4

Describing Data:

Displaying and Exploring Data

GOALS

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

1 Develop and interpret a *dot plot.*

2 Develop and interpret a *stem-and-leaf display.*

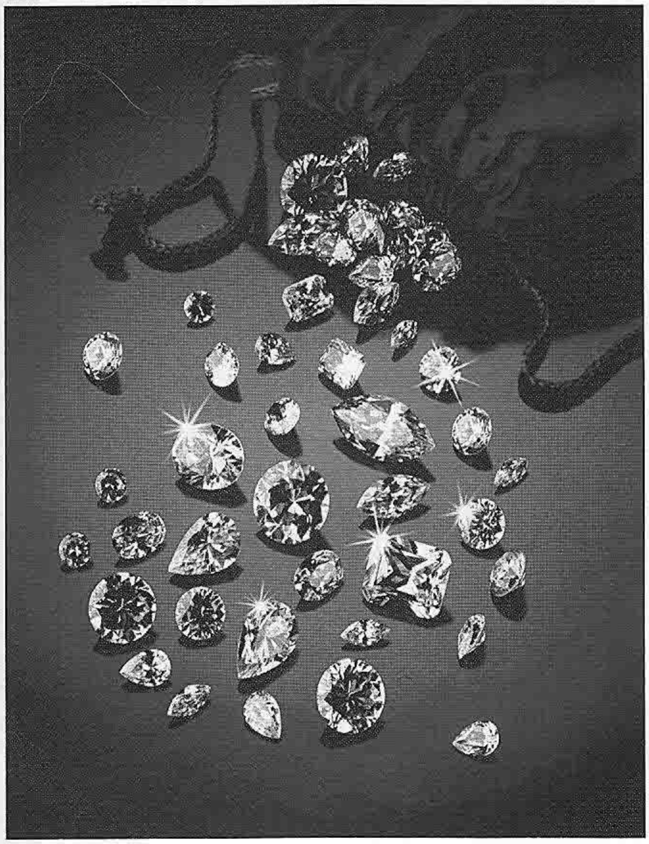
3 Compute and understand *quartiles, deciles,* and *percentiles.*

4 Construct and interpret *box plots.*

5 Compute and understand the *coefficient of skewness.*

6 Draw and interpret a *scatter diagram.*

7 Construct and interpret a *contingency table.*



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. Using the data provided in Exercise 37, develop a box plot of the variable price and comment on the result.

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.

Dot Plots

A histogram groups data into classes. Recall in the Whitner Autoplex data from Table 2–1 that 80 observations were condensed into seven classes. When we organized the data into the seven 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

Recall in Table 2–4 on page 28 we presented data on the selling price of 80 vehicles sold last month at Whitner Autoplex in Raytown, Missouri. Whitner is one of the many dealerships owned by AutoUSA. AutoUSA has many other dealerships located in small towns throughout the United States. Reported below are the number of vehicles sold in the last 24 months at Smith Ford Mercury Jeep, Inc., in Kane, Pennsylvania, and Brophy Honda Volkswagen in Greenville, Ohio. Construct dot plots and report summary statistics for the two small-town AutoUSA lots.

Smith Ford Mercury Jeep, Inc.

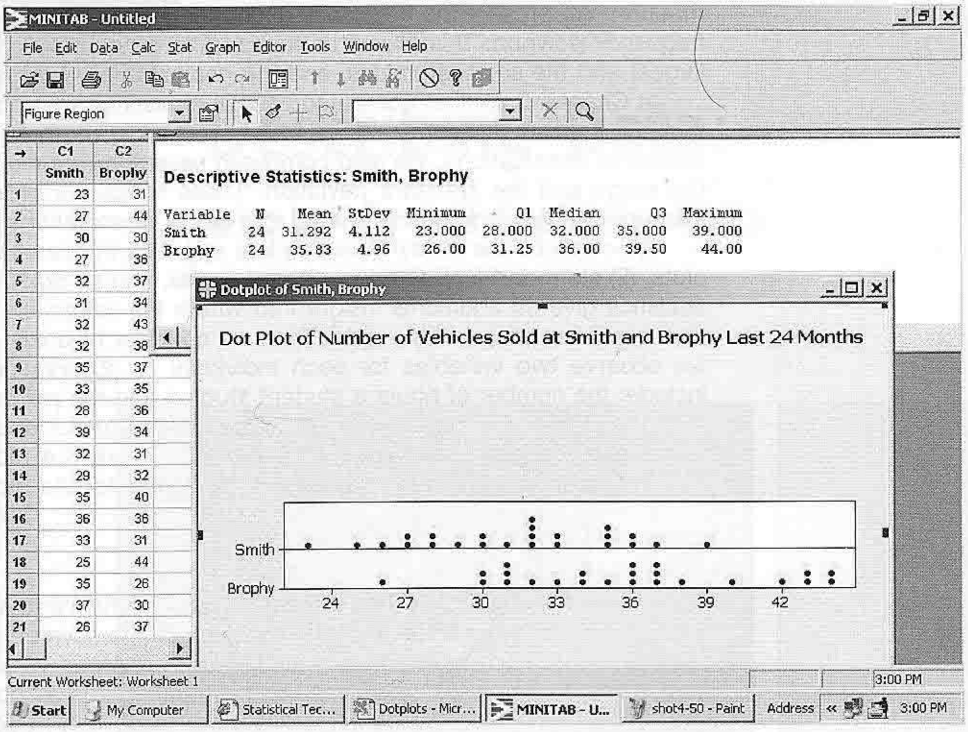
23	27	30	27	32	31	32	32	35	33
28	39	32	29	35	36	33	25	35	37
26	28	36	30						

Brophy Honda Volkswagen

31	44	30	36	37	34	43	38	37	35
36	34	31	32	40	36	31	44	26	30
37	43	42	33						

Solution

The MINITAB system provides a dot plot and calculates the mean, median, maximum, and minimum values, and the standard deviation for the number of cars sold at each of the dealerships over the last 24 months.



→	C1	C2
	Smith	Brophy
1	23	31
2	27	44
3	30	30
4	27	36
5	32	37
6	31	34
7	32	43
8	32	38
9	35	37
10	33	35
11	28	36
12	39	34
13	32	31
14	29	32
15	35	40
16	36	36
17	33	31
18	25	44
19	35	26
20	37	30
21	26	37

Descriptive Statistics: Smith, Brophy

```
Variable   N    Mean  StDev  Minimum      Q1  Median      Q3  Maximum
Smith     24  31.292  4.112   23.000  28.000  32.000  35.000   39.000
Brophy    24   35.83   4.96    26.00   31.25   36.00   39.50    44.00
```

Dot Plot of Number of Vehicles Sold at Smith and Brophy Last 24 Months

From the descriptive statistics we see that Brophy sold a mean of 35.83 vehicles per month and Smith a mean of 31.292. So Brophy typically sells 4.54 more vehicles per month. There is also more dispersion or variation in the monthly Brophy sales than in the Smith sales. How do we know this? The standard deviation is larger at Brophy (4.96 cars per month) than at Smith (4.112 cars per month).

The dot plot, shown in the lower right of the software output, graphically illustrates the distributions for both dealerships. The plots show the difference in the location and dispersion of the observations. By looking at the plots, we can see that Brophy's sales are more widely dispersed and have a larger mean than Smith's sales. Several other features of the monthly sales are apparent:

- Smith sold the fewest cars in any month, 23.
- Brophy sold 26 cars in its lowest month, which is 4 cars less than the next lowest month.
- Smith sold exactly 32 cars in four different months.
- The monthly sales cluster is around 32 for Smith and 36 for Brophy.

Stem-and-Leaf Displays

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 2005. 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.

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.

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:

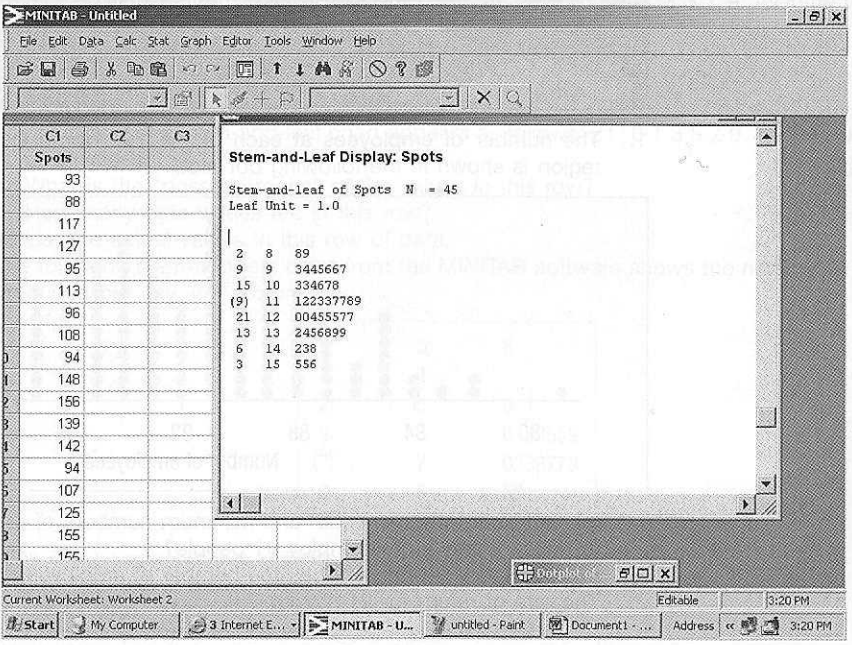
15	5	5	6

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.



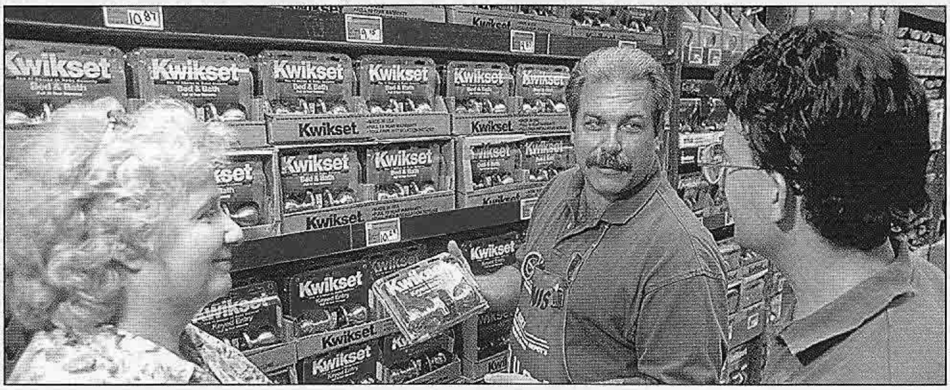
C1	C2	C3
Spots		
93		
88		
117		
127		
95		
113		
96		
108		
94		
148		
156		
139		
142		
94		
107		
125		
155		
155		

```
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
```

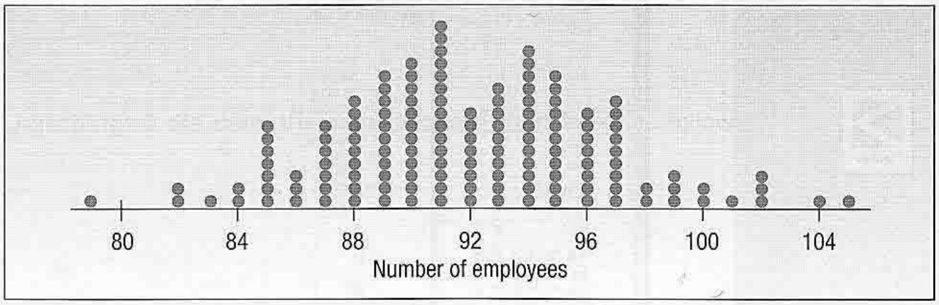
The MINITAB solution provides some additional information regarding cumulative totals. In the column to the left of the stem values are numbers such as 2, 9, 15, and so on. The number 9 indicates that there are 9 observations that have occurred before the value of 100. The number 15 indicates that 15 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. The number 9 in parentheses also tells you there are 9 observations in the middle row.

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?

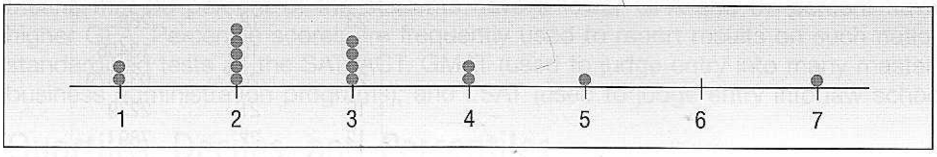
2. The rate of return for 21 stocks is:

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

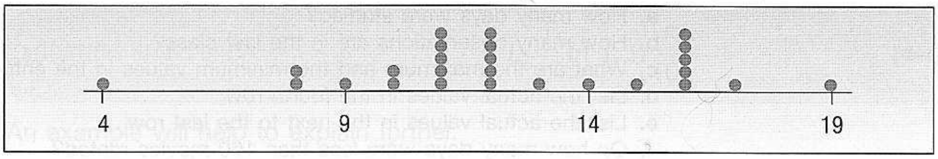
Organize this information into a stem-and-leaf display.
(a) How many rates are less than 9.0?
(b) List the rates in the 10.0 up to 11.0 category.
(c) What is the median?
(d) What are the maximum and the minimum rates of return?

Exercises

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



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



   a. What are the maximum and the minimum number of cell phones sold in a day?
   b. What is a typical number of cell phones sold?
5. The first row of a stem-and-leaf chart appears as follows: 62 | 1 3 3 7 9. Assume whole number values.
   a. What is the "possible range" of the values in this row?
   b. How many data values are in this row?
   c. List the actual values in this row of data.
6. The third row of a stem-and-leaf chart appears as follows: 21 | 0 1 3 5 7 9. Assume whole number values.
   a. What is the "possible range" of the values in this row?
   b. How many data values are in this row?
   c. List the actual values in this row of data.
7. 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
```

   a. How many days were studied?
   b. How many observations are in the first class?
   c. 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 Alltel Wireless 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

Other Measures of Dispersion

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.

In a similar fashion **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 $$L_p = (n + 1)\frac{P}{100} \qquad [4\text{–}1]$$

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.

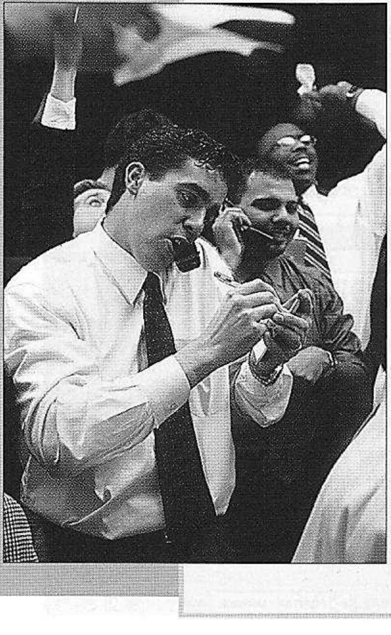
\$2,038	\$1,758	\$1,721	\$1,637	\$2,097	\$2,047	\$2,205	\$1,787	\$2,287
1,940	2,311	2,054	2,406	1,471	1,460			

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

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

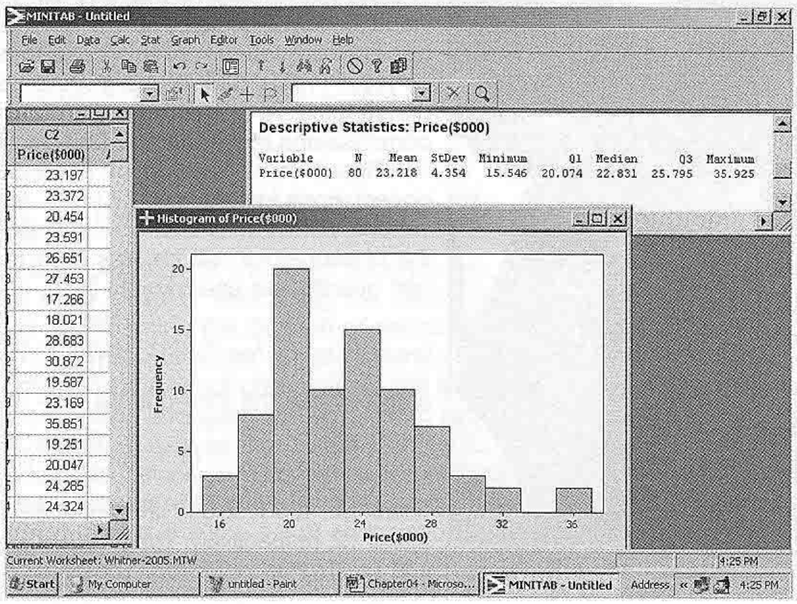
The position formula tells us that the first quartile is located between the first and the second value and that it is .75 of the distance between the first and the second values. The first value is 43 and the second is 61. So the distance between these two values is 18. To locate the first quartile, 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 quite easy to sort the data from smallest to largest and to locate percentiles and deciles. Both MINITAB and Excel output summary statistics. Listed below is the MINITAB output. The data are reported in $000. It includes the first and third quartiles, as well as the mean, median, and standard deviation for the Whitner Autoplex data (see Table 2–4). We conclude that 25 percent of the vehicles sold for less than $20,074 and that 75 percent sold for less than $25,795.

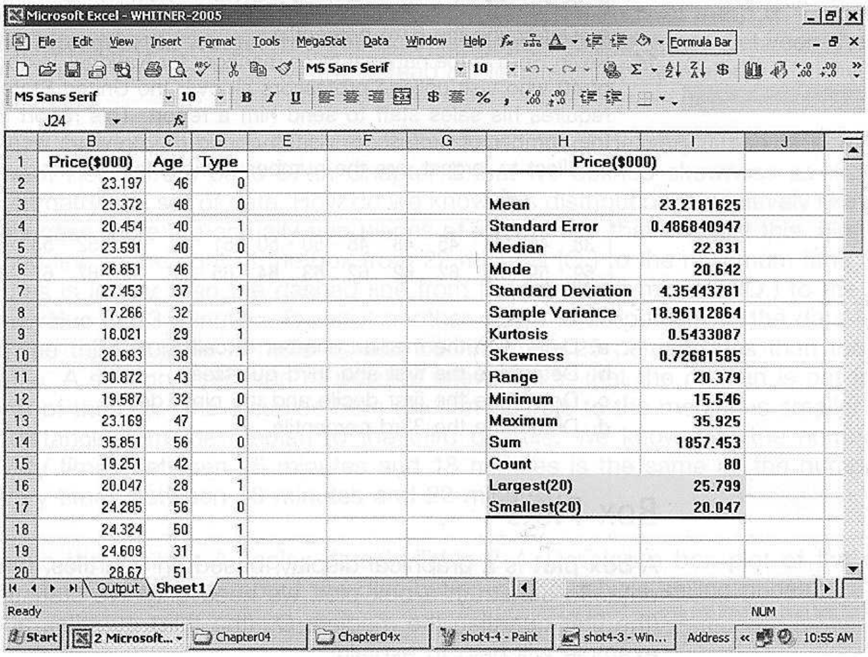


Descriptive Statistics: Price($000)

```
Variable      N    Mean  StDev  Minimum      Q1  Median      Q3  Maximum
Price($000)  80  23.218  4.354   15.546  20.074  22.831  25.795   35.925
```

C2
Price($000)
23.197
23.372
20.454
23.591
26.651
27.453
17.266
18.021
28.683
30.872
19.587
23.169
35.851
19.251
20.047
24.285
24.324

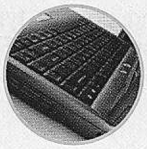
The following Excel output includes the same information regarding the mean, median, and standard deviation. It will also output the quartiles, but the method of calculation is not as precise. To find the quartiles, we multiply the sample size by the desired percentile and report the integer of that value. To explain, in the Whitner Autoplex data there are 80 observations, and we wish to locate the 25th percentile. We multiply $n + 1 = 80 + 1 = 81$ by .25; the result is 20.25. Excel will not allow us to enter a fractional value, so we use 20 and request the location of the largest 20 values and the smallest 20 values. The result is a good approximation of the 25th and 75th percentiles.



	B	C	D
1	Price($000)	Age	Type
2	23.197	46	0
3	23.372	48	0
4	20.454	40	1
5	23.591	40	0
6	26.651	46	1
7	27.453	37	1
8	17.266	32	1
9	18.021	29	1
10	28.683	38	1
11	30.872	43	0
12	19.587	32	0
13	23.169	47	0
14	35.851	56	0
15	19.251	42	1
16	20.047	28	1
17	24.285	56	0
18	24.324	50	1
19	24.609	31	1
20	28.67	51	1

Price($000)	
Mean	23.2181625
Standard Error	0.486840947
Median	22.831
Mode	20.642
Standard Deviation	4.35443781
Sample Variance	18.96112864
Kurtosis	0.5433087
Skewness	0.72681585
Range	20.379
Minimum	15.546
Maximum	35.925
Sum	1857.453
Count	80
Largest(20)	25.799
Smallest(20)	20.047

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

(a) What is the median weight?
(b) Determine the weights corresponding to the first and third quartiles.

Exercises

11. 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

12. 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

13. 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

a. Determine the first and third quartiles.
b. Determine the second decile and the eighth decile.
c. 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

a. Determine the median number of calls.
b. Determine the first and third quartiles.
c. Determine the first decile and the ninth decile.
d. Determine the 33rd percentile.

Box Plots

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

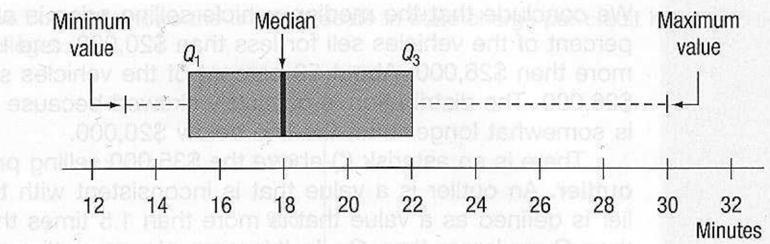
Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following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

$$\begin{aligned} \text{Minimum value} &= 13 \text{ minutes} \\ Q_1 &= 15 \text{ minutes} \\ \text{Median} &= 18 \text{ minutes} \\ Q_3 &= 22 \text{ minutes} \\ \text{Maximum value} &= 30 \text{ minutes} \end{aligned}$$

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.

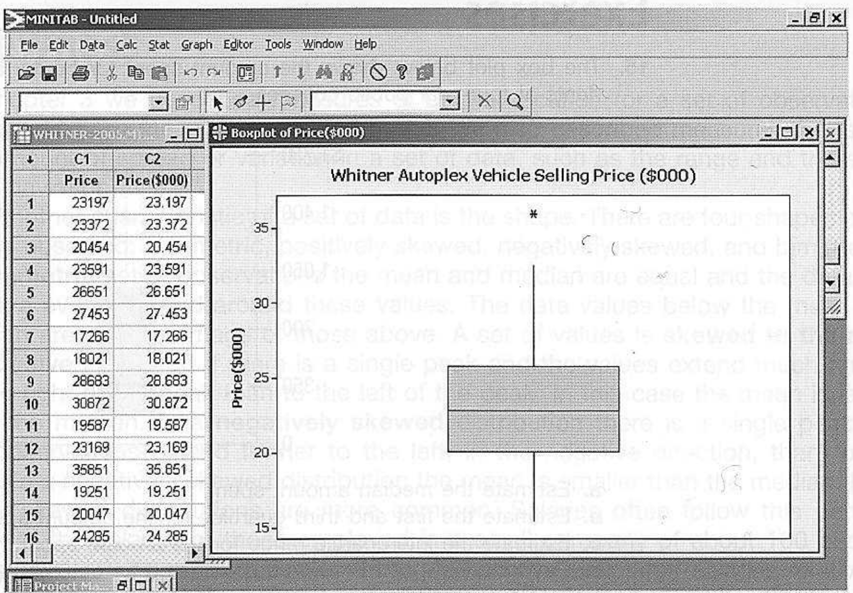
The box plot also reveals that the distribution of delivery times is positively skewed. Recall from page 67 in Chapter 3 that we defined skewness as the lack of symmetry in a set of data. How do we know this distribution is positively skewed? In this case there are actually two pieces of information 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 Whitner Autoplex data in Table 2–4. Develop a box plot of the data. What can we conclude about the distribution of the vehicle selling prices?

Solution

The MINITAB statistical software system was used to develop the following chart.



↓	C1	C2
	Price	Price($000)
1	23197	23.197
2	23372	23.372
3	20454	20.454
4	23591	23.591
5	26651	26.651
6	27453	27.453
7	17266	17.266
8	18021	18.021
9	28683	28.683
10	30872	30.872
11	19587	19.587
12	23169	23.169
13	35851	35.851
14	19251	19.251
15	20047	20.047
16	24285	24.285

We conclude that the median vehicle selling price is about \$23,000, that about 25 percent of the vehicles sell for less than \$20,000, and that about 25 percent sell for more than \$26,000. About 50 percent of the vehicles sell for between \$20,000 and \$26,000. The distribution is positively skewed because the solid line above \$26,000 is somewhat longer than the line below \$20,000.

There is an asterisk (*) above the \$35,000 selling price. An asterisk indicates an **outlier.** An outlier is a value that is inconsistent with the rest of the data. An outlier 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 \$35,000, found by

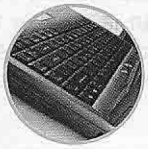
$$\text{Outlier} > Q_3 + 1.5(Q_3 - Q_1) = \$26{,}000 + 1.5(\$26{,}000 - \$20{,}000) = \$35{,}000$$

A value less than \$11,000 is also an outlier.

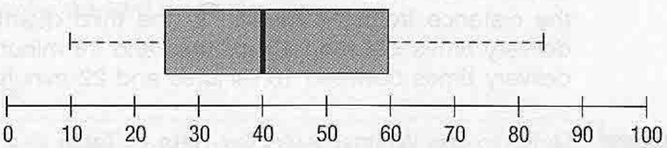
$$\text{Outlier} < Q_1 - 1.5(Q_3 - Q_1) = \$20{,}000 - 1.5(\$26{,}000 - \$20{,}000) = \$11{,}000$$

The MINITAB box plot indicates that there is only one value larger than \$35,000. However, if you look at the actual data in Table 2–4 on page 28 you will notice that there are actually two values (\$35,851 and \$35,925). The software was not able to graph two data points so close together, so it shows only one asterisk.

Self-Review 4–3



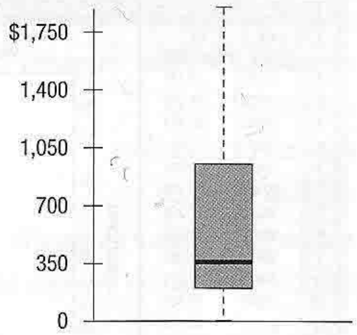
The following box plot shows the assets in millions of dollars for credit unions in Seattle, Washington.



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

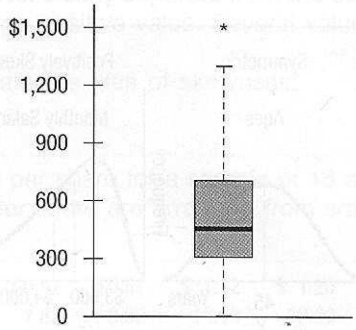
Exercises

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



a. Estimate the median amount spent.
b. Estimate the first and third quartiles for the amount spent.
c. Estimate the interquartile range for the amount spent.
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?

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 2005 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

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 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.

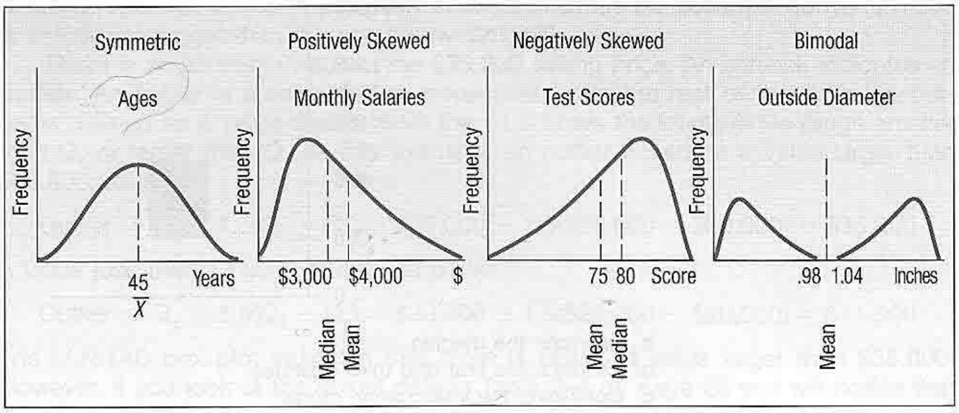


CHART 4–1 Shapes of Frequency Polygons

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
$$sk = \frac{3(\overline{X} - \text{Median})}{s} \qquad [4\text{–}2]$$

Using this relationship the 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
$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{X-\overline{X}}{s}\right)^3\right] \qquad [4\text{–}3]$$

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 $(X - \overline{X})/s$ 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 of the standard deviation. If this difference is positive, the particular value is larger than the mean; if the variation 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.

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% 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."

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 2005. 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

$$\overline{X} = \frac{\Sigma X}{n} = \frac{\$74.26}{15} = \$4.95$$

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 79 to determine the sample standard deviation.

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

Pearson's coefficient of skewness is 1.017, found by

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

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 on the next page. 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:

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

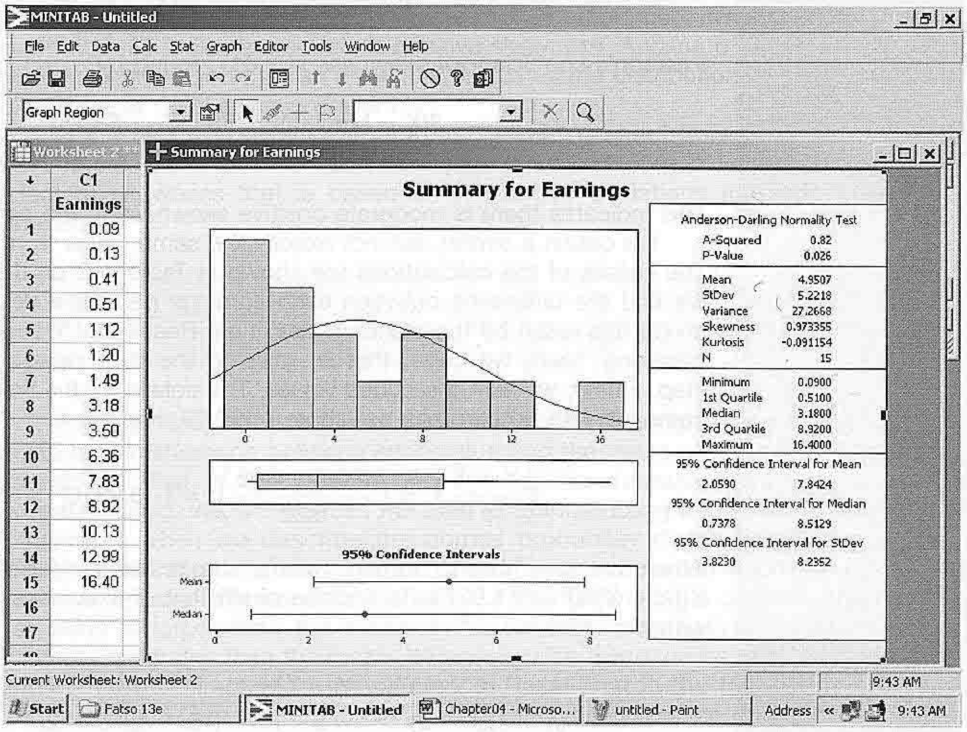
When we sum the 15 cubed values, the result is 11.8274. That is, the term $\Sigma[(X - \overline{X})/s]^3 = 11.8274$. To find the coefficient of skewness, we use formula (4–3), with $n = 15$.

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

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

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.



↓	C1
	Earnings
1	0.09
2	0.13
3	0.41
4	0.51
5	1.12
6	1.20
7	1.49
8	3.18
9	3.50
10	6.36
11	7.83
12	8.92
13	10.13
14	12.99
15	16.40
16	
17	

Summary for Earnings

```
Anderson-Darling Normality Test
   A-Squared          0.82
   P-Value           0.026

   Mean             4.9507
   StDev            5.2218
   Variance        27.2668
   Skewness       0.973355
   Kurtosis      -0.091154
   N                    15

   Minimum          0.0900
   1st Quartile     0.5100
   Median           3.1800
   3rd Quartile     8.9200
   Maximum         16.4000
95% Confidence Interval for Mean
   2.0590           7.8424
95% Confidence Interval for Median
   0.7378           8.5129
95% Confidence Interval for StDev
   3.8230           8.2352
```

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.

(a) Find the mean, median, and the standard deviation.
(b) Compute the coefficient of skewness using Pearson's method.
(c) Calculate the coefficient of skewness using the software method.
(d) What is your conclusion regarding the skewness of the data?

Exercises

For Exercises 19–22:

a. Determine the mean, median, and the standard deviation.
b. Determine the coefficient of skewness using Pearson's method.
c. 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			

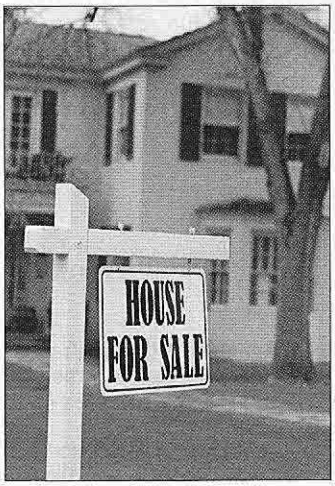
21. Listed below are the commissions earned ($000) last year by the sales representatives at Furniture Patch, Inc.

$ 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 for the New York Yankees for the year 2005. The salary information is reported in $000.

Player	Salary ($000)	Player	Salary ($000)
Rodriguez, Alex	$26,000	Wright, Jaret	$ 5,667
Jeter, Derek	19,600	Stanton, Mike	4,000
Mussina, Mike	19,000	Gordon, Tom	3,750
Johnson, Randy	16,000	Rodriguez, Felix	3,150
Brown, Kevin	15,714	Quantrill, Paul	3,000
Giambi, Jason	13,429	Martinez, Tino	2,750
Sheffield, Gary	13,000	Womack, Tony	2,000
Williams, Bernie	12,357	Sierra, Ruben	1,500
Posada, Jorge	11,000	Sturtze, Tanyon	850
Rivera, Mariano	10,500	Flaherty, John	800
Pavano, Carl	9,000	Sanchez, Rey	600
Matsui, Hideki	8,000	Crosby, Bubba	323
Karsay, Steve	6,000	Phillips, Andy	317

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 prices of vehicles sold at Whitner Autoplex. 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.

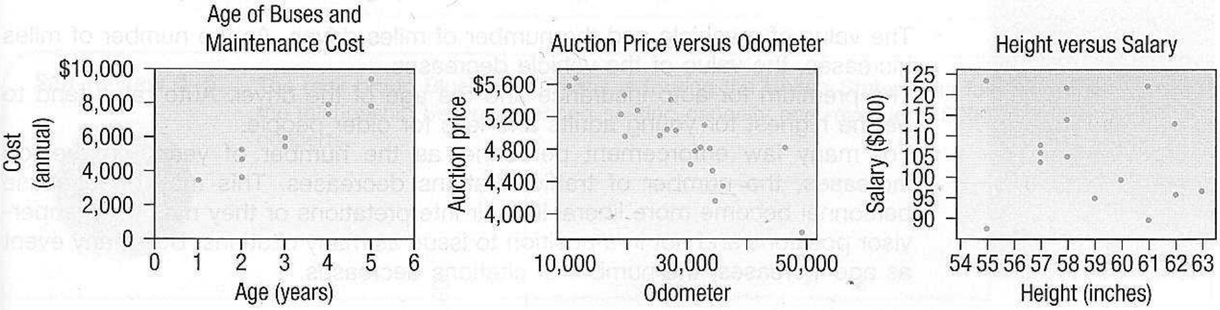


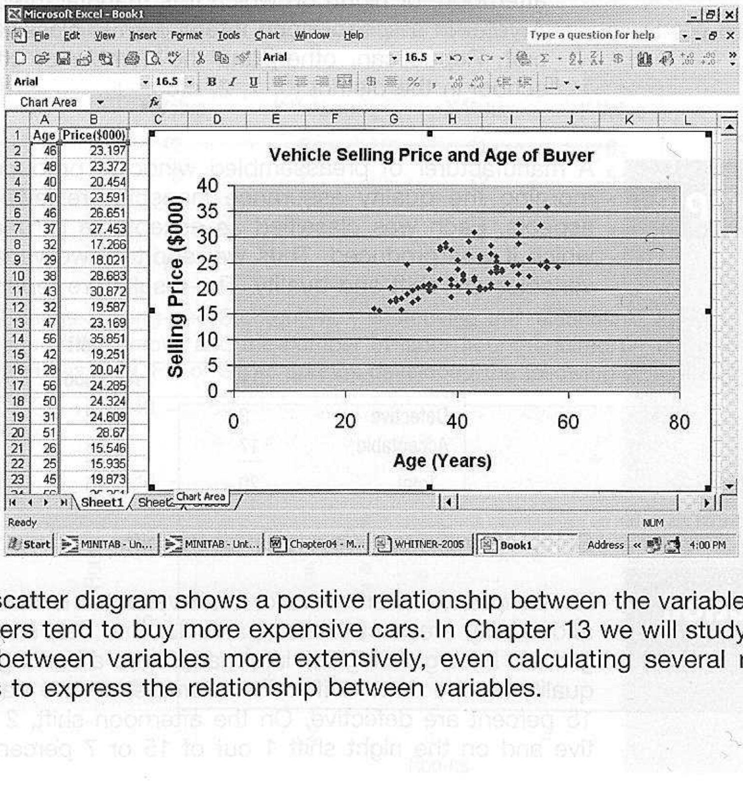
CHART 4–2 Three Examples of Scatter Diagrams.

Example

In the Introduction to Chapter 2 we presented data from AutoUSA. In this case the information concerned the prices of 80 vehicles sold last month at the Whitner Autoplex lot in Raytown, Missouri. The data shown on page 21 include the selling price of the vehicle as well as the age of the purchaser. Is there a relationship between the selling price of a vehicle and the age of the purchaser? Would it be reasonable to conclude that the more expensive vehicles are purchased by older buyers?

Solution

We can investigate the relationship between vehicle selling price and the age of the buyer with a scatter diagram. We scale age on the horizontal, or *X*-axis, and the selling price 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.



	A	B
1	Age	Price($000)
2	46	23.197
3	48	23.372
4	40	20.454
5	40	23.591
6	46	26.651
7	37	27.453
8	32	17.266
9	29	18.021
10	38	28.683
11	43	30.872
12	32	19.587
13	47	23.169
14	56	35.851
15	42	19.251
16	28	20.047
17	56	24.285
18	50	24.324
19	31	24.609
20	51	28.67
21	26	15.546
22	25	15.935
23	45	19.873

The scatter diagram shows a positive relationship between the variables. In fact, older buyers tend to buy more expensive cars. 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 Whitner Autoplex example there is a positive or direct relationship between the variables. That is, as age increased, the vehicle selling price also increased. 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 Whitner Autoplex example both age and selling price 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.**

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

A manufacturer of preassembled windows produced 50 windows yesterday. This morning the quality assurance inspector reviewed each window for all quality aspects. Each was classified as acceptable or unacceptable and by the shift on which it was produced. Thus we reported two variables on a single item. The two variables are shift and quality. The results are reported in the following table.

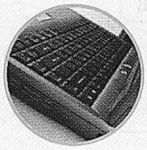
	Shift			
	Day	**Afternoon**	**Night**	**Total**
Defective	3	2	1	6
Acceptable	17	13	14	44
Total	20	15	15	50

Compare the quality levels on each shift.

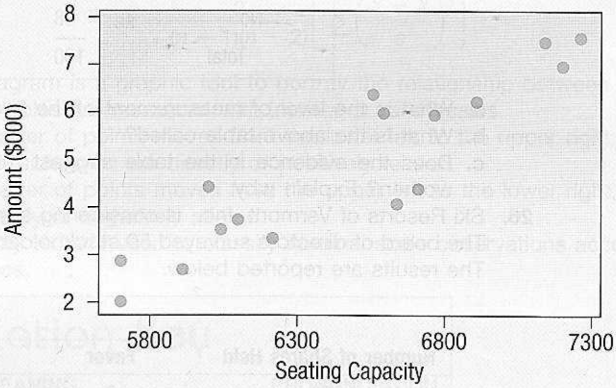
Solution

The level of measurement for both variables is nominal. That is, the variables shift and quality are such that a particular unit can only be classified or assigned into groups. By organizing the information into a contingency table we can compare the quality on the three shifts. For example, on the day shift, 3 out of 20 windows or 15 percent are defective. On the afternoon shift, 2 of 15 or 13 percent are defective and on the night shift 1 out of 15 or 7 percent are defective. Overall 12 percent of the windows are defective. Observe also that 40 percent of the windows are produced on the day shift, found by (20/50)(100). 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.



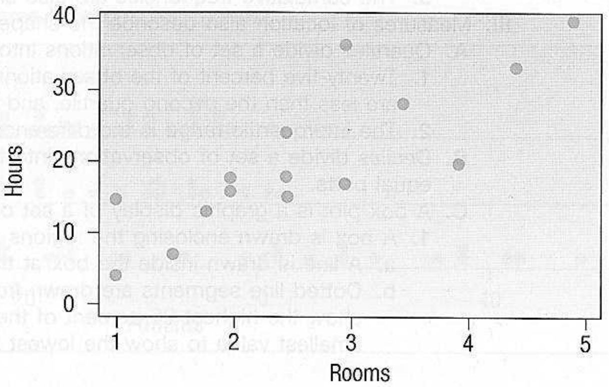
(a) What is the diagram called?
(b) How many concerts were studied?
(c) Estimate the revenue for the concert with the largest seating capacity.
(d) How would you characterize the relationship between revenue and seating capacity? Is it strong or weak, direct or inverse?

Exercises

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.



a. How many moves are in the sample?
b. 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.

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

a. What is the level of measurement of the two variables?
b. What is the above table called?
c. 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, Inc., of Alabama. The board of directors surveyed 50 stockholders concerning their position on the merger. The results are reported below.

	Opinion			
Number of Shares Held	**Favor**	**Oppose**	**Undecided**	**Total**
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

a. What level of measurement is used in this table?
b. What is this table called?
c. 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 a dot is placed above each of the values.
- **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 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 and third quartiles.
    - **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:

$$sk = \frac{3(\overline{X} - \text{Median})}{s} \qquad \textbf{[4–2]}$$

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

$$sk = \frac{n}{(n-1)(n-2)}\left[\sum\left(\frac{X-\overline{X}}{s}\right)^3\right] \qquad \textbf{[4–3]}$$

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.

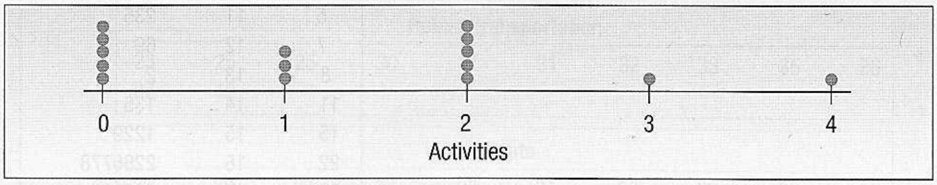
VI. A contingency table is used to classify nominal-scale observations according to two characteristics.

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.

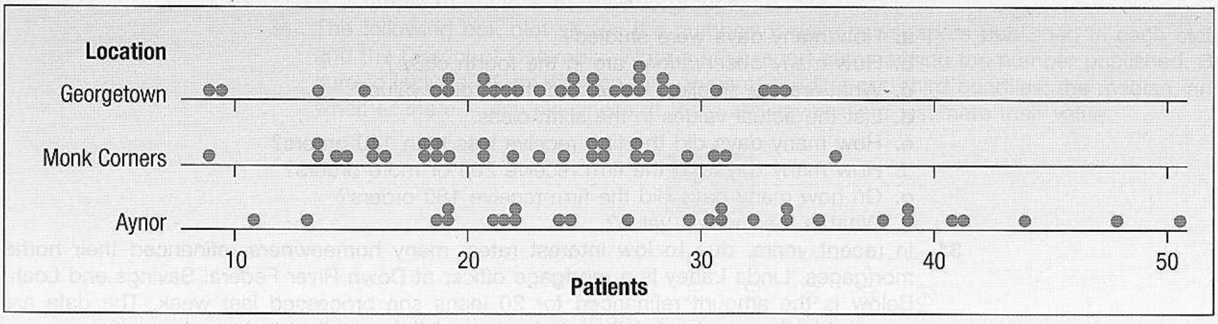


a. What is the name given to this chart?

b. How many students were in the study?

c. 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 following stem-and-leaf display shows the number of minutes of daytime TV viewing for a sample of college students.

2	0	05
3	1	0
6	2	137
10	3	0029
13	4	499
24	5	00155667799
30	6	023468
(7)	7	1366789
33	8	01558
28	9	1122379
21	10	022367899
12	11	2457
8	12	4668
4	13	249
1	14	5

a. How many college students were studied?
b. How many observations are in the second class?
c. What are the smallest value and the largest value?
d. List the actual values in the fourth row.
e. How many students watched less than 60 minutes of TV?
f. How many students watched 100 minutes or more of TV?
g. What is the median value?
h. How many students watched at least 60 minutes but less than 100 minutes?

30. The following stem-and-leaf display reports the number of orders received per day by the Northwest Regional Office of Oriental Trading Co., Inc.

1	9	1
2	10	2
5	11	235
7	12	69
8	13	2
11	14	135
15	15	1229
22	16	2266778
27	17	01599
(11)	18	00013346799
17	19	03346
12	20	4679
8	21	0177
4	22	45
2	23	17

a. How many days were studied?
b. How many observations are in the fourth class?
c. What are the smallest value and the largest value?
d. List the actual values in the sixth class.
e. How many days did the firm receive less than 140 orders?
f. How many days did the firm receive 200 or more orders?
g. On how many days did the firm receive 180 orders?
h. What is the median value?

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							

a. Find the median, first quartile, and third quartile.
b. Find the 26th and 83rd percentiles.
c. 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

a. 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.
b. 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.
c. 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

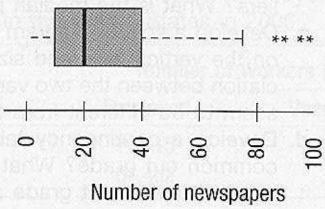
23	25	25	30	31	31	32	33	35	36
37	42								

Private

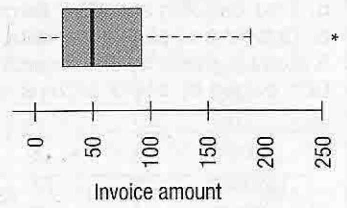
32	32	33	34	37	37	38	38	38	39
40	44								

a. 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.
b. 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.
c. 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 it 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. National Muffler Company claims it will change your muffler in less than 30 minutes. An investigative reporter for WTOL Channel 11 monitored 30 consecutive muffler changes at the National outlet on Liberty Street. The number of minutes to perform changes is reported below.

44	12	22	31	26	22	30	26	18	28	12
40	17	13	14	17	25	29	15	30	10	28
16	33	24	20	29	34	23	13			

a. Develop a box plot for the time to change a muffler.
b. Does the distribution show any outliers?
c. Summarize your findings in a brief report.

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				

a. 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?
b. 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?
c. 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?
d. Develop a contingency table for the variables shape and cut grade. What is most common cut grade? What is the most common shape? What is the most common combination of cut grade and shape?

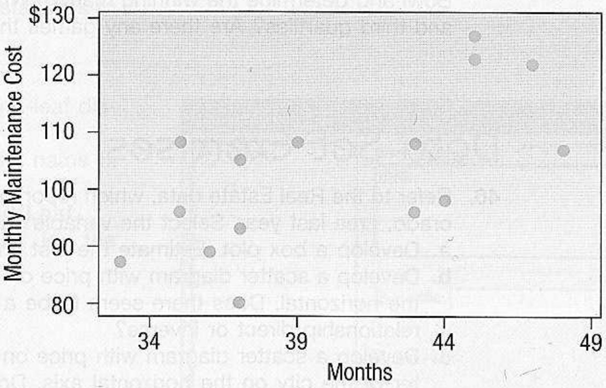
38. Listed below is the amount of commissions earned last month for the eight members of the sales staff at Best Electronics. Calculate the coefficient of skewness using both methods. *Hint:* Use of a spreadsheet will expedite the calculations.

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?

	Age			
Number of Condiments	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 below is a table showing the number of employed and unemployed workers 20 years or older by gender in the United States in 2006.

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

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

exercises.com

44. Refer to Exercise 86 on page 94, which suggests websites to find information on the Dow Jones Industrial Average. One of the websites suggested is Bloomberg, which is an excellent source of business data. The Bloomberg website is: http://bloomberg.com. Click on **Market Data,** then **Stocks,** and **Dow.** You should now have at the bottom of the page a listing of the current selling price of the 30 stocks that make up the Dow Jones Industrial Average. Find the percent change from yesterday for each of the 30 stocks. Develop charts to depict the percent change.

45. The following website gives the Super Bowl results since the game was first played in 1967: http://www.superbowl.com/history/recaps. Download the scores for each Super Bowl and determine the winning margin. What was the typical margin? What are the first and third quartiles? Are there any games that were outliers?

Data Set Exercises

46. Refer to the Real Estate data, which report information on homes sold in the Denver, Colorado, area last year. Select the variable selling price.
a. Develop a box plot. Estimate the first and the third quartiles. Are there any outliers?
b. 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?
c. 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?

47. Refer to the Global Financial Performance data set that reports information on 148 corporations.
a. Select the variable "employees". Develop a box plot. Are there any outliers? List the outliers.
b. Select the variable "sales". Develop a box plot. Are there any outliers? List the outliers. What are the quartiles? Write a brief summary of your analysis.
c. Select the variable "Net income". Develop a box plot. Are there any outliers?
d. Draw a scatter diagram with the sales on the horizontal axis and net income on the vertical axis. Based on diagram, make three statements about the relationship between net income and sales.

48. Refer to the Wage data, which report information on annual wages for a sample of 100 workers. Also included are variables relating to industry, years of education, and gender for each worker.
a. Develop a stem-and-leaf chart for the variable annual wage. Are there any outliers? Write a brief summary of your findings.
b. Develop a stem-and-leaf chart for the variable years of education. Are there any outliers? Write a brief summary of your findings.
c. Draw a bar chart of the variable occupation. Write a brief report summarizing your findings.

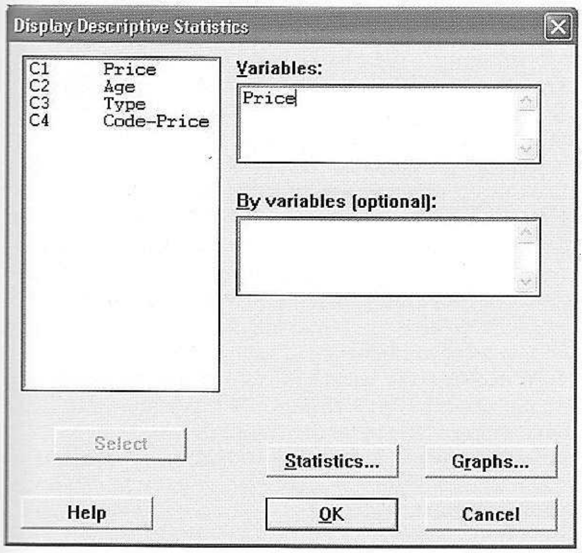
49. Refer to the CIA data, which report demographic and economic information on 62 countries.
a. Select the variable life expectancy. Develop a box plot. Find the first and third quartiles. Are there any outliers? Is the distribution skewed or symmetric? Write a paragraph summarizing your findings.
b. Select the variable GDP/cap. Develop a box plot. Find the first and third quartiles. Are there any outliers? Is the distribution skewed or symmetric? Write a paragraph summarizing your findings.
c. Develop a stem-and-leaf chart for the variable referring to the number of cell phones. Summarize your findings.

Software Commands

1. The MINITAB commands for the dot plot on page 100 are:
   a. Enter the vehicles sold at Smith Ford Mercury Jeep in column C1 and Brophy Honda Volkswagen in C2. Name the variables accordingly.
   b. 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 **Smith** and **Brophy** as the variables to **Graph,** click on **Labels** and write an appropriate title.
   c. To calculate the descriptive statistics shown in the output select **Stat, Basic statistics,** and then **Display Descriptive statistics.** In the dialog box select **Smith** and **Brophy** as the **Variables,** click on **Statistics** and select the desired statistics to be output, and finally click **OK** twice.

2. The MINITAB commands for the stem-and-leaf display on page 103 are:
   a. Import the data from the CD. The file name is **Table4–1.**
   b. Select **Graph,** and click on **Stem-and-Leaf.**
   c. Select the variable **Spots,** enter *10* for the **Increment,** and then click **OK.**

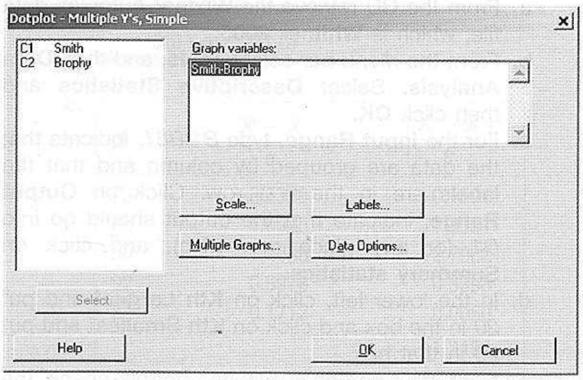
3. The MINITAB commands for the descriptive summary on page 108 are:



C1 Price
C2 Age
C3 Type
C4 Code-Price

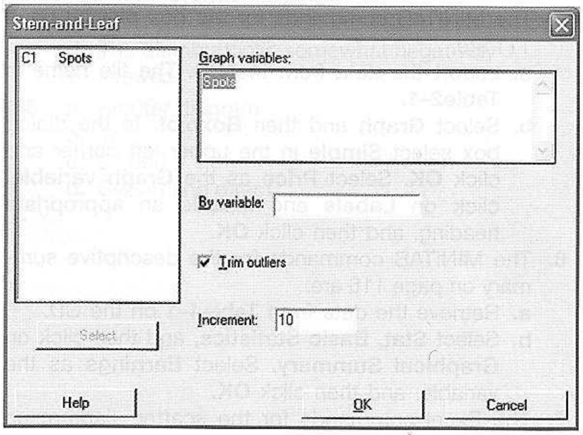
Variables: Price

By variables (optional):



C1 Smith
C2 Brophy

Graph variables: Smith-Brophy



C1 Spots

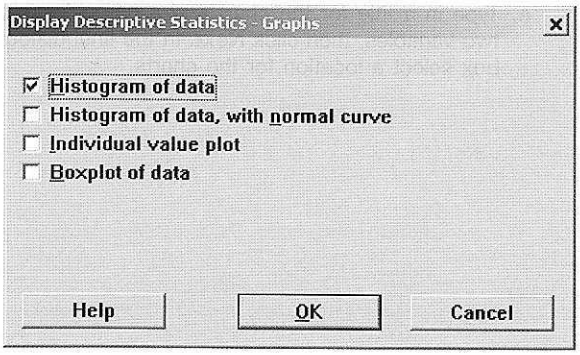
Graph variables: Spots

By variable:

☑ Trim outliers

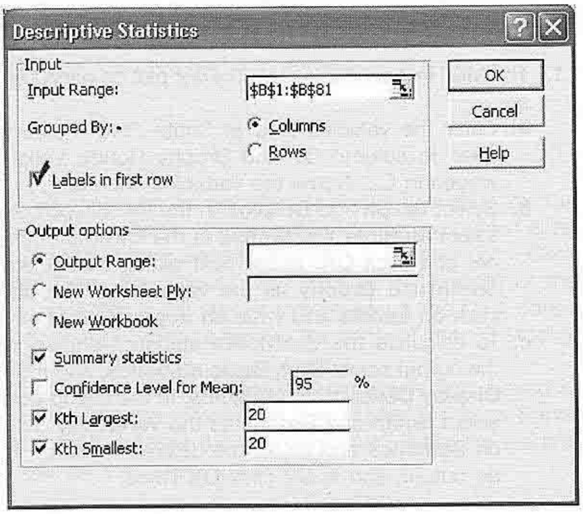
Increment: 10

a. Import the Whitner Autoplex data from the CD. The file name is **Whitner 2005.** Select the variable **Price.**
b. From the toolbar select **Stat, Basic Statistics,** and **Display Descriptive Statistics.** In the dialog box select **Price** as the **Variable,** in the lower right click on **Graphs.** In this box select **Graphs,** click on **Histogram of data,** and then click **OK** twice.



☑ Histogram of data
☐ Histogram of data, with normal curve
☐ Individual value plot
☐ Boxplot of data

4. The Excel Commands for the descriptive statistics on page 109 are:
   a. From the CD retrieve the Whitner Autoplex data file, which is **Whitner 2005.**
   b. From the menu bar select **Tools,** and then **Data Analysis.** Select **Descriptive Statistics** and then click **OK.**
   c. For the **Input Range,** type *B1:B81,* indicate that the data are grouped by column and that the labels are in the first row. Click on **Output Range,** indicate that the output should go into *D1* (or any place you wish), and click on **Summary statistics.**
   d. In the lower left, click on **Kth Largest** and put *20* in the box and click on **Kth Smallest** and put *20* in that box.
   e. After you get your results, double-check the count in the output to be sure it contains the correct number of values.



Descriptive Statistics

Input

Input Range: $B$1:$B$81

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range:

○ New Worksheet Ply:

○ New Workbook

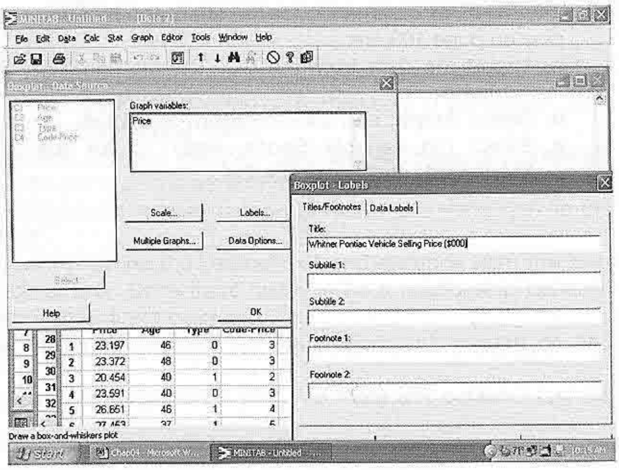
☑ Summary statistics

□ Confidence Level for Mean: 95 %

☑ Kth Largest: 20

☑ Kth Smallest: 20

5. The MINITAB commands for the box plot on page 111 are:
   a. Import the data from the CD. The file name is **Table2–1.**
   b. Select **Graph** and then **Boxplot.** In the dialog box select **Simple** in the upper left corner and click **OK.** Select **Price** as the **Graph variable,** click on **Labels** and include an appropriate heading, and then click **OK.**



MINITAB - Untitled

Boxplot - Data Source

Graph variables: Price

Boxplot - Labels

Title: Whitner Pontiac Vehicle Selling Price ($000)

Subtitle 1:

Subtitle 2:

Footnote 1:

Footnote 2:

	Price	Age	Type	Code-Price
1	23.197	46	0	3
2	23.372	48	0	3
3	20.454	40	1	2
4	23.591	40	0	3
5	26.651	46	1	4

Draw a box-and-whiskers plot

6. The MINITAB commands for the descriptive summary on page 116 are:
   a. Retrieve the data from **Table4–1** on the CD.
   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 119 are:
   a. Retrieve the data from **Whitner 2005** on the CD.
   b. You will need to copy the variables to other columns on the spreadsheet with age in a column and price in the next column. This will allow you to place price on the vertical axis and age on the horizontal axis.
   c. Click on **Chart** under **Insert** to start **Chart Wizard**, select **XY (Scatter)** and the sub-type in the top left, and then click on **Next.**
   d. Select or highlight the variables age followed by price, then click **Next** again.
   e. Type in a title for the chart and a name for the two variables, then click **Next.** In the final dialog box select a location for the charts.

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.

4–4 a. $\overline{X} = \dfrac{407}{5} = 81.4$, Median = 84

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

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

c.

X	$\dfrac{X - \overline{X}}{s}$	$\left[\dfrac{X - \overline{X}}{s}\right]^3$
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 or 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. These may include organizing the values into a frequency distribution and computing measures of location and measures of dispersion and skewness.

Exhaustive Each observation must fall into one of the categories.

Inferential statistics, also called **statistical inference** This facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if 2 out of the 10 hand calculators sampled are 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.

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

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 logically 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 distances between numbers are of a known constant size and *there is a true zero point,* and the ratio of two values is meaningful, the measurement is 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 line charts, bar charts, and pie charts. They are very useful, for example, for depicting the trend in long-term debt or percent changes in profit from last year to this year.

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.

Frequency distribution A grouping of data into classes showing the number of observations in each of the mutually exclusive classes. For example, data are organized into classes such as $1,000 up to $2,000, $2,000 up to $3,000, and so on to summarize the information.

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.

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 $\overline{X}$ 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 1985 and $692 million in 2005, what is the average annual percent increase?

Mean deviation The mean of the deviations from the mean, disregarding signs. It is abbreviated 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 number that pinpoints 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 central 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 A measure of dispersion computed as the maximum value minus the minimum 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 × $10) + (20 × $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 or more nominal characteristics.

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

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 uses 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.

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

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

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 dataset 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.

Exercises

1. Which of the following are not included in the definition of statistics?
- **a.** Collecting.
- **b.** Organizing.
- **c.** Selling.
- **d.** Interpreting.

2. Patrons at a local restaurant are asked to rate the service as excellent, good, fair, or poor. The level of measurement is
- **a.** Nominal.
- **b.** Ordinal.
- **c.** Interval.
- **d.** Ratio.

3. A person's age, income, height, and weight are all examples of
- **a.** Population variables.
- **b.** Qualitative variables.
- **c.** Random variables.
- **d.** Quantitative variables.

4. Which of the following statements are true of a frequency table?
- **a.** It is based on qualitative data.
- **b.** The grouping must be mutually exclusive.
- **c.** The variable is nonnumeric.
- **d.** All of the above are correct.

5. In a bar chart
  a. The frequencies are always reported on the vertical axis.
  b. The classes are reported on the horizontal axis.
  c. The variable of interest is qualitative.
  d. All of the above are correct.

6. In a frequency distribution the number of observations in each class is called
  a. The class midpoint.
  b. The class frequency.
  c. The class interval.
  d. None of the above.

7. A set of data includes 75 observations. How many classes would you recommend?
  a. 2
  b. 7
  c. 9
  d. 8

A sample of five of the vice presidents at Midlands Federal Savings Bank is selected. They have been with the company 11, 4, 9, 16, and 10 years. Use this information to answer questions 8 through 12.

8. What is the mean number of years with the bank? _____

9. What is the median number of years with the bank? _____

10. What is the range of the number of years with the bank? _____

11. What is the standard deviation of the number of years with the bank? _____

12. What is the 80th percentile? _____

13. A useful measure to observe the lack of symmetry in a set of data is called the
  a. Coefficient of skewness.
  b. Coefficient of normalcy.
  c. Coefficient of variation.
  d. Variance.

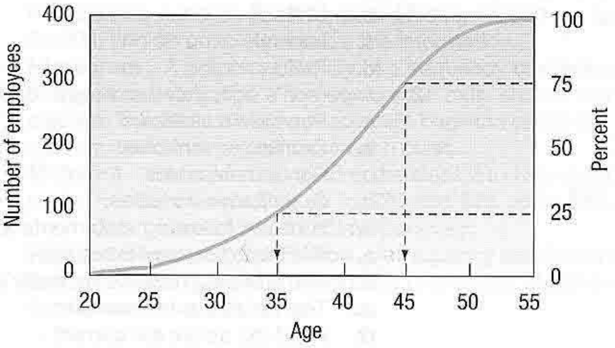
14. For a set of data the mean, median, and the mode are all 100. The standard deviation is 4. About 95 percent of the observations lie between
  a. 92 and 108.
  b. 96 and 104.
  c. 95 and 105.
  d. Cannot be estimated.

15. Fine Furniture Inc. produced 2,460 desks in 1995 and 6,520 in 2005. What is the geometric mean annual rate of increase for the period? _____

16. A graph that shows the relationship between two interval or ratio variables is called a
  a. Contingency table.
  b. Scatter diagram.
  c. Stem-and-leaf display.
  d. Dot plot.

17. A summary of data measured on two nominal variables is called a
  a. Scatter diagram.
  b. Contingency table.
  c. Frequency distribution.
  d. Histogram.

Refer to the adjacent graph to answer questions 18 through 20.



18. The graph is called a
  a. Frequency distribution.
  b. Cumulative frequency distribution.
  c. Frequency polygon.
  d. Histogram.

19. The interquartile range is
  a. 5
  b. 10
  c. 15
  d. 35

20. Which of the following statements is true?
- **a.** About 300 employees are younger than 30.
- **b.** 25 percent of the employees are older than 45.
- **c.** The interquartile range represents 60 percent of the employees.
- **d.** 75 percent of the employees are younger than 35.

21. 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

Using the preceding raw data and a statistical package (such as MINITAB):
- **a.** Organize the data into a frequency distribution.
- **b.** Calculate the mean, median, and other descriptive measures. Include a dot plot, stem-and-leaf display, and a box plot. You decide on the class interval.
- **c.** Interpret the computer output; that is, describe the central tendency, spread, skewness, and other measures.

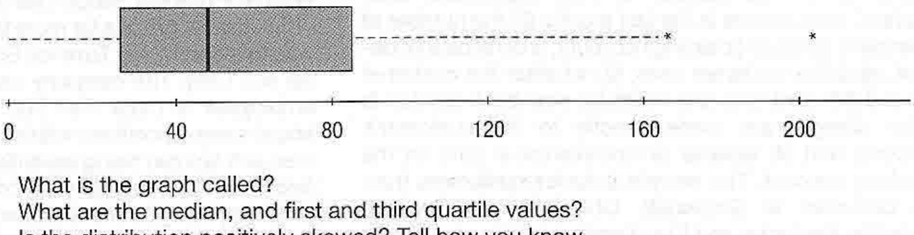
22. A sample of 12 homes sold last week in St. Paul, Minnesota, revealed the following information. Draw a scatter diagram. Can we conclude that, as the size of the home (reported below in thousands of square feet) increases, the selling price (reported in $ thousands) also increases?

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

23. Following are the ages when the 43 U.S. presidents began their terms in office. Organize the data into a stem-and-leaf chart. Also construct a dot plot. Determine a typical age at the time of inauguration. Comment on the variation in age.

57	61	57	57	58	57	61	54	68	51
49	64	50	48	65	52	56	46	54	49
50	47	55	55	54	42	51	56	55	51
54	51	60	62	43	55	56	61	52	69
65	46	54							

24. Refer to the following diagram.



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

25. The per capita personal income by state (including the District of Columbia), in thousands of dollars, follows.

11.1	17.7	13.2	10.7	16.8	15.1	19.2	15.1
18.9	14.3	13.2	14.7	11.4	15.4	12.9	13.2
14.4	11.1	11.2	12.7	16.6	17.5	14.1	14.7
9.5	13.6	11.9	13.8	15.1	15.9	18.3	11.1
17.1	12.2	12.3	13.7	12.4	12.2	13.9	14.7
11.1	11.9	11.8	13.5	10.7	12.8	15.4	14.5
10.5	13.8	13.2					

a. Organize these data into a frequency distribution.
b. What is a "typical" per capita income for a state?
c. How much variation in the income data is there?
d. Is the distribution symmetrical?
e. Summarize your findings.

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 relatively new 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.

1. 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?
2. 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.
3. 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