16. In each of the following data sets, tell whether the outlier seems certain to be due to an error, or whether it could conceivably be correct.

a. The length of a rod is measured five times. The readings in centimeters are 48.5, 47.2, 4.91, 49.5, 46.3.

b. The prices of five cars on a dealer's lot are $25,000, $30,000, $42,000, $110,000, $31,000.

1.3 Graphical Summaries

Stem-and-Leaf Plots

The mean, median, and standard deviation are numerical summaries of a sample or of a population. Graphical summaries are used as well to help visualize a list of numbers. The graphical summary that we will discuss first is the **stem-and-leaf plot**. A stem-and-leaf plot is a simple way to summarize a data set.

As an example, the data in Table 1.3 concern the geyser Old Faithful in Yellowstone National Park. This geyser alternates periods of eruption, which typically last from 1.5 to 4 minutes, with periods of dormancy, which are considerably longer. Table 1.3 presents the durations, in minutes, of 60 dormant periods. The list has been sorted into numerical order.

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

Figure 1.5 presents a stem-and-leaf plot of the geyser data. Each item in the sample is divided into two parts: a **stem**, consisting of the leftmost one or two digits, and the **leaf**, which consists of the next digit. In Figure 1.5, the stem consists of the tens digit and the leaf consists of the ones digit. Each line of the stem-and-leaf plot contains all of the sample items with a given stem. The stem-and-leaf plot is a compact way to represent the data. It also gives some indication of its shape. For the geyser data, we can see that there are relatively few durations in the 60–69 minute interval, compared with the 50–59, 70–79, or 80–89 minute intervals.

```
Stem    Leaf
   4    259
   5    0111133556678
   6    067789
   7    01233455556666699
   8    000012223344456668
   9    013
```

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

```
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.

When there are a great many sample items with the same stem, it is often necessary to assign more than one row to that stem. As an example, Figure 1.6 presents a computer-generated stem-and-leaf plot, produced by MINITAB, for the PM data in Table 1.2 in Section 1.2. The middle column, consisting of 0s, 1s, and 2s, contains the stems, which are the tens digits. To the right of the stems are the leaves, consisting of the ones digits for each of the sample items. Since many numbers are less than 10, the 0 stem must be assigned several lines, five in this case. Specifically, the first line contains the sample items whose ones digits are either 0 or 1, the next line contains the items whose ones digits are either 2 or 3, and so on. For consistency, all the stems are assigned several lines in the same way, even though there are few enough values for the 1 and 2 stems that they could have fit on fewer lines.

The output in Figure 1.6 contains a cumulative frequency column to the left of the stem-and-leaf plot. The upper part of this column provides a count of the number of items at or above the current line, and the lower part of the column provides a count of the number of items at or below the current line. Next to the line that contains the median is the count of items in that line, shown in parentheses.

A good feature of stem-and-leaf plots is that they display all the sample values. One can reconstruct the sample in its entirety from a stem-and-leaf plot—with one important exception: The order in which the items were sampled cannot be determined.

Dotplots

A **dotplot** is a graph that can be used to give a rough impression of the shape of a sample. It is useful when the sample size is not too large and when the sample contains some repeated values. Figure 1.7 (page 27) presents a dotplot for the geyser data in Table 1.3. For each value in the sample a vertical column of dots is drawn, with the number of dots in the column equal to the number of times the value appears in the sample. The dotplot gives a good indication of where the sample values are concentrated and where the gaps are. For example, it is immediately apparent from Figure 1.7 that the sample contains no dormant periods between 61 and 65 minutes in length.

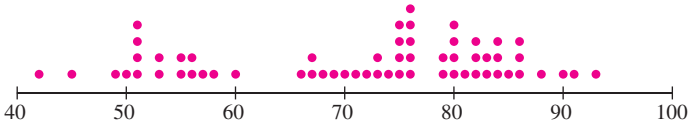


FIGURE 1.7 Dotplot for the geyser data in Table 1.3.

Stem-and-leaf plots and dotplots are good methods for informally examining a sample, and they can be drawn fairly quickly with pencil and paper. They are rarely used in formal presentations, however. Graphics more commonly used in formal presentations include the histogram and the boxplot, which we will now discuss.

Histograms

A **histogram** is a graphic that gives an idea of the "shape" of a sample, indicating regions where sample points are concentrated and regions where they are sparse. We will construct a histogram for the PM emissions of 62 vehicles driven at high altitude, as presented in Table 1.2 (Section 1.2). The sample values range from a low of 1.11 to a high of 23.38, in units of grams of emissions per gallon of fuel. The first step is to construct a **frequency table**, shown in Table 1.4.

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

The intervals in the left-hand column are called **class intervals**. They divide the sample into groups. For most histograms, the class intervals all have the same width. In Table 1.4, all classes have width 2. The notation 1–< 3, 3–< 5, and so on, indicates that a point on the boundary will go into the class on its right. For example, a sample value equal to 3 will go into the class 3–< 5, not 1–< 3.

There is no hard-and-fast rule as to how to choose the endpoints of the class intervals. In general, it is good to have more intervals rather than fewer, but it is also good to have large numbers of sample points in the intervals. Striking the proper balance is a matter of judgment and of trial and error. When the number of observations n is large (several hundred or more), some have suggested that reasonable starting points for the number

of classes may be $\log_2 n$ or $2n^{1/3}$. When the number of observations is smaller, more classes than these are often needed.

The column labeled "Frequency" in Table 1.4 presents the numbers of data points that fall into each of the class intervals. The column labeled "Relative Frequency" presents the frequencies divided by the total number of data points, which for these data is 62. The relative frequency of a class interval is the proportion of data points that fall into the interval. Note that since every data point is in exactly one class interval, the relative frequencies must sum to 1. Finally, the column labeled "Density" presents the relative frequency divided by the class width. In this case all classes have width 2, so the densities are found by dividing the relative frequencies by 2. Note that when the classes are of equal width, the frequencies, relative frequencies, and densities are proportional to one another.

Figure 1.8 presents a histogram for Table 1.4. The units on the horizontal axis are the units of the data, in this case grams per gallon. Each class interval is represented by a rectangle. When the class intervals are of equal width, the heights of the rectangles may be set equal to the frequencies, the relative frequencies, or the densities. Since these three quantities are proportional, the shape of the histogram will be the same in each case. For the histogram in Figure 1.8, the heights of the rectangles are the relative frequencies.

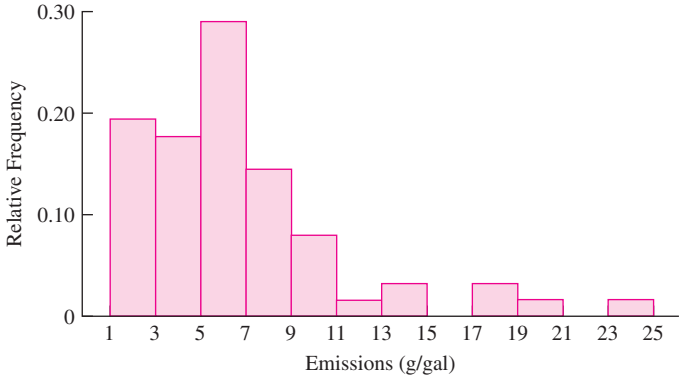


FIGURE 1.8 Histogram for the data in Table 1.4. In this histogram the heights of the rectangles are the relative frequencies. Since the class widths are all the same, the frequencies, relative frequencies, and densities are proportional to one another, so it would have been equally appropriate to set the heights equal to the frequencies or to the densities.

Unequal Class Widths

In some cases, histograms are drawn with class intervals of differing widths. This may be done when it is desired for the histogram to have a smoother appearance, or when the data come in the form of a frequency table in which the classes have unequal widths. Table 1.5 presents the PM data of Table 1.4 with the last seven classes collapsed into two.

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

It is important to note that because the class widths vary in size, the densities are no longer proportional to the relative frequencies. Instead, the densities adjust the relative frequency for the width of the class. Other things being equal, wider classes tend to contain more sample items than the narrower classes, and thus tend to have larger relative frequencies. Dividing the relative frequency by the class width to obtain the density adjusts for this tendency. For this reason, when the classes have unequal widths, *the heights of the rectangles must be set equal to the densities*. The *areas* of the rectangles then represent the relative frequencies.

Figure 1.9 presents the histogram for Table 1.5. Comparing this histogram to the one in Figure 1.8 shows that the string of small rectangles on the right has been smoothed out.

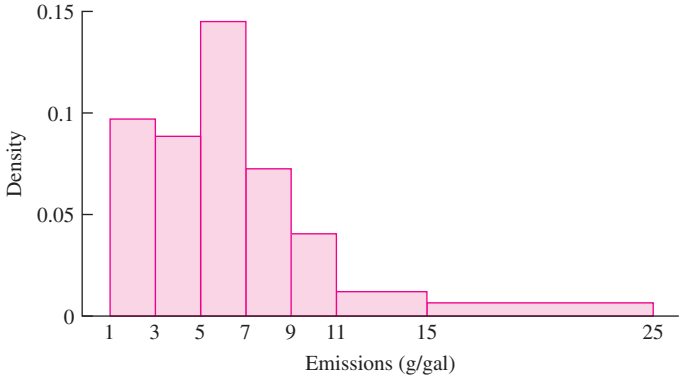


FIGURE 1.9 Histogram for the PM emissions for high-altitude vehicles. The frequency table is presented in Table 1.5. Since the classes have differing widths, the heights of the rectangles *must* be set equal to the densities. The areas of the rectangles are then equal to the relative frequencies. Compare with the equal-class-width histogram in Figure 1.8.

Summary

When the class intervals are of unequal widths, the heights of the rectangles must be set equal to the densities. The areas of the rectangles will then be the relative frequencies.

Example 1.17

Use the histogram in Figure 1.8 to determine the proportion of the vehicles in the sample with emissions between 7 and 11 g/gal.

Solution

The proportion is the sum of the relative frequencies of the classes spanning the range between 7 and 11. This is found by adding the heights of the rectangles for the two class intervals covered. The result is $0.1452 + 0.0806 = 0.2258$. Note that this result can also be obtained from the frequency table. The proportion of data points with values between 7 and 9 is 0.1452, and the proportion between 9 and 11 is 0.0806. The proportion between 7 and 11 is therefore equal to $0.1452 + 0.0806 = 0.2258$.

Example 1.18

Use the histogram in Figure 1.9 to determine the proportion of the vehicles in the sample with emissions between 9 and 15 g/gal.

Solution

The proportion is the sum of the relative frequencies of the two classes spanning the range between 9 and 15. Since the heights of the rectangles represent densities, the areas of the rectangles represent relative frequencies. The sum of the areas of the rectangles is $(2)(0.0403) + (4)(0.0121) = 0.129$. Note that this result can also be obtained from the frequency table. The proportion of data points with values between 9 and 11 is 0.0806, and the proportion between 11 and 15 is 0.0484. The proportion between 9 and 15 is therefore equal to $0.0806 + 0.0484 = 0.129$.

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.

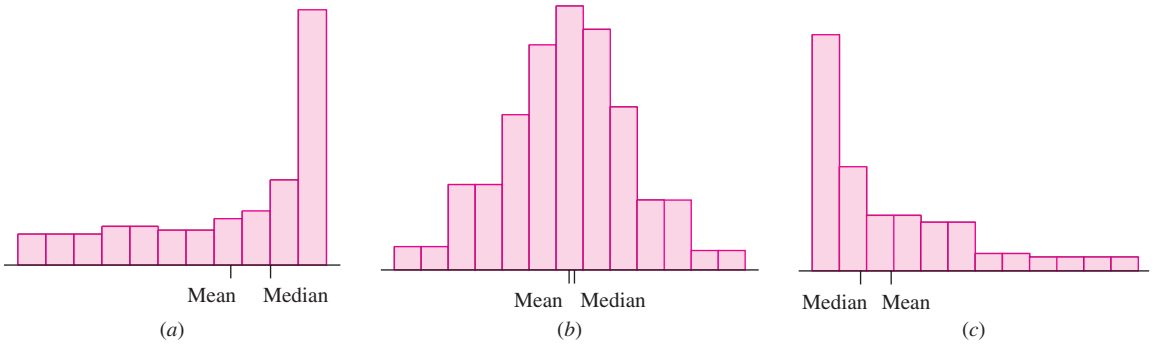


FIGURE 1.10 *(a)* A histogram skewed to the left. The mean is less than the median. *(b)* A nearly symmetric histogram. The mean and median are approximately equal. *(c)* A histogram skewed to the right. The mean is greater than the median.

Symmetry and Skewness

A histogram is perfectly **symmetric** if its right half is a mirror image of its left half. Histograms that are not symmetric are referred to as **skewed**. In practice, virtually no sample has a perfectly symmetric histogram; all exhibit some degree of skewness. In a skewed histogram, one side, or tail, is longer than the other. A histogram with a long right-hand tail is said to be **skewed to the right**, or **positively skewed**. A histogram with a long left-hand tail is said to be **skewed to the left**, or **negatively skewed**. While there is a formal mathematical method for measuring the skewness of a histogram, it is rarely used; instead people judge the degree of skewness informally by looking at the histogram. Figure 1.10 presents some histograms for hypothetical samples. Note that for a histogram that is skewed to the right (Figure 1.10c), the mean is greater than the median. The reason for this is that the mean is near the center of mass of the histogram, that is, it is near the point where the histogram would balance if supported there. For a histogram skewed to the right, more than half the data will be to the left of the center of mass. Similarly, the mean is less than the median for a histogram that is skewed to the left (Figure 1.10a). The histogram for the PM data (Figure 1.8) is skewed to the right. The sample mean is 6.596, which is greater than the sample median of 5.75.

Unimodal and Bimodal Histograms

We have used the term "mode" to refer to the most frequently occurring value in a sample. This term is also used in regard to histograms and other curves to refer to a peak, or local maximum. A histogram is **unimodal** if it has only one peak, or mode, and **bimodal** if it has two clearly distinct modes. In principle, a histogram can have more than two modes, but this does not happen often in practice. The histograms in Figure 1.10 are all unimodal. Figure 1.11 (page 32) presents a bimodal histogram for a hypothetical sample.

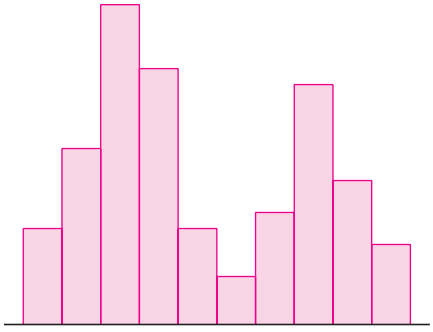


FIGURE 1.11 A bimodal histogram.

In some cases, a bimodal histogram indicates that the sample can be divided into two subsamples that differ from each other in some scientifically important way. Each sample corresponds to one of the modes. As an example, Table 1.6 presents the durations of 60 dormant periods of the geyser Old Faithful (originally presented in Table 1.3). Along with the durations of the dormant period, in minutes, the duration of the eruption immediately preceding the dormant period is classified either as short (less than 3 minutes) or long (more than 3 minutes).

Figure 1.12a presents a histogram for all 60 durations. Figures 1.12b and 1.12c present histograms for the durations following short and long eruptions, respectively. The histogram for all the durations is clearly bimodal. The histograms for the durations following short or long eruptions are both unimodal, and their modes form the two modes of the histogram for the full sample.

TABLE 1.6 Durations of dormant periods (in minutes) and of the previous eruptions of the geyser Old Faithful

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

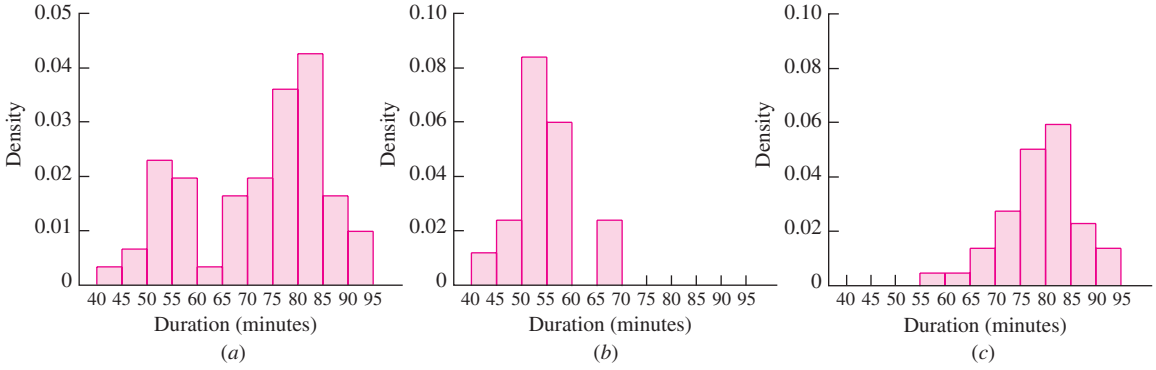


FIGURE 1.12 *(a)* Histogram for all 60 durations in Table 1.6. This histogram is bimodal. *(b)* Histogram for the durations in Table 1.6 that follow short eruptions. *(c)* Histogram for the durations in Table 1.6 that follow long eruptions. The histograms for the durations following short eruptions and for those following long eruptions are both unimodal, but the modes are in different places. When the two samples are combined, the histogram is bimodal.

Boxplots

A **boxplot** is a graphic that presents the median, the first and third quartiles, and any outliers that are present in a sample. Boxplots are easy to understand, but there is a bit of terminology that goes with them. The **interquartile range** is the difference between the third quartile and the first quartile. Note that since 75% of the data is less than the third quartile, and 25% of the data is less than the first quartile, it follows that 50%, or half, of the data are between the first and third quartiles. The interquartile range is therefore the distance needed to span the middle half of the data.

We have defined outliers as points that are unusually large or small. If IQR represents the interquartile range, then for the purpose of drawing boxplots, any point that is more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile, is considered an outlier. Some texts define a point that is more than 3 IQR from the first or third quartile as an **extreme outlier**. These definitions of outliers are just conventions for drawing boxplots and need not be used in other situations.

Figure 1.13 (page 34) presents a boxplot for some hypothetical data. The plot consists of a box whose bottom side is the first quartile and whose top side is the third quartile. A horizontal line is drawn at the median. 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.

Apart from any outliers, a boxplot can be thought of as having four pieces: the two parts of the box separated by the median line, and the two whiskers. Again apart from outliers, each of these four parts represents one-quarter of the data. The boxplot therefore indicates how large an interval is spanned by each quarter of the data, and in this way it can be used to determine the regions in which the sample values are more densely crowded and the regions in which they are more sparse.

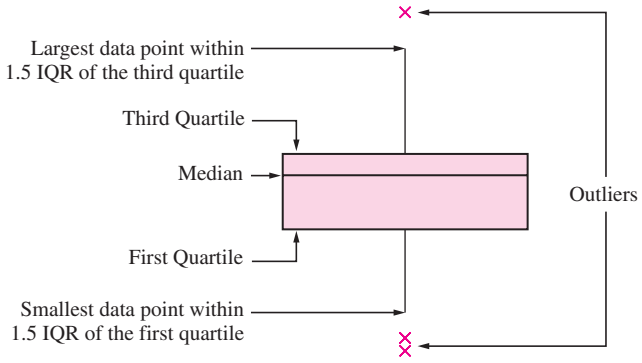


FIGURE 1.13 Anatomy of a boxplot.

Steps in the Construction of a Boxplot

- Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box.
- 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.
- 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.

Figure 1.14 presents a boxplot for the geyser data presented in Table 1.6. First note that there are no outliers in these data. Comparing the four pieces of the boxplot, we can tell that the sample values are comparatively densely packed between the median and

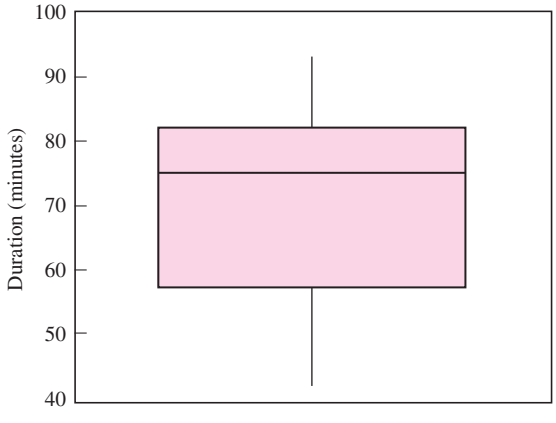


FIGURE 1.14 Boxplot for the Old Faithful dormant period data presented in Table 1.6.

the third quartile, and more sparse between the median and the first quartile. The lower whisker is a bit longer than the upper one, indicating that the data has a slightly longer lower tail than an upper tail. Since the distance between the median and the first quartile is greater than the distance between the median and the third quartile, and since the lower quarter of the data produces a longer whisker than the upper quarter, this boxplot suggests that the data are skewed to the left.

A histogram for these data was presented in Figure 1.12a. The histogram presents a more general impression of the spread of the data. Importantly, the histogram indicates that the data are bimodal, which a boxplot cannot do.

Comparative Boxplots

A major advantage of boxplots is that several of them may be placed side by side, allowing for easy visual comparison of the features of several samples. Tables 1.1 and 1.2 (in Section 1.2) presented PM emissions data for vehicles driven at high and low altitudes. Figure 1.15 presents a side-by-side comparison of the boxplots for these two samples.

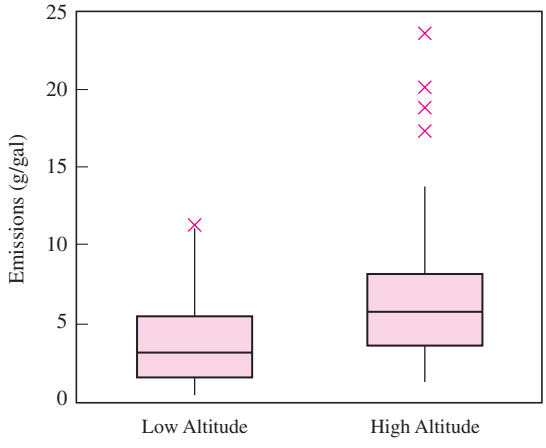


FIGURE 1.15 Comparative boxplots for PM emissions data for vehicles driven at high versus low altitudes.

The comparative boxplots in Figure 1.15 show that vehicles driven at low altitude tend to have lower emissions. In addition, there are several outliers among the data for high-altitude vehicles whose values are much higher than any of the values for the low-altitude vehicles (there is also one low-altitude value that barely qualifies as an outlier). We conclude that at high altitudes, vehicles have somewhat higher emissions in general, and that a few vehicles have much higher emissions. The box for the high-altitude vehicles is a bit taller, and the lower whisker a bit longer, than that for the low-altitude vehicles. We conclude that apart from the outliers, the spread in values is slightly larger for the high-altitude vehicles and is much larger when the outliers are considered.

In Figure 1.4 (in Section 1.2) we compared the values of some numerical descriptive statistics for these two samples, and reached some conclusions similar to the previous

ones. The visual nature of the comparative boxplots in Figure 1.15 makes comparing the features of samples much easier.

We have mentioned that it is important to scrutinize outliers to determine whether they have resulted from errors, in which case they may be deleted. By identifying outliers, boxplots can be useful in this regard. The following example provides an illustration.

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

The 12 wafers in each run were of several different types and were processed in several different furnace locations. The purpose in collecting the data was to determine whether the thickness of the oxide layer was affected either by the type of wafer or the furnace location. This was therefore a factorial experiment, with wafer type and furnace location as the factors, and oxide layer thickness as the outcome. The experiment was designed so that there was not supposed to be any systematic difference in the thicknesses between one run and another. The first step in the analysis was to construct a boxplot for

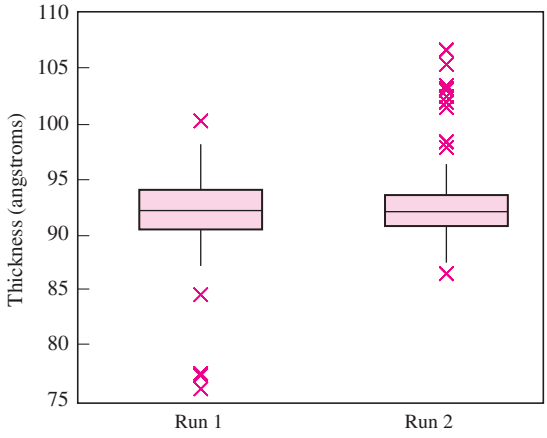


FIGURE 1.16 Comparative boxplots for oxide layer thickness data.

the data in each run to help determine if this condition was in fact met, and whether any of the observations should be deleted. The results are presented in Figure 1.16.

The boxplots show that there were several outliers in each run. Note that apart from these outliers, there are no striking differences between the samples, and therefore no evidence of any systematic difference between the runs. The next task is to inspect the outliers, to determine which, if any, should be deleted. By examining the data in Table 1.7, it can be seen that the eight largest measurements in run 2 occurred on a single wafer: number 10.

It was then determined that this wafer had been contaminated with a film residue, which caused the large thickness measurements. It would therefore be appropriate to delete these measurements. In the actual experiment, the engineers had data from several other runs available, and for technical reasons, decided to delete the entire run, rather than to analyze a run that was missing one wafer. In run 1, the three smallest measurements were found to have been caused by a malfunctioning gauge, and were therefore appropriately deleted. No cause could be determined for the remaining two outliers in run 1, so they were included in the analysis.

Multivariate Data

Sometimes the items in a population may have several values associated with them. For example, imagine choosing a random sample of days and determining the average temperature and humidity on each day. Each day in the population provides two values, temperature and humidity. The random sample therefore would consist of pairs of numbers. If the precipitation were measured on each day as well, the sample would consist of triplets. In principle, any number of quantities could be measured on each day, producing a sample in which each item is a list of numbers.

Data for which each item consists of more than one value is called **multivariate data**. 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**. If the data

consist of ordered pairs $(x_1, y_1), \ldots, (x_n, y_n)$, then a scatterplot is constructed simply by plotting each point on a two-dimensional coordinate system. Scatterplots can also be used to summarize multivariate data when each item consists of more than two values. One simply constructs separate scatterplots for each pair of values.

The following example illustrates the usefulness of scatterplots. The article "Advances in Oxygen Equivalence Equations for Predicting the Properties of Titanium Welds" (D. Harwig, W. Ittiwattana, and H. Castner, *The Welding Journal*, 2001: 126s–136s) presents data concerning the chemical composition and strength characteristics of a number of titanium welds. Figure 1.17 presents two scatterplots. Figure 1.17a is a plot of the yield strength [in thousands of pounds per square inch (ksi)] versus carbon content (in percent) for some of these welds. Figure 1.17b is a plot of the yield strength (in ksi) versus nitrogen content (in percent) for the same welds.

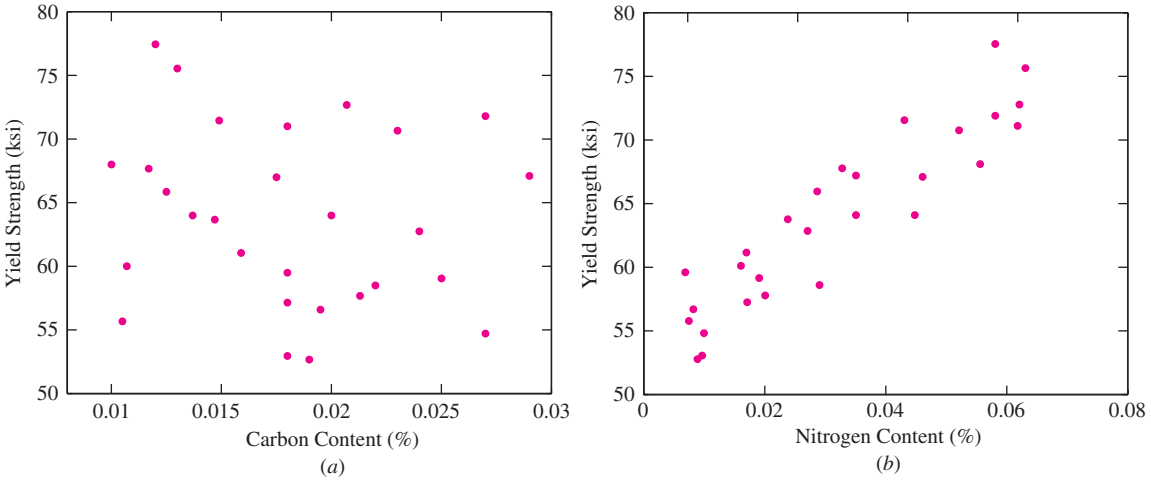


FIGURE 1.17 *(a)* A scatterplot showing that there is not much of a relationship between carbon content and yield strength for a certain group of welds. *(b)* A scatterplot showing that for these same welds, higher nitrogen content is associated with higher yield strength.

The plot of yield strength versus nitrogen content (Figure 1.17b) shows some clear structure—the points seem to be following a line from lower left to upper right. In this way, the plot illustrates a relationship between nitrogen content and yield strength: Welds with higher nitrogen content tend to have higher yield strength. This scatterplot might lead investigators to try to predict strength from nitrogen content or to try to increase nitrogen content to increase strength. (The fact that there is a relationship on a scatterplot does not guarantee that these attempts will be successful, as we will discuss in Section 7.1.) In contrast, there does not seem to be much structure to the scatterplot of yield strength versus carbon content, and thus there is no evidence of a relationship between these two quantities. This scatterplot would discourage investigators from trying to predict strength from carbon content.

Exercises for Section 1.3

1. The weather in Los Angeles is dry most of the time, but it can be quite rainy in the winter. The rainiest month of the year is February. The following table presents the annual rainfall in Los Angeles, in inches, for each February from 1965 to 2006.

0.2	3.7	1.2	13.7	1.5	0.2	1.7
0.6	0.1	8.9	1.9	5.5	0.5	3.1
3.1	8.9	8.0	12.7	4.1	0.3	2.6
1.5	8.0	4.6	0.7	0.7	6.6	4.9
0.1	4.4	3.2	11.0	7.9	0.0	1.3
2.4	0.1	2.8	4.9	3.5	6.1	0.1

a. Construct a stem-and-leaf plot for these data.
b. Construct a histogram for these data.
c. Construct a dotplot for these data.
d. Construct a boxplot for these data. Does the boxplot show any outliers?

2. Forty-five specimens of a certain type of powder were analyzed for sulfur trioxide content. Following are the results, in percent. The list has been sorted into numerical order.

14.1	14.4	14.7	14.8	15.3	15.6	16.1	16.6	17.3
14.2	14.4	14.7	14.9	15.3	15.7	16.2	17.2	17.3
14.3	14.4	14.8	15.0	15.4	15.7	16.4	17.2	17.8
14.3	14.4	14.8	15.0	15.4	15.9	16.4	17.2	21.9
14.3	14.6	14.8	15.2	15.5	15.9	16.5	17.2	22.4

a. Construct a stem-and-leaf plot for these data.
b. Construct a histogram for these data.
c. Construct a dotplot for these data.
d. Construct a boxplot for these data. Does the boxplot show any outliers?

3. Refer to Table 1.2 (in Section 1.2). Construct a stem-and-leaf plot with the ones digit as the stem (for values greater than or equal to 10 the stem will have two digits) and the tenths digit as the leaf. How many stems are there (be sure to include leafless stems)? What are some advantages and disadvantages of this plot, compared to the one in Figure 1.6 (page 26)?

4. Following are measurements of soil concentrations (in mg/kg) of chromium (Cr) and nickel (Ni) at 20 sites in the area of Cleveland, Ohio. These data are taken from the article "Variation in North American Regulatory Guidance for Heavy Metal Surface Soil Contamination at Commercial and Industrial Sites" (A. Jennings and J. Ma, *J Environment Eng*, 2007:587–609).

Cr:	34	1	511	2	574	496	322	424
	269	140	244	252	76	108	24	38
	18	34	30	191				
Ni:	23	22	55	39	283	34	159	37
	61	34	163	140	32	23	54	837
	64	354	376	471				

a. Construct a histogram for each set of concentrations.
b. Construct comparative boxplots for the two sets of concentrations.
c. Using the boxplots, what differences can be seen between the two sets of concentrations?

5. A certain reaction was run several times using each of two catalysts, A and B. The catalysts were supposed to control the yield of an undesirable side product. Results, in units of percentage yield, for 24 runs of catalyst A and 20 runs of catalyst B are as follows:

Catalyst A			
4.4	3.4	2.6	3.8
4.9	4.6	5.2	4.7
4.1	2.6	6.7	4.1
3.6	2.9	2.6	4.0
4.3	3.9	4.8	4.5
4.4	3.1	5.7	4.5

Catalyst B			
3.4	1.1	2.9	5.5
6.4	5.0	5.8	2.5
3.7	3.8	3.1	1.6
3.5	5.9	6.7	5.2
6.3	2.6	4.3	3.8

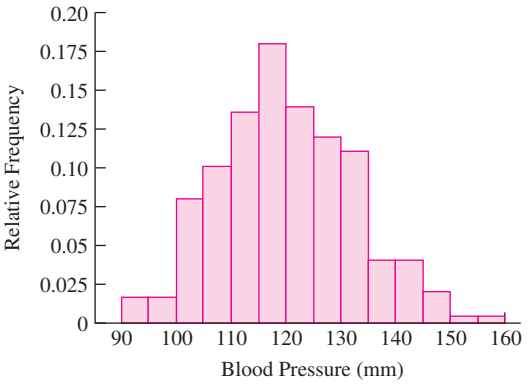
a. Construct a histogram for the yields of each catalyst.
b. Construct comparative boxplots for the yields of the two catalysts.
c. Using the boxplots, what differences can be seen between the results of the yields of the two catalysts?

6. Sketch a histogram for which
a. The mean is greater than the median.
b. The mean is less than the median.
c. The mean is approximately equal to the median.

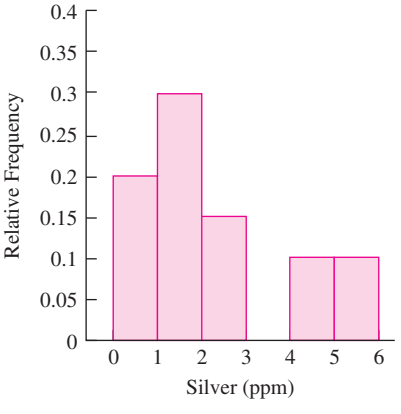
7. The following histogram presents the distribution of systolic blood pressure for a sample of women. Use it to answer the following questions.

a. Is the percentage of women with blood pressures above 130 mm closest to 25%, 50%, or 75%?

b. In which interval are there more women: 130–135 or 140–150 mm?



8. The following histogram presents the amounts of silver [in parts per million (ppm)] found in a sample of rocks. One rectangle from the histogram is missing. What is its height?



9. Refer to Table 1.4 (in Section 1.3).

a. Using the class intervals in the table, construct a histogram in which the heights of the rectangles are equal to the frequencies.

b. Using the class intervals in the table, construct a histogram in which the heights of the rectangles are equal to the densities.

c. Compare the histograms in parts (a) and (b) with the histogram in Figure 1.8, for which the heights are the relative frequencies. Are the shapes of the histograms the same?

10. Refer to Table 1.5 (in Section 1.3).

a. Using the class intervals in the table, construct a histogram in which the heights of the rectangles are equal to the relative frequencies.

b. Compare the histogram in part (a) with the histogram in Figure 1.9, for which the heights are the densities. Are the shapes of the histograms the same?

c. Explain why the heights should not be set equal to the relative frequencies in this case.

d. Which classes are visually exaggerated by making the heights equal to the relative frequencies?

11. The following table presents the number of students absent in a middle school in northwestern Montana for each school day in January 2008.

Date	Number Absent	Date	Number Absent	Date	Number Absent
Jan. 2	65	Jan. 14	59	Jan. 23	42
Jan. 3	67	Jan. 15	49	Jan. 24	45
Jan. 4	71	Jan. 16	42	Jan. 25	46
Jan. 7	57	Jan. 17	56	Jan. 28	100
Jan. 8	51	Jan. 18	45	Jan. 29	59
Jan. 9	49	Jan. 21	77	Jan. 30	53
Jan. 10	44	Jan. 22	44	Jan. 31	51
Jan. 11	41				

a. Construct a boxplot.

b. There was a snowstorm on January 27. Was the number of absences the next day an outlier?

12. Which of the following statistics *cannot* be determined from a boxplot?

i. The median

ii. The mean

iii. The first quartile

iv. The third quartile

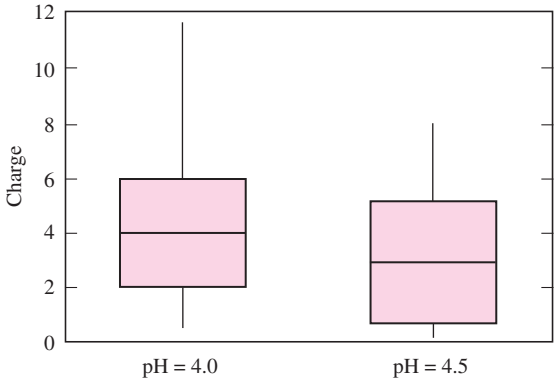
v. The interquartile range

13. A sample of 100 men has average height 70 in. and standard deviation 2.5 in. A sample of 100 women has average height 64 in. and standard deviation 2.5 in. If both samples are combined, the standard deviation of all 200 heights will be _____

i. less than 2.5 in.
ii. greater than 2.5 in.
iii. equal to 2.5 in.
iv. can't tell from the information given.

(*Hint:* Don't do any calculations. Just try to sketch, very roughly, histograms for each sample separately, and then one for the combined sample.)

14. Following are boxplots comparing the charge [in coulombs per mole (C/mol) $\times$ 10^{-25}] at pH 4.0 and pH 4.5 for a collection of proteins (from the article "Optimal Synthesis of Protein Purification Processes," E. Vasquez-Alvarez, M. Leinqueo, and J. Pinto, *Biotechnology Progress* 2001:685–695). True or false:



a. The median charge for the pH of 4.0 is greater than the 75th percentile of charge for the pH of 4.5.
b. Approximately 25% of the charges for pH 4.5 are less than the smallest charge at pH 4.0.
c. About half the sample values for pH 4.0 are between 2 and 4.
d. There is a greater proportion of values outside the box for pH 4.0 than for pH 4.5.
e. Both samples are skewed to the right.
f. Both samples contain outliers.

15. Following are summary statistics for two data sets, A and B.

	A	B
Minimum	0.066	−2.235
1st Quartile	1.42	5.27
Median	2.60	8.03
3rd Quartile	6.02	9.13
Maximum	10.08	10.51

a. Compute the interquartile ranges for both A and B.
b. Do the summary statistics for A provide enough information to construct a boxplot? If so, construct the boxplot. If not, explain why.
c. Do the summary statistics for B provide enough information to construct a boxplot? If so, construct the boxplot. If not, explain why.

16. Match each histogram to the boxplot that represents the same data set.