

4

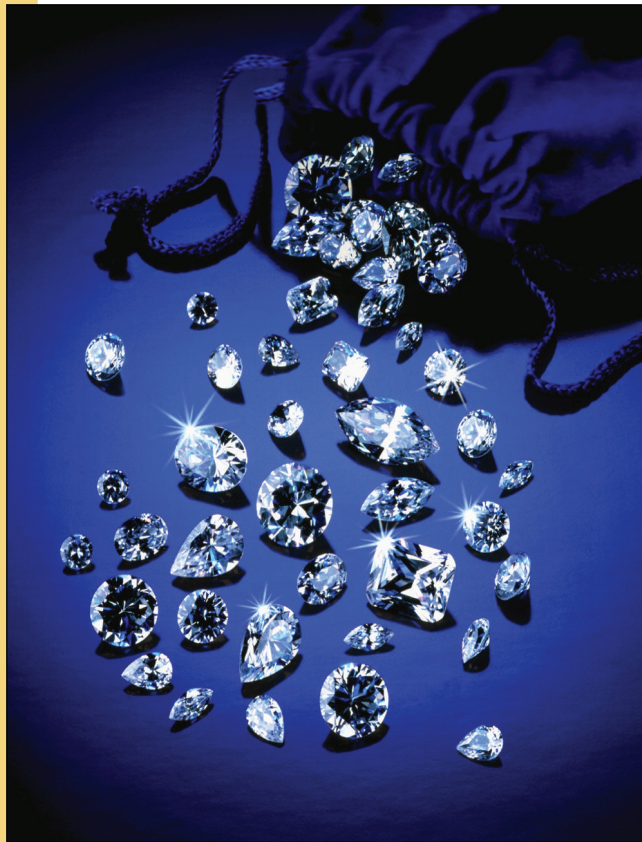
Learning Objectives

When you have completed this chapter, you will be able to:

- L01** Construct and interpret a dot plot.
- L02** Construct and describe a stem-and-leaf display.
- L03** Identify and compute measures of position.
- L04** Construct and analyze a box plot.
- L05** Compute and describe the coefficient of skewness.
- L06** Create and interpret a scatter diagram.
- L07** Develop and explain a contingency table.

Describing Data:

Displaying and Exploring Data



McGivern Jewelers recently ran an advertisement in the local newspaper reporting the shape, size, price, and cut grade for 33 of its diamonds in stock. Develop a box plot of the variable price and comment on the result. (See Exercise 37 and L04.)

4.1 Introduction

Chapter 2 began our study of descriptive statistics. In order to transform raw or ungrouped data into a meaningful form, we organize the data into a frequency distribution. We present the frequency distribution in graphic form as a histogram or a frequency polygon. This allows us to visualize where the data tends to cluster, the largest and the smallest values, and the general shape of the data.

In Chapter 3, we first computed several measures of location, such as the mean and the median. These measures of location allow us to report a typical value in the set of observations. We also computed several measures of dispersion, such as the range and the standard deviation. These measures of dispersion allow us to describe the variation or the spread in a set of observations.

We continue our study of descriptive statistics in this chapter. We study (1) dot plots, (2) stem-and-leaf displays, (3) percentiles, and (4) box plots. These charts and statistics give us additional insight into where the values are concentrated as well as the general shape of the data. Then we consider bivariate data. In bivariate data, we observe two variables for each individual or observation selected. Examples include: the number of hours a student studied and the points earned on an examination; whether a sampled product is acceptable or not and the shift on which it is manufactured; and the amount of electricity used in a month by a homeowner and the mean daily high temperature in the region for the month.

4.2 Dot Plots

L01 Construct and interpret a dot plot.

Dot plots give a visual idea of the spread and concentration of the data.

Recall for the Applewood Auto Group data, we summarized the profit earned on the 180 vehicles sold into eight classes. When we organized the data into the eight classes, we lost the exact value of the observations. A **dot plot**, on the other hand, groups the data as little as possible, and we do not lose the identity of an individual observation. To develop a dot plot, we simply display a dot for each observation along a horizontal number line indicating the possible values of the data. If there are identical observations or the observations are too close to be shown individually, the dots are “piled” on top of each other. This allows us to see the shape of the distribution, the value about which the data tend to cluster, and the largest and smallest observations. Dot plots are most useful for smaller data sets, whereas histograms tend to be most useful for large data sets. An example will show how to construct and interpret dot plots.

Example

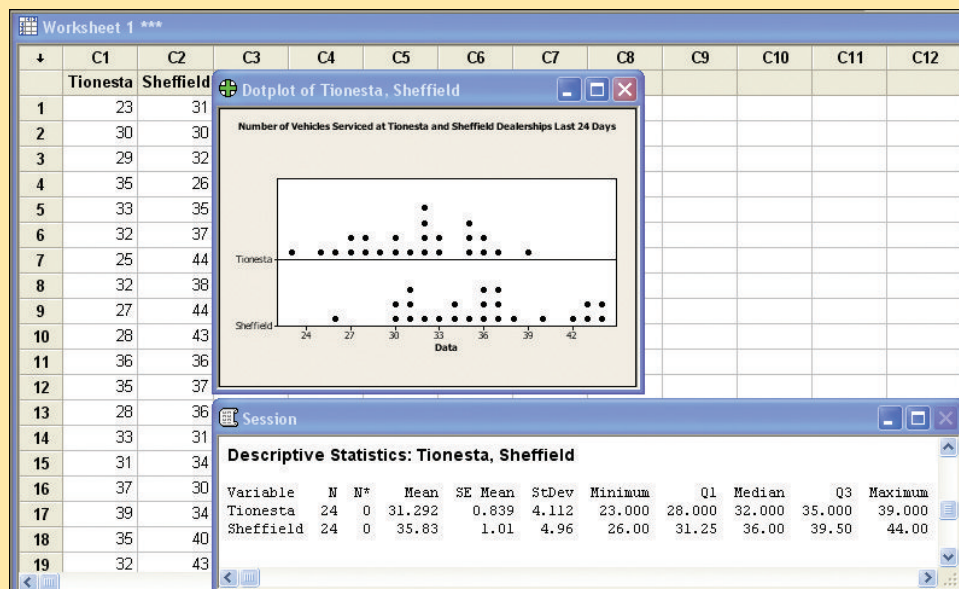
The service departments at Tionesta Ford Lincoln Mercury and Sheffield Motors Inc., two of the four Applewood Auto Group dealerships, were both open 24 working days last month. Listed below is the number of vehicles serviced last month at the two dealerships. Construct dot plots and report summary statistics to compare the two dealerships.

Tionesta Ford Lincoln Mercury					
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
23	33	27	28	39	26
30	32	28	33	35	32
29	25	36	31	32	27
35	32	35	37	36	30

Solution

Sheffield Motors Inc.					
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
31	35	44	36	34	37
30	37	43	31	40	31
32	44	36	34	43	36
26	38	37	30	42	33

The Minitab system provides a dot plot and outputs the mean, median, maximum, and minimum values, and the standard deviation for the number of cars serviced at both of the dealerships over the last 24 working days.



The dot plots, shown in the center of the software output, graphically illustrate the distributions for both dealerships. The plots show the difference in the location and dispersion of the observations. By looking at the dot plots, we can see that the number of vehicles serviced at the Sheffield dealership is more widely dispersed and has a larger mean than at the Tionesta dealership. Several other features of the number of vehicles serviced are:

- Tionesta serviced the fewest cars in any day, 23.
- Sheffield serviced 26 cars during their slowest day, which is 4 cars less than the next lowest day.
- Tionesta serviced exactly 32 cars on four different days.
- The numbers of cars serviced cluster around 36 for Sheffield and 32 for Tionesta.

From the descriptive statistics, we see that Sheffield serviced a mean of 35.83 vehicles per day. Tionesta serviced a mean of 31.292 vehicles per day during the same period. So Sheffield typically services 4.54 more vehicles per day. There is also more dispersion, or variation, in the daily number of vehicles serviced at Sheffield than at Tionesta. How do we know this? The standard deviation is larger at Sheffield (4.96 vehicles per day) than at Tionesta (4.112 cars per day).

4.3 Stem-and-Leaf Displays

L02 Construct and describe a stem-and-leaf display.



Statistics in Action

John W. Tukey (1915–2000) received a PhD in mathematics from Princeton in 1939. However, when he joined the Fire Control Research Office during World War II, his interest in abstract mathematics shifted to applied statistics. He developed effective numerical and graphical methods for studying patterns in data. Among the graphics he developed are the stem-and-leaf diagram and the box-and-whisker plot or box plot. From 1960 to 1980, Tukey headed the statistical division of NBC's election night vote projection team. He became renowned in 1960 for preventing an early call of victory for Richard Nixon in the presidential election won by John F. Kennedy.

In Chapter 2, we showed how to organize data into a frequency distribution so we could summarize the raw data into a meaningful form. The major advantage to organizing the data into a frequency distribution is that we get a quick visual picture of the shape of the distribution without doing any further calculation. To put it another way, we can see where the data are concentrated and also determine whether there are any extremely large or small values. There are two disadvantages, however, to organizing the data into a frequency distribution: (1) we lose the exact identity of each value and (2) we are not sure how the values within each class are distributed. To explain, the following frequency distribution shows the number of advertising spots purchased by the 45 members of the Greater Buffalo Automobile Dealers Association in the year 2010. We observe that 7 of the 45 dealers purchased at least 90 but less than 100 spots. However, are the spots purchased within this class clustered about 90, spread evenly throughout the class, or clustered near 99? We cannot tell.

Number of Spots Purchased	Frequency
80 up to 90	2
90 up to 100	7
100 up to 110	6
110 up to 120	9
120 up to 130	8
130 up to 140	7
140 up to 150	3
150 up to 160	3
Total	45

One technique that is used to display quantitative information in a condensed form is the **stem-and-leaf display**. An advantage of the stem-and-leaf display over a frequency distribution is that we do not lose the identity of each observation. In the above example, we would not know the identity of the values in the 90 up to 100 class. To illustrate the construction of a stem-and-leaf display using the number of advertising spots purchased, suppose the seven observations in the 90 up to 100 class are: 96, 94, 93, 94, 95, 96, and 97. The **stem** value is the leading digit or digits, in this case 9. The **leaves** are the trailing digits. The stem is placed to the left of a vertical line and the leaf values to the right.

The values in the 90 up to 100 class would appear as follows:

9		6	4	3	4	5	6	7
---	--	---	---	---	---	---	---	---

It is also customary to sort the values within each stem from smallest to largest. Thus, the second row of the stem-and-leaf display would appear as follows:

9		3	4	4	5	6	6	7
---	--	---	---	---	---	---	---	---

With the stem-and-leaf display, we can quickly observe that there were two dealers that purchased 94 spots and that the number of spots purchased ranged from 93 to 97. A stem-and-leaf display is similar to a frequency distribution with more information, that is, the identity of the observations is preserved.

STEM-AND-LEAF DISPLAY A statistical technique to present a set of data. Each numerical value is divided into two parts. The leading digit(s) becomes the stem and the trailing digit the leaf. The stems are located along the vertical axis, and the leaf values are stacked against each other along the horizontal axis.

The following example will explain the details of developing a stem-and-leaf display.

Example

Listed in Table 4–1 is the number of 30-second radio advertising spots purchased by each of the 45 members of the Greater Buffalo Automobile Dealers Association last year. Organize the data into a stem-and-leaf display. Around what values do the number of advertising spots tend to cluster? What is the fewest number of spots purchased by a dealer? The largest number purchased?

TABLE 4–1 Number of Advertising Spots Purchased by Members of the Greater Buffalo Automobile Dealers Association

96	93	88	117	127	95	113	96	108	94	148	156
139	142	94	107	125	155	155	103	112	127	117	120
112	135	132	111	125	104	106	139	134	119	97	89
118	136	125	143	120	103	113	124	138			

Solution

From the data in Table 4–1, we note that the smallest number of spots purchased is 88. So we will make the first stem value 8. The largest number is 156, so we will have the stem values begin at 8 and continue to 15. The first number in Table 4–1 is 96, which will have a stem value of 9 and a leaf value of 6. Moving across the top row, the second value is 93 and the third is 88. After the first 3 data values are considered, your chart is as follows.

Stem	Leaf
8	8
9	6 3
10	
11	
12	
13	
14	
15	

Organizing all the data, the stem-and-leaf chart looks as follows.

Stem	Leaf
8	8 9
9	6 3 5 6 4 4 7
10	8 7 3 4 6 3
11	7 3 2 7 2 1 9 8 3
12	7 5 7 0 5 5 0 4
13	9 5 2 9 4 6 8
14	8 2 3
15	6 5 5

The usual procedure is to sort the leaf values from the smallest to largest. The last line, the row referring to the values in the 150s, would appear as:

15	5	5	6
----	---	---	---

The final table would appear as follows, where we have sorted all of the leaf values.

Stem	Leaf
8	8 9
9	3 4 4 5 6 6 7
10	3 3 4 6 7 8
11	1 2 2 3 3 7 7 8 9
12	0 0 4 5 5 5 7 7
13	2 4 5 6 8 9 9
14	2 3 8
15	5 5 6

You can draw several conclusions from the stem-and-leaf display. First, the minimum number of spots purchased is 88 and the maximum is 156. Two dealers purchased less than 90 spots, and three purchased 150 or more. You can observe, for example, that the three dealers who purchased more than 150 spots actually purchased 155, 155, and 156 spots. The concentration of the number of spots is between 110 and 130. There were nine dealers who purchased between 110 and 119 spots and eight who purchased between 120 and 129 spots. We can also tell that within the 120 to 129 group the actual number of spots purchased was spread evenly throughout. That is, two dealers purchased 120 spots, one dealer purchased 124 spots, three dealers purchased 125 spots, and two purchased 127 spots.

We can also generate this information on the Minitab software system. We have named the variable *Spots*. The Minitab output is below. You can find the Minitab commands that will produce this output at the end of the chapter.

The screenshot shows a Minitab worksheet titled 'Worksheet 3 ***' with a column 'C1' containing the variable 'Spots'. The data values are: 96, 93, 88, 117, 127, 95, 113, 96, 108, 94, 148, 156, 139. An overlaid 'Session' window displays the following output:

```

Stem-and-Leaf Display: Spots

Stem-and-leaf of Spots N = 45
Leaf Unit = 1.0

 2   8   89
 9   9   3445667
15  10   334678
(9) 11  122337789
21  12  00455577
13  13  2456899
 6  14   238
 3  15   556

```

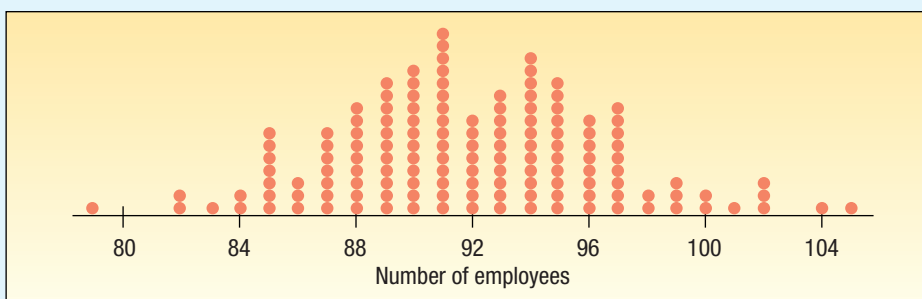
The Minitab solution provides some additional information regarding cumulative totals. In the column to the left of the stem values are numbers such as 2, 9, 15, and so on. The number 9 indicates that there are 9 observations that have occurred before the value of 100. The number 15 indicates that 15 observations have occurred prior to 110. About halfway down the column the number 9 appears in parentheses. The parentheses indicate that the middle value or median appears in that row and that there are nine values in this group. In this case, we describe the middle value as the value below which half of the observations occur. There are a total of 45 observations, so the middle value, if the data were arranged from smallest to largest, would be the 23rd observation; its value is 118. After the median, the values begin to decline. These values represent the “more than” cumulative totals. There are 21 observations of 120 or more, 13 of 130 or more, and so on.

Which is the better choice, a dot plot or a stem-and-leaf chart? This is really a matter of personal choice and convenience. For presenting data, especially with a large number of observations, you will find dot plots are more frequently used. You will see dot plots in analytical literature, marketing reports, and occasionally in annual reports. If you are doing a quick analysis for yourself, stem-and-leaf tallies are handy and easy, particularly on a smaller set of data.

Self-Review 4-1



1. The number of employees at each of the 142 Home Depot Stores in the Southeast region is shown in the following dot plot.



- (a) What are the maximum and minimum numbers of employees per store?
 - (b) How many stores employ 91 people?
 - (c) Around what values does the number of employees per store tend to cluster?
2. The rate of return for 21 stocks is:

8.3	9.6	9.5	9.1	8.8	11.2	7.7	10.1	9.9	10.8	
10.2	8.0	8.4	8.1	11.6	9.6	8.8	8.0	10.4	9.8	9.2

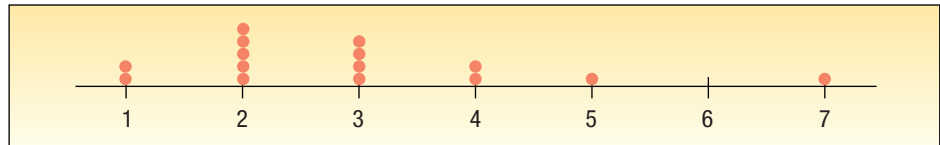
Organize this information into a stem-and-leaf display.

- (a) How many rates are less than 9.0?
- (b) List the rates in the 10.0 up to 11.0 category.
- (c) What is the median?
- (d) What are the maximum and the minimum rates of return?

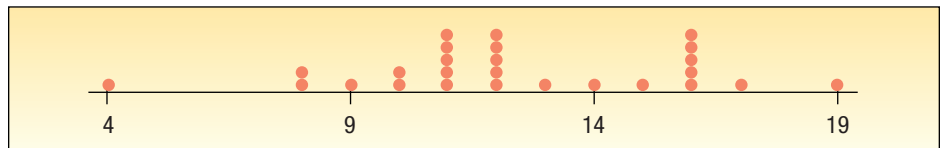
Exercises

connect™

- Describe the differences between a histogram and a dot plot. When might a dot plot be better than a histogram?
- Describe the differences between a histogram and a stem-and-leaf display.
- Consider the following chart.



- What is this chart called?
 - How many observations are in the study?
 - What are the maximum and the minimum values?
 - Around what values do the observations tend to cluster?
- The following chart reports the number of cell phones sold at Radio Shack for the last 26 days.




- What are the maximum and the minimum number of cell phones sold in a day?
 - What is a typical number of cell phones sold?
- The first row of a stem-and-leaf chart appears as follows: 62 | 1 3 3 7 9. Assume whole number values.
 - What is the “possible range” of the values in this row?
 - How many data values are in this row?
 - List the actual values in this row of data.
 - The third row of a stem-and-leaf chart appears as follows: 21 | 0 1 3 5 7 9. Assume whole number values.
 - What is the “possible range” of the values in this row?
 - How many data values are in this row?
 - List the actual values in this row of data.
 - The following stem-and-leaf chart from the Minitab software shows the number of units produced per day in a factory.

1	3	8
1	4	
2	5	6
9	6	0133559
(7)	7	0236778
9	8	59
7	9	00156
2	10	36


- How many days were studied?
- How many observations are in the first class?
- What are the minimum value and the maximum value?

- d. List the actual values in the fourth row.
 - e. List the actual values in the second row.
 - f. How many values are less than 70?
 - g. How many values are 80 or more?
 - h. What is the median?
 - i. How many values are between 60 and 89, inclusive?
8. The following stem-and-leaf chart reports the number of movies rented per day at Video Connection on the corner of Fourth and Main Streets.

3	12	689
6	13	123
10	14	6889
13	15	589
15	16	35
20	17	24568
23	18	268
(5)	19	13456
22	20	034679
16	21	2239
12	22	789
9	23	00179
4	24	8
3	25	13
1	26	
1	27	0

- a. How many days were studied?
 - b. How many observations are in the last class?
 - c. What are the maximum and the minimum values in the entire set of data?
 - d. List the actual values in the fourth row.
 - e. List the actual values in the next to the last row.
 - f. On how many days were less than 160 movies rented?
 - g. On how many days were 220 or more movies rented?
 - h. What is the middle value?
 - i. On how many days were between 170 and 210 movies rented?
9. A survey of the number of cell phone calls made by a sample of Verizon subscribers last week revealed the following information. Develop a stem-and-leaf chart. How many calls did a typical subscriber make? What were the maximum and the minimum number of calls made? 

52	43	30	38	30	42	12	46	39
37	34	46	32	18	41	5		

10. Aloha Banking Co. is studying ATM use in suburban Honolulu. A sample of 30 ATMs showed they were used the following number of times yesterday. Develop a stem-and-leaf chart. Summarize the number of times each ATM was used. What was the typical, minimum, and maximum number of times each ATM was used? 

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

4.4 Measures of Position

L03 Identify and compute measures of position.

Quartiles divide a set of data into four parts.

The standard deviation is the most widely used measure of dispersion. However, there are other ways of describing the variation or spread in a set of data. One method is to determine the *location* of values that divide a set of observations into equal parts. These measures include **quartiles**, **deciles**, and **percentiles**.

Quartiles divide a set of observations into four equal parts. To explain further, think of any set of values arranged from smallest to largest. In Chapter 3, we called the middle value of a set of data arranged from smallest to largest the median. That is, 50 percent of the observations are larger than the median and 50 percent are smaller. The median is a measure of location because it pinpoints the center of the data. In a similar fashion, **quartiles** divide a set of observations into four equal parts. The first quartile, usually labeled Q_1 , is the value below which 25 percent of the observations occur, and the third quartile, usually labeled Q_3 , is the value below which 75 percent of the observations occur. Logically, Q_2 is the median. Q_1 can be thought of as the “median” of the lower half of the data and Q_3 the “median” of the upper half of the data.

Similarly, **deciles** divide a set of observations into 10 equal parts and **percentiles** into 100 equal parts. So if you found that your GPA was in the 8th decile at your university, you could conclude that 80 percent of the students had a GPA lower than yours and 20 percent had a higher GPA. A GPA in the 33rd percentile means that 33 percent of the students have a lower GPA and 67 percent have a higher GPA. Percentile scores are frequently used to report results on such national standardized tests as the SAT, ACT, GMAT (used to judge entry into many master of business administration programs), and LSAT (used to judge entry into law school).

Quartiles, Deciles, and Percentiles

To formalize the computational procedure, let L_p refer to the location of a desired percentile. So if we want to find the 33rd percentile we would use L_{33} and if we wanted the median, the 50th percentile, then L_{50} . The number of observations is n , so if we want to locate the median, its position is at $(n + 1)/2$, or we could write this as $(n + 1)(P/100)$, where P is the desired percentile.

LOCATION OF A PERCENTILE

$$L_p = (n + 1) \frac{P}{100}$$

[4-1]

An example will help to explain further.

Example

Listed below are the commissions earned last month by a sample of 15 brokers at Salomon Smith Barney's Oakland, California office. Salomon Smith Barney is an investment company with offices located throughout the United States.

\$2,038	\$1,758	\$1,721	\$1,637	\$2,097	\$2,047	\$2,205	\$1,787	\$2,287
1,940	2,311	2,054	2,406	1,471	1,460			

Locate the median, the first quartile, and the third quartile for the commissions earned.

Solution

The first step is to sort the data from the smallest commission to the largest.

\$1,460	\$1,471	\$1,637	\$1,721	\$1,758	\$1,787	\$1,940	\$2,038
2,047	2,054	2,097	2,205	2,287	2,311	2,406	

The median value is the observation in the center. The center value or L_{50} is located at $(n + 1)(50/100)$, where n is the number of observations. In this case, that is position number 8, found by $(15 + 1)(50/100)$. The eighth largest commission is \$2,038. So we conclude this is the median and that half the brokers earned commissions more than \$2,038 and half earned less than \$2,038.

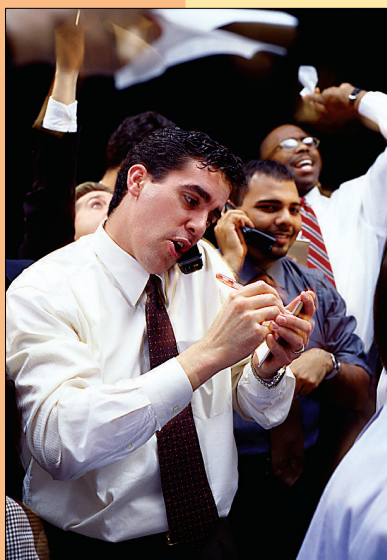
Recall the definition of a quartile. Quartiles divide a set of observations into four equal parts. Hence 25 percent of the observations will be less than the first quartile. Seventy-five percent of the observations will be less than the third quartile. To locate the first quartile, we use formula (4-1), where $n = 15$ and $P = 25$:

$$L_{25} = (n + 1) \frac{P}{100} = (15 + 1) \frac{25}{100} = 4$$

and to locate the third quartile, $n = 15$ and $P = 75$:

$$L_{75} = (n + 1) \frac{P}{100} = (15 + 1) \frac{75}{100} = 12$$

Therefore, the first and third quartile values are located at positions 4 and 12, respectively. The fourth value in the ordered array is \$1,721 and the twelfth is \$2,205. These are the first and third quartiles.



In the above example, the location formula yielded a whole number. That is, we wanted to find the first quartile and there were 15 observations, so the location formula indicated we should find the fourth ordered value. What if there were 20 observations in the sample, that is $n = 20$, and we wanted to locate the first quartile? From the location formula (4-1):

$$L_{25} = (n + 1) \frac{P}{100} = (20 + 1) \frac{25}{100} = 5.25$$

We would locate the fifth value in the ordered array and then move .25 of the distance between the fifth and sixth values and report that as the first quartile. Like the median, the quartile does not need to be one of the actual values in the data set.

To explain further, suppose a data set contained the six values: 91, 75, 61, 101, 43, and 104. We want to locate the first quartile. We order the values from smallest to largest: 43, 61, 75, 91, 101, and 104. The first quartile is located at

$$L_{25} = (n + 1) \frac{P}{100} = (6 + 1) \frac{25}{100} = 1.75$$

The position formula tells us that the first quartile is located between the first and the second value and that it is .75 of the distance between the first and the second values. The first value is 43 and the second is 61. So the distance between these two values is 18. To locate the first quartile, we need to move .75 of the distance between the first and second values, so $.75(18) = 13.5$. To complete the procedure, we add 13.5 to the first value and report that the first quartile is 56.5.

We can extend the idea to include both deciles and percentiles. To locate the 23rd percentile in a sample of 80 observations, we would look for the 18.63 position.

$$L_{23} = (n + 1) \frac{P}{100} = (80 + 1) \frac{23}{100} = 18.63$$

To find the value corresponding to the 23rd percentile, we would locate the 18th value and the 19th value and determine the distance between the two values. Next, we would multiply this difference by 0.63 and add the result to the smaller value. The result would be the 23rd percentile.

With a statistical software package, it is easy to sort the data from smallest to largest and to locate percentiles and deciles. Both Minitab and Excel provide summary statistics. Listed below is the output from the Minitab system for the Smith Barney commission data. Included are the first and third quartiles, mean, median, and standard deviation. We conclude that 25 percent of the commissions earned were less than \$1,721 and 75 percent were less than \$2,205. The same values were reported in the Example on the previous page.

The screenshot shows a Minitab worksheet titled 'Worksheet 1 ***' with columns C1 through C12. Column C1 is labeled 'Commissions' and contains the following values: 1460, 1471, 1637, 1721, 1758, 1787, 1940. A 'Session' window is open, displaying 'Descriptive Statistics: Commissions'.

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Commissions	15	0	1947.9	77.1	298.8	1460.0	1721.0	2038.0	2205.0	2406.0

Excel and MegaStat, which is based on Excel, will also calculate quartiles and output the results. However, the method of solution used is slightly different. To simplify the issues, assume the data set contains an odd number of values. The method described in the Example, and supported by Minitab, for the first quartile is:

1. Find the median of the set of n observations.
2. Focus only on the observations *below* the median and find the median of these values. That is, do not consider the median as part of the new data set.
3. Report this value as the first quartile.

So in our Smith Barney commissions data, the median commission is the 8th observation in the set of 15 observations. This commission is \$2,038, so we focus on the seven observations less than \$2,038. The median of these seven observations is located in position 4 and that value is \$1,721, the value found in our Example and in the Minitab output.

Below is the Excel spreadsheet. Also shown are the first and third quartiles for the Smith Barney commission data. Notice the results differ. Again, to simplify the situation, assume there are an odd number of values. Excel finds the median according to the following method:

1. Find the median of the set of n observations.
2. Focus on all the observation equal to or less than the median. That is, include the median in the new subset of data.
3. Find the median of this set of values.
4. Report this value as the first quartile.

In our Smith Barney commission data, the median of the original 15 observations is \$2,038. So our new set of values is the eight ordered observations between \$1,460 and \$2,038. The median is halfway between \$1,721 and \$1,758, or \$1,739 as reported by Excel.

	A	B	C	D
1	\$1,460.00			
2	\$1,471.00			
3	\$1,637.00			
4	\$1,721.00		Quartile 1	\$1,739.50
5	\$1,758.00			
6	\$1,787.00		Quartile 3	\$2,151.00
7	\$1,940.00			
8	\$2,038.00			
9	\$2,047.00			
10	\$2,054.00			
11	\$2,097.00			
12	\$2,205.00			
13	\$2,287.00			
14	\$2,311.00			
15	\$2,406.00			

So the essential difference between the two methods is:

- In the Minitab method, the median is not included in the subset of data.
- In the Excel method, the median is included in the subset of data.

In this example, there was an odd number of observations. What happens in the Excel method if there is an even number of observations? Instead of using formula 4-1 to find the location, it uses $0.25n + 0.75$ to locate the position of the first quartile and $0.75n + 0.25$ to locate the position of the third quartile.

Is this difference important? No, usually it is just a nuisance. Statisticians usually prefer the first method discussed. When the sample is large, the difference in the results from the two methods is small. For example, recall the Applewood Auto Group data in which the profit data on the sale of 180 vehicles is reported. Below are the Minitab and Excel results. Not much difference, only \$7.00 over 180 vehicles! Reporting either value would make little difference in the interpretation.

Session										
Descriptive Statistics: Profit										
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Profit	180	0	1843.2	48.0	643.6	294.0	1415.5	1882.5	2275.5	3292.0

APPLEWOOD AUTO GROUP				
	A	B	C	D
1	Age	Profit		
2	44	\$294		
3	40	\$323		
4	42	\$335	Quartile 1	1422.50
5	40	\$352	Quartile 3	2268.50
6	46	\$369		
7	53	\$377		
8	30	\$443		
9	40	\$482		
10	37	\$732		
11	30	\$754		
12	62	\$783		
13	45	\$820		
14	50	\$842		

Self-Review 4–2

The Quality Control department of Plainsville Peanut Company is responsible for checking the weight of the 8-ounce jar of peanut butter. The weights of a sample of nine jars produced last hour are:

7.69 7.72 7.8 7.86 7.90 7.94 7.97 8.06 8.09


- What is the median weight?
- Determine the weights corresponding to the first and third quartiles.

Exercises

connect™

- Determine the median and the values corresponding to the first and third quartiles in the following data. 


46 47 49 49 51 53 54 54 55 55 59


- Determine the median and the values corresponding to the first and third quartiles in the following data. 

5.24 6.02 6.67 7.30 7.59 7.99 8.03 8.35 8.81 9.45
9.61 10.37 10.39 11.86 12.22 12.71 13.07 13.59 13.89 15.42

- The Thomas Supply Company Inc. is a distributor of gas-powered generators. As with any business, the length of time customers take to pay their invoices is important. Listed below, arranged from smallest to largest, is the time, in days, for a sample of The Thomas Supply Company Inc. invoices.

13 13 13 20 26 27 31 34 34 34 35 35 36 37 38
41 41 41 45 47 47 47 50 51 53 54 56 62 67 82

- Determine the first and third quartiles.
- Determine the second decile and the eighth decile.
- Determine the 67th percentile. 

14. Kevin Horn is the national sales manager for National Textbooks Inc. He has a sales staff of 40 who visit college professors all over the United States. Each Saturday morning he requires his sales staff to send him a report. This report includes, among other things, the number of professors visited during the previous week. Listed below, ordered from smallest to largest, are the number of visits last week. 

38	40	41	45	48	48	50	50	51	51	52	52	53	54	55	55	55	56	56	57
59	59	59	62	62	62	63	64	65	66	66	67	67	69	69	71	77	78	79	79

- Determine the median number of calls.
- Determine the first and third quartiles.
- Determine the first decile and the ninth decile.
- Determine the 33rd percentile.

Box Plots

L04 Construct and analyze a box plot.

A **box plot** is a graphical display, based on quartiles, that helps us picture a set of data. To construct a box plot, we need only five statistics: the minimum value, Q_1 (the first quartile), the median, Q_3 (the third quartile), and the maximum value. An example will help to explain.

Example

Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

Minimum value = 13 minutes

Q_1 = 15 minutes

Median = 18 minutes

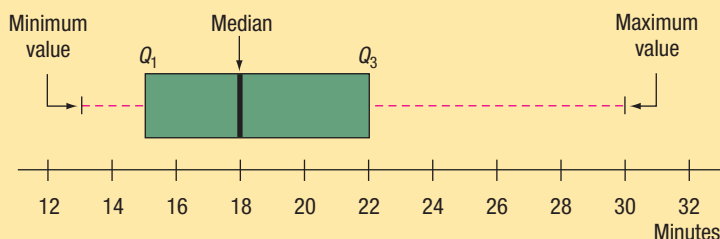
Q_3 = 22 minutes

Maximum value = 30 minutes

Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

Solution

The first step in drawing a box plot is to create an appropriate scale along the horizontal axis. Next, we draw a box that starts at Q_1 (15 minutes) and ends at Q_3 (22 minutes). Inside the box we place a vertical line to represent the median (18 minutes). Finally, we extend horizontal lines from the box out to the minimum value (13 minutes) and the maximum value (30 minutes). These horizontal lines outside of the box are sometimes called “whiskers” because they look a bit like a cat’s whiskers.



The box plot shows that the middle 50 percent of the deliveries take between 15 minutes and 22 minutes. The distance between the ends of the box, 7 minutes, is the **interquartile range**. The interquartile range is the distance between the first and the third quartile. It shows the spread or dispersion of the majority of deliveries.

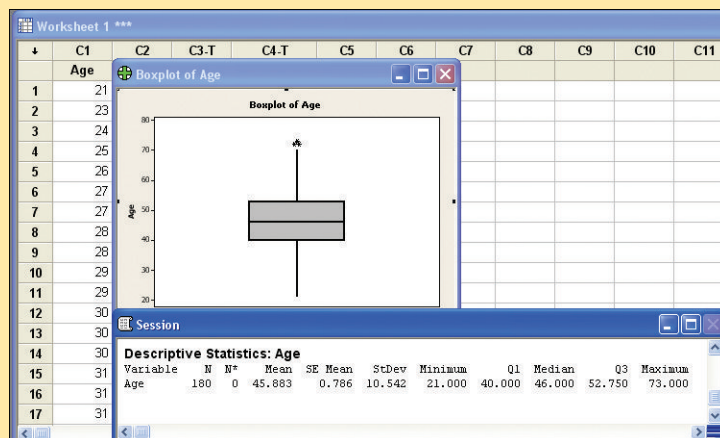
The box plot also reveals that the distribution of delivery times is positively skewed. Recall from page 70 in Chapter 3 that we defined skewness as the lack of symmetry in a set of data. How do we know this distribution is positively skewed? In this case, there are actually two pieces of information that suggest this. First, the dashed line to the right of the box from 22 minutes (Q_3) to the maximum time of 30 minutes is longer than the dashed line from the left of 15 minutes (Q_1) to the minimum value of 13 minutes. To put it another way, the 25 percent of the data larger than the third quartile is more spread out than the 25 percent less than the first quartile. A second indication of positive skewness is that the median is not in the center of the box. The distance from the first quartile to the median is smaller than the distance from the median to the third quartile. We know that the number of delivery times between 15 minutes and 18 minutes is the same as the number of delivery times between 18 minutes and 22 minutes.

Example

Refer to the Applewood Auto Group data. Develop a box plot for the variable age of the buyer. What can we conclude about the distribution of the age of the buyer?

Solution

The Minitab statistical software system was used to develop the following chart and summary statistics.



The median age of the purchaser was 46 years, 25 percent of the purchasers were less than 40 years of age, and 25 percent were more than 52.75 years of age. Based on the summary information and the box plot, we conclude:

- Fifty percent of the purchasers were between the ages of 40 and 52.75 years.
- The distribution of ages is symmetric. There are two reasons for this conclusion. The length of the whisker above 52.75 years (Q_3) is about the same length as the whisker below 40 years (Q_1). Also, the area in the box between 40 years and the median of 46 years is about the same as the area between the median and 52.75.

There are three asterisks (*) above 70 years. What do they indicate? In a box plot, an asterisk identifies an **outlier**. An outlier is a value that is inconsistent with the rest of the data. It is defined as a value that is more than 1.5 times the interquartile range smaller than Q_1 or larger than Q_3 . In this example, an outlier would be a value larger than 71.875 years, found by:

$$\text{Outlier} > Q_3 + 1.5(Q_3 - Q_1) = 52.75 + 1.5(52.75 - 40) = 71.875$$

An outlier would also be a value less than 20.875 years.

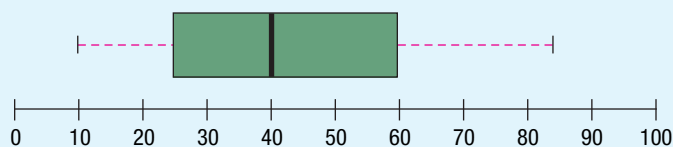
$$\text{Outlier} < Q_1 - 1.5(Q_3 - Q_1) = 40 - 1.5(52.75 - 40) = 20.875$$

From the box plot, we conclude that there are three purchasers 72 years of age or older and none less than 21 years of age. Technical note: In some cases, a single asterisk may represent more than one observation, because of the limitations of the software and space available. It is a good idea to check the actual data. In this instance, there are three purchasers 72 years old or older; two are 72 and one is 73.

Self-Review 4–3



The following box plot shows the assets in millions of dollars for credit unions in Seattle, Washington.

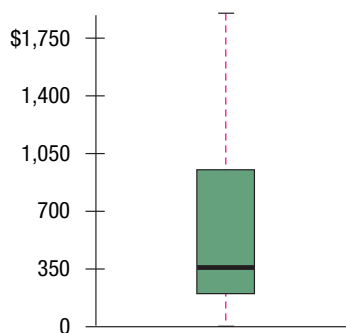


What are the smallest and largest values, the first and third quartiles, and the median? Would you agree that the distribution is symmetrical? Are there any outliers?

Exercises

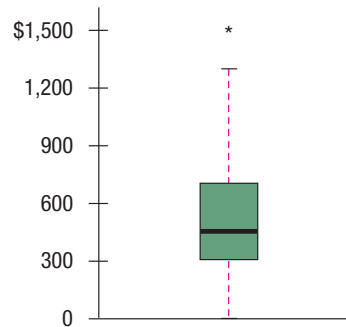



15. The box plot below shows the amount spent for books and supplies per year by students at four-year public colleges.



- Estimate the median amount spent.
- Estimate the first and third quartiles for the amount spent.
- Estimate the interquartile range for the amount spent.
- Beyond what point is a value considered an outlier?

- e. Identify any outliers and estimate their value.
 f. Is the distribution symmetrical or positively or negatively skewed?
16. The box plot shows the undergraduate in-state charge per credit hour at four-year public colleges.



- a. Estimate the median.
 b. Estimate the first and third quartiles.
 c. Determine the interquartile range.
 d. Beyond what point is a value considered an outlier?
 e. Identify any outliers and estimate their value.
 f. Is the distribution symmetrical or positively or negatively skewed?
17. In a study of the gasoline mileage of model year 2011 automobiles, the mean miles per gallon was 27.5 and the median was 26.8. The smallest value in the study was 12.70 miles per gallon, and the largest was 50.20. The first and third quartiles were 17.95 and 35.45 miles per gallon, respectively. Develop a box plot and comment on the distribution. Is it a symmetric distribution?
18. A sample of 28 time shares in the Orlando, Florida, area revealed the following daily charges for a one-bedroom suite. For convenience, the data are ordered from smallest to largest. Construct a box plot to represent the data. Comment on the distribution. Be sure to identify the first and third quartiles and the median. 

\$116	\$121	\$157	\$192	\$207	\$209	\$209
229	232	236	236	239	243	246
260	264	276	281	283	289	296
307	309	312	317	324	341	353

4.5 Skewness

In Chapter 3, we described measures of central location for a set of observations by reporting the mean, median, and mode. We also described measures that show the amount of spread or variation in a set of data, such as the range and the standard deviation.

Another characteristic of a set of data is the shape. There are four shapes commonly observed: symmetric, positively skewed, negatively skewed, and bimodal. In a **symmetric** set of observations the mean and median are equal and the data values are evenly spread around these values. The data values below the mean and median are a mirror image of those above. A set of values is **skewed to the right** or **positively skewed** if there is a single peak and the values extend much further to the right of the peak than to the left of the peak. In this case, the mean is larger than the median. In a **negatively skewed** distribution there is a single peak but the observations extend further to the left, in the negative direction, than to the right. In a negatively skewed distribution, the mean is smaller than the median. Positively skewed

L05 Compute and describe the coefficient of skewness.

Skewness shows the lack of symmetry in a set of observations.

distributions are more common. Salaries often follow this pattern. Think of the salaries of those employed in a small company of about 100 people. The president and a few top executives would have very large salaries relative to the other workers and hence the distribution of salaries would exhibit positive skewness. A **bimodal distribution** will have two or more peaks. This is often the case when the values are from two or more populations. This information is summarized in Chart 4–1.

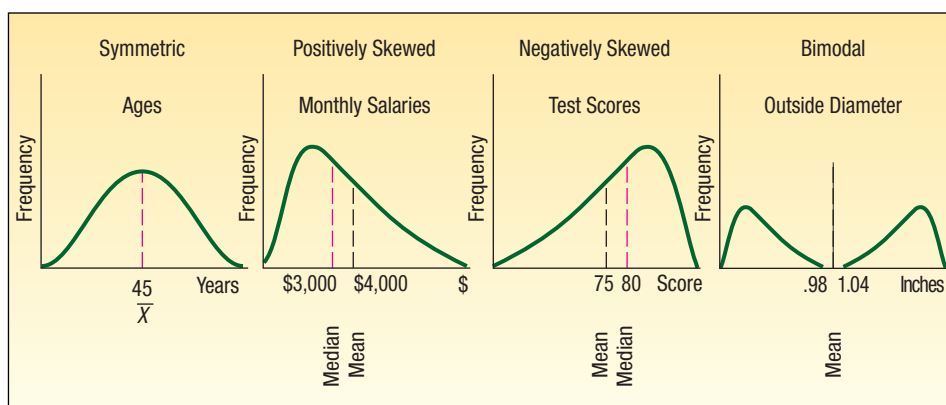


CHART 4–1 Shapes of Frequency Polygons

There are several formulas in statistical literature used to calculate skewness. The simplest, developed by Professor Karl Pearson (1857–1936), is based on the difference between the mean and the median.

PEARSON'S COEFFICIENT OF SKEWNESS

$$sk = \frac{3(\bar{X} - \text{Median})}{s} \quad [4-2]$$

Using this relationship, coefficient of skewness can range from -3 up to 3 . A value near -3 , such as -2.57 , indicates considerable negative skewness. A value such as 1.63 indicates moderate positive skewness. A value of 0 , which will occur when the mean and median are equal, indicates the distribution is symmetrical and that there is no skewness present.

In this text, we present output from the statistical software packages Minitab and Excel. Both of these software packages compute a value for the coefficient of skewness that is based on the cubed deviations from the mean. The formula is:

SOFTWARE COEFFICIENT OF SKEWNESS

$$sk = \frac{n}{(n-1)(n-2)} \left[\sum \left(\frac{X - \bar{X}}{s} \right)^3 \right] \quad [4-3]$$

Formula (4–3) offers an insight into skewness. The right-hand side of the formula is the difference between each value and the mean, divided by the standard deviation. That is the portion $(X - \bar{X})/s$ of the formula. This idea is called **standardizing**. We will discuss the idea of standardizing a value in more detail in Chapter 7 when we describe the normal probability distribution. At this point, observe that the result is to report the difference between each value and the mean in units



Statistics in Action

The late Stephen Jay Gould (1941–2002) was a professor of zoology and professor of geology at Harvard University. In 1982, he was diagnosed with cancer and had an expected survival time of eight months. However, never to be discouraged, his research showed that the distribution of survival time is dramatically skewed to the right and showed that not only do 50 percent of similar cancer patients survive more than 8 months, but that the survival time could be years rather than months! Based on his experience, he wrote a widely published essay titled, “The Median Is not the Message.”

of the standard deviation. If this difference is positive, the particular value is larger than the mean; if the value is negative, the standardized quantity is smaller than the mean. When we cube these values, we retain the information on the direction of the difference. Recall that in the formula for the standard deviation [see formula (3-11)] we squared the difference between each value and the mean, so that the result was all non-negative values.

If the set of data values under consideration is symmetric, when we cube the standardized values and sum over all the values, the result would be near zero. If there are several large values, clearly separate from the others, the sum of the cubed differences would be a large positive value. Several values much smaller will result in a negative cubed sum.

An example will illustrate the idea of skewness.

Example

Following are the earnings per share for a sample of 15 software companies for the year 2010. The earnings per share are arranged from smallest to largest.

\$0.09	\$0.13	\$0.41	\$0.51	\$ 1.12	\$ 1.20	\$ 1.49	\$3.18
3.50	6.36	7.83	8.92	10.13	12.99	16.40	

Compute the mean, median, and standard deviation. Find the coefficient of skewness using Pearson's estimate and the software methods. What is your conclusion regarding the shape of the distribution?

Solution

These are sample data, so we use formula (3-2) to determine the mean

$$\bar{X} = \frac{\sum X}{n} = \frac{\$74.26}{15} = \$4.95$$

The median is the middle value in a set of data, arranged from smallest to largest. In this case, the middle value is \$3.18, so the median earnings per share is \$3.18.

We use formula (3-11) on page 84 to determine the sample standard deviation.

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{(\$0.09 - \$4.95)^2 + \cdots + (\$16.40 - \$4.95)^2}{15 - 1}} = \$5.22$$

Pearson's coefficient of skewness is 1.017, found by

$$sk = \frac{3(\bar{X} - \text{Median})}{s} = \frac{3(\$4.95 - \$3.18)}{\$5.22} = 1.017$$

This indicates there is moderate positive skewness in the earnings per share data.

We obtain a similar, but not exactly the same, value from the software method. The details of the calculations are shown in Table 4-2. To begin, we find the difference between each earnings per share value and the mean and divide this result by the standard deviation. Recall that we referred to this as standardizing. Next, we cube, that is, raise to the third power, the result of the first step. Finally, we sum the cubed values. The details for the first company, that is, the company with an earnings per share of \$0.09, are:

$$\left(\frac{X - \bar{X}}{s}\right)^3 = \left(\frac{0.09 - 4.95}{5.22}\right)^3 = (-0.9310)^3 = -0.8070$$

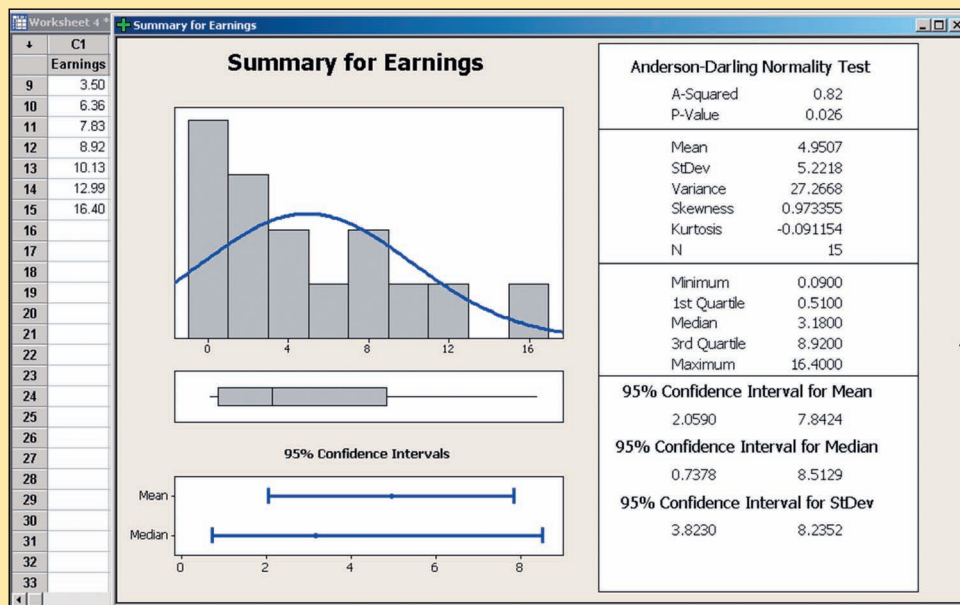
TABLE 4-2 Calculation of the Coefficient of Skewness

Earnings per Share	$\frac{(X - \bar{X})}{s}$	$\left(\frac{X - \bar{X}}{s}\right)^3$
0.09	-0.9310	-0.8070
0.13	-0.9234	-0.7873
0.41	-0.8697	-0.6579
0.51	-0.8506	-0.6154
1.12	-0.7337	-0.3950
1.20	-0.7184	-0.3708
1.49	-0.6628	-0.2912
3.18	-0.3391	-0.0390
3.50	-0.2778	-0.0214
6.36	0.2701	0.0197
7.83	0.5517	0.1679
8.92	0.7605	0.4399
10.13	0.9923	0.9772
12.99	1.5402	3.6539
16.40	2.1935	10.5537
		11.8274

When we sum the 15 cubed values, the result is 11.8274. That is, the term $\sum[(X - \bar{X})/s]^3 = 11.8274$. To find the coefficient of skewness, we use formula (4-3), with $n = 15$.

$$sk = \frac{n}{(n-1)(n-2)} \sum \left(\frac{X - \bar{X}}{s} \right)^3 = \frac{15}{(15-1)(15-2)} (11.8274) = 0.975$$

We conclude that the earnings per share values are somewhat positively skewed. The following chart, from Minitab, reports the descriptive measures, such as the mean, median, and standard deviation of the earnings per share data. Also included are the coefficient of skewness and a histogram with a bell-shaped curve superimposed.



Self-Review 4–4




A sample of five data entry clerks employed in the Horry County Tax Office revised the following number of tax records last hour: 73, 98, 60, 92, and 84.

- Find the mean, median, and the standard deviation.
- Compute the coefficient of skewness using Pearson's method.
- Calculate the coefficient of skewness using the software method.
- What is your conclusion regarding the skewness of the data?


Exercises

connect™

For Exercises 19–22:

- Determine the mean, median, and the standard deviation.
 - Determine the coefficient of skewness using Pearson's method.
 - Determine the coefficient of skewness using the software method.
19. The following values are the starting salaries, in \$000, for a sample of five accounting graduates who accepted positions in public accounting last year. 


36.0	26.0	33.0	28.0	31.0
------	------	------	------	------

20. Listed below are the salaries, in \$000, for a sample of 15 chief financial officers in the electronics industry. 

\$516.0	\$548.0	\$566.0	\$534.0	\$586.0	\$529.0
546.0	523.0	538.0	523.0	551.0	552.0
486.0	558.0	574.0			

21. Listed below are the commissions earned (\$000) last year by the sales representatives at Furniture Patch Inc. 

\$ 3.9	\$ 5.7	\$ 7.3	\$10.6	\$13.0	\$13.6	\$15.1	\$15.8	\$17.1
17.4	17.6	22.3	38.6	43.2	87.7			

22. Listed below are the salaries in \$000 of the 25 players on the opening day roster of the 2010 New York Yankees Major League Baseball team. 

Player	Salary (\$000)	Position	Player	Salary (\$000)	Position
Aceves, Alfredo	435.7	Pitcher	Pena, Ramiro	412.1	Infielder
Burnett, A.J.	16,500.0	Pitcher	Pettitte, Andy	11,750.0	Pitcher
Cano, Robinson	9,000.0	Second Baseman	Posada, Jorge	13,100.0	Catcher
Cervelli, Francisco	410.8	Catcher	Rivera, Mariano	15,000.0	Pitcher
Chamberlain, Joba	488.0	Pitcher	Robertson, David	426.7	Pitcher
Gardner, Brett	452.5	Outfielder	Rodriguez, Alex	33,000.0	Third Baseman
Granderson, Curtis	5,500.0	Outfielder	Sabathia, CC	24,285.7	Pitcher
Hughes, Phil	447.0	Pitcher	Swisher, Nick	6,850.0	Outfielder
Jeter, Derek	22,600.0	Shortstop	Teixeira, Mark	20,625.0	First Baseman
Johnson, Nick	5,500.0	First Baseman	Thames, Marcus	900.0	Outfielder
Marte, Damaso	4,000.0	Pitcher	Vazquez, Javier	11,500.0	Pitcher
Mitre, Sergio	850.0	Pitcher	Winn, Randy	1,100.0	Outfielder
Park, Chan Ho	1,200.0	Pitcher			

4.6 Describing the Relationship between Two Variables



In Chapter 2 and the first section of this chapter we presented graphical techniques to summarize the distribution of a single variable. We used a histogram in Chapter 2 to summarize the profit on vehicles sold by the Applewood Auto Group. Earlier in this chapter we used dot plots and stem-and-leaf displays to visually summarize a set of data. Because we are studying a single variable, we refer to this as **univariate** data.

There are situations where we wish to study and visually portray the relationship between two variables. When we study the relationship between two variables, we refer to the data as **bivariate**. Data analysts frequently wish to understand the relationship between two variables. Here are some examples:

- Tybo and Associates is a law firm that advertises extensively on local TV. The partners are considering increasing their advertising budget. Before doing so, they would like to know the relationship between the amount spent per month on advertising and the total amount of billings for that month. To put it another way, will increasing the amount spent on advertising result in an increase in billings?
- Coastal Realty is studying the selling prices of homes. What variables seem to be related to the selling price of homes? For example, do larger homes sell for more than smaller ones? Probably. So Coastal might study the relationship between the area in square feet and the selling price.
- Dr. Stephen Givens is an expert in human development. He is studying the relationship between the height of fathers and the height of their sons. That is, do tall fathers tend to have tall children? Would you expect Shaquille O'Neal, the 7'1", 335-pound professional basketball player, to have relatively tall sons?

One graphical technique we use to show the relationship between variables is called a **scatter diagram**.

To draw a scatter diagram we need two variables. We scale one variable along the horizontal axis (X-axis) of a graph and the other variable along the vertical axis (Y-axis). Usually one variable depends to some degree on the other. In the third example above, the height of the son *depends* on the height of the father. So we scale the height of the father on the horizontal axis and that of the son on the vertical axis.

We can use statistical software, such as Excel, to perform the plotting function for us. *Caution:* You should always be careful of the scale. By changing the scale of either the vertical or the horizontal axis, you can affect the apparent visual strength of the relationship.

Following are three scatter diagrams (Chart 4–2). The one on the left shows a rather strong positive relationship between the age in years and the maintenance cost last year for a sample of 10 buses owned by the city of Cleveland, Ohio. Note that as the age of the bus increases, the yearly maintenance cost also increases. The example in the center, for a sample of 20 vehicles, shows a rather strong indirect relationship between the odometer reading and the auction price. That is, as the number of miles driven increases, the auction price decreases. The example on the right depicts the relationship between the height and yearly salary for a sample of 15 shift supervisors. This graph indicates there is little relationship between their height and yearly salary.

L06 Create and interpret a scatter diagram.

A scatter diagram is used as a way to understand the relationship between two variables.

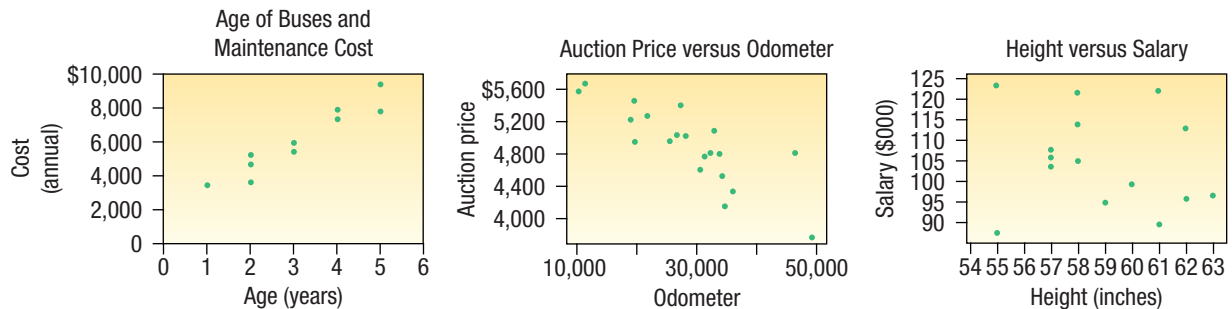


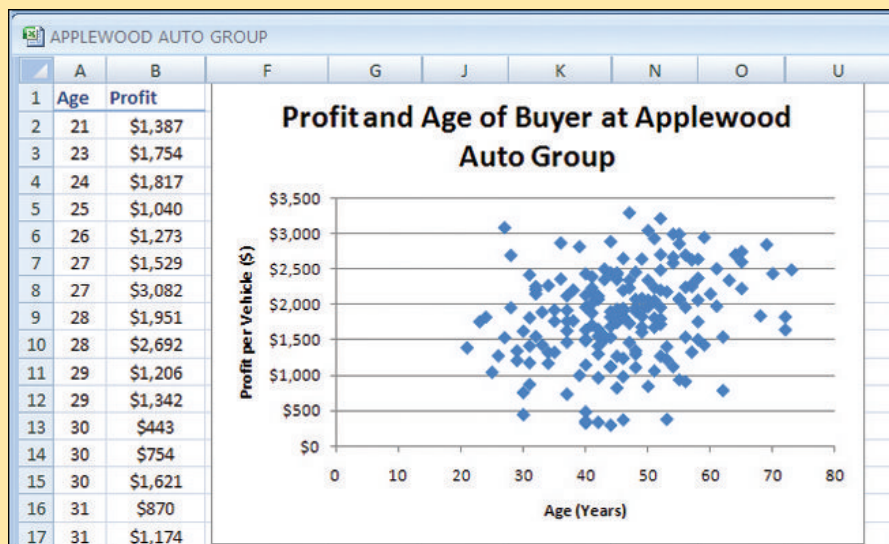
CHART 4-2 Three Examples of Scatter Diagrams.

Example

In the introduction to Chapter 2, we presented data from the Applewood Auto Group. We gathered information concerning several variables, including the profit earned from the sale of 180 vehicles sold last month. In addition to the amount of profit on each sale, one of the other variables is the age of the purchaser. Is there a relationship between the profit earned on a vehicle sale and the age of the purchaser? Would it be reasonable to conclude that more profit is made on vehicles purchased by older buyers?

Solution

We can investigate the relationship between vehicle profit and the age of the buyer with a scatter diagram. We scale age on the horizontal, or X-axis, and the profit on the vertical, or Y-axis. We use Microsoft Excel to develop the scatter diagram. The Excel commands necessary for the output are shown in the **Software Commands** section at the end of the chapter.



The scatter diagram shows a rather weak positive relationship between the two variables. It does not appear there is much relationship between the vehicle profit and the age of the buyer. In Chapter 13, we will study the relationship between variables more extensively, even calculating several numerical measures to express the relationship between variables.

In the preceding example, there is a weak positive, or direct, relationship between the variables. There are, however, many instances where there is a relationship between the variables, but that relationship is inverse or negative. For example:

- The value of a vehicle and the number of miles driven. As the number of miles increases, the value of the vehicle decreases.
- The premium for auto insurance and the age of the driver. Auto rates tend to be the highest for young adults and less for older people.
- For many law enforcement personnel, as the number of years on the job increases, the number of traffic citations decreases. This may be because personnel become more liberal in their interpretations or they may be in supervisor positions and not in a position to issue as many citations. But in any event, as age increases, the number of citations decreases.

A scatter diagram requires that both of the variables be at least interval scale. In the Applewood Auto Group example, both age and vehicle profit are ratio scale variables. Height is also ratio scale as used in the discussion of the relationship between the height of fathers and the height of their sons. What if we wish to study the relationship between two variables when one or both are nominal or ordinal scale? In this case, we tally the results in a **contingency table**.

L07 Develop and explain a contingency table.

CONTINGENCY TABLE A table used to classify observations according to two identifiable characteristics.

A contingency table is a cross-tabulation that simultaneously summarizes two variables of interest. For example:

- Students at a university are classified by gender and class rank.
- A product is classified as acceptable or unacceptable and by the shift (day, afternoon, or night) on which it is manufactured.
- A voter in a school bond referendum is classified as to party affiliation (Democrat, Republican, other) and the number of children that voter has attending school in the district (0, 1, 2, etc.).

Example

There are four dealerships in the Applewood Auto Group. Suppose we want to compare the profit earned on each vehicle sold by the particular dealership. To put it another way, is there a relationship between the amount of profit earned and the dealership?

Solution

The level of measurement for the variable dealership is nominal and ratio for the variable profit. To effectively use a contingency table, both variables need to be either of the nominal or ordinal scale. To make the variables compatible, we classify the variable profit into two categories, those cases where the profit earned is more than the median and those cases where it is less. On page 69 we calculated the median profit for all sales last month at Applewood Auto Group to be \$1,882.50.

Contingency Table Showing the Relationship between Profit and Dealership					
Above/Below Median Profit	Kane	Olean	Sheffield	Tionesta	Total
Above	25	20	19	26	90
Below	27	20	26	17	90
Total	52	40	45	43	180

By organizing the information into a contingency table, we can compare the profit at the four dealerships. We observe the following:

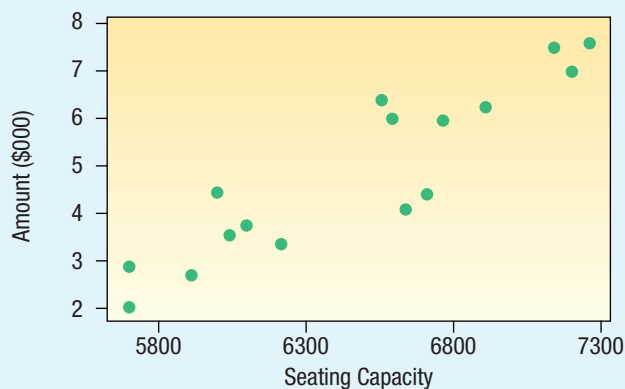
- From the Total column on the right, 90 of the 180 cars sold had a profit above the median and half below. From the definition of the median, this is expected.
- For the Kane dealership 25 out of the 52, or 48 percent, of the cars sold were sold for a profit more than the median.
- The percentage of profits above the median for the other dealerships are 50 percent for Olean, 42 percent for Sheffield, and 60 percent for Tionesta.

We will return to the study of contingency tables in Chapter 5 during the study of probability and in Chapter 17 during the study of nonparametric methods of analysis.

Self-Review 4–5




The rock group Blue String Beans is touring the United States. The following chart shows the relationship between concert seating capacity and revenue in \$000 for a sample of concerts.



- What is the diagram called?
- How many concerts were studied?
- Estimate the revenue for the concert with the largest seating capacity.
- How would you characterize the relationship between revenue and seating capacity? Is it strong or weak, direct or inverse?

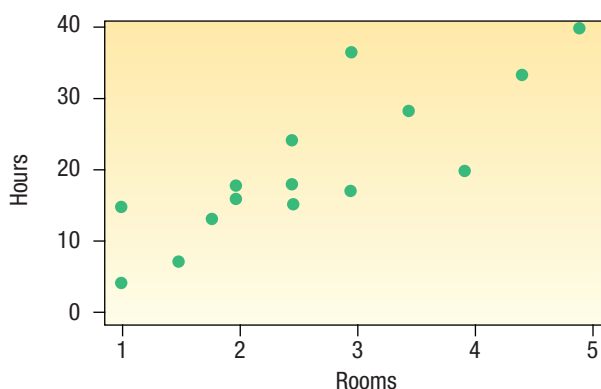
Exercises

connect™

23. Develop a scatter diagram for the following sample data. How would you describe the relationship between the values? 

X-Value	Y-Value	X-Value	Y-Value
10	6	11	6
8	2	10	5
9	6	7	2
11	5	7	3
13	7	11	7

24. Silver Springs Moving and Storage Inc. is studying the relationship between the number of rooms in a move and the number of labor hours required for the move. As part of the analysis, the CFO of Silver Springs developed the following scatter diagram.



- How many moves are in the sample?
 - Does it appear that more labor hours are required as the number of rooms increases, or do labor hours decrease as the number of rooms increases?
25. The Director of Planning for Devine Dining Inc. wishes to study the relationship between the gender of a guest and whether the guest orders dessert. To investigate the relationship, the manager collected the following information on 200 recent customers.

Dessert Ordered	Gender		Total
	Male	Female	
Yes	32	15	47
No	68	85	153
Total	100	100	200

- What is the level of measurement of the two variables?
 - What is the above table called?
 - Does the evidence in the table suggest men are more likely to order dessert than women? Explain why.
26. Ski Resorts of Vermont Inc. is considering a merger with Gulf Shores Beach Resorts Inc. of Alabama. The board of directors surveyed 50 stockholders concerning their position on the merger. The results are reported below.

Number of Shares Held	Opinion			Total
	Favor	Oppose	Undecided	
Under 200	8	6	2	16
200 up to 1,000	6	8	1	15
Over 1,000	6	12	1	19
Total	20	26	4	50

- What level of measurement is used in this table?
- What is this table called?
- What group seems most strongly opposed to the merger?

Chapter Summary

- I. A dot plot shows the range of values on the horizontal axis and the number of observations for each value on the vertical axis.
 - A. Dot plots report the details of each observation.
 - B. They are useful for comparing two or more data sets.
- II. A stem-and-leaf display is an alternative to a histogram.
 - A. The leading digit is the stem and the trailing digit the leaf.
 - B. The advantages of a stem-and-leaf display over a histogram include:
 1. The identity of each observation is not lost.
 2. The digits themselves give a picture of the distribution.
 3. The cumulative frequencies are also shown.
- III. Measures of location also describe the shape of a set of observations.
 - A. Quartiles divide a set of observations into four equal parts.
 1. Twenty-five percent of the observations are less than the first quartile, 50 percent are less than the second quartile, and 75 percent are less than the third quartile.
 2. The interquartile range is the difference between the third quartile and the first quartile.
 - B. Deciles divide a set of observations into ten equal parts and percentiles into 100 equal parts.
 - C. A box plot is a graphic display of a set of data.
 1. A box is drawn enclosing the regions between the first quartile and the third quartile.
 - a. A line is drawn inside the box at the median value.
 - b. Dotted line segments are drawn from the third quartile to the largest value to show the highest 25 percent of the values and from the first quartile to the smallest value to show the lowest 25 percent of the values.
 2. A box plot is based on five statistics: the maximum and minimum values, the first and third quartiles, and the median.
- IV. The coefficient of skewness is a measure of the symmetry of a distribution.
 - A. There are two formulas for the coefficient of skewness.
 1. The formula developed by Pearson is:

$$sk = \frac{3(\bar{X} - \text{Median})}{s} \quad [4-2]$$

2. The coefficient of skewness computed by statistical software is:

$$sk = \frac{n}{(n-1)(n-2)} \left[\sum \left(\frac{X - \bar{X}}{s} \right)^3 \right] \quad [4-3]$$

- V. A scatter diagram is a graphic tool to portray the relationship between two variables.
 - A. Both variables are measured with interval or ratio scales.
 - B. If the scatter of points moves from the lower left to the upper right, the variables under consideration are directly or positively related.
 - C. If the scatter of points moves from the upper left to the lower right, the variables are inversely or negatively related.
- VI. A contingency table is used to classify nominal-scale observations according to two characteristics.

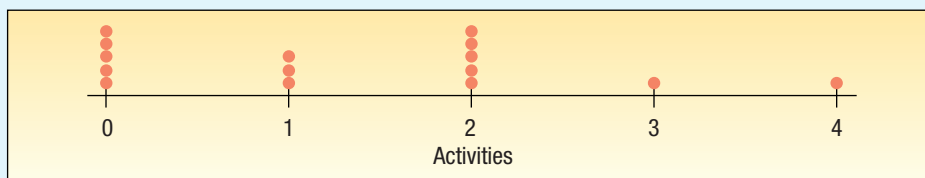
Pronunciation Key

SYMBOL	MEANING	PRONUNCIATION
L_p	Location of percentile	L sub p
Q_1	First quartile	Q sub 1
Q_3	Third quartile	Q sub 3

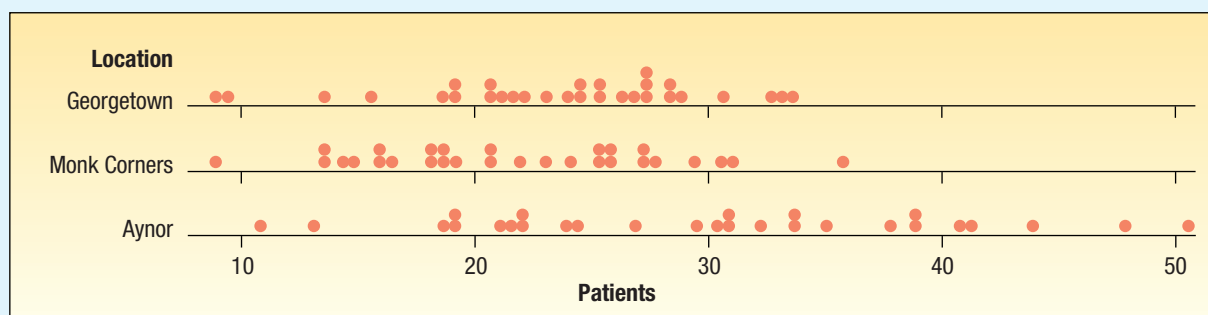


Chapter Exercises

27. A sample of students attending Southeast Florida University is asked the number of social activities in which they participated last week. The chart below was prepared from the sample data.



- What is the name given to this chart?
 - How many students were in the study?
 - How many students reported attending no social activities?
28. Doctor's Care is a walk-in clinic, with locations in Georgetown, Monks Corners, and Aynor, at which patients may receive treatment for minor injuries, colds, and flu, as well as physical examinations. The following charts report the number of patients treated in each of the three locations last month.



Describe the number of patients served at the three locations each day. What are the maximum and minimum numbers of patients served at each of the locations?

29. The screen size for 23 LCD televisions is given below. Make a stem-and-leaf display of this variable.



46	52	46	40	42	46	40	37	46	40	52	32	37	32	52
40	32	52	40	52	46	46	52							


30. The top 25 companies (by market capitalization) operating in the Washington, DC, area along with the year they were founded and the number of employees are given below. Make a stem-and-leaf display of each of these variables and write a short description of your findings.



Company Name	Year Founded	Employees
AES Corp.	1981	30000
American Capital Strategies Ltd.	1986	484
AvalonBay Communities Inc.	1978	1767
Capital One Financial Corp.	1995	31800
Constellation Energy Group Inc.	1816	9736
Coventry Health Care Inc.	1986	10250
Danaher Corp.	1984	45000
Dominion Resources Inc.	1909	17500
Fannie Mae	1938	6450
Freddie Mac	1970	5533

(continued)

Company Name	Year Founded	Employees
Gannett Co.	1906	49675
General Dynamics Corp.	1952	81000
Genworth Financial Inc.	2004	7200
Harman International Industries Inc.	1980	11246
Host Hotels & Resorts Inc.	1927	229
Legg Mason Inc.	1899	3800
Lockheed Martin Corp.	1995	140000
Marriott International Inc.	1927	151000
MedImmune Inc.	1988	2516
NII Holdings Inc.	1996	7748
Norfolk Southern Corp.	1982	30594
Pepco Holdings Inc.	1896	5057
Sallie Mae	1972	11456
Sprint Nextel Corp.	1899	64000
T. Rowe Price Group Inc.	1937	4605
The Washington Post Co.	1877	17100


31. In recent years, due to low interest rates, many homeowners refinanced their home mortgages. Linda Lahey is a mortgage officer at Down River Federal Savings and Loan. Below is the amount refinanced for 20 loans she processed last week. The data are reported in thousands of dollars and arranged from smallest to largest. 


59.2	59.5	61.6	65.5	66.6	72.9	74.8	77.3	79.2
83.7	85.6	85.8	86.6	87.0	87.1	90.2	93.3	98.6
100.2	100.7							

- Find the median, first quartile, and third quartile.
 - Find the 26th and 83rd percentiles.
 - Draw a box plot of the data.
32. A study is made by the recording industry in the United States of the number of music CDs owned by senior citizens and young adults. The information is reported below.

Seniors									
28	35	41	48	52	81	97	98	98	99
118	132	133	140	145	147	153	158	162	174
177	180	180	187	188					

Young Adults									
81	107	113	147	147	175	183	192	202	209
233	251	254	266	283	284	284	316	372	401
417	423	490	500	507	518	550	557	590	594

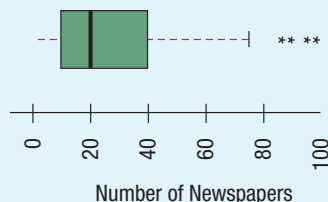
- Find the median and the first and third quartiles for the number of CDs owned by senior citizens. Develop a box plot for the information.
 - Find the median and the first and third quartiles for the number of CDs owned by young adults. Develop a box plot for the information.
 - Compare the number of CDs owned by the two groups. 
33. The corporate headquarters of *Bank.com*, a new Internet company that performs all banking transactions via the Internet, is located in downtown Philadelphia. The director of human resources is making a study of the time it takes employees to get to work. The city is planning to offer incentives to each downtown employer if they will encourage their

employees to use public transportation. Below is a listing of the time to get to work this morning according to whether the employee used public transportation or drove a car. 

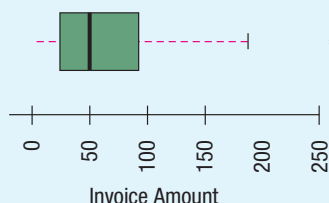
Public Transportation									
23	25	25	30	31	31	32	33	35	36
37	42								

Private									
32	32	33	34	37	37	38	38	38	39
40	44								

- Find the median and the first and third quartiles for the time it took employees using public transportation. Develop a box plot for the information.
 - Find the median and the first and third quartiles for the time it took employees who drove their own vehicle. Develop a box plot for the information.
 - Compare the times of the two groups.
34. The following box plot shows the number of daily newspapers published in each state and the District of Columbia. Write a brief report summarizing the number published. Be sure to include information on the values of the first and third quartiles, the median, and whether there is any skewness. If there are any outliers, estimate their value.



35. Walter Gogel Company is an industrial supplier of fasteners, tools, and springs. The amounts of its invoices vary widely, from less than \$20.00 to more than \$400.00. During the month of January the company sent out 80 invoices. Here is a box plot of these invoices. Write a brief report summarizing the invoice amounts. Be sure to include information on the values of the first and third quartiles, the median, and whether there is any skewness. If there are any outliers, approximate the value of these invoices.



36. The American Society of PeriAnesthesia Nurses (ASPAN; www.aspan.org) is a national organization serving nurses practicing in ambulatory surgery preanesthesia and postanesthesia care. The organization consists of 40 components, which are listed below.

State/Region	Membership	State/Region	Membership
Alabama	95	Illinois	562
Arizona	399	Indiana	270
Maryland, Delaware, DC	531	Iowa	117
Connecticut	239	Kentucky	197
Florida	631	Louisiana	258
Georgia	384	Michigan	411
Hawaii	73	Massachusetts	480

(continued)

State/Region	Membership	State/Region	Membership
Maine	97	California	1,165
Minnesota, Dakotas	289	New Mexico	79
Missouri, Kansas	282	Pennsylvania	575
Mississippi	90	Rhode Island	53
Nebraska	115	Colorado	409
North Carolina	542	South Carolina	237
Nevada	106	Texas	1,026
New Jersey, Bermuda	517	Tennessee	167
Alaska, Idaho, Montana,		Utah	67
Oregon, Washington	708	Virginia	414
New York	891	Vermont,	
Ohio	708	New Hampshire	144
Oklahoma	171	Wisconsin	311
Arkansas	68	West Virginia	62

Use statistical software to answer the following questions.

- Find the mean, median, and standard deviation of the number of members per component.
- Find the coefficient of skewness, using the software. What do you conclude about the shape of the distribution of component size?
- Determine the first and third quartiles. Do *not* use the method described by Excel.
- Develop a box plot. Are there any outliers? Which components are outliers? What are the limits for outliers?




37. McGivern Jewelers is located in the Levis Square Mall just south of Toledo, Ohio. Recently it ran an advertisement in the local newspaper reporting the shape, size, price, and cut grade for 33 of its diamonds currently in stock. The information is reported below.


Shape	Size (carats)	Price	Cut Grade	Shape	Size (carats)	Price	Cut Grade
Princess	5.03	\$44,312	Ideal cut	Round	0.77	\$2,828	Ultra ideal cut
Round	2.35	20,413	Premium cut	Oval	0.76	3,808	Premium cut
Round	2.03	13,080	Ideal cut	Princess	0.71	2,327	Premium cut
Round	1.56	13,925	Ideal cut	Marquise	0.71	2,732	Good cut
Round	1.21	7,382	Ultra ideal cut	Round	0.70	1,915	Premium cut
Round	1.21	5,154	Average cut	Round	0.66	1,885	Premium cut
Round	1.19	5,339	Premium cut	Round	0.62	1,397	Good cut
Emerald	1.16	5,161	Ideal cut	Round	0.52	2,555	Premium cut
Round	1.08	8,775	Ultra ideal cut	Princess	0.51	1,337	Ideal cut
Round	1.02	4,282	Premium cut	Round	0.51	1,558	Premium cut
Round	1.02	6,943	Ideal cut	Round	0.45	1,191	Premium cut
Marquise	1.01	7,038	Good cut	Princess	0.44	1,319	Average cut
Princess	1.00	4,868	Premium cut	Marquise	0.44	1,319	Premium cut
Round	0.91	5,106	Premium cut	Round	0.40	1,133	Premium cut
Round	0.90	3,921	Good cut	Round	0.35	1,354	Good cut
Round	0.90	3,733	Premium cut	Round	0.32	896	Premium cut
Round	0.84	2,621	Premium cut				

- Develop a box plot of the variable price and comment on the result. Are there any outliers? What is the median price? What is the value of the first and the third quartile?
- Develop a box plot of the variable size and comment on the result. Are there any outliers? What is the median price? What is the value of the first and the third quartile?
- Develop a scatter diagram between the variables price and size. Be sure to put price on the vertical axis and size on the horizontal axis. Does there seem to be an association between the two variables? Is the association direct or indirect? Does any point seem to be different from the others?
- Develop a contingency table for the variables shape and cut grade. What is the most common cut grade? What is the most common shape? What is the most common combination of cut grade and shape?



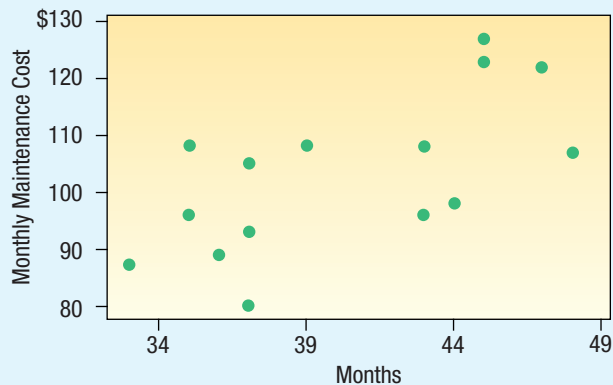
38. Listed below is the amount of commissions earned last month for the eight members of the sales staff at Best Electronics. Calculate the coefficient of skewness using both methods. *Hint:* Use of a spreadsheet will expedite the calculations. 


980.9	1,036.5	1,099.5	1,153.9	1,409.0	1,456.4	1,718.4	1,721.2
-------	---------	---------	---------	---------	---------	---------	---------

39. Listed below is the number of car thefts in a large city over the last week. Calculate the coefficient of skewness using both methods. *Hint:* Use of a spreadsheet will expedite the calculations. 

3	12	13	7	8	3	8
---	----	----	---	---	---	---

40. The manager of Information Services at Wilkin Investigations, a private investigation firm, is studying the relationship between the age (in months) of a combination printer, copy, and fax machine and its monthly maintenance cost. For a sample of 15 machines, the manager developed the following chart. What can the manager conclude about the relationship between the variables?



41. An auto insurance company reported the following information regarding the age of a driver and the number of accidents reported last year. Develop a scatter diagram for the data and write a brief summary. 

Age	Accidents	Age	Accidents
16	4	23	0
24	2	27	1
18	5	32	1
17	4	22	3

42. Wendy's offers eight different condiments (mustard, catsup, onion, mayonnaise, pickle, lettuce, tomato, and relish) on hamburgers. A store manager collected the following information on the number of condiments ordered and the age group of the customer. What can you conclude regarding the information? Who tends to order the most or least number of condiments?

Number of Condiments	Age			
	Under 18	18 up to 40	40 up to 60	60 or older
0	12	18	24	52
1	21	76	50	30
2	39	52	40	12
3 or more	71	87	47	28

43. Listed at the top of the next page is a table showing the number of employed and unemployed workers 20 years or older by gender in the United States.

Gender	Number of Workers (000)	
	Employed	Unemployed
Men	70,415	4,209
Women	61,402	3,314

- How many workers were studied?
- What percent of the workers were unemployed?
- Compare the percent unemployed for the men and the women.

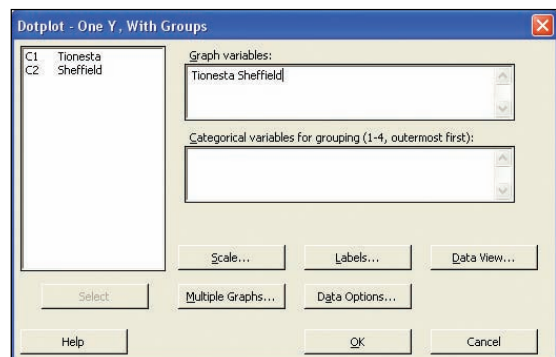
Data Set Exercises

- Refer to the Real Estate data, which reports information on homes sold in the Goodyear, Arizona, area during the last year. Prepare a report on the selling prices of the homes. Be sure to answer the following questions in your report.
 - Develop a box plot. Estimate the first and the third quartiles. Are there any outliers?
 - Develop a scatter diagram with price on the vertical axis and the size of the home on the horizontal. Does there seem to be a relationship between these variables? Is the relationship direct or inverse?
 - Develop a scatter diagram with price on the vertical axis and distance from the center of the city on the horizontal axis. Does there seem to be a relationship between these variables? Is the relationship direct or inverse?
- Refer to the Baseball 2009 data, which reports information on the 30 Major League Baseball teams for the 2009 season. Refer to the variable team salary.
 - Select the variable that refers to the year in which the stadium was built. (*Hint:* Subtract the year in which the stadium was built from the current year to find the age of the stadium and work this variable.) Develop a box plot. Are there any outliers? Which stadiums are outliers?
 - Select the variable team salary and draw a box plot. Are there any outliers? What are the quartiles? Write a brief summary of your analysis. How do the salaries of the New York Yankees compare with the other teams?
 - Draw a scatter diagram with the number of games won on the vertical axis and the team salary on the horizontal axis. What are your conclusions?
 - Select the variable wins. Draw a dot plot. What can you conclude from this plot?
- Refer to the Buena School District bus data.
 - Refer to the maintenance cost variable. Develop a box plot. What are the first and third quartiles? Are there any outliers?
 - Determine the median maintenance cost. Based on the median, develop a contingency table with bus manufacturer as one variable and whether the maintenance cost was above or below the median as the other variable. What are your conclusions?

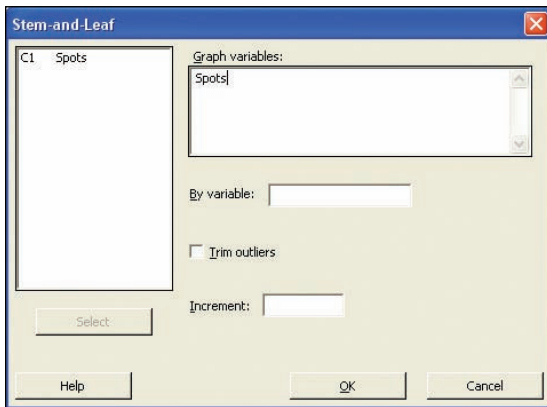
Software Commands

- The Minitab commands for the dot plot on page 104 are:
 - Enter the number of vehicles serviced at Tionesta Ford Lincoln Mercury in column C1 and Sheffield Motors in C2. Name the variables accordingly.
 - Select **Graph** and **Dotplot**. In the first dialog box, select **Multiple Y's, Simple** in the lower left corner, and click **OK**. In the next dialog box select **Tionesta** and **Sheffield** as the variables to **Graph**, click on **Labels** and write an appropriate title. Then click **OK**.
 - To calculate the descriptive statistics shown in the output, select **Stat**, **Basic statistics**, and then **Display Descriptive statistics**. In the dialog box, select **Tionesta** and **Sheffield** as the Variables, click on **Statistics**, select the

desired statistics to be output, and finally click **OK** twice.

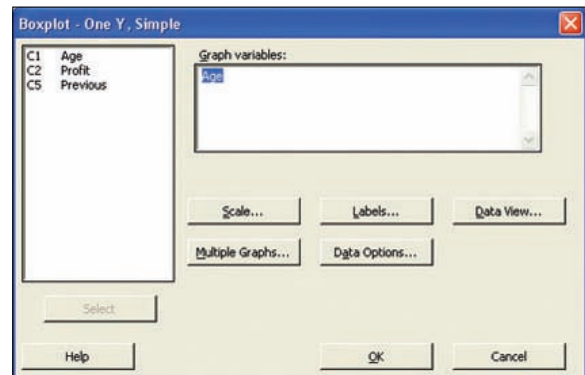


2. The Minitab commands for the stem-and-leaf display on page 107 are:
 - a. Import the data for **Table 4-1**.
 - b. Select **Graph**, and click on **Stem-and-Leaf**.
 - c. Select the variable **Spots**, enter **10** for the **Increment**, and then click **OK**.



3. The Minitab commands for the descriptive summary on page 113 are:
 - a. Input the data on the Smith Barney commissions from the Example on page 111.
 - b. From the toolbar, select **Stat, Basic Statistics**, and **Display Descriptive Statistics**. In the dialog box, select **Commissions** as the **Variable**, and then click **OK**.
4. The Excel commands for the descriptive statistics on page 114 are:
 - a. Input the data on the Smith Barney commissions from the Example on page 111.
 - b. In cell C4 write **Quartile 1** and in C6 write **Quartile 3**.
 - c. In cell D4 write “**=QUARTILE(A1:A16,1)**” and hit Enter. In cell D6 write “**=QUARTILE(A1:A16,1)**” and hit Enter.
5. The Minitab commands for the box plot on page 117 are:
 - a. Import the Applewood Auto Group data.

- b. Select **Graph** and then **Boxplot**. In the dialog box, select **Simple** in the upper left corner and click **OK**. Select **Age** as the **Graph Variable**, click on **Labels** and include an appropriate heading, and then click **OK**.



6. The Minitab commands for the descriptive summary on page 122 are:
 - a. Enter the data in the first column. In the cell below C1, enter the variable **Earnings**.
 - b. Select **Stat, Basic Statistics**, and then click on **Graphical Summary**. Select **Earnings** as the variable, and then click **OK**.
7. The Excel commands for the scatter diagram on page 125 are:
 - a. Retrieve the Applewood Auto data.
 - b. Using the mouse, highlight the column of age and profit. Include the first row.
 - c. Select the **Insert** tab. Select **Scatter** from the **Chart** options. Select the top left option. The scatter plot will appear.
 - d. With **Chart Tools** displayed at the top, select the **Layout** tab. Select **Chart Title** and type in a title for the plot. Next, under the same **Layout** tab, select **AxisTitles**. Using **Primary Vertical Axis Title**, name the vertical axis **Profit**. Using the **Primary Horizontal Axis Title**, name the horizontal axis **Age**. Next, select **Legend** and select **None**.

Chapter 4 Answers to Self-Review



- 4-1 1. a. 79, 105
 b. 15
 c. From 88 to 97; 75 percent of the stores are in this range.

2.

7	7
8	0013488
9	1256689
10	1248
11	26

- a. 8
 b. 10.1, 10.2, 10.4, 10.8
 c. 9.5
 d. 11.6, 7.7

- 4-2 a. 7.9
 b. $Q_1 = 7.76$, $Q_3 = 8.015$

- 4-3 The smallest value is 10 and the largest 85; the first quartile is 25 and the third 60. About 50 percent of the values are between 25 and 60. The median value is 40. The distribution is positively skewed. There are no outliers.

4-4 a. $\bar{X} = \frac{407}{5} = 81.4$, Median = 84

$$s = \sqrt{\frac{923.2}{5-1}} = 15.19$$

b. $sk = \frac{3(81.4 - 84.0)}{15.19} = -0.51$

c.

X	$\frac{X - \bar{X}}{s}$	$\left[\frac{X - \bar{X}}{s} \right]^3$
73	-0.5530	-0.1691
98	1.0928	1.3051
60	-1.4088	-2.7962
92	0.6978	0.3398
84	0.1712	0.0050
		-1.3154

$$sk = \frac{5}{(4)(3)} [-1.3154] \\ = -0.5481$$

d. The distribution is somewhat negatively skewed.

- 4-5 a. Scatter diagram
b. 16
c. \$7,500
d. Strong and direct

A Review of Chapters 1–4

This section is a review of the major concepts and terms introduced in Chapters 1–4. Chapter 1 began by describing the meaning and purpose of statistics. Next we described the different types of variables and the four levels of measurement. Chapter 2 was concerned with describing a set of observations by organizing it into a frequency distribution and then portraying the frequency distribution as a histogram or a frequency polygon. Chapter 3 began by describing measures of location, such as the mean, weighted mean, median, geometric mean, and mode. This chapter also included measures of dispersion, or spread. Discussed in this section were the range, mean deviation, variance, and standard deviation. Chapter 4 included several graphing techniques such as dot plots, box plots, and scatter diagrams. We also discussed the coefficient of skewness, which reports the lack of symmetry in a set of data.

Throughout this section we stressed the importance of statistical software, such as Excel and Minitab. Many computer outputs in these chapters demonstrated how quickly and effectively a large data set can be organized into a frequency distribution, several of the measures of location or measures of variation calculated, and the information presented in graphical form.

Glossary

Chapter 1

Descriptive statistics The techniques used to describe the important characteristics of a set of data. This includes organizing the data values into a frequency distribution, computing measures of location, and computing measures of dispersion and skewness.

Inferential statistics, also called **statistical inference** This facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if a sample of 10 TI-36X solar calculators revealed 2 to be defective, we might infer that 20 percent of the production is defective.

Interval measurement If one observation is greater than another by a certain amount, and the zero point is arbitrary, the measurement is on an interval scale. For example, the difference between temperatures of 70 degrees and 80 degrees is 10 degrees. Likewise, a temperature of

90 degrees is 10 degrees more than a temperature of 80 degrees, and so on.

Nominal measurement The “lowest” level of measurement. If data are classified into categories and the order of those categories is not important, it is the nominal level of measurement. Examples are gender (male, female) and political affiliation (Republican, Democrat, Independent, all others). If it makes no difference whether male or female is listed first, the data are nominal level.

Ordinal measurement Data that can be ranked are referred to as ordinal measures. For example, consumer response to the sound of a new speaker might be excellent, very good, fair, or poor.

Population The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Ratio measurement If the distance between numbers is a constant size, there is a true zero point, and the ratio of two values is meaningful, then the data are ratio scale. For example, the distance between \$200 and \$300 is \$100, and in the case of money there is a true zero point. If you have zero dollars, there is an absence of money (you have none). Also the ratio between \$200 and \$300 is meaningful.

Sample A portion, or subset, of the population being studied.

Statistics The science of collecting, organizing, analyzing, and interpreting numerical data for the purpose of making more effective decisions.

Chapter 2

Charts Special graphical formats used to portray a frequency distribution, including histograms, frequency polygons, and cumulative frequency polygons. Other graphical devices used to portray data are bar charts and pie charts.

Class The interval in which the data are tallied. For example, \$4 up to \$7 is a class; \$7 up to \$11 is another class.

Class frequency The number of observations in each class. If there are 16 observations in the \$4 up to \$6 class, 16 is the class frequency.

Exhaustive Each observation must fall into one of the categories.

Frequency distribution A grouping of data into classes showing the number of observations in each of the mutually exclusive classes.

Histogram A graphical display of a frequency or relative frequency distribution. The horizontal axis shows the classes. The vertical height of adjacent bars shows the frequency or relative frequency of each class.

Midpoint The value that divides the class into two equal parts. For the classes \$10 up to \$20 and \$20 up to \$30, the midpoints are \$15 and \$25, respectively.

Mutually exclusive A property of a set of categories such that an individual, object, or measurement is included in only one category.

Relative frequency distribution A frequency distribution that shows the fraction or proportion of the total observations in each class.

Chapter 3

Arithmetic mean The sum of the values divided by the number of values. The symbol for the mean of a sample is \bar{X} and the symbol for a population mean is μ .

Geometric mean The n th root of the product of all the values. It is especially useful for averaging rates of change and index numbers. It minimizes the importance of extreme values. A second use of the geometric mean is to find the mean annual percent change over a period of time. For example, if gross sales were \$245 million in 1990 and \$692 million in 2010, the average annual rate of return is 5.33 percent.

Mean deviation The mean of the deviations from the mean, disregarding signs. It is identified as MD .

Measure of dispersion A value that shows the spread of a data set. The range, variance, and standard deviation are measures of dispersion.

Measure of location A single value that is typical of the data. It pinpoints the center of a distribution. The arithmetic mean, weighted mean, median, mode, and geometric mean are measures of location.

Median The value of the middle observation after all the observations have been arranged from low to high. For example, if observations 6, 9, 4 are rearranged to read 4, 6, 9, the median is 6, the middle value.

Mode The value that appears most frequently in a set of data. For grouped data, it is the *midpoint* of the class containing the largest number of values.

Range It is a measure of dispersion. The range is found by subtracting the minimum value from the maximum value.

Standard deviation The square root of the variance.

Variance A measure of dispersion based on the average squared differences from the arithmetic mean.

Weighted mean Each value is weighted according to its relative importance. For example, if 5 shirts cost \$10 each and 20 shirts cost \$8 each, the weighted mean price is \$8.40: $[(5 \times \$10) + (20 \times \$8)]/25 = \$210/25 = \8.40 .

Chapter 4

Box plot A graphic display that shows the general shape of a variable's distribution. It is based on five descriptive statistics: the maximum and minimum values, the first and third quartiles, and the median.

Coefficient of skewness A measure of the lack of symmetry in a distribution. For a symmetric distribution there is no skewness, so the coefficient of skewness is zero. Otherwise, it is either positive or negative, with the limits of ± 3.0 .

Contingency table A table used to classify observations according to two characteristics.

Deciles Values of an ordered (minimum to maximum) data set that divide the data into ten equal parts.

Dot plot A dot plot summarizes the distribution of one variable by stacking dots at points on a number line that shows the values of the variable. A dot plot shows all values.

Interquartile range The absolute numerical difference between the first and third quartiles. Fifty percent of a distribution's values occur in this range.

Outlier A data point that is usually far from the others. An accepted rule is to classify an observation as an outlier if it is 1.5 times the interquartile range above the third quartile or below the first quartile.

Percentiles Values of an ordered (minimum to maximum) data set that divide the data into one hundred intervals.

Quartiles Values of an ordered (minimum to maximum) data set that divide the data into four intervals.

Scatter diagram Graphical technique used to show the relationship between two variables measured with interval or ratio scales.


Stem-and-leaf display A method to display a variable's distribution using every value. Values are classified by the data's leading digit. For example, if a data set contains values between 13 and 84, eight classes based on the 10s digit would be used for the stems. The 1s digits would be the leaves.

Problems

1. A sample of the funds deposited in First Federal Savings Bank's MCA (miniature checking account) revealed the following amounts.


\$124	\$14	\$150	\$289	\$52	\$156	\$203	\$82	\$27	\$248
39	52	103	58	136	249	110	298	251	157
186	107	142	185	75	202	119	219	156	78
116	152	206	117	52	299	58	153	219	148
145	187	165	147	158	146	185	186	149	140

Use a statistical software package such as Excel or Minitab to help answer the following questions.

- Determine the mean, median, and standard deviation.
 - Determine the first and third quartiles.
 - Develop a box plot. Are there any outliers? Do the amounts follow a symmetric distribution or are they skewed? Justify your answer.
 - Organize the distribution of funds into a frequency distribution.
 - Write a brief summary of the results in parts a to d. 
2. Listed below are the 44 U.S. presidents and their age as they began their terms in office.

Number	Name	Age	Number	Name	Age
1	Washington	57	23	B. Harrison	55
2	J. Adams	61	24	Cleveland	55
3	Jefferson	57	25	McKinley	54
4	Madison	57	26	T. Roosevelt	42
5	Monroe	58	27	Taft	51
6	J. Q. Adams	57	28	Wilson	56
7	Jackson	61	29	Harding	55
8	Van Buren	54	30	Coolidge	51
9	W. H. Harrison	68	31	Hoover	54
10	Tyler	51	32	F. D. Roosevelt	51
11	Polk	49	33	Truman	60
12	Taylor	64	34	Eisenhower	62
13	Fillmore	50	35	Kennedy	43
14	Pierce	48	36	L. B. Johnson	55
15	Buchanan	65	37	Nixon	56
16	Lincoln	52	38	Ford	61
17	A. Johnson	56	39	Carter	52
18	Grant	46	40	Reagan	69
19	Hayes	54	41	G.H.W. Bush	64
20	Garfield	49	42	Clinton	46
21	Arthur	50	43	G. W. Bush	54
22	Cleveland	47	44	Obama	47


Use a statistical software package such as Excel or Minitab to help answer the following questions.

- Determine the mean, median, and standard deviation.
- Determine the first and third quartiles.
- Develop a box plot. Are there any outliers? Do the amounts follow a symmetric distribution or are they skewed? Justify your answer.
- Organize the distribution of ages into a frequency distribution.
- Write a brief summary of the results in parts a to d. 

3. Listed below is the per capita income for the 50 states and the District of Columbia.

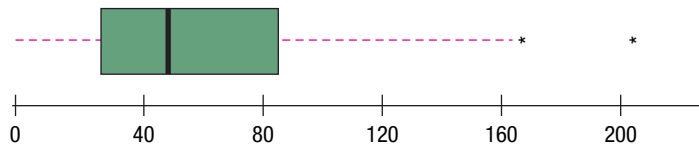
State	Amount	State	Amount
Alabama	\$30,894	Montana	\$30,790
Alaska	38,138	Nebraska	34,440
Arizona	31,936	Nevada	38,994
Arkansas	28,473	New Hampshire	39,753
California	39,626	New Jersey	46,763
Colorado	39,491	New Mexico	29,929
Connecticut	50,762	New York	44,027
Delaware	39,131	North Carolina	32,247
DC	57,746	North Dakota	32,763
Florida	36,720	Ohio	33,320
Georgia	32,095	Oklahoma	32,391
Hawaii	37,023	Oregon	33,299
Idaho	29,920	Pennsylvania	36,825
Illinois	38,409	Rhode Island	37,523
Indiana	32,288	South Carolina	29,767
Iowa	33,038	South Dakota	32,030
Kansas	34,799	Tennessee	32,172
Kentucky	29,729	Texas	35,166
Louisiana	31,821	Utah	29,406
Maine	32,095	Vermont	34,871
Maryland	43,788	Virginia	39,540
Massachusetts	46,299	Washington	38,212
Michigan	33,788	West Virginia	28,206
Minnesota	38,859	Wisconsin	34,405
Mississippi	27,028	Wyoming	40,655
Missouri	32,789		

Use a statistical software package such as Excel or Minitab to help answer the following questions. Determine the first and third quartiles.

- Determine the mean, median, and standard deviation.
 - Determine the first and third quartiles.
 - Develop a box plot. Are there any outliers? Do the amounts follow a symmetric distribution or are they skewed? Justify your answer.
 - Organize the distribution of funds into a frequency distribution.
 - Write a brief summary of the results in parts a to d. 
4. A sample of 12 homes sold last week in St. Paul, Minnesota, revealed the following information. Draw a scatter diagram. Can we conclude that, as the size of the home (reported below in thousands of square feet) increases, the selling price (reported in \$ thousands) also increases?

Home Size (thousands of square feet)	Selling Price (\$ thousands)	Home Size (thousands of square feet)	Selling Price (\$ thousands)
1.4	100	1.3	110
1.3	110	0.8	85
1.2	105	1.2	105
1.1	120	0.9	75
1.4	80	1.1	70
1.0	105	1.1	95

5. Refer to the following diagram.



- What is the graph called?
- What are the median, and first and third quartile values?
- Is the distribution positively skewed? Tell how you know.
- Are there any outliers? If yes, estimate these values.
- Can you determine the number of observations in the study?

Cases

A. Century National Bank

The following case will appear in subsequent review sections. Assume that you work in the Planning Department of the Century National Bank and report to Ms. Lamberg. You will need to do some data analysis and prepare a short written report. Remember, Mr. Selig is the president of the bank, so you will want to ensure that your report is complete and accurate. A copy of the data appears in Appendix A.6.

Century National Bank has offices in several cities in the Midwest and the southeastern part of the United States. Mr. Dan Selig, president and CEO, would like to know the characteristics of his checking account customers. What is the balance of a typical customer?

How many other bank services do the checking account customers use? Do the customers use the ATM service and, if so, how often? What about debit cards? Who uses them, and how often are they used?

To better understand the customers, Mr. Selig asked Ms. Wendy Lamberg, director of planning, to select a sample of customers and prepare a report. To begin, she has appointed a team from her staff. You are the head of the team and responsible for preparing the report. You select a random sample of 60 customers. In addition to the balance in each account at the end of last month, you determine: (1) the number of ATM (automatic teller machine) transactions in the last month; (2) the number of other bank services (a savings account, a certificate of deposit, etc.) the customer uses; (3) whether the customer has a debit card (this is a bank service in which charges are made directly to the customer's account); and (4) whether or not interest is paid on the checking account. The sample includes customers from the branches in Cincinnati, Ohio; Atlanta, Georgia; Louisville, Kentucky; and Erie, Pennsylvania.

- Develop a graph or table that portrays the checking balances. What is the balance of a typical customer? Do many customers have more than \$2,000 in their accounts? Does it appear that there is a difference in the distribution of the accounts among the four branches? Around what value do the account balances tend to cluster?
- Determine the mean and median of the checking account balances. Compare the mean and the

median balances for the four branches. Is there a difference among the branches? Be sure to explain the difference between the mean and the median in your report.

- Determine the range and the standard deviation of the checking account balances. What do the first and third quartiles show? Determine the coefficient of skewness and indicate what it shows. Because Mr. Selig does not deal with statistics daily, include a brief description and interpretation of the standard deviation and other measures.

B. Wildcat Plumbing Supply Inc.: Do We Have Gender Differences?

Wildcat Plumbing Supply has served the plumbing needs of Southwest Arizona for more than 40 years. The company was founded by Mr. Terrence St. Julian and is run today by his son Cory. The company has grown from a handful of employees to more than 500 today. Cory is concerned about several positions within the company where he has men and women doing essentially the same job but at different pay. To investigate, he collected the information below. Suppose you are a student intern in the Accounting Department and have been given the task to write a report summarizing the situation.

Yearly Salary (\$000)	Women	Men
Less than 30	2	0
30 up to 40	3	1
40 up to 50	17	4
50 up to 60	17	24
60 up to 70	8	21
70 up to 80	3	7
80 or more	0	3

To kick off the project, Mr. Cory St. Julian held a meeting with his staff and you were invited. At this meeting, it was suggested that you calculate several measures of location, draw charts, such as a cumulative

frequency distribution, and determine the quartiles for both men and women. Develop the charts and write the report summarizing the yearly salaries of employees at Wildcat Plumbing Supply. Does it appear that there are pay differences based on gender?

C. Kimble Products: Is There a Difference In the Commissions?

At the January national sales meeting, the CEO of Kimble Products was questioned extensively regarding the company policy for paying commissions to its sales representatives. The company sells sporting goods to two major markets. There are 40 sales representatives who

call directly on large volume customers, such as the athletic departments at major colleges and universities and professional sports franchises. There are 30 sales representatives who represent the company to retail stores located in shopping malls and large discounters such as Kmart and Target.

Upon his return to corporate headquarters, the CEO asked the sales manager for a report comparing the commissions earned last year by the two parts of the sales team. The information is reported below. Write a brief report. Would you conclude that there is a difference? Be sure to include information in the report on both the central tendency and dispersion of the two groups.

Commissions Earned by Sales Representatives Calling on Athletic Departments (\$)									
354	87	1,676	1,187	69	3,202	680	39	1,683	1,106
883	3,140	299	2,197	175	159	1,105	434	615	149
1,168	278	579	7	357	252	1,602	2,321	4	392
416	427	1,738	526	13	1,604	249	557	635	527

Commissions Earned by Sales Representatives Calling on Large Retailers (\$)									
1,116	681	1,294	12	754	1,206	1,448	870	944	1,255
1,213	1,291	719	934	1,313	1,083	899	850	886	1,556
886	1,315	1,858	1,262	1,338	1,066	807	1,244	758	918

Practice Test

There is a practice test at the end of each review section. The tests are in two parts. The first part contains several objective questions, usually in a fill-in-the-blank format. The second part is problems. In most cases, it should take 30 to 45 minutes to complete the test. The problems require a calculator. Check the answers in the Answer Section in the back of the book.

Part 1—Objective

- The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making effective decisions is called _____. **1.** _____
- Methods of organizing, summarizing, and presenting data in an informative way is called _____. **2.** _____
- The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest is called the _____. **3.** _____
- List the two types of variables. **4.** _____
- The number of bedrooms in a house is an example of a _____. (discrete variable, continuous variable, qualitative variable—pick one) **5.** _____
- The jersey numbers of Major League Baseball players is an example of what level of measurement? **6.** _____
- The classification of students by eye color is an example of what level of measurement? **7.** _____
- The sum of the differences between each value and the mean is always equal to what value? **8.** _____
- A set of data contained 70 observations. How many classes would you suggest in order to construct a frequency distribution? **9.** _____
- What percent of the values in a data set are always larger than the median? **10.** _____
- The square of the standard deviation is the _____. **11.** _____
- The standard deviation assumes a negative value when _____. (All the values are negative, when at least half the values are negative, or never—pick one.) **12.** _____
- Which of the following is least affected by an outlier? (mean, median, or range—pick one) **13.** _____

Part 2—Problems

- The Russell 2000 index of stock prices increased by the following amounts over the last three years.

18%	4%	2%
-----	----	----

What is the geometric mean increase for the three years?

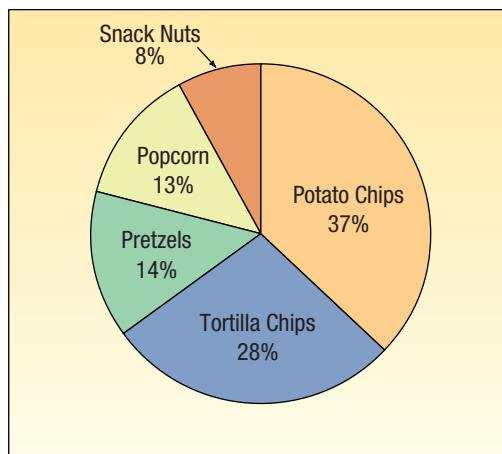
2. The information below refers to the selling prices (\$000) of homes sold in Warren, PA, during 2010.

Selling Price (\$000)	Frequency
120.0 up to 150.0	4
150.0 up to 180.0	18
180.0 up to 210.0	30
210.0 up to 240.0	20
240.0 up to 270.0	17
270.0 up to 300.0	10
300.0 up to 330.0	6

- What is the class interval?
 - How many homes were sold in 2010?
 - How many homes sold for less than \$210,000?
 - What is the relative frequency of the 210 up to 240 class?
 - What is the midpoint of the 150 up to 180 class?
 - The selling prices range between what two amounts?
3. A sample of eight college students revealed they owned the following number of CDs.

52	76	64	79	80	74	66	69
----	----	----	----	----	----	----	----

- What is the mean number of CDs owned?
 - What is the median number of CDs owned?
 - What is the 40th percentile?
 - What is the range of the number of CDs owned?
 - What is the standard deviation of the number of CDs owned?
4. An investor purchased 200 shares of the Blair Company for \$36 each in July of 2010, 300 shares at \$40 each in September 2010, and 500 shares at \$50 each in January 2011. What is the investor's weighted mean price per share?
5. During the 2008 Super Bowl, 30 million pounds of snack food was eaten. The chart below depicts this information.



- What is the name given to this graph?
- Estimate, in millions of pounds, the amount of potato chips eaten during the game.
- Estimate the relationship of potato chips to popcorn. (twice as much, half as much, three times, none of these—pick one)
- What percent of the total do potato chips and tortilla chips comprise?