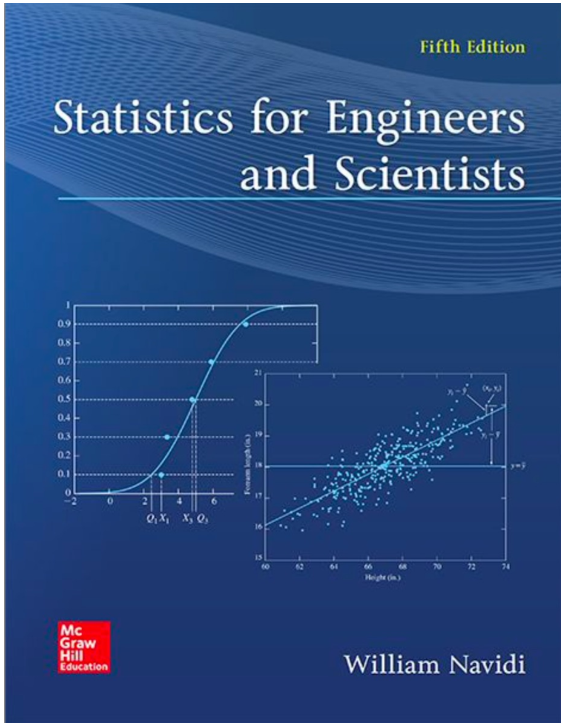


Chapter 1

Sampling and Descriptive Statistics (part 3)

Graphical Summaries

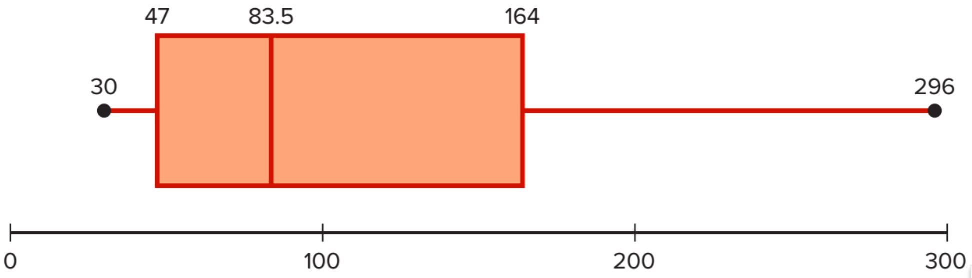
- Statistical graphs can be used to describe the data set or to analyze it.
- Graphs are also useful in getting the audience’s attention in a publication or a speaking presentation.
- They can be used to discuss an issue, reinforce a critical point, or summarize a data set.
- They can also be used to discover a trend or pattern in a situation over a period of time.

Exploratory Data Analysis (cont.)

- The **Five-Number Summary** is composed of the following numbers: Lowest (min), Q_1, MD, Q_3, Highest (max)
- The Five-Number Summary can be graphically represented using a **Boxplot** (also called a box and whisker plot).

Constructing Boxplots

1. Find the five-number summary.
2. Draw a horizontal axis with a scale that includes the maximum and minimum data values.
3. Draw a box with vertical sides through Q_1 *and* Q_3, *and draw a vertical line* though the median.
4. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.



Example: Meteorites

The number of meteorites found in 10 U.S. states is shown. Construct a boxplot for the data.

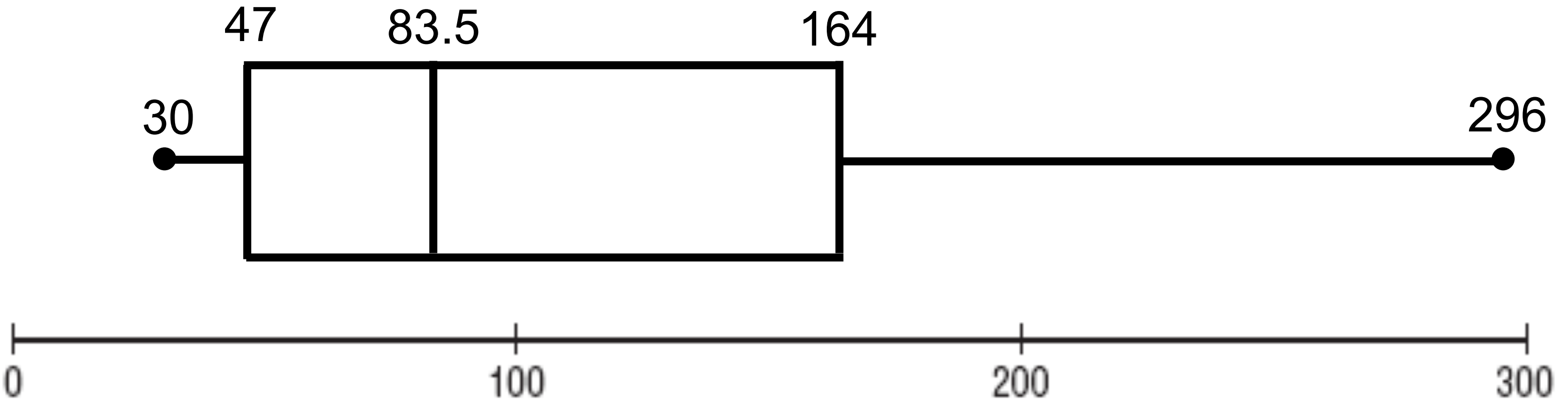
89, 47, 164, 296, 30, 215, 138, 78, 48, 39

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

↑ ↑ ↑ ↑ ↑

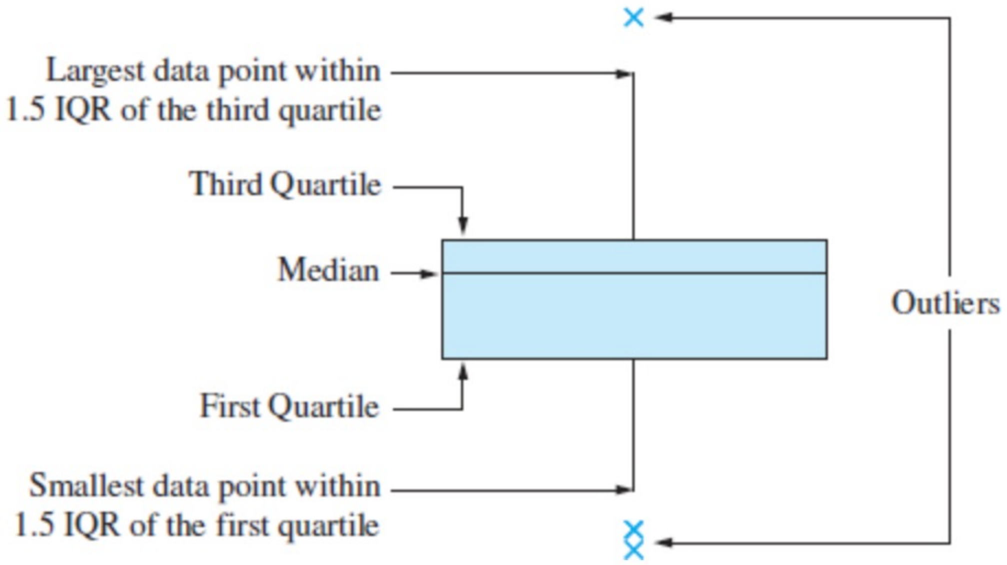
Low Q_1 MD Q_3 High

Five-Number Summary: 30-47-83.5-164-296



Boxplots and Outliers

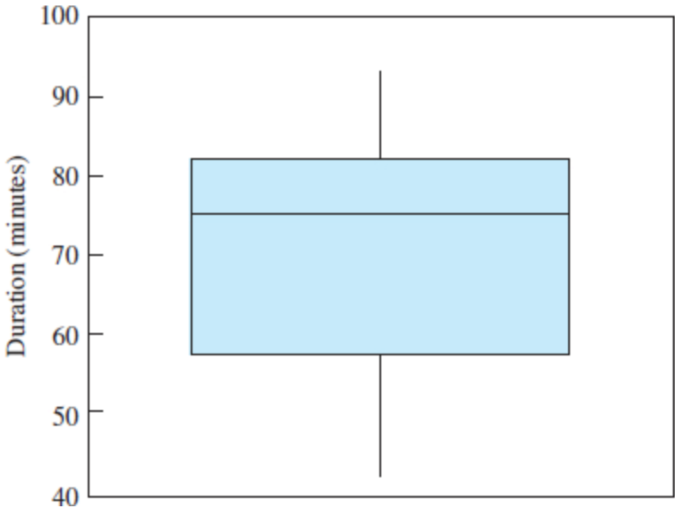
- Outliers can also be shown on boxplots.
- The "outliers" are plotted individually and are indicated by crosses in the figure.
- Extending from the top and bottom of the box are vertical lines called "whiskers."
- The whiskers end at the most extreme data point that is not an outlier.


Largest data point within 1.5 IQR of the third quartile
Third Quartile
Median
First Quartile
Smallest data point within 1.5 IQR of the first quartile
Outliers

Boxplots and Outliers (cont.)

Steps in the Construction of a Boxplot

1. Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box.
2. Find the largest sample value that is no more than 1.5 IQR above the third quartile, and the smallest sample value that is no more than 1.5 IQR below the first quartile. Extend vertical lines (whiskers) from the quartile lines to these points.
3. Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile, are designated as outliers. Plot each outlier individually.



Boxplot for the Old Faithful dormant period data presented in Table 1.6.

Comparative Boxplots

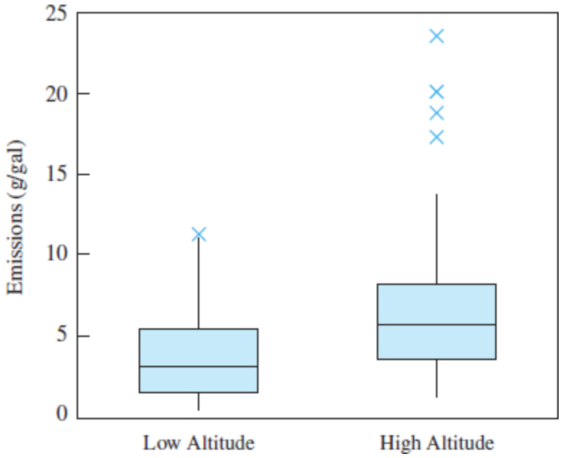
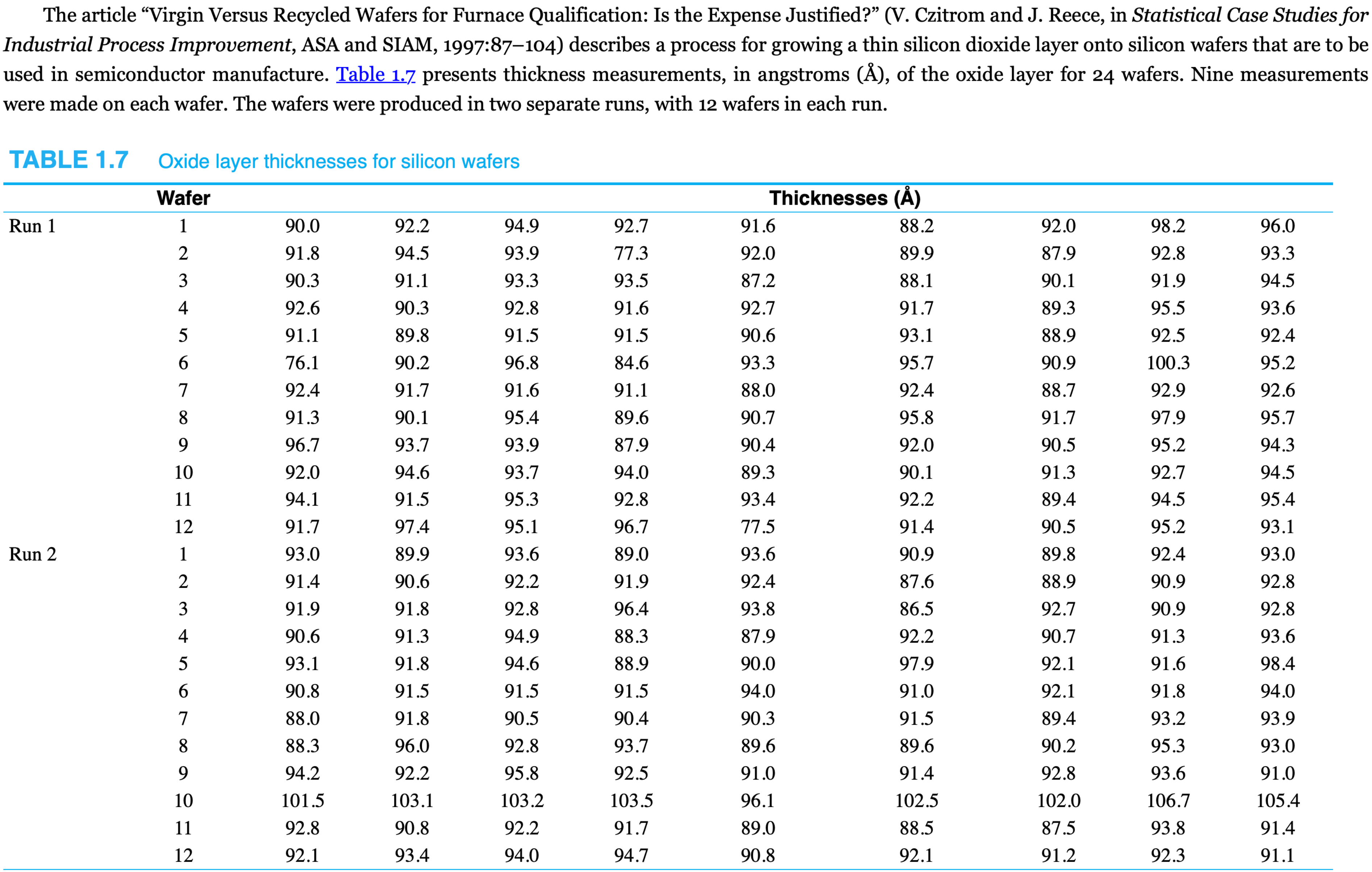


FIGURE 1.15 Comparative boxplots for PM emissions data for vehicles driven at high versus low altitudes.



The article “Virgin Versus Recycled Wafers for Furnace Qualification: Is the Expense Justified?” (V. Czitrom and J. Reece, in *Statistical Case Studies for Industrial Process Improvement*, ASA and SIAM, 1997:87–104) describes a process for growing a thin silicon dioxide layer onto silicon wafers that are to be used in semiconductor manufacture. Table 1.7 presents thickness measurements, in angstroms (Å), of the oxide layer for 24 wafers. Nine measurements were made on each wafer. The wafers were produced in two separate runs, with 12 wafers in each run.

TABLE 1.7 Oxide layer thicknesses for silicon wafers

	Wafer	Thicknesses (Å)								
Run 1	1	90.0	92.2	94.9	92.7	91.6	88.2	92.0	98.2	96.0
	2	91.8	94.5	93.9	77.3	92.0	89.9	87.9	92.8	93.3
	3	90.3	91.1	93.3	93.5	87.2	88.1	90.1	91.9	94.5
	4	92.6	90.3	92.8	91.6	92.7	91.7	89.3	95.5	93.6
	5	91.1	89.8	91.5	91.5	90.6	93.1	88.9	92.5	92.4
	6	76.1	90.2	96.8	84.6	93.3	95.7	90.9	100.3	95.2
	7	92.4	91.7	91.6	91.1	88.0	92.4	88.7	92.9	92.6
	8	91.3	90.1	95.4	89.6	90.7	95.8	91.7	97.9	95.7
	9	96.7	93.7	93.9	87.9	90.4	92.0	90.5	95.2	94.3
	10	92.0	94.6	93.7	94.0	89.3	90.1	91.3	92.7	94.5
	11	94.1	91.5	95.3	92.8	93.4	92.2	89.4	94.5	95.4
	12	91.7	97.4	95.1	96.7	77.5	91.4	90.5	95.2	93.1
Run 2	1	93.0	89.9	93.6	89.0	93.6	90.9	89.8	92.4	93.0
	2	91.4	90.6	92.2	91.9	92.4	87.6	88.9	90.9	92.8
	3	91.9	91.8	92.8	96.4	93.8	86.5	92.7	90.9	92.8
	4	90.6	91.3	94.9	88.3	87.9	92.2	90.7	91.3	93.6
	5	93.1	91.8	94.6	88.9	90.0	97.9	92.1	91.6	98.4
	6	90.8	91.5	91.5	91.5	94.0	91.0	92.1	91.8	94.0
	7	88.0	91.8	90.5	90.4	90.3	91.5	89.4	93.2	93.9
	8	88.3	96.0	92.8	93.7	89.6	89.6	90.2	95.3	93.0
	9	94.2	92.2	95.8	92.5	91.0	91.4	92.8	93.6	91.0
	10	101.5	103.1	103.2	103.5	96.1	102.5	102.0	106.7	105.4
	11	92.8	90.8	92.2	91.7	89.0	88.5	87.5	93.8	91.4
	12	92.1	93.4	94.0	94.7	90.8	92.1	91.2	92.3	91.1

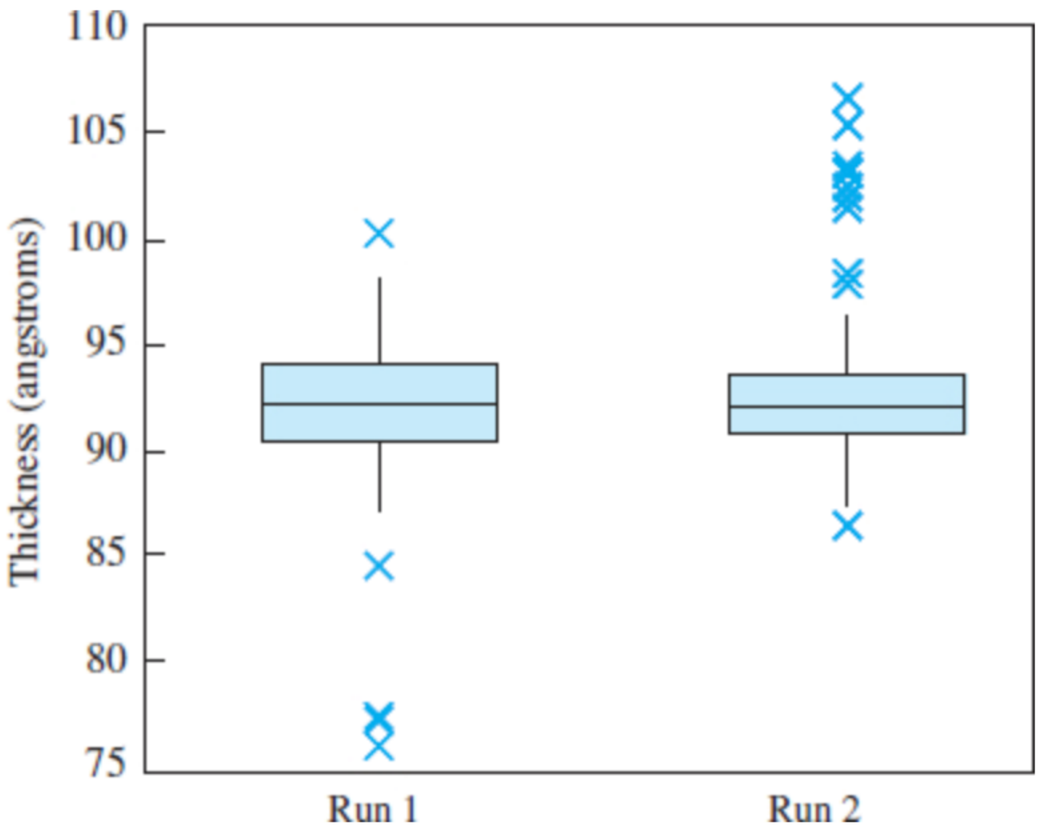
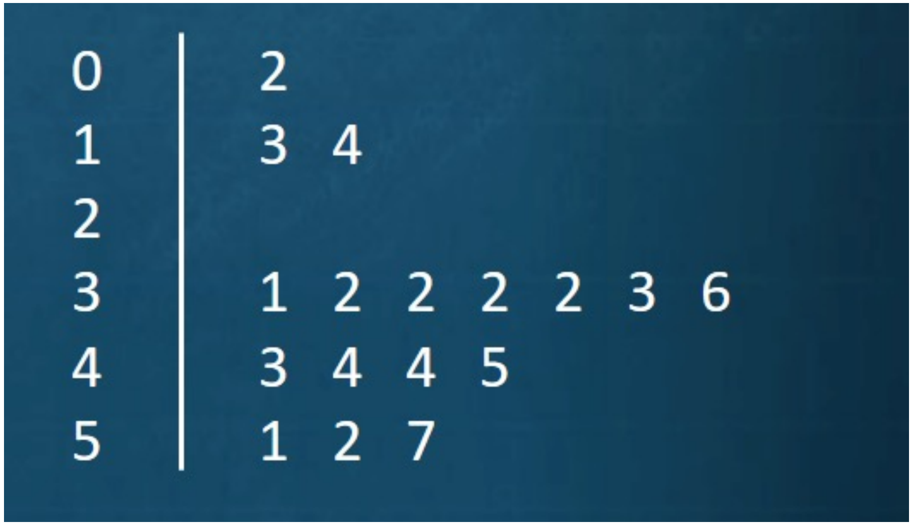


FIGURE 1.16 Comparative boxplots for oxide layer thickness data.

Stem and Leaf Plot

- A **stem and leaf plot** is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.
- The stem and leaf plot is a method of organizing data and is a combination of sorting and graphing.
- It has the advantage over a grouped frequency distribution of retaining the actual data while showing them in graphical form.



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

Example

At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	43	33	51
14	32	44	13
36	52	32	23
32	20	57	44
31	2	32	45

Example…

Step 1 is to arrange the data in order

2	25	32	44
13	31	33	45
14	32	36	51
20	32	43	52
23	32	44	57

Example…

Step 2 is to separate the data according to the first digit.

The smallest number is 2. Which would have a 0 in the tens place.

02

This is significant in the process of constructing a stem and leaf plot.

The next two values are 13 and 14. Which have a 1 in the tens place.

13, 14

Example…

We will continue this process until we have listed each value in the data set.

02

13, 14

20, 23, 25

31, 32, 32, 32, 32, 33, 36

43, 44, 44, 45

51, 52, 57

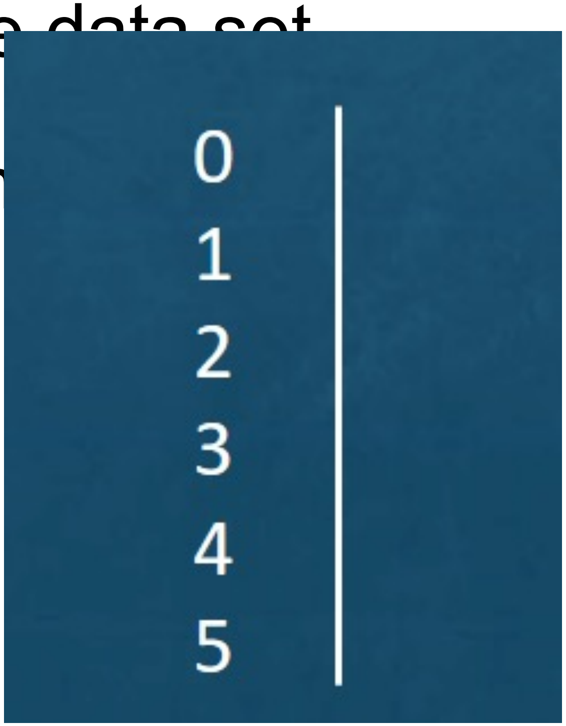
Example…

Step 3 is to create a display using the leading digit as the stem and the trailing digit as the leaf. We will identify *the stem unit as representing the ten's place*. Then we will list each of the ten's place digits that are represented in the data set.

We will separate the stems from the leaves using a vertical line.

Stem unit = 10

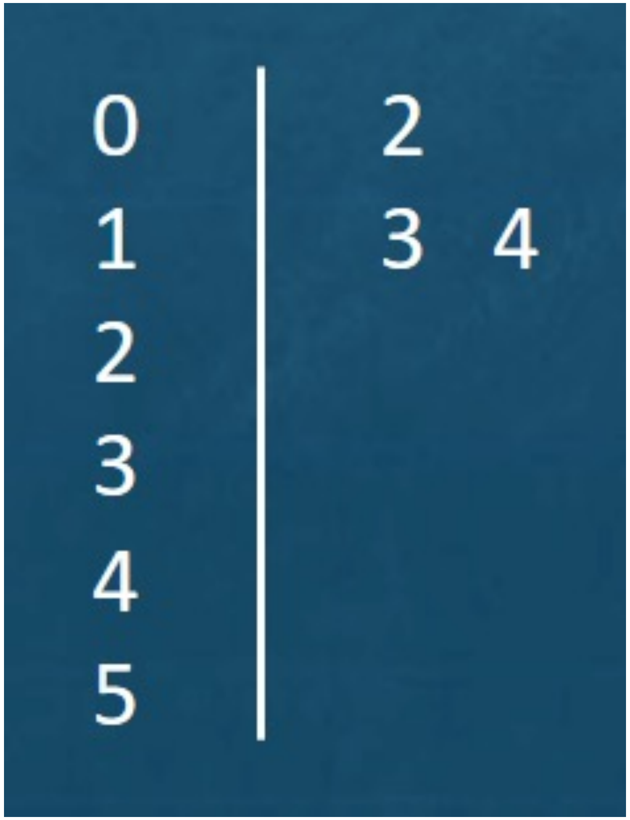
Key 1 I 0 means 10



0	
1	
2	
3	
4	
5	

Example…

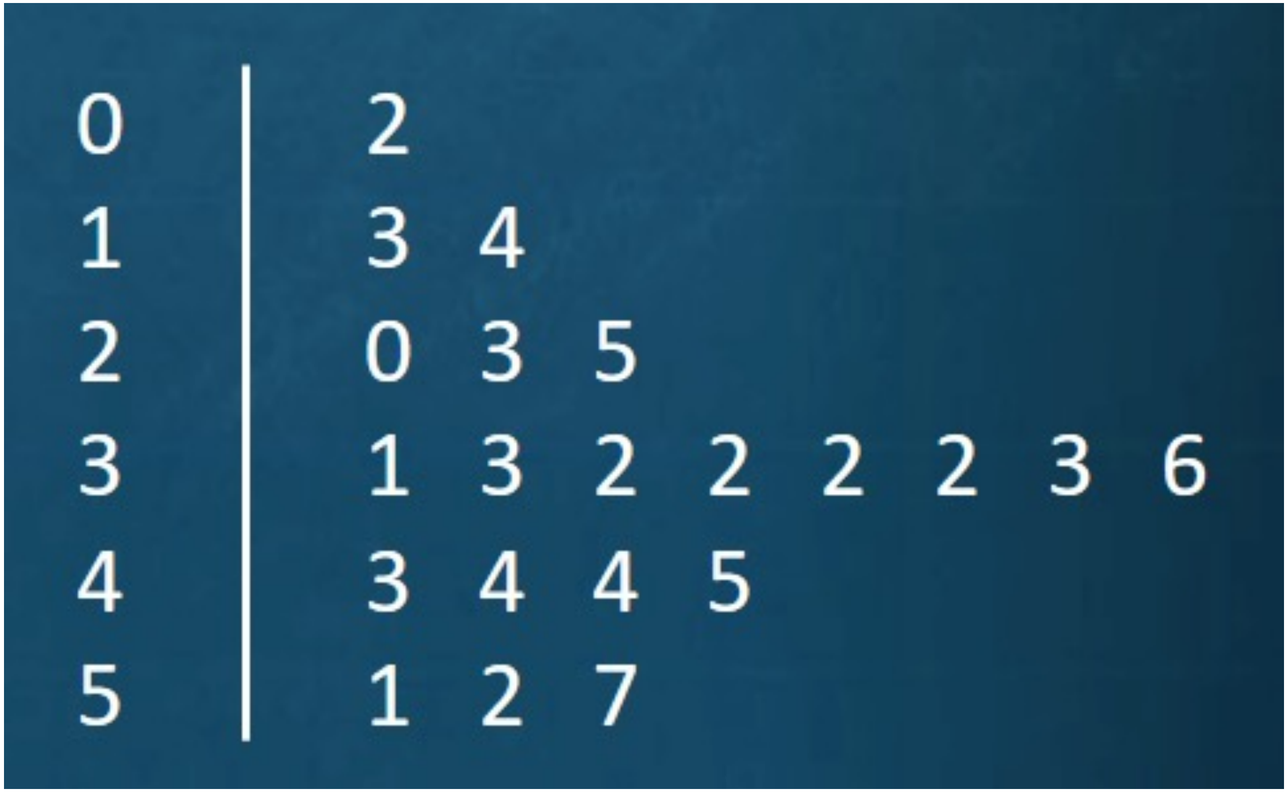
List the corresponding one's place values. Our first data value will be 2. In the next row we will place a 3 and a 4 to represent the data values 13 and 14.



0	2	
1	3	4
2		
3		
4		
5		

Example…

Continue this process until each value in the data set is represented.

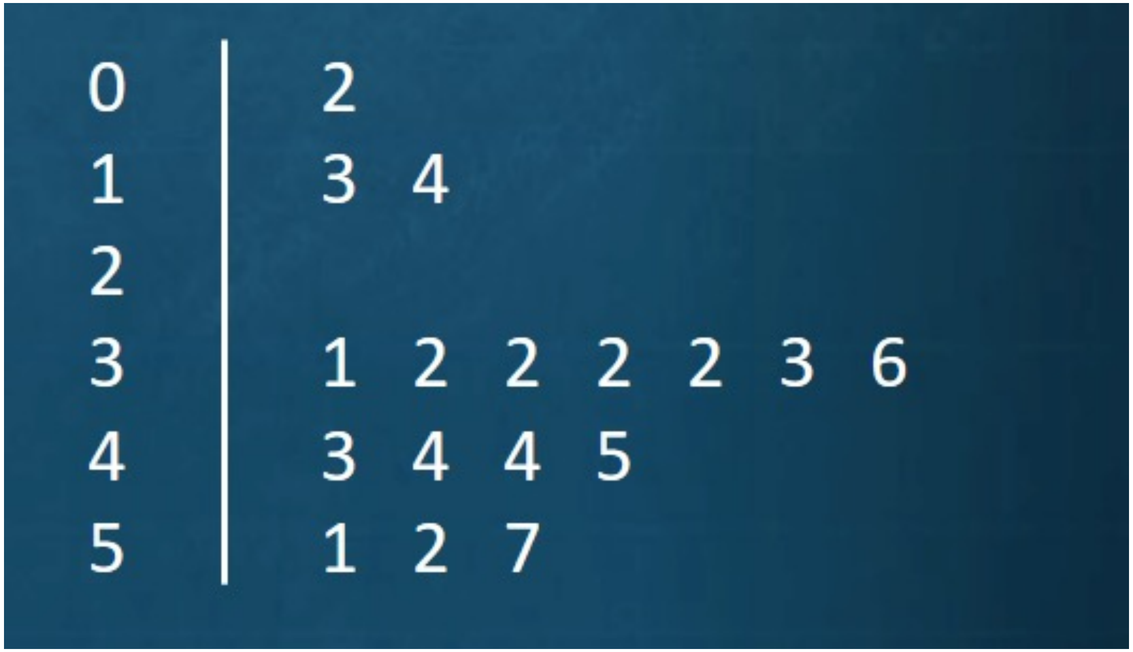


0	2
1	3 4
2	0 3 5
3	1 3 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

No Data Values

In a case where there are **no data values** in a class you should write the stem number and leave the leaf row blank.

For instance, if this data set had contained no values in the 20 to 29 range, then we would have listed the 2 representing 20s but left that row blank.

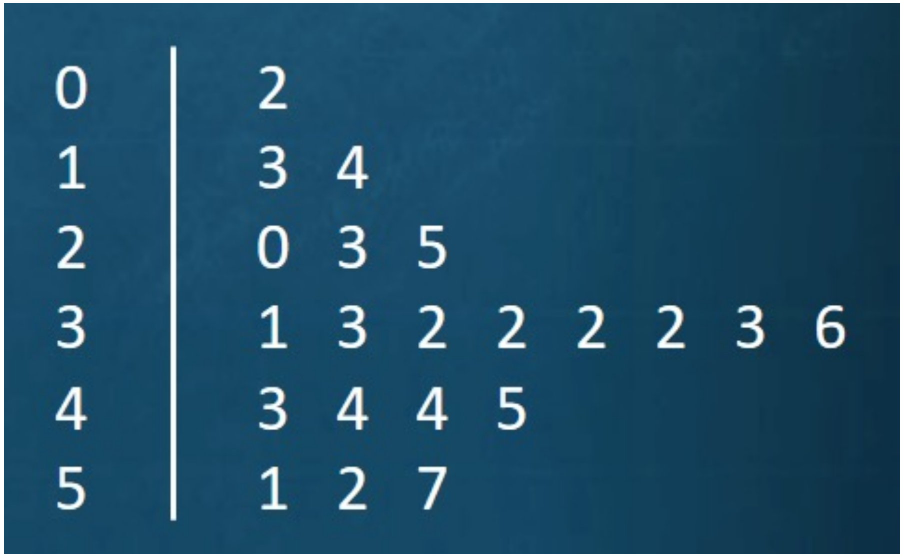


0	2
1	3 4
2	
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

Benefits

Some benefits of a stem and leaf display are the following:

- Easily determine the minimum and maximum values
- See the clustering of data
- Idea of how the data values are distributed.



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

Exercise: Stem and Leaf Plot

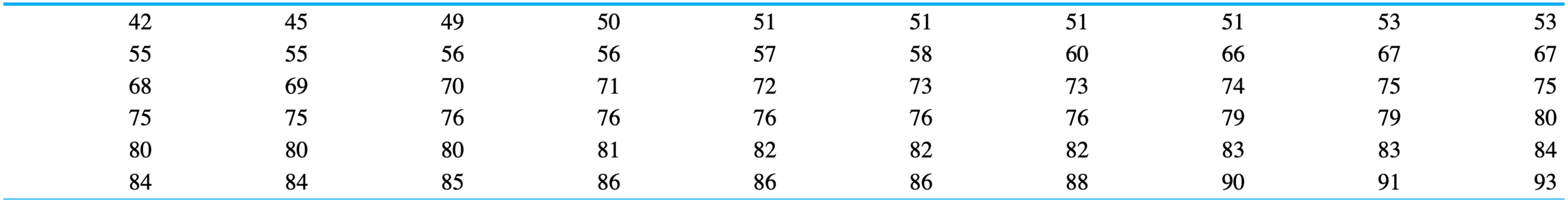
Consider the percentage scores of 15 students in a Statistics exam, given in the following list:

58, 55, 58, 61, 72, 79, 97, 67, 61, 77, 92, 64, 69, 62, 53.

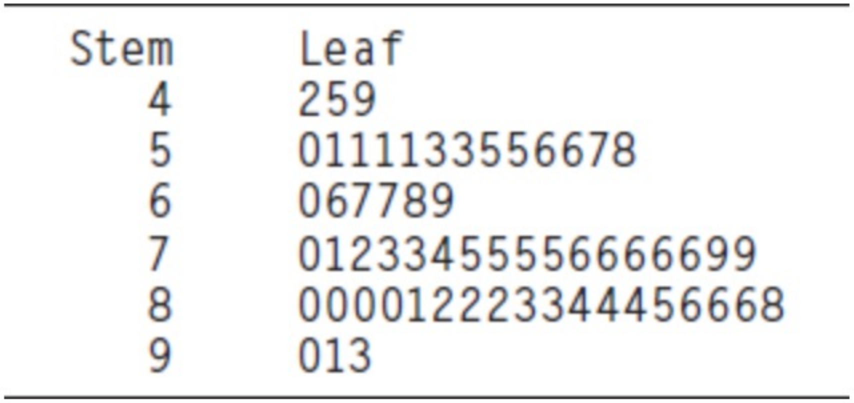
Construct a stem and leaf plot for this data.

Stem and Leaf Plot (cont.)

TABLE 1.3 Durations (in minutes) of dormant periods of the geyser Old Faithful



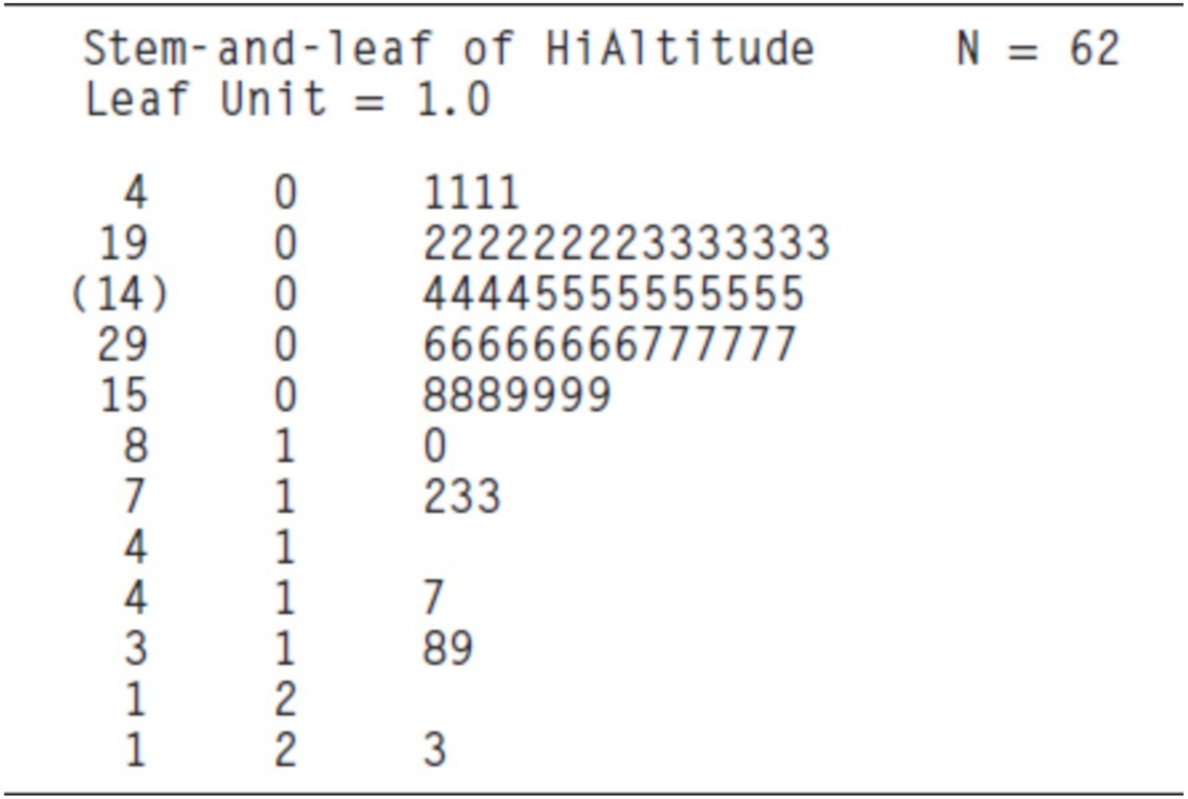
42	45	49	50	51	51	51	51	53	53
55	55	56	56	57	58	60	66	67	67
68	69	70	71	72	73	73	74	75	75
75	75	76	76	76	76	76	79	79	80
80	80	80	81	82	82	82	83	83	84
84	84	85	86	86	86	88	90	91	93



Stem	Leaf
4	259
5	011113355678
6	067789
7	012334555566666699
8	00001222334445666 8
9	013

FIGURE 1.5 Stem-and-leaf plot for the geyser data in Table 1.3.

Stem and Leaf Plot (cont.)

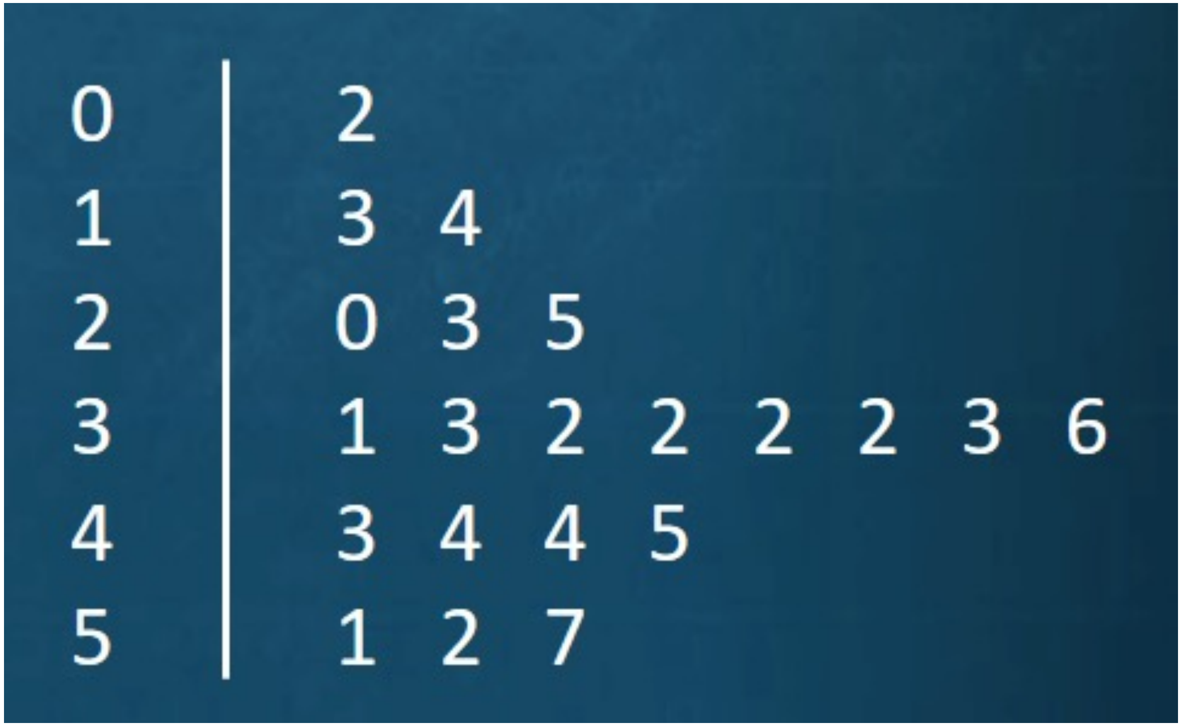


```
Stem-and-leaf of HiAltitude      N = 62
Leaf Unit = 1.0

  4     0    1111
 19     0    222222223333333
(14)    0    44445555555555
 29     0    66666666777777
 15     0    8889999
  8     1    0
  7     1    233
  4     1
  4     1    7
  3     1    89
  1     2
  1     2    3
```

FIGURE 1.6 Stem-and-leaf plot of the PM data in Table 1.2 in Section 1.2 as produced by MINITAB.

Summary - Stem and Leaf Plot



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

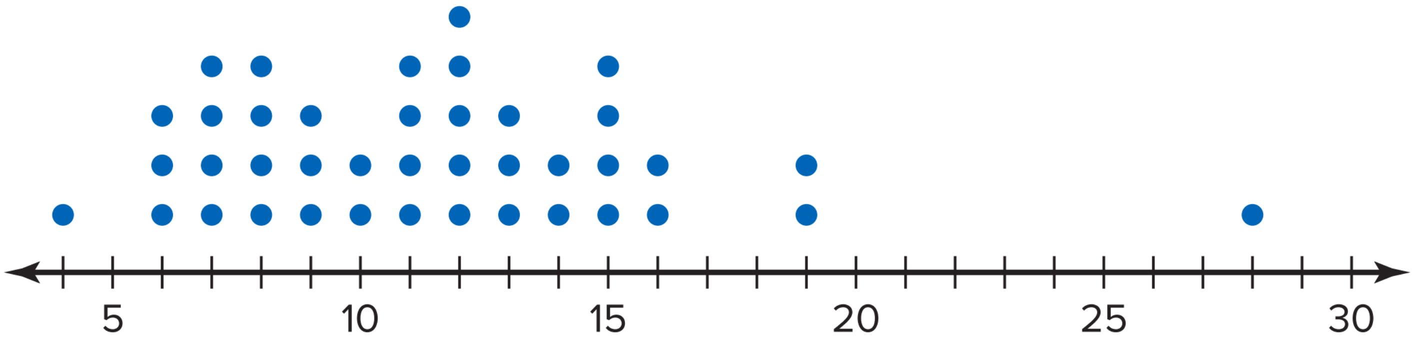
Stem unit = 10 *or* **Key: 1 I 0 means 10**

Dotplot

A dotplot is a statistical graph in which each data value is plotted as a point (dot) above the horizontal axis.

Dotplots are used to show how the data values are distributed and to see if there are any extremely high or low data values.

Number of Named Tropical Storms Each Year for the Years 1971–2010



Dotplot (cont.)

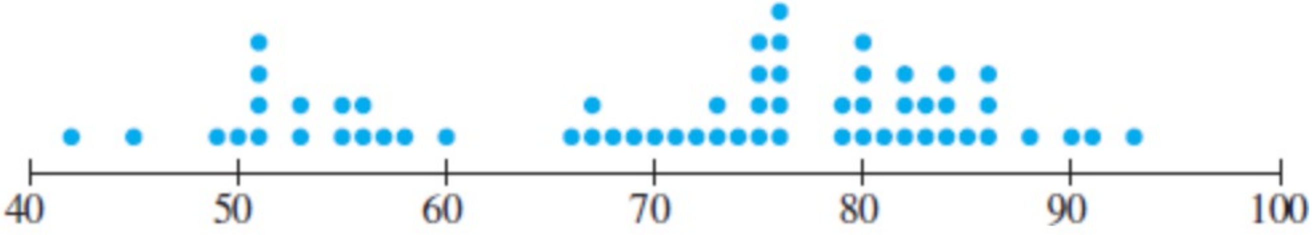
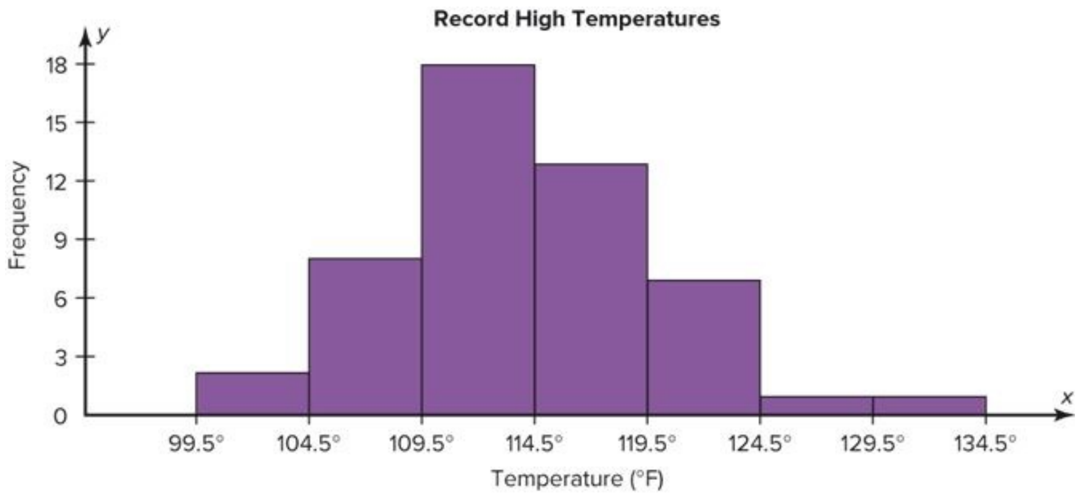


FIGURE 1.7 Dotplot for the geyser data in Table 1.3.

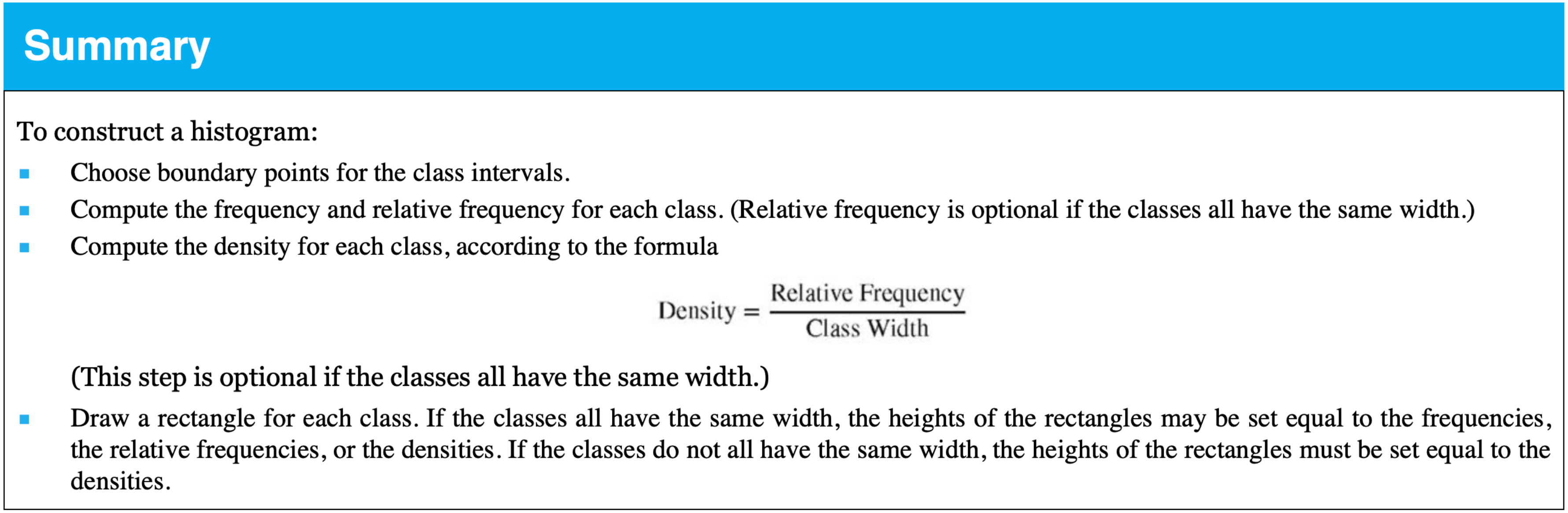
Histogram

Histogram is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is zero) of various heights to represent the frequencies of the classes.

Its purpose is to provide us with a graphical display of the distribution of a data set.



Histogram (cont.)



Summary

To construct a histogram:

- Choose boundary points for the class intervals.
- Compute the frequency and relative frequency for each class. (Relative frequency is optional if the classes all have the same width.)
- Compute the density for each class, according to the formula

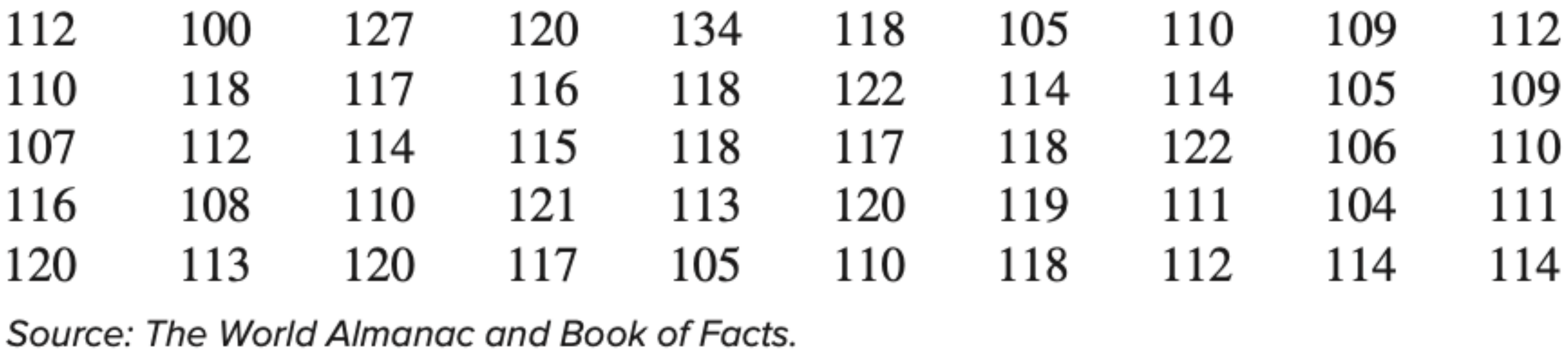
$$\text{Density} = \frac{\text{Relative Frequency}}{\text{Class Width}}$$

(This step is optional if the classes all have the same width.)

- Draw a rectangle for each class. If the classes all have the same width, the heights of the rectangles may be set equal to the frequencies, the relative frequencies, or the densities. If the classes do not all have the same width, the heights of the rectangles must be set equal to the densities.

Example – Record High Temperatures

Construct a histogram to represent the data shown for the record high temperatures (^{o}F) for each of the 50 U.S. states.



112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Source: The World Almanac and Book of Facts.

Example – Record High Temperatures

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Source: The World Almanac and Book of Facts.

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

Example – Record High Temperatures

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

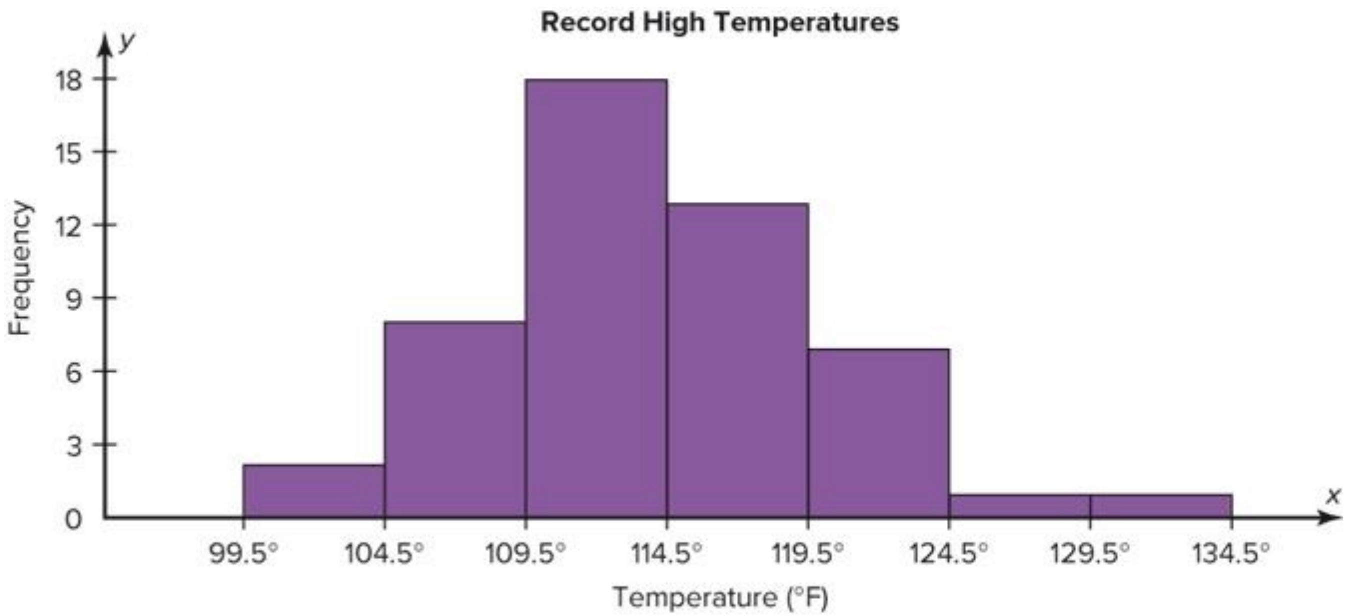


TABLE 1.4 Frequency table for PM emissions of 62 vehicles driven at high altitude

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 13	1	0.0161	0.0081
13–< 15	2	0.0323	0.0161
15–< 17	0	0.0000	0.0000
17–< 19	2	0.0323	0.0161
19–< 21	1	0.0161	0.0081
21–< 23	0	0.0000	0.0000
23–< 25	1	0.0161	0.0081

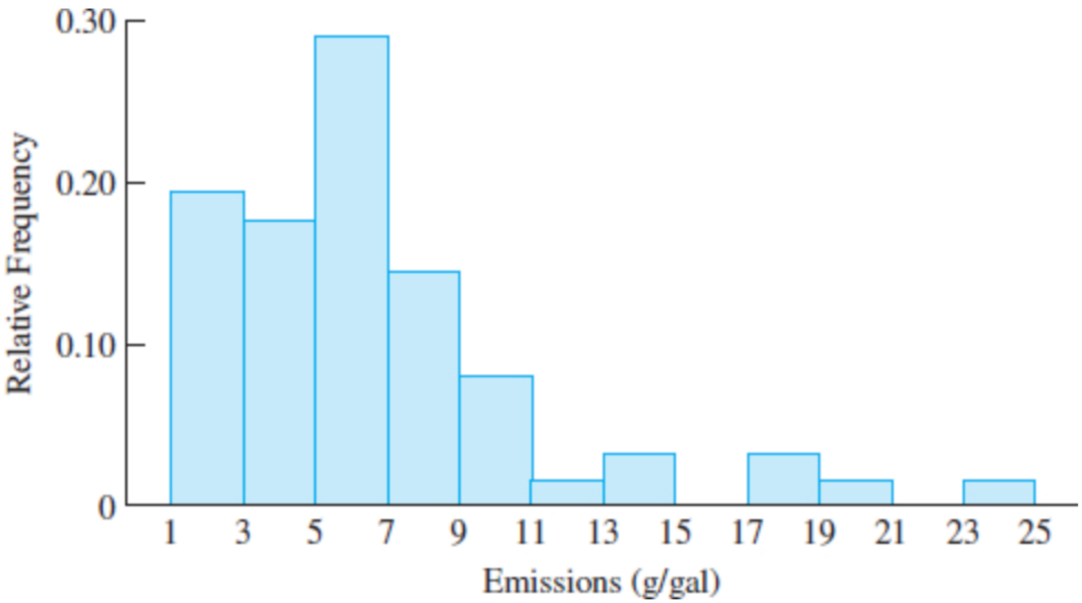


FIGURE 1.8 Histogram for the data in Table 1.4.

TABLE 1.5 Frequency table, with unequal class widths, for PM emissions of 62 vehicles driven at high altitude

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 15	3	0.0484	0.0121
15–< 25	4	0.0645	0.0065

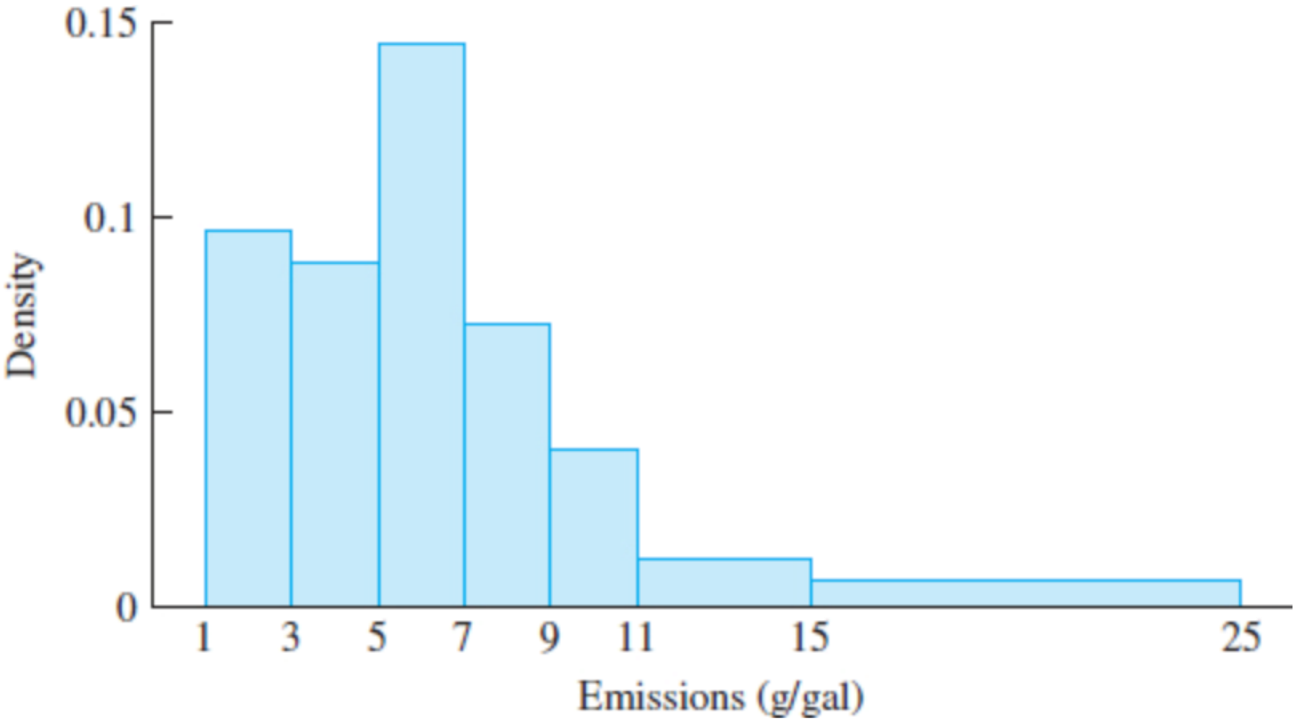
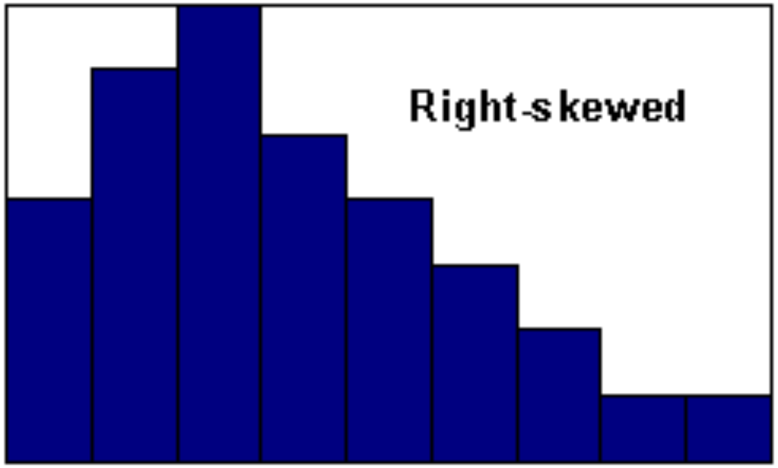


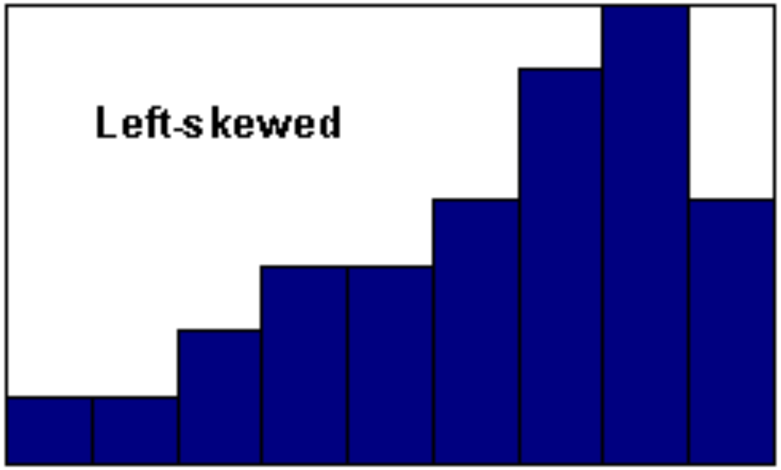
FIGURE 1.9 Histogram for the PM emissions for high-altitude vehicles.

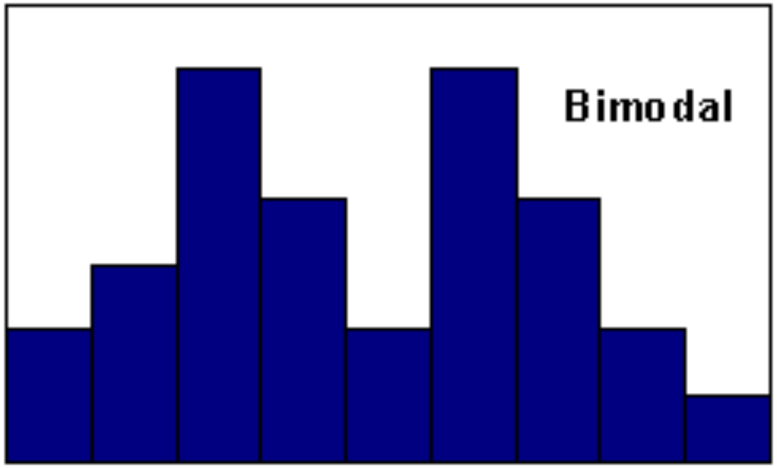
Distribution Shapes

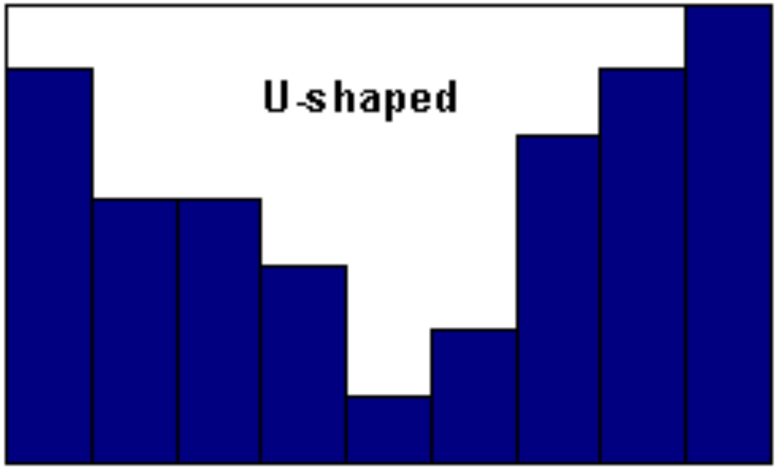
- When one is describing data, it is important to be able to recognize the shapes of the distribution values.
- In later chapters, you will see that the shape of a distribution also determines the appropriate statistical methods used to analyze the data.
- A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution.
- Distributions are most often not perfectly shaped, so it is not necessary to have an exact shape but rather to identify an overall pattern.

Distribution Shapes (cont.)

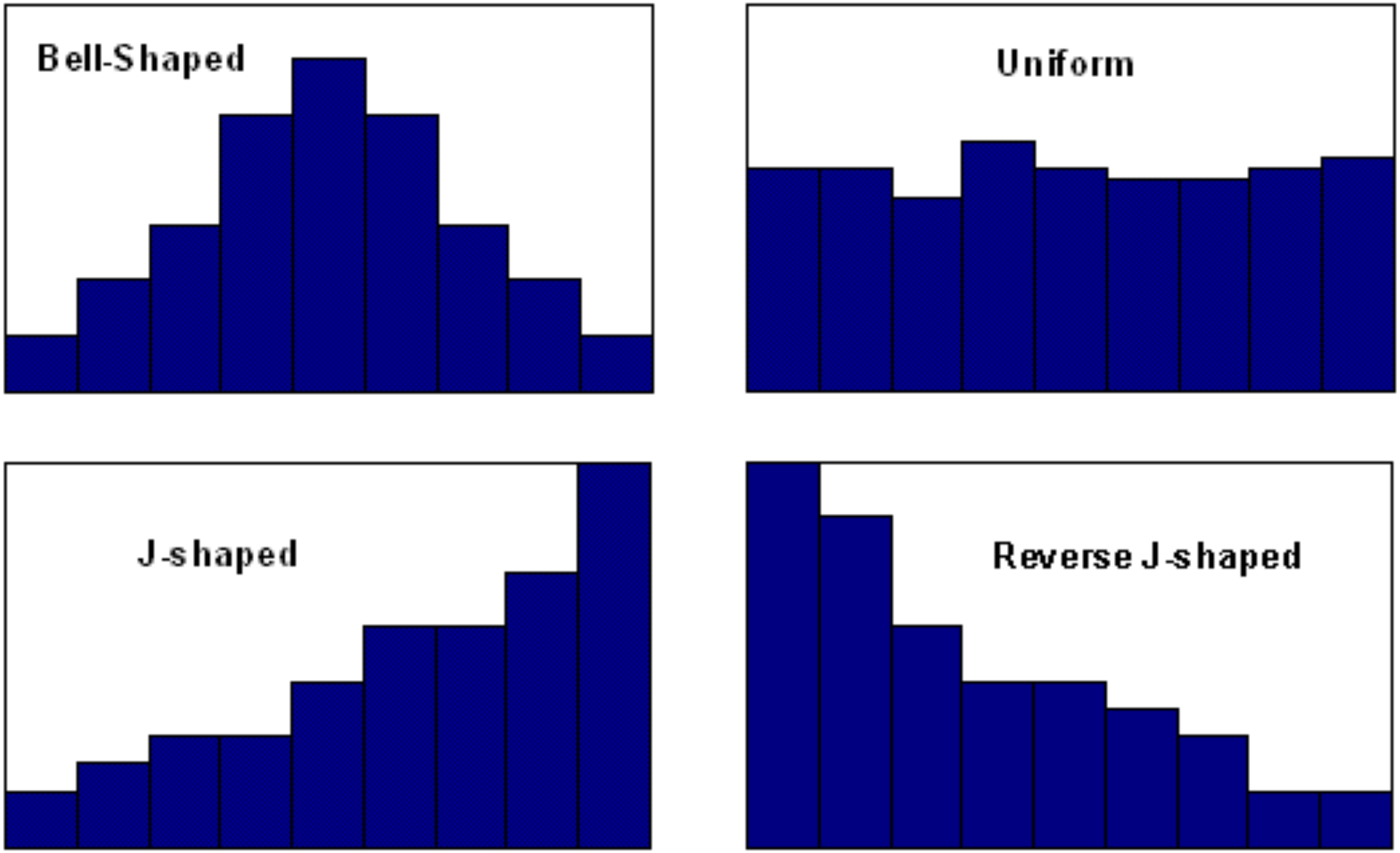








Distribution Shapes



Distribution Shapes (cont.)

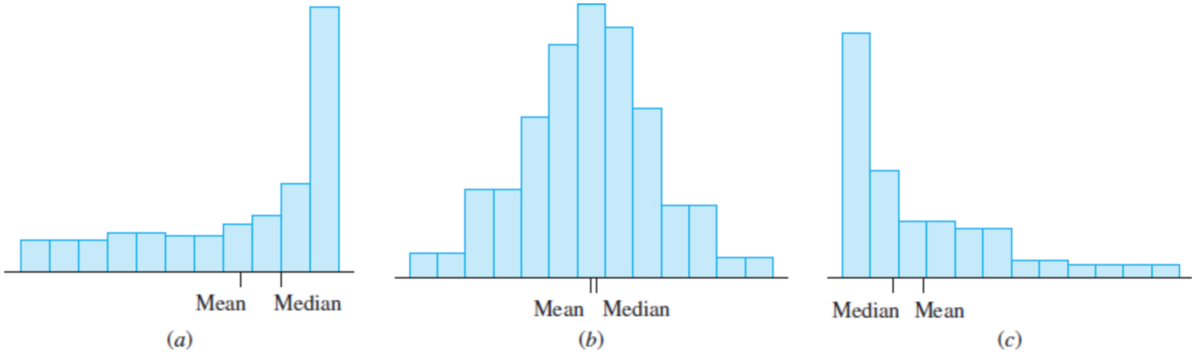
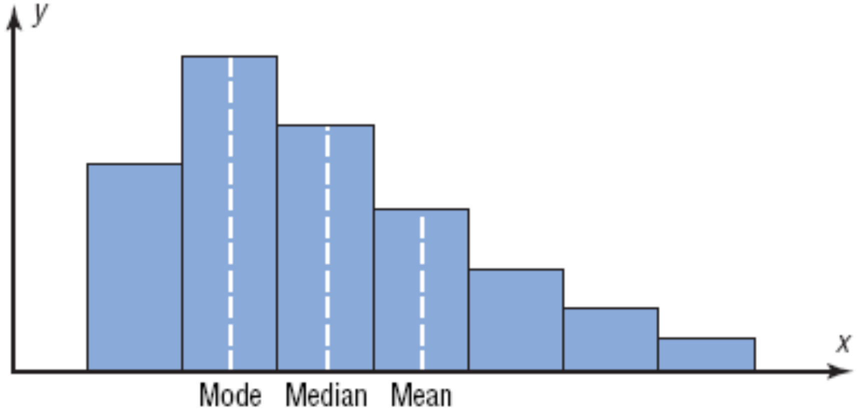
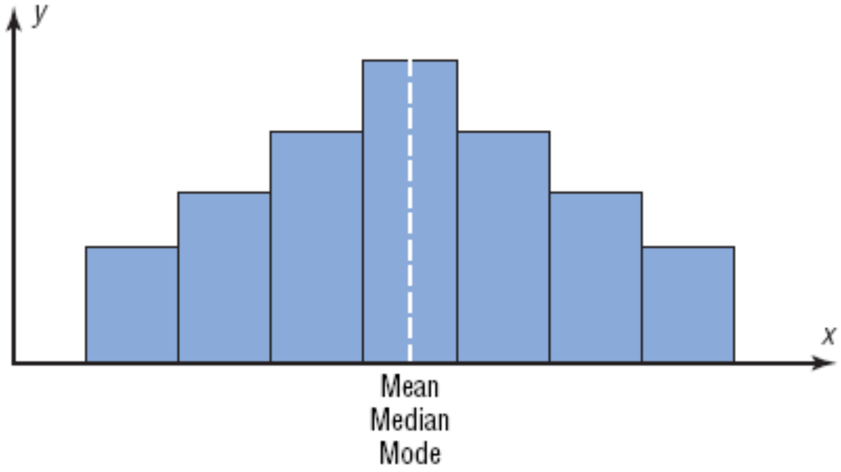
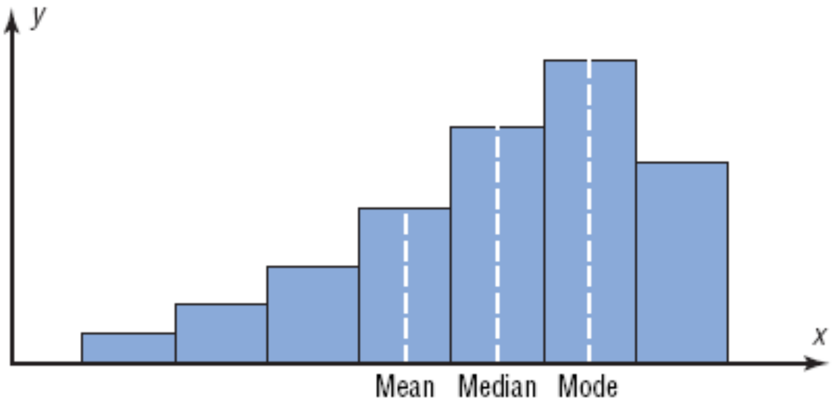


FIGURE 1.10 *(a)* A histogram skewed to the left. The mean is less than the median. *(b)* A nearly symmetric histogram. The mean and median are approximately equal. *(c)* A histogram skewed to the right. The mean is greater than the median.

Distribution Shapes (cont.)

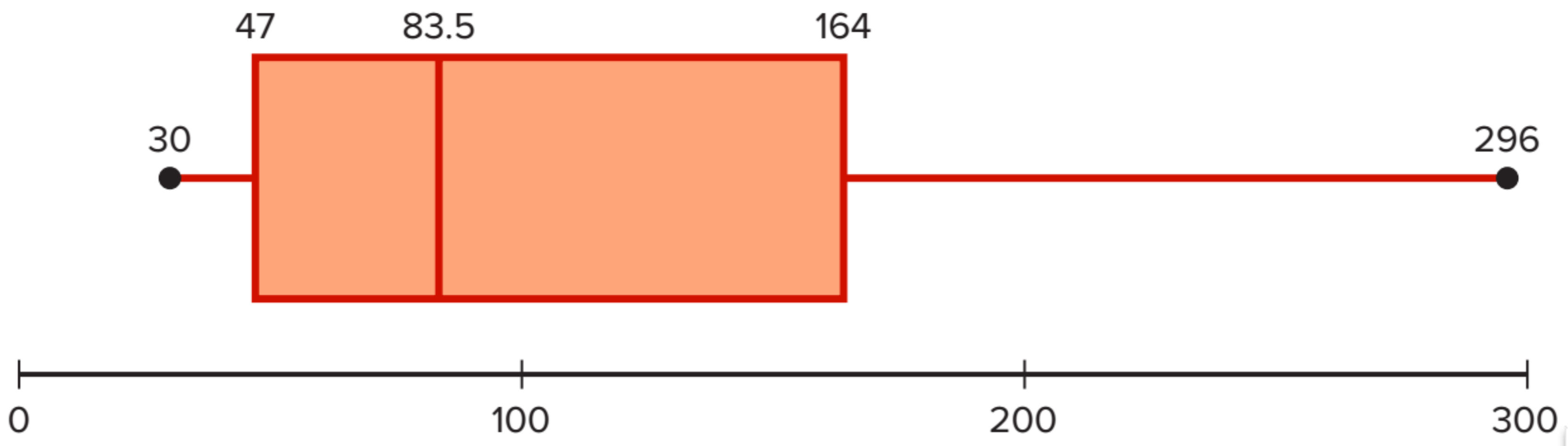






Information Obtained from a Boxplot

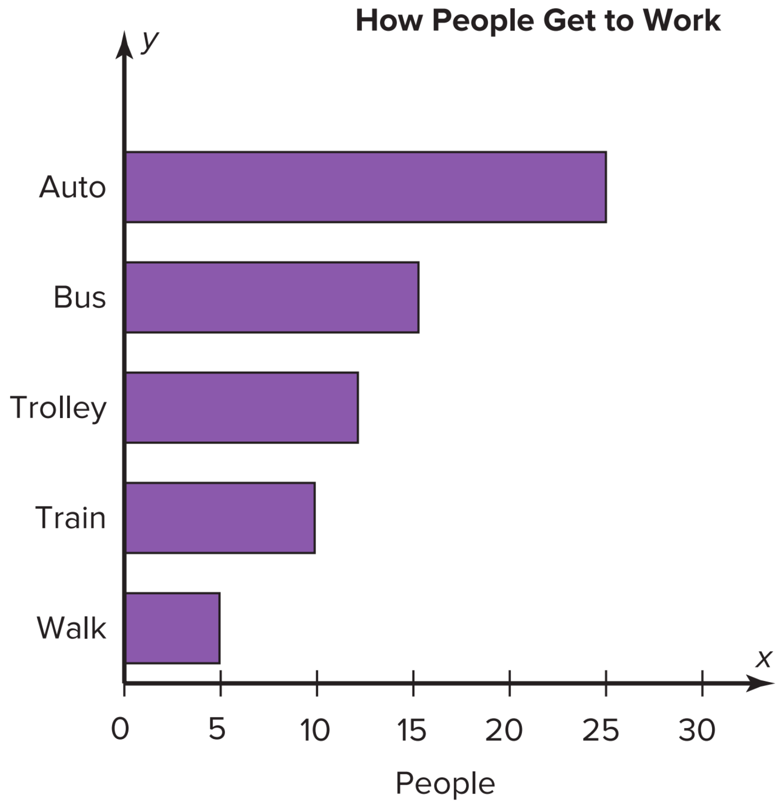
1. *a*. If the median is near the center of the box, the distribution is approximately symmetric.
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center, the distribution is negatively skewed.
2. *a*. If the lines are about the same length, the distribution is approximately symmetric.
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line, the distribution is negatively skewed.



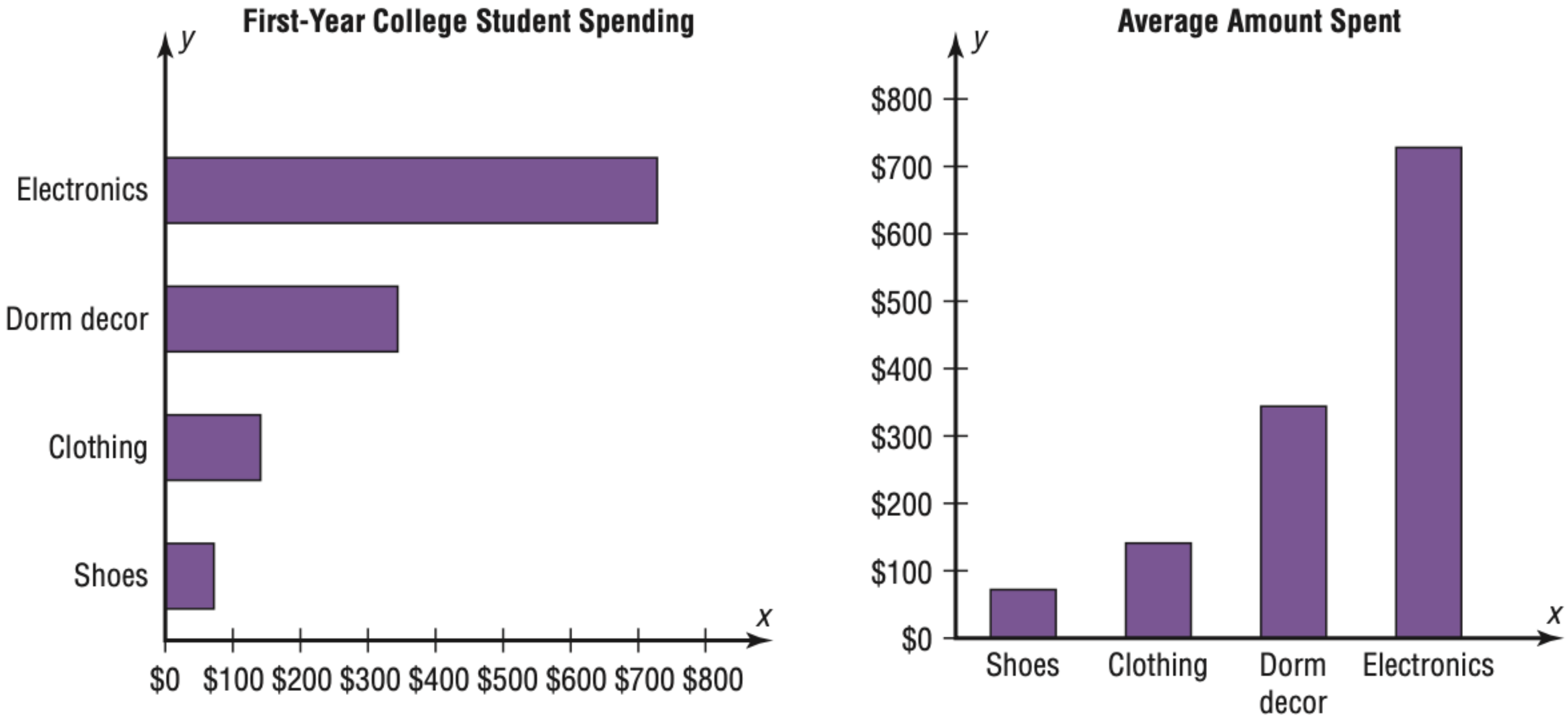
This boxplot indicates that the distribution is **positively (right) skewed**.

Bar Graph

A bar graph represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

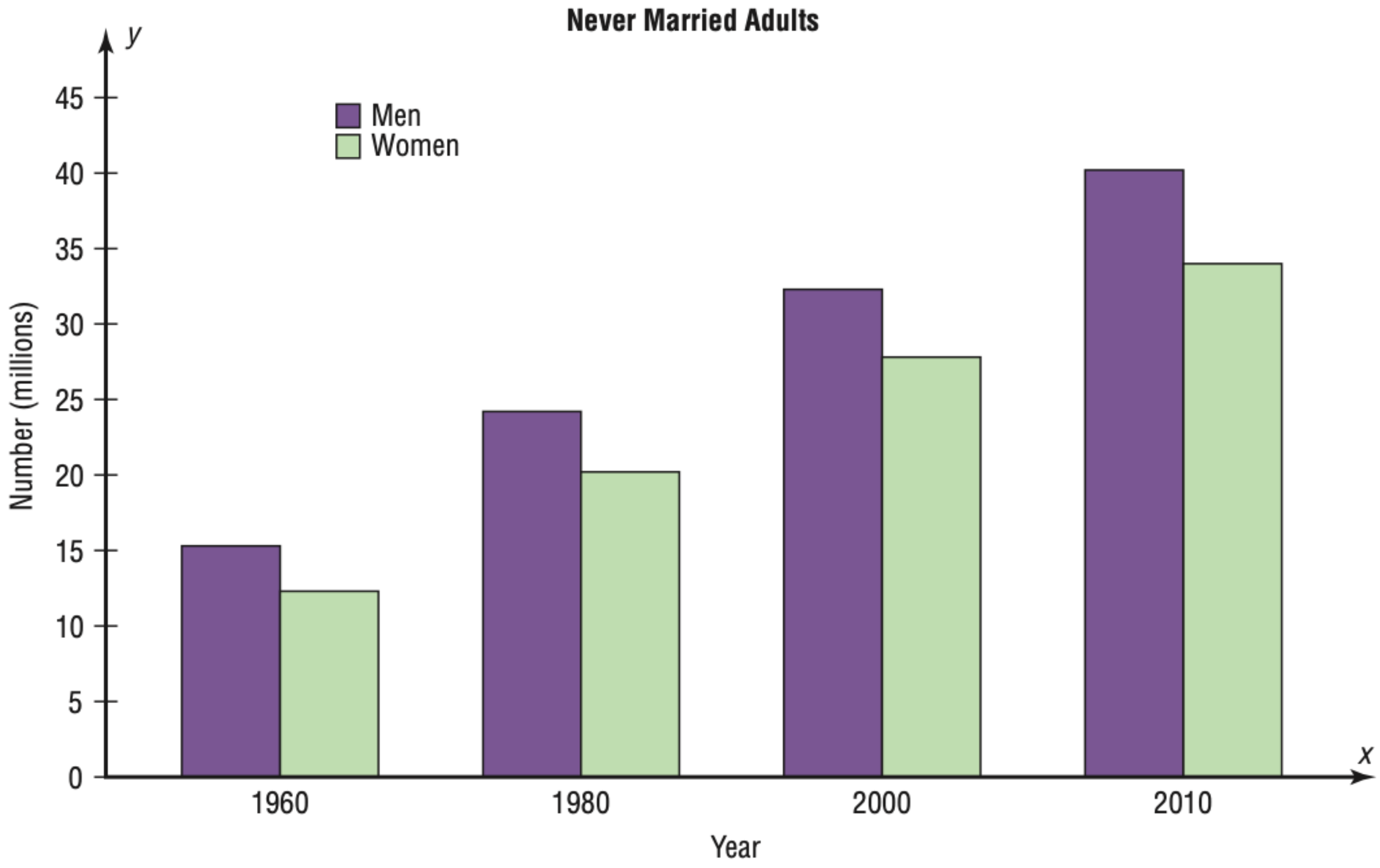


Bar Graph (cont.)



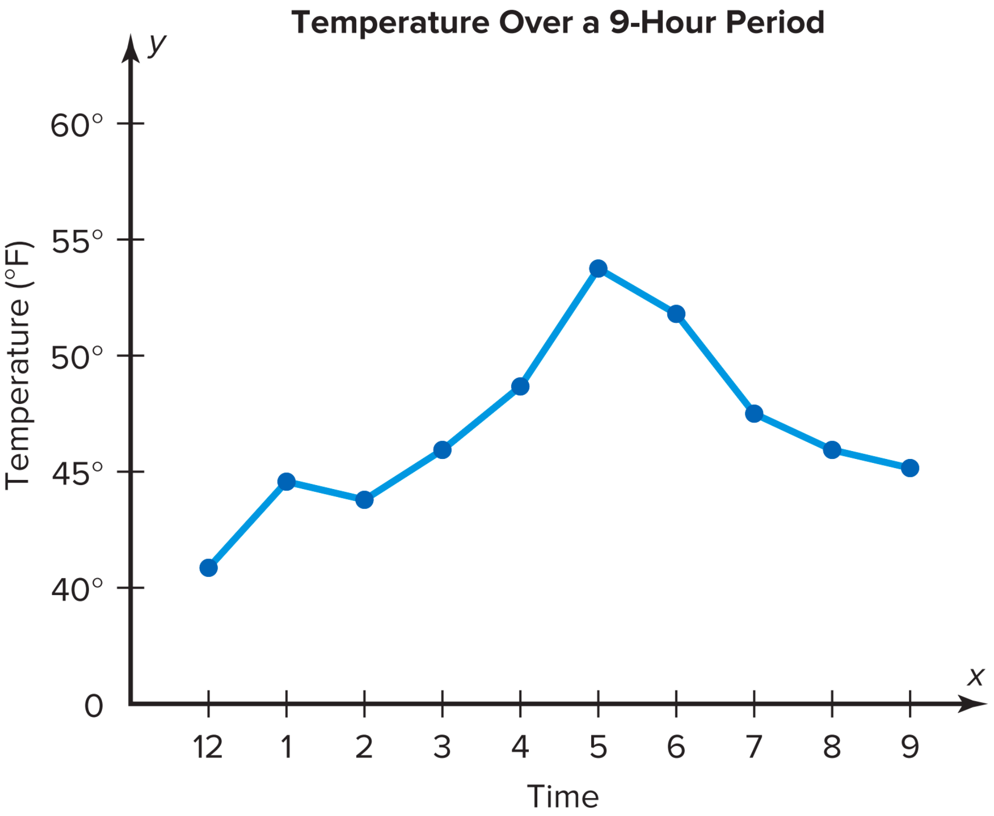
The graphs show that first-year college students spend the most on electronic equipment.

Bar Graph (cont.)

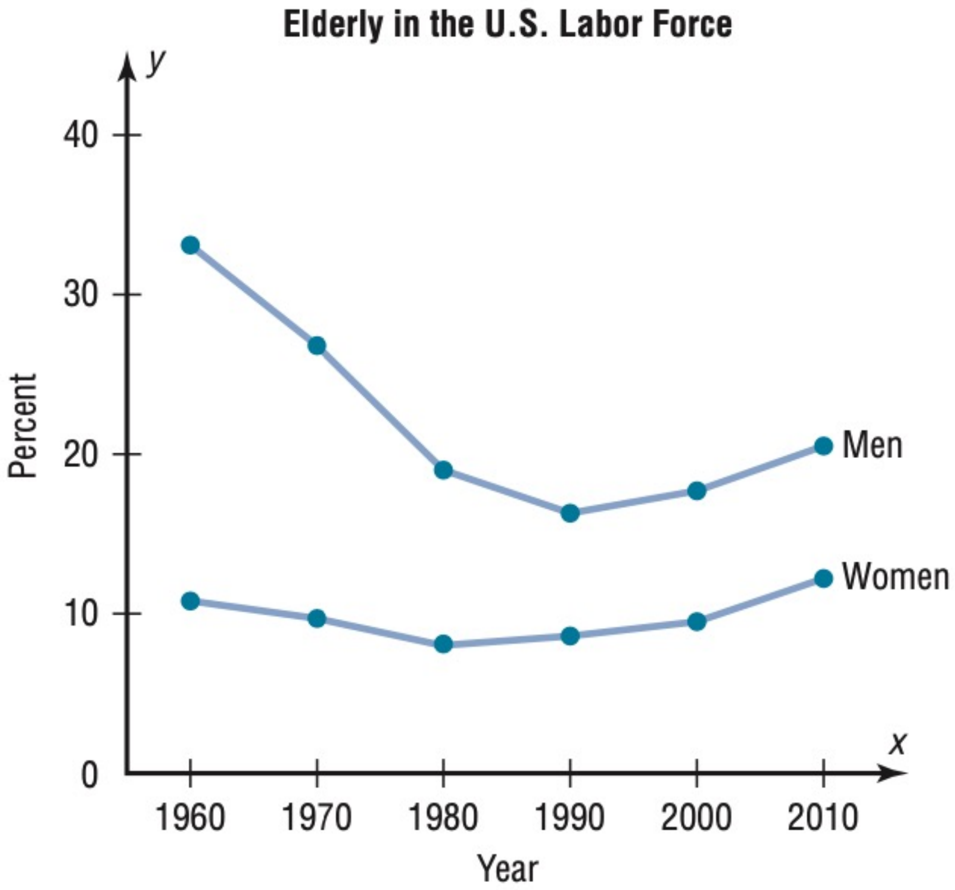


Time Series Graph

A time series graph represents data that occur over a specific period of time.



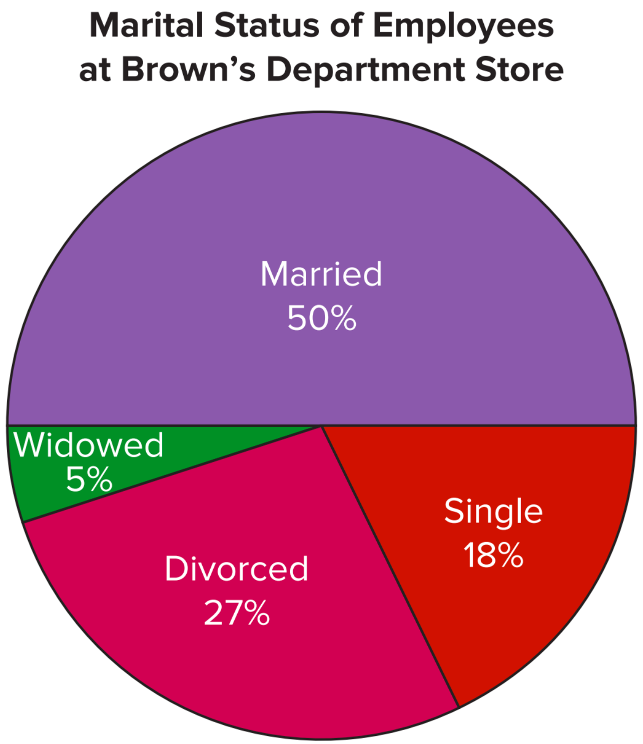
Time Series Graph (cont.)



Source: Bureau of Census, U.S. Department of Commerce.

Pie Graph

A pie graph (or pie chart) is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.



Scatter Plot

- When each item is a pair of values, the data are said to be bivariate.
- One of the most useful graphical summaries for numerical bivariate data is the scatterplot.

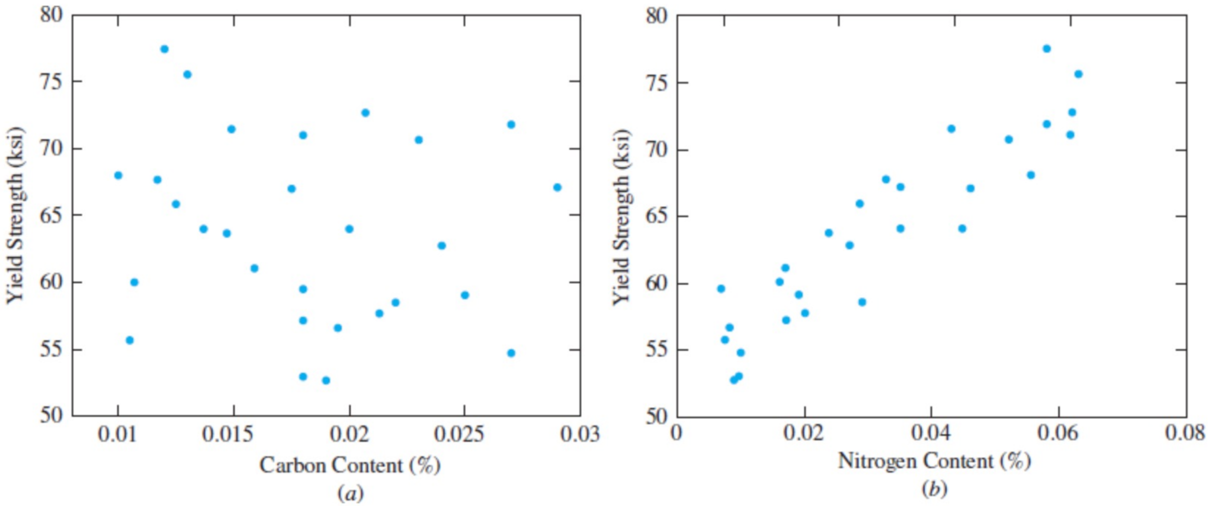
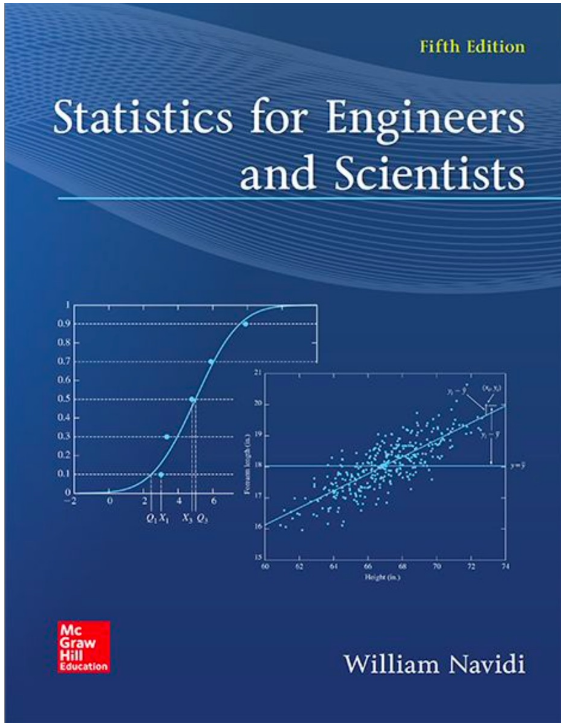


FIGURE 1.17 *(a)* A scatterplot showing that there is not much of a relationship between carbon content and yield strength for a certain group of welds. *(b)* A scatterplot showing that for these same welds, higher nitrogen content is associated with higher yield strength.



Chapter 1

Sampling and Descriptive Statistics

(End of part 3)