4. A set of 200 data points was broken up into 8 classes each of size (in the units of the data) 3, and the frequency of values in each class was determined. A frequency table was then constructed. However, some of the entries of this table were lost. Suppose that the part of the frequency table that remains is as follows:

Class interval	Frequency	Relative frequency
		0.05
	14	
	18	
15–18	38	
		0.10
	42	
	11	

Fill in the missing numbers and draw a relative frequency histogram.

5. The following is the ozone concentration (measured in parts per 100 million) of air in the downtown Los Angeles area during 25 consecutive summer days in 2004:

6.2, 9.1, 2.4, 3.6, 1.9, 1.7, 4.5, 4.2, 3.3, 5.1, 6.0, 1.8, 2.3,
4.9, 3.7, 3.8, 5.5, 6.4, 8.6, 9.3, 7.7, 5.4, 7.2, 4.9, 6.2

(a) Construct a frequency histogram for this data set having 3 to 5 as a class interval.
(b) Construct a frequency histogram for this data set having 2 to 3 as a class interval.
(c) Which frequency histogram do you find more informative?

6. The following is the 2002 meat production, in thousands of metric tons, for 11 different countries.

Country	Production	Country	Production
Argentina	2,748	Japan	520
Australia	2,034	Mexico	1,450
Brazil	7,150	Spain	592
China	5,616	United Kingdom	1,390
France	1,666	United States	12,424
Italy	1,161		

(a) Represent the given data in a frequency histogram.
(b) A data value that is far removed from the others is called an *outlier*. Is there an outlier in the given data?

7. Consider the blood cholesterol levels of the first 100 students in the data set presented in App. A. Divide these students by gender groupings, and construct a class relative frequency table for each. Plot, on the same chart, separate class relative frequency polygons for the female and male students. Can any conclusions be drawn about the relationship between gender and cholesterol level?
8. The following is a frequency table of daily travel times (in minutes):

Travel time	Frequency
15	6
18	5
22	4
23	3
26	4
32	3
48	1

(a) How many days are reported in the frequency table.
(b) Find the sum of the travel times of all those days.

9. Construct a relative frequency histogram of yearly death rates due to motor vehicles.
10. Construct a relative frequency histogram of yearly death rates due to falls.
11. Construct a relative frequency histogram of total yearly death rates due to all listed causes.
12. Would you say that the accidental death rates are remaining relatively steady?
13. Using the table described prior to Prob. 12 in Sec. 2.2, construct a histogram for the average yearly number of rainy days for the cities listed.
14. Consider the following table.

Age of driver, years	Percentage of all drivers	Percentage of all drivers in fatal accidents
15–20	9	18
20–25	13	21
25–30	13	14
30–35	11	11
35–40	9	7
40–45	8	6
45–50	8	5
50–55	7	5
55–60	6	4
60–65	6	3
65–70	4	2
70–75	3	2
Over 75	3	2

By the left-end convention, 13 percent of all drivers are at least 25 but less than 30 years old, and 11 percent of drivers killed in car accidents are at least 30 but less than 35 years old.

(a) Draw a relative frequency histogram for the age breakdown of drivers.

(b) Draw a relative frequency histogram for the age breakdown of those drivers who are killed in car accidents.

(c) Which age group accounts for the largest number of fatal accidents?

(d) Which age group should be charged the highest insurance premiums? Explain your reasoning.

15. A cumulative relative frequency table gives, for an increasing sequence of values, the percentage of data values that are less than that value. It can be constructed from a relative frequency table by simply adding the relative frequencies in a cumulative fashion. The following table is the beginning of such a table for the two data sets shown in Table 2.9. It says, for instance, that 5.44 percent of men aged 30 to 40 years have blood pressures below 110, as opposed to only 3.56 percent of those aged 50 to 60 years.

A Cumulative Relative Frequency Table for the Data Sets of Table 2.9

	Percentage of workers	
Blood pressure less than	Aged 30–40	Aged 50–60
90	0.12	0.14
100	0.79	0.41
110	5.44	3.56
120		
130		
.		
.		
.		
240	100	100

(a) Explain why the cumulative relative frequency for the last class must be 100.
(b) Complete the table.
(c) What does the table tell you about the two data sets? (That is, which one tends to have smaller values?)
(d) Graph, on the same chart, cumulative relative frequency polygons for the given data. Such graphs are called *ogives* (pronounced "OH jives").

2.4 STEM-AND-LEAF PLOTS

A very efficient way of displaying a small-to-moderate size data set is to utilize a *stem-and-leaf plot*. Such a plot is obtained by dividing each data value into two parts—its stem and its leaf. For instance, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit. That is, the value 84 is expressed as

Stem	Leaf
8	4

and the two data values 84 and 87 are expressed as

Stem	Leaf
8	4, 7

■ Example 2.5

Table 2.11 presents the per capita personal income for each of the 50 states and the District of Columbia. The data are for 2002.

The data presented in Table 2.11 are represented in the following stem-and-leaf plot. Note that the values of the leaves are put in the plot in increasing order.

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575, 579
26	183, 894, 982
27	671, 711, 744
28	240, 280, 551, 731, 821, 936
29	141, 405, 567, 596, 771, 923
30	001, 180, 296, 578
31	319, 727
32	151, 677, 779, 922, 996
33	276, 404
34	071, 334
36	043, 298
39	244, 453
42	120, 706

Table 2.11 Per Capita Personal Income (Dollars per Person), 2002

State name		State name		State name	
United States	30,941	Kentucky	25,579	Ohio	29,405
Alabama	25,128	Louisiana	25,446	Oklahoma	25,575
Alaska	32,151	Maine	27,744	Oregon	28,731
Arizona	26,183	Maryland	36,298	Pennsylvania	31,727
Arkansas	23,512	Massachusetts	39,244	Rhode Island	31,319
California	32,996	Michigan	30,296	South Carolina	25,400
Colorado	33,276	Minnesota	34,071	South Dakota	26,894
Connecticut	42,706	Mississippi	22,372	Tennessee	27,671
Delaware	32,779	Missouri	28,936	Texas	28,551
District of Columbia	42,120	Montana	25,020	Utah	24,306
Florida	29,596	Nebraska	29,771	Vermont	29,567
Georgia	28,821	Nevada	30,180	Virginia	32,922
Hawaii	30,001	New Hampshire	34,334	Washington	32,677
Idaho	25,057	New Jersey	39,453	West Virginia	23,688
Illinois	33,404	New Mexico	23,941	Wisconsin	29,923
Indiana	28,240	New York	36,043	Wyoming	30,578
Iowa	28,280	North Carolina	27,711		
Kansas	29,141	North Dakota	26,982		

The choice of stems should always be made so that the resultant stem-and-leaf plot is informative about the data. For instance, consider Example 2.6. ■

■ Example 2.6

The following data represent the proportion of public elementary school students that are classified as minority in each of 18 cities.

55.2, 47.8, 44.6, 64.2, 61.4, 36.6, 28.2, 57.4, 41.3,
44.6, 55.2, 39.6, 40.9, 52.2, 63.3, 34.5, 30.8, 45.3

If we let the stem denote the tens digit and the leaf represent the remainder of the value, then the stem-and-leaf plot for the given data is as follows:

$$\begin{array}{r|l}
2 & 8.2 \\
3 & 0.8,\ 4.5,\ 6.6,\ 9.6 \\
4 & 0.9,\ 1.3,\ 4.6,\ 4.6,\ 5.3,\ 7.8 \\
5 & 2.2,\ 5.2,\ 5.2,\ 7.4 \\
6 & 1.4,\ 3.3,\ 4.2
\end{array}$$

We could have let the stem denote the integer part and the leaf the decimal part of the value, so that the value 28.2 would be represented as

28	.2

However, this would have resulted in too many stems (with too few leaves each) to clearly illustrate the data set. ■

■ Example 2.7

The following stem-and-leaf plot represents the weights of 80 attendees at a sporting convention. The stem represents the tens digit, and the leaves are the ones digit.

10	2, 3, 3, 4, 7	(5)
11	0, 1, 2, 2, 3, 6, 9	(7)
12	1, 2, 4, 4, 6, 6, 6, 7, 9	(9)
13	1, 2, 2, 5, 5, 6, 6, 8, 9	(9)
14	0, 4, 6, 7, 7, 9, 9	(7)
15	1, 1, 5, 6, 6, 6, 7	(7)
16	0, 1, 1, 1, 2, 4, 5, 6, 8, 8	(10)
17	1, 1, 3, 5, 6, 6, 6	(7)
18	1, 2, 2, 5, 5, 6, 6, 9	(8)
19	0, 0, 1, 2, 4, 5	(6)
20	9, 9	(2)
21	7	(1)
22	1	(1)
23		(0)
24	9	(1)

The numbers in parentheses on the right represent the number of values in each stem class. These summary numbers are often useful. They tell us, for instance, that there are 10 values having stem 16; that is, 10 individuals have weights between 160 and 169. Note that a stem without any leaves (such as stem value 23) indicates that there are no occurrences in that class.

It is clear from this plot that almost all the data values are between 100 and 200, and the spread is fairly uniform throughout this region, with the exception of fewer values in the intervals between 100 and 110 and between 190 and 200. ■

Stem-and-leaf plots are quite useful in showing all the data values in a clear representation that can be the first step in describing, summarizing, and learning from the data. It is most helpful in moderate-size data sets. (If the size of the data set were very large, then, from a practical point of view, the values of all the leaves might be too overwhelming and a stem-and-leaf plot might not be any more informative than a histogram.) Physically this plot looks like a histogram turned on its side, with the additional plus that it presents the original within-group data values. These within-group values can be quite valuable to help you discover patterns in the data, such as that all the data values are multiples of some common value, or find out which values occur most frequently within a stem group.

Sometimes a stem-and-leaf plot appears to have too many leaves per stem line and as a result looks cluttered. One possible solution is to double the number of stems by having two stem lines for each stem value. On the top stem line in the pair we could include all leaves having values 0 through 4, and on the bottom stem line all leaves having values 5 through 9. For instance, suppose one line of a stem-and-leaf plot is as follows:

$$6 \mid 0, 0, 1, 2, 2, 3, 4, 4, 4, 4, 5, 5, 6, 6, 7, 7, 7, 7, 8, 9, 9$$

This could be broken into two lines:

$$\begin{array}{r|l} 6 & 0, 0, 1, 2, 2, 3, 4, 4, 4, 4 \\ 6 & 5, 5, 6, 6, 7, 7, 7, 7, 8, 9, 9 \end{array}$$

PROBLEMS

1. For the following data, draw stem-and-leaf plots having (a) 4 stems and (b) 8 stems.

   124, 129, 118, 135, 114, 139, 127, 141, 111, 144, 133, 127, 122, 119, 132, 137, 146, 122, 119, 115, 125, 132, 118, 126, 134, 147, 122, 119, 116, 125, 128, 130, 127, 135, 122, 141

2. The following table gives the maximal marginal 2016 tax rates of a variety of states. Represent the data in a stem and leaf plot.

   State Individual Income Taxes
   (Tax rates for tax year 2016)

State	Maximal rate	State	Maximal rate
Alabama	5	Idaho	7.4
Alaska	0	Illinois	3.75
Arizona	4.54	Indiana	3.3
Arkansas	6.9	Iowa	8.98
California	13.3	Kansas	4.6
Colorado	4.63	Kentucky	6.0
Connecticut	6.99	Louisiana	6.0
Delaware	6.6	Maine	5.1
Florida	0	Maryland	5.75
Georgia	6.0	Massachusetts	5.1
Hawaii	8.25		

Source: Statistical Abstract of the United States.

3. The following are the ages, to the nearest year, of 43 patients admitted to the emergency ward of a certain adult hospital:

23, 18, 31, 79, 44, 51, 24, 19, 17, 25, 27, 19, 44, 61, 22, 18,
14, 17, 29, 31, 22, 17, 15, 40, 55, 16, 17, 19, 20, 32, 20, 45,
53, 27, 16, 19, 22, 20, 18, 30, 20, 33, 21

Draw a stem-and-leaf plot for this data set. Use this plot to determine the 5-year interval of ages that contains the largest number of data points.

4. A psychologist recorded the following 48 reaction times (in seconds) to a certain stimulus.

1.1, 2.1, 0.4, 3.3, 1.5, 1.3, 3.2, 2.0, 1.7, 0.6, 0.9, 1.6, 2.2, 2.6, 1.8, 0.9,
2.5, 3.0, 0.7, 1.3, 1.8, 2.9, 2.6, 1.8, 3.1, 2.6, 1.5, 1.2, 2.5, 2.8, 0.7, 2.3,
0.6, 1.8, 1.1, 2.9, 3.2, 2.8, 1.2, 2.4, 0.5, 0.7, 2.4, 1.6, 1.3, 2.8, 2.1, 1.5

(a) Construct a stem-and-leaf plot for these data.
(b) Construct a second stem-and-leaf plot, using additional stems.
(c) Which one seems more informative?
(d) Suppose a newspaper article stated, "The typical reaction time was ______ seconds." Fill in your guess as to the missing word.

5. The following data represent New York City's daily revenue from parking meters (in units of $5000) during 30 days in 2002.

108, 77, 58, 88, 65, 52, 104, 75, 80, 83, 74, 68, 94, 97, 83,
71, 78, 83, 90, 79, 84, 81, 68, 57, 59, 32, 75, 93, 100, 88

(a) Represent this data set in a stem-and-leaf plot.
(b) Do any of the data values seem "suspicious"? Why?

6. The volatility of a stock is an important property in the theory of stock options pricing. It is an indication of how much change there tends to be in the day-to-day price of the stock. A volatility of 0 means that the price of the stock always remains the same. The higher the volatility, the more the stock's price tends to change. The following is a list of the volatility of 32 companies whose stock is traded on the American Stock Exchange:

0.26, 0.31, 0.45, 0.30, 0.26, 0.17, 0.33, 0.32, 0.37, 0.38, 0.35, 0.28, 0.37,
0.35, 0.29, 0.20, 0.33, 0.19, 0.31, 0.26, 0.24, 0.50, 0.22, 0.33, 0.51,
0.44, 0.63, 0.30, 0.28, 0.48, 0.42, 0.37

(a) Represent these data in a stem-and-leaf plot.
(b) What is the largest data value?
(c) What is the smallest data value?
(d) What is a "typical" data value?

7. The following table gives the scores of the first 25 Super Bowl games in professional football. Use it to construct a stem-and-leaf plot of
(a) The winning scores

(b) The losing scores
(c) The amounts by which the winning teams outscored the losing teams

Super Bowls I-XXV

Game	Date	Winner	Loser
XXV	Jan. 27, 1991	New York (NFC) 20	Buffalo (AFC) 19
XXIV	Jan. 28, 1990	San Francisco (NFC) 55	Denver (AFC) 10
XXIII	Jan. 22, 1989	San Francisco (NFC) 20	Cincinnati (AFC) 16
XXII	Jan. 31, 1988	Washington (NFC) 42	Denver (AFC) 10
XXI	Jan. 25, 1987	New York (NFC) 39	Denver (AFC) 20
XX	Jan. 26, 1986	Chicago (NFC) 46	New England (AFC) 10
XIX	Jan. 20, 1985	San Francisco (NFC) 38	Miami (AFC) 16
XVIII	Jan. 22, 1984	Los Angeles Raiders (AFC) 38	Washington (NFC) 9
XVII	Jan. 30, 1983	Washington (NFC) 27	Miami (AFC) 17
XVI	Jan. 24, 1982	San Francisco (NFC) 26	Cincinnati (AFC) 21
XV	Jan. 25, 1981	Oakland (AFC) 27	Philadelphia (NFC) 10
XIV	Jan. 20, 1980	Pittsburgh (AFC) 31	Los Angeles (NFC) 19
XIII	Jan. 21, 1979	Pittsburgh (AFC) 35	Dallas (NFC) 31
XII	Jan. 15, 1978	Dallas (NFC) 27	Denver (AFC) 10
XI	Jan. 9, 1977	Oakland (AFC) 32	Minnesota (NFC) 14
X	Jan. 18, 1976	Pittsburgh (AFC) 21	Dallas (NFC) 17
IX	Jan. 12, 1975	Pittsburgh (AFC) 16	Minnesota (NFC) 6
VIII	Jan. 13, 1974	Miami (AFC) 24	Minnesota (NFC) 7
VII	Jan. 14, 1973	Miami (AFC) 14	Washington (NFC) 7
VI	Jan. 16, 1972	Dallas (NFC) 24	Miami (AFC) 3
V	Jan. 17, 1971	Baltimore (AFC) 16	Dallas (NFC) 13
IV	Jan. 11, 1970	Kansas City (AFL) 23	Minnesota (NFL) 7
III	Jan. 12, 1969	New York (AFL) 16	Baltimore (NFL) 7
II	Jan. 14, 1968	Green Bay (NFL) 33	Oakland (AFL) 14
I	Jan. 15, 1967	Green Bay (NFL) 35	Kansas City (AFL) 10

8. Consider the following stem-and-leaf plot and histogram concerning the same set of data.

2	1,1,4,7	2–3	x, x, x, x
3	0, 0, 3, 3, 6, 9, 9, 9	3–4	x, x, x, x, x, x, x, x
4	2, 2, 5, 8, 8, 8	4–5	x, x, x, x, x, x
5	1, 1, 7, 7	5–6	x, x, x, x
6	3, 3, 3, 6	6–7	x, x, x, x
7	2, 2, 5, 5, 5, 8	7–8	x, x, x, x, x, x

What can you conclude from the stem-and-leaf plot that would not have been apparent from the histogram?

9. Use the data represented in the stem-and-leaf plot in Prob. 8 to answer the following questions.

(a) How many data values are in the 40s?
(b) What percentage of values is greater than 50?
(c) What percentage of values has the ones digit equal to 1?

10. A useful way of comparing two data sets is to put their stem-and-leaf plots side by side. The following represents the scores of students in two different schools on a standard examination. In both schools 24 students took the examination.

School A Leaves	Stem	School B Leaves
0	5	3, 5, 7
8, 5	6	2, 5, 8, 9, 9
9, 7, 4, 2, 0	7	3, 6, 7, 8, 8, 9
9, 8, 8, 7, 7, 6, 5, 3	8	0, 2, 3, 5, 6, 6
8, 8, 6, 6, 5, 5, 3, 0	9	0, 1, 5
	10	0

(a) Which school had the "high scorer"?
(b) Which school had the "low scorer"?
(c) Which school did better on the examination?
(d) Combine the two schools, and draw a stem-and-leaf plot for all 48 values.

2.5 SETS OF PAIRED DATA

Sometimes a data set consists of pairs of values that have some relationship to each other. Each member of the data set is thought of as having an x value and a y value. We often express the ith pair by the notation $(x_i, y_i), i = 1, \ldots, n$. For instance, in the data set presented in Table 2.12, x_i represents the score on an intelligence quotient (IQ) test, and y_i represents the annual salary (to the nearest \$1000) of the ith chosen worker in a sample of 30 workers from a particular company. In this section, we show how to effectively display data sets of paired values.

One approach to representing such a data set is to first consider each part of the paired data separately and then plot the relevant histograms or stem-and-leaf plots for each. For instance, Figs. 2.11 and 2.12 are stem-and-leaf plots of, respectively, the IQ test scores and the annual salaries for the data presented in Table 2.12.

However, although Figs. 2.11 and 2.12 tell us a great deal about the individual IQ scores and worker salaries, they tell us nothing about the relationship between these two variables. Thus, for instance, by themselves they would not be useful in helping us learn whether higher IQ scores tend to go along with higher income at this company. To learn about how the data relate to such questions, it is necessary to consider the paired values of each data point simultaneously.

Table 2.12 Salaries versus IQ

Worker i	IQ score x_i	Six month salary y_i (in units of $1000)	Worker i	IQ score x_i	Six month salary y_i (in units of $1000)
1	110	68	16	84	19
2	107	30	17	83	16
3	83	13	18	112	52
4	87	24	19	80	11
5	117	40	20	91	13
6	104	22	21	113	29
7	110	25	22	124	71
8	118	62	23	79	19
9	116	45	24	116	43
10	94	70	25	113	44
11	93	15	26	94	17
12	101	22	27	95	15
13	93	18	28	104	30
14	76	20	29	115	63
15	91	14	30	90	16

12	4	(1)
11	0,0,2,3,3,5,6,6,7,8	(10)
10	1,4,4,7	(4)
9	0,1,1,3,3,4,4,5	(8)
8	0,3,3,4,7	(5)
7	6,9	(2)

FIGURE 2.11 Stem-and-leaf plot for IQ scores.

7	0,1	(2)
6	2,3,8	(3)
5	2	(1)
4	0,3,4,5	(4)
3	0,0	(2)
2	0,2,2,4,5,9	(6)
1	1,3,3,4,5,5,6,6,7,8,9,9	(12)

FIGURE 2.12 Stem-and-leaf plot for annual salaries (in $1000).

A useful way of portraying a data set of paired values is to plot the data on a two-dimensional rectangular plot with the x axis representing the x value of the data and the y axis representing the y value. Such a plot is called a *scatter diagram*. Figure 2.13 presents a scatter diagram for the data of Table 2.12.

It is clear from Fig. 2.13 that higher incomes appear to go along with higher scores on the IQ test. That is, while not every worker with a high IQ score receives a larger salary than another worker with a lower score (compare worker 5 with worker 29), it appears to be generally true.

The scatter diagram of Fig. 2.13 also appears to have some predictive uses. For instance, suppose we wanted to predict the salary of a worker, similar to the

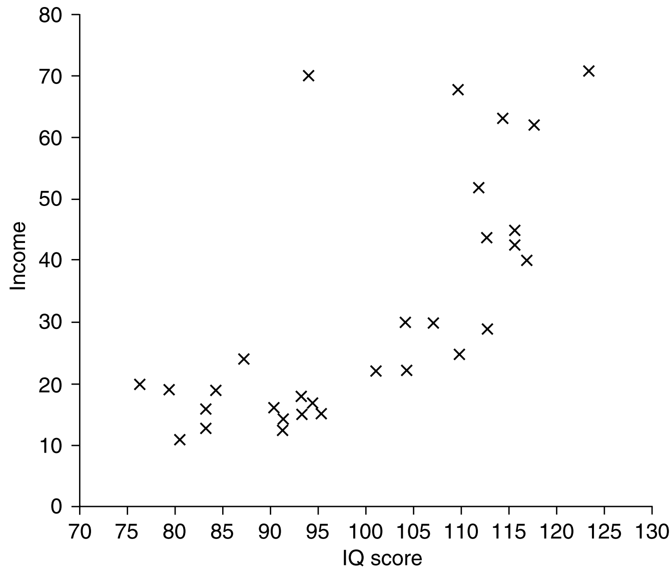


FIGURE 2.13 Scatter diagram of IQ versus income data.

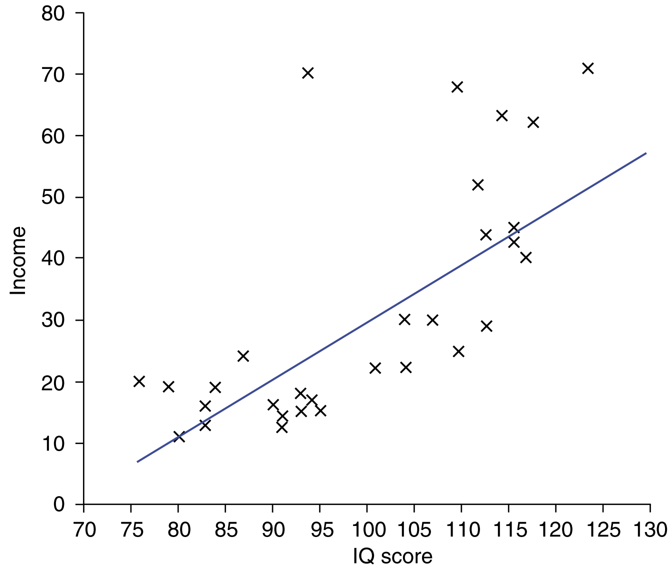


FIGURE 2.14 Scatter diagram for IQ versus income: fitting a straight line by eye.

ones just considered, whose IQ test score is 120. One way to do this is to "fit by eye" a line to the data set, as is done in Fig. 2.14. Since the y value on the line corresponding to the x value of 120 is about 45, this seems like a reasonable prediction for the annual salary of a worker whose IQ is 120.

In addition to displaying joint patterns of two variables and guiding predictions, a scatter diagram is useful in detecting *outliers*, which are data points that do not appear to follow the pattern of the other data points. (For example, the point (94, 70) in Fig. 2.13 does not appear to follow the general trend.) Having noted the outliers, we can then decide whether the data pair is meaningful or is caused by an error in data collection.

PROBLEMS

1. In an attempt to determine the relationship between the daily midday temperature (measured in degrees Celsius) and the number of defective parts produced during that day, a company recorded the following data over 22 workdays.

Temperature	Number of defective parts	Temperature	Number of defective parts
24.2	25	24.8	23
22.7	31	20.6	20
30.5	36	25.1	25
28.6	33	21.4	25
25.5	19	23.7	23
32.0	24	23.9	27
28.6	27	25.2	30
26.5	25	27.4	33
25.3	16	28.3	32
26.0	14	28.8	35
24.4	22	26.6	24

   (a) Draw a scatter diagram.
   (b) What can you conclude from the scatter diagram?
   (c) If tomorrow's midday temperature reading were 24.0, what would your best guess be as to the number of defective parts produced?

2. The following are the heights and starting salaries of 12 law school classmates whose law school examination scores were roughly the same.

Height	Salary
64	91
65	94
66	88
67	103
69	77
70	96
72	105
72	88
74	122
74	102
75	90
76	114

   Represent these data in a scatter diagram.

3. The following are the ages and weights of a random sample of 14 high school male students

Age	Weight
14	129
16	173
18	188
15	212
17	190
16	166
14	133
16	155
15	152
14	115
18	194
16	144
15	140
16	160

(a) Represent these data in a scatter diagram.
(b) What conclusions can be drawn.

4. The following table gives the number of days in each year from 1993 to 2002 that did not meet acceptable air quality standards in a selection of U.S. metropolitan areas.

Air Quality of Selected U.S. Metropolitan Areas, 1993–2002

Metropolitan statistical area	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Atlanta, GA	36	15	36	28	33	52	67	34	18	24
Bakersfield, CA	97	105	107	110	58	78	144	132	125	152
Baltimore, MD	48	40	36	28	30	51	40	19	32	42
Boston, MA–NH	2	6	7	4	7	8	10	1	12	16
Chicago, IL	4	13	24	7	10	12	19	2	22	21
Dallas, TX	12	24	29	10	27	33	25	22	16	15
Denver, CO	6	3	5	2	0	9	5	3	8	8
Detroit, MI	5	11	14	13	11	17	20	15	27	26
El Paso, TX	7	6	3	6	2	6	5	4	9	13
Fresno, CA	59	55	61	70	75	67	133	131	138	152
Houston, TX	27	41	66	28	47	38	52	42	29	23
Las Vegas, NV–AZ	3	3	3	14	4	5	8	2	1	6
Los Angeles–Long Beach, CA	134	139	113	94	60	56	56	87	88	80
Miami, FL	6	1	2	1	3	8	7	2	1	1
Minneapolis–St. Paul, MN–WI	0	2	5	0	0	1	1	2	2	1
New Haven–Meriden, CT	12	13	14	8	19	9	19	9	15	25

(*Continued*)

(*Continued*)

Metropolitan statistical area	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
New York, NY	11	16	21	14	23	18	25	19	19	31
Orange County, CA	25	15	9	9	3	6	14	31	31	19
Philadelphia, PA–NJ	62	37	38	38	38	37	32	22	29	33
Phoenix–Mesa, AZ	14	10	22	15	12	14	10	10	8	8
Pittsburgh, PA	14	22	27	12	21	39	40	29	52	53
Riverside–San Bernardino, CA	168	150	125	118	107	96	123	145	155	145
Sacramento, CA	20	37	41	44	17	29	69	45	49	69
St. Louis, MO–IL	9	33	38	23	15	24	31	18	17	34
Salt Lake City–Ogden, UT	5	17	5	14	2	19	8	15	15	18
San Diego, CA	59	46	48	31	14	33	33	31	31	20
San Francisco, CA	0	0	2	0	0	0	10	4	12	17
Seattle–Bellevue–Everett, WA	0	3	2	6	1	3	6	7	3	6
Ventura, CA	43	63	66	62	45	29	24	31	25	11
Washington, DC–MD–VA–WV	52	22	32	18	30	47	39	11	22	34

Note: Data indicate the number of days metropolitan statistical areas failed to meet acceptable air quality standards. All figures were revised based on new standards set in 1998. Includes fine particles less than or equal to 2.5 mm in diameter.
Source: U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards.

(a) Draw a scatter diagram relating the 2000 and 2002 entries for each city.
(b) Do higher values in 2002 tend to go with higher values in 2000?

5. The following data relate the attention span (in minutes) to a score on an IQ examination of 18 preschool-age children.

Attention span	IQ score	Attention span	IQ score	Attention span	IQ score
2.0	82	6.3	105	5.5	118
3.0	88	5.4	108	3.6	128
4.4	86	6.6	112	5.4	128
5.2	94	7.0	116	3.8	130
4.9	90	6.5	122	2.7	140
6.1	99	7.2	110	2.2	142

(a) Draw a scatter diagram.
(b) Give a plausible inference concerning the relation of attention span to IQ score.

6. The following data relate prime lending rates and the corresponding inflation rate during 8 years in the 1970s.

Inflation rate	Prime lending rate	Inflation rate	Prime lending rate
3.3	5.2	5.8	6.8
6.2	8.0	6.5	6.9
11.0	10.8	7.6	9.0
9.1	7.9		

(a) Draw a scatter diagram.
(b) Fit a straight line drawn "by hand" to the data pairs.
(c) Using your straight line, predict the prime lending rate in a year whose inflation rate is 7.2 percent.

7. A random group of 12 high school juniors were asked to estimate the average number of hours they study each week. The grade point averages of these students were then determined, with the resulting data being as given in the following. Use it to represent these data in a scatter diagram.

Hours reported working and GPA

Hours	GPA	Hours	GPA
6	2.8	11	3.3
14	3.2	12	3.4
3	3.1	5	2.7
22	3.6	24	3.8
9	3.0	15	3.0

8. Problem 7 of Sec. 2.4 gives the scores of the first 25 Super Bowl football games. For each game, let y denote the score of the winning team, and let x denote the number of points by which that team won. Draw a scatter diagram relating x and y. Do high values of one tend to go with high values of the other?

2.6 SOME HISTORICAL COMMENTS

Probably the first recorded instance of statistical graphics—that is, the representation of data by tables or graphs—was Sir Edmund Halley's graphical analysis of barometric pressure as a function of altitude, published in 1686. Using the rectangular coordinate system introduced by the French scientist René Descartes in his study of analytic geometry, Halley plotted a scatter diagram and was then able to fit a curve to the plotted data.

In spite of Halley's demonstrated success with graphical plotting, almost all the applied scientists until the latter part of the 18th century emphasized tables rather than graphs in presenting their data. Indeed, it was not until 1786, when William Playfair invented the bar graph to represent a frequency table, that graphs began to be regularly employed. In 1801 Playfair invented the pie chart and a short time later originated the use of histograms to display data.

The use of graphs to represent continuous data—that is, data in which all the values are distinct—did not regularly appear until the 1830s. In 1833 the Frenchman A. M. Guerry applied the bar chart form to continuous crime data, by first breaking up the data into classes, to produce a histogram. Systematic development of the histogram was carried out by the Belgian statistician and social scientist Adolphe Quetelet about 1846. Quetelet and his students demonstrated the usefulness of graphical analysis in their development of the social sciences. In doing so, Quetelet popularized the practice, widely followed today, of initiating a research study by first gathering and presenting numerical data. Indeed, along with the additional steps of summarizing the data and then utilizing the methods of statistical inference to draw conclusions, this has become the accepted paradigm for research in all fields connected with the social sciences. It has also become an important technique in other fields, such as medical research (the testing of new drugs and therapies), as well as in such traditionally nonnumerical fields as literature (in deciding authorship) and history (particularly as developed by the French historian Fernand Braudel).


(Princeton University)

John Tukey

The term *histogram* was first used by Karl Pearson in his 1895 lectures on statistical graphics. The stem-and-leaf plot, which is a variant of the histogram, was introduced by the U.S. statistician John Tukey in 1970. In the words of Tukey, "Whereas a histogram uses a nonquantitative mark to indicate a data value, clearly the best type of mark is a digit."

KEY TERMS

Frequency: The number of times that a given value occurs in a data set.

Frequency table: A table that presents, for a given set of data, each distinct data value along with its frequency.

Line graph: A graph of a frequency table. The abscissa specifies a data value, and the frequency of occurrence of that value is indicated by the height of a vertical line.

Bar chart (or bar graph): Similar to a line graph, except now the frequency of a data value is indicated by the height of a bar.

Frequency polygon: A plot of the distinct data values and their frequencies that connects the plotted points by straight lines.

Symmetric data set: A data set is symmetric about a given value x_0 if the frequencies of the data values $x_0 - c$ and $x_0 + c$ are the same for all values of c.

Relative frequency: The frequency of a data value divided by the number of pieces of data in the set.

Pie chart: A chart that indicates relative frequencies by slicing up a circle into distinct sectors.

Histogram: A graph in which the data are divided into class intervals, whose frequencies are shown in a bar graph.

Relative frequency histogram: A histogram that plots relative frequencies for each data value in the set.

Stem-and-leaf plot: Similar to a histogram except that the frequency is indicated by stringing together the last digits (the leaves) of the data.

Scatter diagram: A two-dimensional plot of a data set of paired values.

SUMMARY

This chapter presented various ways to graphically represent data sets. For instance, consider the following set of 13 data values:

1, 2, 3, 1, 4, 2, 6, 2, 4, 3, 5, 4, 2

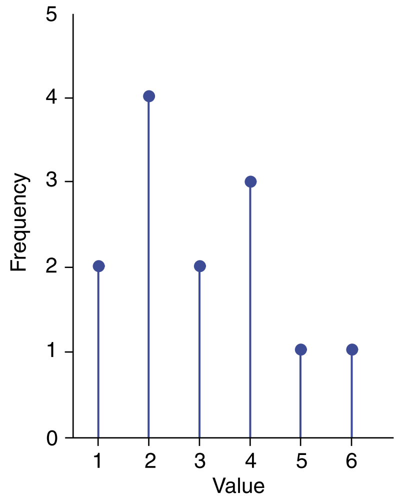
These values can be represented in a *frequency table*, which lists each value and the number of times it occurs in the data, as follows:

A Frequency Table

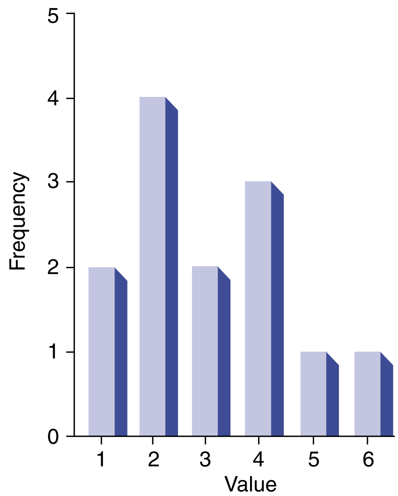
Value	Frequency	Value	Frequency
1	2	4	3
2	4	5	1
3	2	6	1

The data also can be graphically pictured by either a *line graph* or a *bar chart*. Sometimes the frequencies of the different data values are plotted on a graph, and then the resulting points are connected by straight lines. This gives rise to a *frequency polygon*.

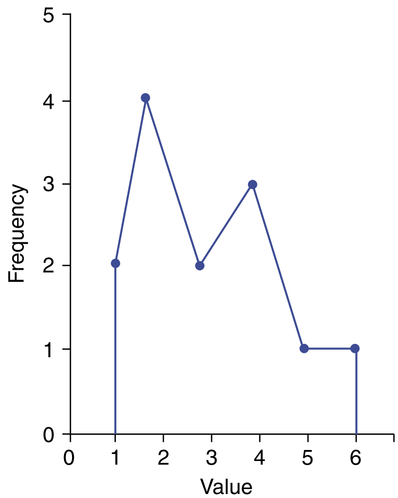
When there are a large number of data values, often we break them up into class intervals. A bar chart plot relating each class interval to the number of data values falling in the interval is called a *histogram*. The y axis of this plot can represent either the class frequency (that is, the number of data values in the interval) or the proportion of all the data that lies in the class. In the former case we call the plot a *frequency histogram* and in the latter case a *relative frequency histogram*.



A line graph.



A bar graph.



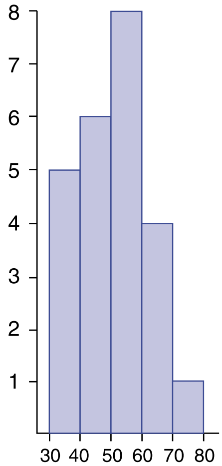
A frequency polygon.

Consider this data set:

41, 38, 44, 47, 33, 35, 55, 52, 41, 66, 64, 50, 49, 56, 55, 48, 52, 63, 59, 57, 75, 63, 38, 37

Using the five class intervals

30–40, 40–50, 50–60, 60–70, 70–80



A histogram.

along with the left-end inclusion convention (which signifies that the interval contains all points greater than or equal to its left-end member and less than its right-end member), we have the histogram above to represent this data set.

Data sets can also be graphically displayed in a *stem-and-leaf plot*. The following stem-and-leaf plot is for the preceding data set.

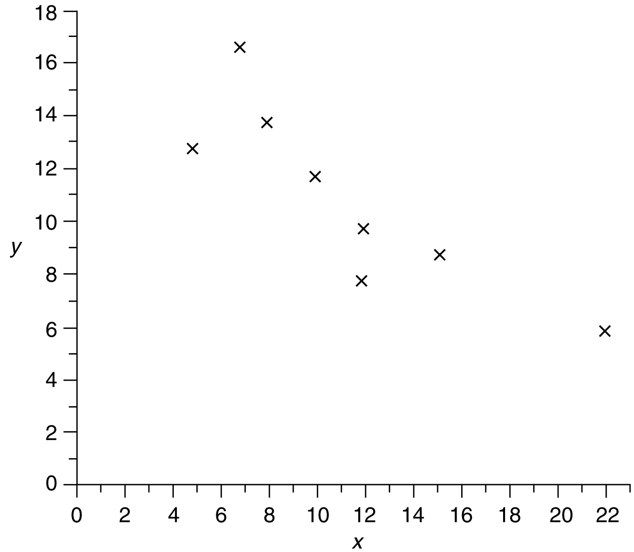
7	5
6	3,3,4,6
5	0,2,2,5,5,6,7,9
4	1,1,4,7,8,9
3	3,5,7,8,8

A stem-and-leaf plot.

Often data come in pairs. That is, for each element of the data set there is an x value and a y value. A plot of the x and y values is called a *scatter diagram*. A scatter diagram can be quite useful in ascertaining such things as whether high x values appear to go along with high y values, or whether high x values tend to go along with low y values, or whether there appears to be no particular association between the x and y values of a pair.

The following data set of pairs

i	1	2	3	4	5	6	7	8
x_i	8	12	7	15	5	12	10	22
y_i	14	10	17	9	13	8	12	6



A scatter diagram.

is represented in the scatter diagram above. The diagram indicates that high values of one member of the pair appear to be generally associated with low values of the other member of the pair.

Using these graphical tools, often we can communicate pertinent features of a data set at a glance. As a result, we can learn things about the data that are not immediately evident in the raw numbers themselves. The choice of which display to use depends on such things as the size of the data set, the type of data, and the number of distinct values.

REVIEW PROBLEMS

1. The following data are the blood types of 50 volunteers at a blood plasma donation clinic:

   O A O AB A A O O B A O A AB B O O O A B A A O A A O
   B A O AB A O O A B A A A O B O O A O A B O AB A O B

   (a) Represent these data in a frequency table.
   (b) Represent them in a relative frequency table.
   (c) Represent them in a pie chart.
2. The following is a sample of prices, rounded to the nearest cent, charged per gallon of gasoline in the Los Angeles area in September 2016.

   298, 311, 330, 307, 359, 339, 288, 325, 277, 294,
   312, 378, 323, 316, 289, 342, 282, 329, 305, 320

(a) Construct a frequency histogram for this data set.
(b) Construct a stem-and-leaf plot.

3. The following frequency table presents the number of female suicides that took place in eight German states over 14 years.

Number of suicides per year	0	1	2	3	4	5	6	7	8	9	10
Frequency	9	19	17	20	15	11	8	2	3	5	3

Thus, for instance, there were a total of 20 cases in which states had 3 suicides in a year.
(a) How many suicides were reported over the 14 years?
(b) Represent the above data in a histogram.

4. The following table gives the 1991 crime rate (per 100,000 population) in each state. Use it to construct a
(a) Frequency histogram of the total violent crime rates in the northeastern states
(b) Relative frequency histogram of the total property crime rates in the southern states
(c) Stem-and-leaf plot of the murder rates in the western states
(d) Stem-and-leaf plot of the burglary rates in the midwestern states.

		Violent crime					Property crime			
Region, Division, and State	**Total**	**Total**	**Murder**	**Forcible rape**	**Robbery**	**Aggravated assault**	**Total**	**Burglary**	**Larceny—theft**	**Motor vehicle theft**
United States	**5,898**	**758**	**9.8**	**42**	**273**	**433**	**5,140**	**1,252**	**3,229**	**659**
Northeast	5,155	752	8.4	29	352	363	4,403	1,010	2,598	795
New England	4,950	532	4.1	30	159	338	4,419	1,103	2,600	716
Maine	3,768	132	1.2	22	23	86	3,636	903	2,570	163
New Hampshire	3,448	119	3.6	30	33	53	3,329	735	2,373	220
Vermont	3,955	117	2.1	31	12	72	3,838	1,020	2,674	144
Massachusetts	5,332	736	4.2	32	195	505	4,586	1,167	2,501	919
Rhode Island	5,039	462	3.7	31	123	304	4,577	1,127	2,656	794
Connecticut	5,364	540	5.7	29	224	280	4,824	1,191	2,838	796
Middle Atlantic	5,227	829	9.9	29	419	372	4,398	978	2,598	823
New York	6,245	1,164	14.2	28	622	499	5,081	1,132	2,944	1,004
New Jersey	5,431	635	5.2	29	293	307	4,797	1,016	2,855	926
Pennsylvania	3,559	450	6.3	29	194	221	3,109	720	1,907	482
Midwest	5,257	631	7.8	45	223	355	4,626	1,037	3,082	507
East north central	5,482	704	8.9	50	263	383	4,777	1,056	3,151	570
Ohio	5,033	562	7.2	53	215	287	4,471	1,055	2,916	500
Indiana	4,818	505	7.5	41	116	340	4,312	977	2,871	465
Illinois	6,132	1,039	11.3	40	456	532	5,093	1,120	3,318	655
Michigan	6,138	803	10.8	79	243	470	5,335	1,186	3,469	680
Wisconsin	4,466	277	4.8	25	119	128	4,189	752	3,001	436

(Continued)

(*Continued*)

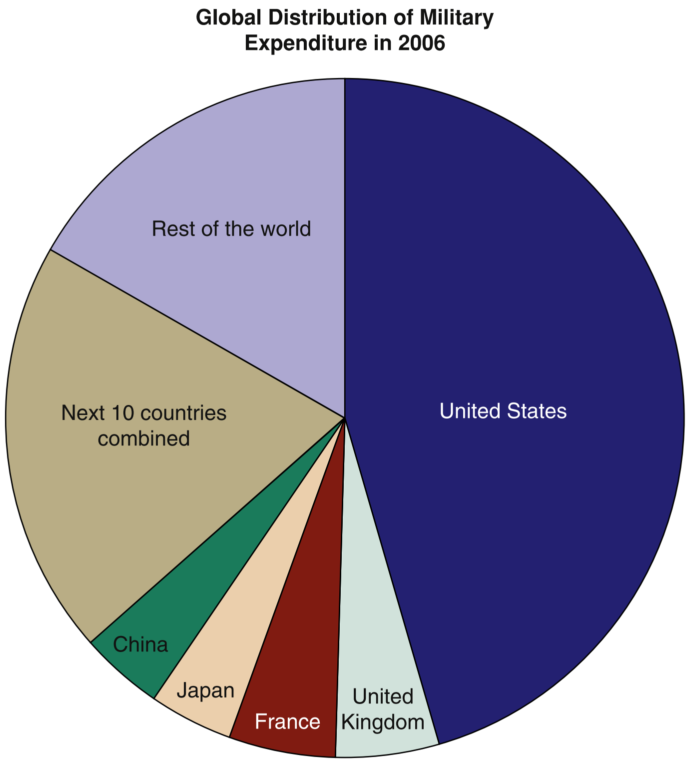
Region, Division, and State	Total	Violent crime: Total	Murder	Forcible rape	Robbery	Aggravated assault	Property crime: Total	Burglary	Larceny—theft	Motor vehicle theft
West north central	4,722	457	5.4	34	129	288	4,265	991	2,918	356
Minnesota	4,496	316	3.0	40	98	175	4,180	854	2,963	363
Iowa	4,134	303	2.0	21	45	235	3,831	832	2,828	171
Missouri	5,416	763	10.5	34	251	467	4,653	1,253	2,841	558
North Dakota	2,794	65	1.1	18	8	38	2,729	373	2,229	127
South Dakota	3,079	182	1.7	40	19	122	2,897	590	2,192	115
Nebraska	4,354	335	3.3	28	54	249	4,020	727	3,080	213
Kansas	5,534	500	6.1	45	138	310	5,035	1,307	3,377	351
South	6,417	798	12.1	45	252	489	5,618	1,498	3,518	603
South Atlantic	6,585	851	11.4	44	286	510	5,734	1,508	3,665	561
Delaware	5,869	714	5.4	86	215	408	5,155	1,128	3,652	375
Maryland	6,209	956	11.7	46	407	492	5,253	1,158	3,365	731
District of Columbia	10,768	2,453	80.6	36	1,216	1,121	8,315	2,074	4,880	1,360
Virginia	4,607	373	9.3	30	138	196	4,234	783	3,113	339
West Virginia	2,663	191	6.2	23	43	119	2,472	667	1,631	175
North Carolina	5,889	658	11.4	35	178	434	5,230	1,692	3,239	299
South Carolina	6,179	973	11.3	59	171	731	5,207	1,455	3,365	387
Georgia	6,493	738	12.8	42	268	415	5,755	1,515	3,629	611
Florida	8,547	1,184	9.4	52	400	723	7,363	2,006	4,573	784
East south central	4,687	631	10.4	41	149	430	4,056	1,196	2,465	395
Kentucky	3,358	438	6.8	35	83	313	2,920	797	1,909	215
Tennessee	5,367	726	11.0	46	213	456	4,641	1,365	2,662	614
Alabama	5,366	844	11.5	36	153	644	4,521	1,269	2,889	363
Mississippi	4,221	389	12.8	46	116	214	3,832	1,332	2,213	286
West south central	7,118	806	14.2	50	254	488	6,312	1,653	3,871	788
Arkansas	5,175	593	11.1	45	136	402	4,582	1,227	3,014	341
Louisiana	6,425	951	16.9	41	279	614	5,473	1,412	3,489	573
Oklahoma	5,669	584	7.2	51	129	397	5,085	1,478	3,050	557
Texas	7,819	840	15.3	53	286	485	6,979	1,802	4,232	944
West	6,478	841	9.6	46	287	498	5,637	1,324	3,522	791
Mountain	6,125	544	6.5	44	122	371	5,581	1,247	3,843	491
Montana	3,648	140	2.6	20	19	99	3,508	524	2,778	206
Idaho	4,196	290	1.8	29	21	239	3,905	826	2,901	178
Wyoming	4,389	310	3.3	26	17	264	4,079	692	3,232	155
Colorado	6,074	559	5.9	47	107	399	5,515	1,158	3,930	426
New Mexico	6,679	835	10.5	52	120	652	5,845	1,723	3,775	346
Arizona	7,406	671	7.8	42	166	455	6,735	1,607	4,266	861
Utah	5,608	287	2.9	46	55	183	5,321	840	4,240	241
Nevada	6,299	677	11.8	66	312	287	5,622	1,404	3,565	652
Pacific	6,602	945	10.7	47	345	542	5,656	1,351	3,409	896
Washington	6,304	523	4.2	70	146	303	5,781	1,235	4,102	444
Oregon	5,755	506	4.6	53	150	298	5,249	1,176	3,598	474
California	6,773	1,090	12.7	42	411	624	5,683	1,398	3,246	1,039
Alaska	5,702	614	7.4	92	113	402	5,088	979	3,575	534
Hawaii	5,970	242	4.0	33	87	118	5,729	1,234	4,158	336

Source: U.S. Federal Bureau of Investigation, *Crime in the United States*, annual.

5. Construct a frequency table for a data set of 10 values that is symmetric and has (a) 5 distinct values and (b) 4 distinct values. (c) About what values are the data sets in parts (a) and (b) symmetric?
6. The following are the estimated oil reserves, in billions of barrels, for four regions in the western hemisphere. Represent the data in a pie chart.

United States	38.7
South America	22.6
Canada	8.8
Mexico	60.0

7. The following pie chart represents the percentages of the world's 2006 total military spending by countries and regions of the world. Use it to estimate the percentages of all military expenditures spent by (a) the United States, and (b) China.



Source: Stockholm International Peace Research Institute Yearbook 2007.

8. The following data refer to the ages (to the nearest year) at which patients died at a large inner-city (nonbirthing) hospital:

1, 1, 1, 1, 3, 3, 4, 9, 17, 18, 19, 20, 20, 22, 24, 26, 28, 34,
45, 52, 56, 59, 63, 66, 68, 68, 69, 70, 74, 77, 81, 90

(a) Represent this data set in a histogram.
(b) Represent it in a frequency polygon.
(c) Represent it in a cumulative frequency polygon.
(d) Represent it in a stem-and-leaf plot.

Problems 9 to 11 refer to the last 50 student entries in App. A.

9. (a) Draw a histogram of the weights of these students.
(b) Comment on this histogram.
10. Draw a scatter diagram relating weight and cholesterol level. Comment on what the scatter diagram indicates.
11. Draw a scatter diagram relating weight and blood pressure. What does this diagram indicate?

Problems 12 and 13 refer to the following table concerning the mathematics and verbal SAT scores of a graduating class of high school seniors.

Student	Verbal score	Mathematics score	Student	Verbal score	Mathematics score
1	520	505	8	620	576
2	605	575	9	604	622
3	528	672	10	720	704
4	720	780	11	490	458
5	630	606	12	524	552
6	504	488	13	646	665
7	530	475	14	690	550

12. Draw side-by-side stem-and-leaf plots of the student scores on the mathematics and verbal SAT examinations. Did the students, as a group, perform better on one examination? If so, which one?
13. Draw a scatter diagram of student scores on the two examinations. Do high scores on one tend to go along with high scores on the other?
14. The following table gives information about the age of the population in both the United States and Mexico.

	Proportion of population (percent)	
Age, years	Mexico	United States
0–9	32.5	17.5
10–19	24	20
20–29	14.5	14.5
30–39	11	12
40–49	7.5	12.5
50–59	4.5	10.5
60–69	3.5	7
70–79	1.5	4
Over 80	1	2

(a) What percentage of the Mexican population is less than 30 years old?
(b) What percentage of the U.S. population is less than 30 years old?

(c) Draw two relative frequency polygons on the same graph. Use different colors for Mexican and for U.S. data.

(d) In general, how do the age distributions compare for the two countries?

15. The following data relate to the normal monthly and annual precipitation (in inches) for various cities.

State	City	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual
AL	Mobile	4.59	4.91	6.48	5.35	5.46	5.07	7.74	6.75	6.56	2.62	3.67	5.44	64.64
AK	Juneau	3.69	3.74	3.34	2.92	3.41	2.98	4.13	5.02	6.40	7.71	5.15	4.66	53.15
AZ	Phoenix	0.73	0.59	0.81	0.27	0.14	0.17	0.74	1.02	0.64	0.63	0.54	0.83	7.11
AR	Little Rock	3.91	3.83	4.69	5.41	5.29	3.67	3.63	3.07	4.26	2.84	4.37	4.23	49.20
CA	Los Angeles	3.06	2.49	1.76	0.93	0.14	0.04	0.01	0.10	0.15	0.26	1.52	1.62	12.08
	Sacramento	4.03	2.88	2.06	1.31	0.33	0.11	0.05	0.07	0.27	0.86	2.23	2.90	17.10
	San Diego	2.11	1.43	1.60	0.78	0.24	0.06	0.01	0.11	0.19	0.33	1.10	1.36	9.32
	San Francisco	4.65	3.23	2.64	1.53	0.32	0.11	0.03	0.05	0.19	1.06	2.35	3.55	19.71
CO	Denver	0.51	0.69	1.21	1.81	2.47	1.58	1.93	1.53	1.23	0.98	0.82	0.55	15.31
CT	Hartford	3.53	3.19	4.15	4.02	3.37	3.38	3.09	4.00	3.94	3.51	4.05	4.16	44.39
DE	Wilmington	3.11	2.99	3.87	3.39	3.23	3.51	3.90	4.03	3.59	2.89	3.33	3.54	41.38
DC	Washington	2.76	2.62	3.46	2.93	3.48	3.35	3.88	4.40	3.22	2.90	2.82	3.18	39.00
FL	Jacksonville	3.07	3.48	3.72	3.32	4.91	5.37	6.54	7.15	7.26	3.41	1.94	2.59	52.76
	Miami	2.08	2.05	1.89	3.07	6.53	9.15	5.98	7.02	8.07	7.14	2.71	1.86	57.55
GA	Atlanta	4.91	4.43	5.91	4.43	4.02	3.41	4.73	3.41	3.17	2.53	3.43	4.23	48.61
HI	Honolulu	3.79	2.72	3.48	1.49	1.21	0.49	0.54	0.60	0.62	1.88	3.22	3.43	23.47
ID	Boise	1.64	1.07	1.03	1.19	1.21	0.95	0.26	0.40	0.58	0.75	1.29	1.34	11.71
IL	Chicago	1.60	1.31	2.59	3.66	3.15	4.08	3.63	3.53	3.35	2.28	2.06	2.10	33.34
	Peoria	1.60	1.41	2.86	3.81	3.84	3.88	3.99	3.39	3.63	2.51	1.96	2.01	34.89
IN	Indianapolis	2.65	2.46	3.61	3.68	3.66	3.99	4.32	3.46	2.74	2.51	3.04	3.00	39.12
IA	Des Moines	1.01	1.12	2.20	3.21	3.96	4.18	3.22	4.11	3.09	2.16	1.52	1.05	30.83
KS	Wichita	0.68	0.85	2.01	2.30	3.91	4.06	3.62	2.80	3.45	2.47	1.47	0.99	28.61
KY	Louisville	3.38	3.23	4.73	4.11	4.15	3.60	4.10	3.31	3.35	2.63	3.49	3.48	43.56
LA	New Orleans	4.97	5.23	4.73	4.50	5.07	4.63	6.73	6.02	5.87	2.66	4.06	5.27	59.74

Source: U.S. National Oceanic and Atmospheric Administration, *Climatography of the United States*, September 1982.

(a) Represent the normal precipitation amounts for April in a stem-and-leaf plot.

(b) Represent the annual amounts in a histogram.

(c) Draw a scatter diagram relating the April amount to the annual amount.

16. A data value that is far away from the other values is called an *outlier*. In the following data sets, specify which, if any, of the data values are outliers.

(a) 14, 22, 17, 5, 18, 22, 10, −17, 25, 28, 33, 12

(b) 5, 2, 13, 16, 9, 12, 7, 10, 54, 22, 18, 15, 12

(c) 18, 52, 14, 20, 24, 27, 43, 17, 25, 28, 3, 22, 6