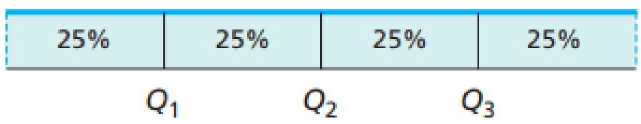
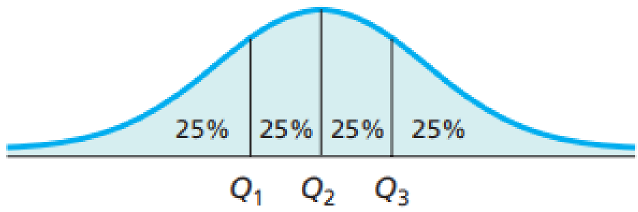
Box-and-Whisker Plot

Quartiles:

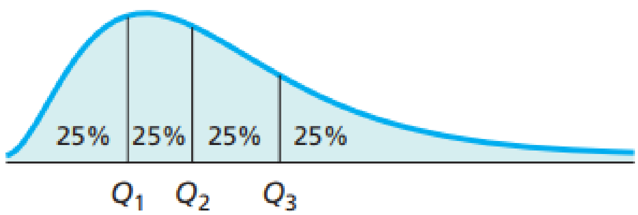
Quartiles are the most commonly used percentiles. A data set has three quartiles, which we denote Q1, Q2, and Q3. Roughly speaking, the first quartile, Q1, is the number that divides the bottom 25% of the data from the top 75%; the second quartile, Q2, is the median, which, as you know, is the number that divides the bottom 50% of the data from the top 50%; and the third quartile, Q3, is the number that divides the bottom 75% of the data from the top 25%. Note that the first and third quartiles are the 25th and 75th percentiles, respectively



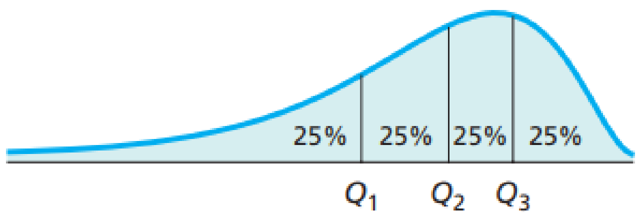
(a) Uniform



(b) Bell shaped



(c) Right skewed



(d) Left skewed

Quartiles

Arrange the data in increasing order and determine the median.

- The **first quartile** is the median of the part of the entire data set that lies at or below the median of the entire data set.
- The **second quartile** is the median of the entire data set.
- The **third quartile** is the median of the part of the entire data set that lies at or above the median of the entire data set.

Box-and-Whisker Plot:

A **box-and-whisker plot** gives a graphic presentation of data using five measures: the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences. (The inner fences are explained in Example 3–27.) A box-and-whisker plot can help us visualize the center, the spread, and the skewness of a data set. It also helps detect outliers. We can compare different distributions by making box-and-whisker plots for each of them.

Box-and-Whisker Plot A plot that shows the center, spread, and skewness of a data set. It is constructed by drawing a box and two whiskers that use the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

Note that a box-and-whisker plot is made using the following five-number summary values:

Minimum value Q_1 Median Q_3 Maximum value

The minimum and the maximum values used here must be within the lower and the upper inner fences that will be explained below. Making a box-and-whisker plot involves five steps that are

To Construct a Boxplot

Step 1 **Determine the quartiles.**

Step 2 **Determine potential outliers and the adjacent values.**

Step 3 **Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.**

Step 4 **Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.**

Step 5 **Plot each potential outlier with an asterisk.**

Interquartile Range

The **interquartile range,** or **IQR,** is the difference between the first and third quartiles; that is, $IQR = Q_3 - Q_1$.

Lower and Upper Limits

The **lower limit** and **upper limit** of a data set are

$$\text{Lower limit} = Q_1 - 1.5 \cdot \text{IQR};$$
$$\text{Upper limit} = Q_3 + 1.5 \cdot \text{IQR}.$$

Note:

- In a boxplot, the two lines emanating from the box are called whiskers.
- Boxplots are frequently drawn vertically instead of horizontally.
- Symbols other than an asterisk are often used to plot potential outlier

Example: Incomes of Households

The following data are the incomes (in thousands of dollars) for a sample of 12 households.

75 69 84 112 74 104 81 90 94 144 79 98

Construct a box-and-whisker plot for these data.

Solution The following five steps are performed to construct a box-and-whisker plot.

Step 1. First, rank the data in increasing order and calculate the values of the median, the first quartile, the third quartile, and the interquartile range. The ranked data are

69 74 75 79 81 84 90 94 98 104 112 144

For these data,

$$\text{Median} = (84 + 90)/2 = 87$$
$$Q_1 = (75 + 79)/2 = 77$$
$$Q_3 = (98 + 104)/2 = 101$$
$$\text{IQR} = Q_3 - Q_1 = 101 - 77 = 24$$

Step 2. Find the points that are $1.5 \times \text{IQR}$ below Q_1 and $1.5 \times \text{IQR}$ above Q_3. These two points are called the **lower** and the **upper inner fences**, respectively.

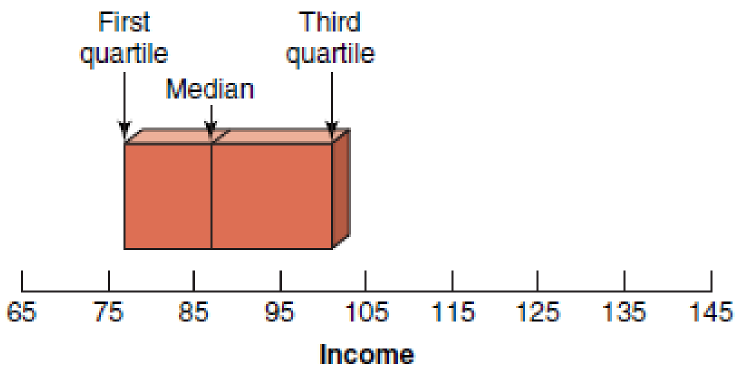
$$1.5 \times \text{IQR} = 1.5 \times 24 = 36$$
$$\text{Lower inner fence} = Q_1 - 36 = 77 - 36 = 41$$
$$\text{Upper inner fence} = Q_3 + 36 = 101 + 36 = 137$$

Step 3. Determine the smallest and the largest values in the given data set within the two inner fences. These two values for our example are as follows:

$$\text{Smallest value within the two inner fences} = 69$$
$$\text{Largest value within the two inner fences} = 112$$

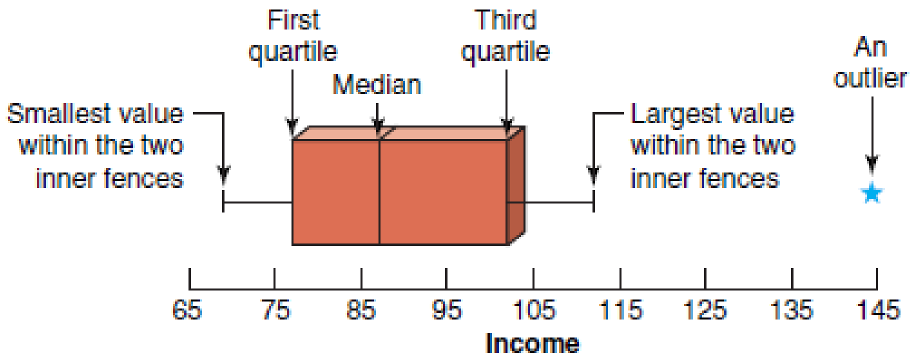
Step 4. Draw a horizontal line and mark the income levels on it such that all the values in the given data set are covered. Above the horizontal line, draw a box with its left side at the position of the first quartile and the right side at the position of the third quartile. Inside the box, draw a vertical line at the position of the median. The result of this step is shown in Figure 3.13.

Figure 3.13



Step 5. By drawing two lines, join the points of the smallest and the largest values within the two inner fences to the box. These values are 69 and 112 in this example as listed in Step 3. The two lines that join the box to these two values are called **whiskers**. A value that falls outside the two inner fences is shown by marking an asterisk and is called an outlier. This completes the box-and-whisker plot, as shown in Figure 3.14.

Figure 3.14



In Figure 3.14, about 50% of the data values fall within the box, about 25% of the values fall on the left side of the box, and about 25% fall on the right side of the box. Also, 50% of the values fall on the left side of the median and 50% lie on the right side of the median. The data of this example are skewed to the right because the lower 50% of the values are spread over a smaller range than the upper 50% of the values. ■

Note:

The observations that fall outside the two inner fences are called outliers. These outliers can be classified into two kinds of outliers—mild and extreme outliers. To do so, we define two outer fences—a **lower outer fence** at $3.0 \times$ IQR below the first quartile and an **upper outer fence** at $3.0 \times$ IQR above the third quartile. If an observation is outside either of the two inner fences but within the two outer fences, it is called a **mild outlier**. An observation that is outside either of the two outer fences is called an **extreme outlier**. For Example 3–27, the two outer fences are calculated as follows.

$$3 \times \text{IQR} = 3 \times 24 = 72$$
$$\text{Lower outer fence} = Q_1 - 72 = 77 - 72 = 5$$
$$\text{Upper outer fence} = Q_3 + 72 = 101 + 72 = 173$$

Because 144 is outside the upper inner fence but inside the upper outer fence, it is a mild outlier.

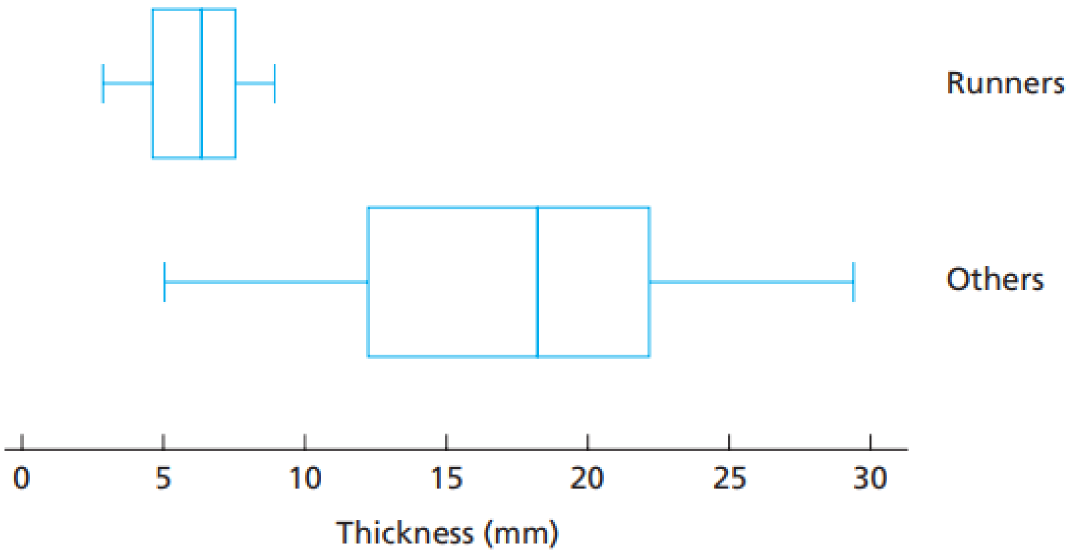
Using the box-and-whisker plot, we can conclude whether the distribution of our data is symmetric, skewed to the right, or skewed to the left. If the line representing the median is in the middle of the box and the two whiskers are of about the same length, then the data have a symmetric distribution. If the line representing the median is not in the middle of the box and/or the two whiskers are not of the same length, then the distribution of data values is skewed. The distribution is skewed to the right if the median is to the left of the center of the box with the right side whisker equal to or longer than the whisker on the left side, or if the median is in the center of the box but the whisker on the right side is longer than the one on the left side. The distribution is skewed to the left if the median is to the right of the center of the box with the left side whisker equal to or longer than the whisker on the right side, or if the median is in the center of the box but the whisker on the left side is longer than the one on the right side.

Example: Comparing Data Sets by Using Boxplots

Skinfold Thickness A study titled "Body Composition of Elite Class Distance Runners" was conducted by M. Pollock et al. to determine whether elite distance runners are actually thinner than other people. Their results were published in The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies (P. Milvey (ed.), New York: New York Academy of Sciences, p. 366). The researchers measured skinfold thickness, an indirect indicator of body fat, of samples of runners and nonrunners in the same age group. The sample data, in millimeters (mm), presented in Table 3.13 are based on their results. Use boxplots to compare these two data sets, paying special attention to center and variation

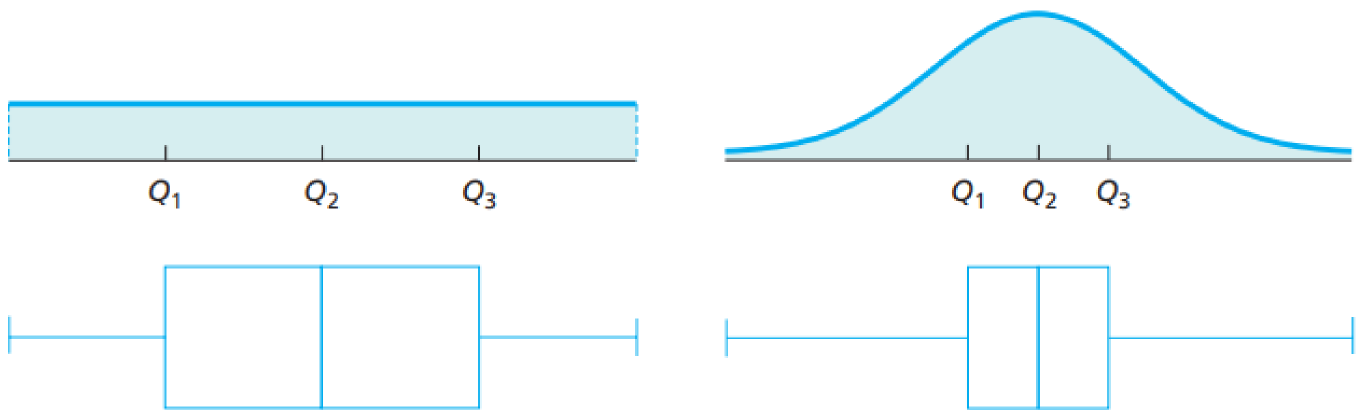
Runners			Others			
7.3	6.7	8.7	24.0	19.9	7.5	18.4
3.0	5.1	8.8	28.0	29.4	20.3	19.0
7.8	3.8	6.2	9.3	18.1	22.8	24.2
5.4	6.4	6.3	9.6	19.4	16.3	16.3
3.7	7.5	4.6	12.4	5.2	12.2	15.6

Solution Figure 3.10 displays boxplots for the two data sets, using the same scale.

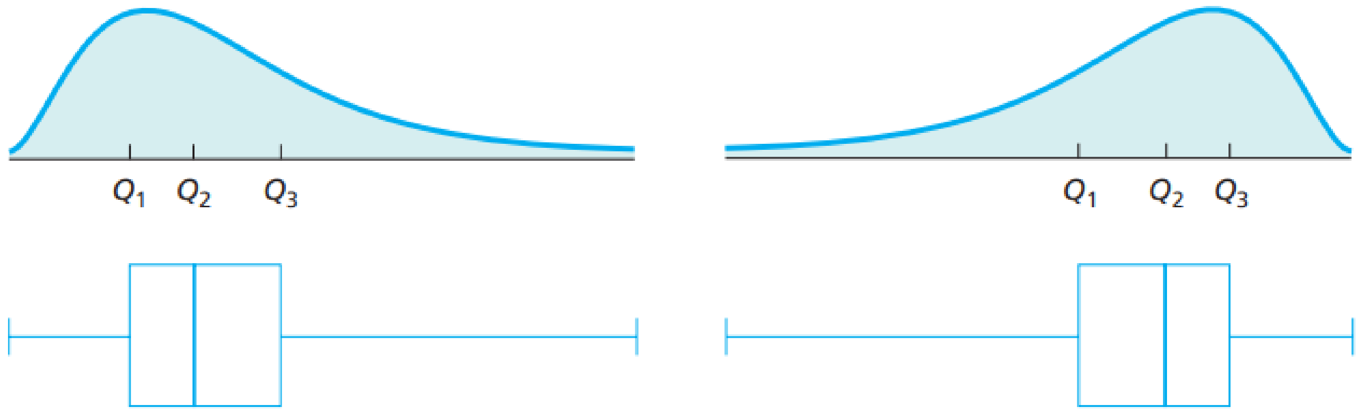


From Fig. 3.10, it is apparent that, on average, the elite runners sampled have smaller skinfold thickness than the other people sampled. Furthermore, there is much less variation in skinfold thickness among the elite runners sampled than among the other people sampled. By the way, when you study inferential statistics, you will be able to decide whether these descriptive properties of the samples can be extended to the populations from which the samples were drawn.

Shape of the distribution:



(a) Uniform (b) Bell shaped



(c) Right skewed (d) Left skewed

You can also use a boxplot to identify the approximate shape of the distribution of a data set displays some common distribution shapes and their corresponding boxplots. Pay particular attention to how box width and whisker length relate to skewness and symmetry. Employing boxplots to identify the shape of a distribution is most useful with large data sets. For small data sets, boxplots can be unreliable in identifying distribution shape; using a stem-and-leaf diagram or a dotplot is generally better.

Practice Problems from text book

3.75 Prepare a box-and-whisker plot for the following data:

36	43	28	52	41	59	47	61
24	55	63	73	32	25	35	49
31	22	61	42	58	65	98	34

Does this data set contain any outliers?

3.76 The following data give the time (in minutes) that each of 20 students selected from a university waited in line at their bookstore to pay for their textbooks in the beginning of the Fall 2015 semester.

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

Prepare a box-and-whisker plot. Comment on the skewness of these data.

3.77 The following data give the 2015 bonuses (in thousands of dollars) of 15 randomly selected Wall Street managers.

107	122	175	89	53	361	67	258
61	781	136	208	391	247	71	

Prepare a box-and-whisker plot. Are these data skewed in any direction?

3.78 The following data give the total food expenditures (in dollars) for the past one month for a sample of 20 families.

1125	530	1234	595	427	872	1480	699	1274	1187
933	1127	716	1065	934	1630	1046	2199	1353	441

Prepare a box-and-whisker plot. Is the distribution of these data symmetric or skewed? Are there any outliers? If so, classify them as mild or extreme.

3.79 The following data give the annual salaries (in thousand dollars) of 20 randomly selected health care workers.

50	71	57	39	45	64	38	53	35	62
74	40	67	44	77	61	58	55	64	59

Prepare a box-and-whisker plot. Are these data skewed in any direction?

3.80 The following data give the number of patients who visited a walk-in clinic on each of 24 randomly selected days.

23	37	26	19	33	22	30	42	24	26	64	8
28	32	37	29	38	24	35	20	34	38	28	16

Prepare a box-and-whisker plot. Comment on the skewness of these data.