Stem-and-Leaf Plot and Histogram

Statistical data, generated in large masses, can be very useful for studying the behavior of the distribution if presented in a combined tabular and graphic display called a stem-and-leaf plot.

The method that is developed in the 1960s by the late Professor John Tukey of Princeton University, is called a stem-and-leaf diagram, or stem plot. This ingenious diagram is often easier to construct than either a frequency distribution or a histogram and generally displays more information. With a stem-and-leaf diagram, we think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit. In general, stems may use as many digits as required, but each leaf must contain only one digit.

To Construct a Stem-and-Leaf Diagram

Step 1 **Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.**

Step 2 **Write the stems from smallest to largest in a vertical column to the left of a vertical rule.**

Step 3 **Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.**

Step 4 **Arrange the leaves in each row in ascending order.**

Example:

Days to Maturity for Short-Term Investments Table 2.12 repeats the data on the number of days to maturity for 40 short-term investments. Previously, we grouped these data with a frequency distribution (Table 2.7 on page 52) and graphed them with a frequency histogram (Fig. 2.5(a) on page 56). Now let's construct a stem-and-leaf diagram, which simultaneously groups the data and provides a graphical display similar to a histogram.

TABLE 2.12

Days to maturity for 40 short-term investments

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

Solution:

Step 1 Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.

Referring to Table 2.12, we note that these observations are two-digit numbers. Thus, in this case, we use the first digit of each observation as the stem and the second digit as the leaf.

Step 2 Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

Referring again to Table 2.12, we see that the stems consist of the numbers 3, 4, . . . , 9. See the numbers to the left of the vertical rule in Fig. 2.8(a).

Step 3 Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

The first number in Table 2.12 is 70, which calls for a 0 to the right of the stem 7. Reading down the first column of Table 2.12, we find that the second number is 62, which calls for a 2 to the right of the stem 6. We continue in this manner until we account for all of the observations in Table 2.12. The result is the diagram displayed in Fig. 2.8(a).

Step 4 Arrange the leaves in each row in ascending order.

The first row of leaves in Fig. 2.8(a) is 8, 6, and 9. Arranging these numbers in ascending order, we get the numbers 6, 8, and 9, which we write in the first row to the right of the vertical rule in Fig. 2.8(b). We continue in this manner until the leaves in each row are in ascending order, as shown in Fig. 2.8(b), which is the stem-and-leaf diagram for the days-to-maturity data.

FIGURE 2.8
Constructing a stem-and-leaf diagram for the days-to-maturity data

3	8 6 9
4	7
5	7 1 6 3 5 1 0 5
6	2 4 7 3 6 4 0 9 8 5
7	0 5 1 0 9 8 0
8	5 9 1 7 0 3 6
9	9 9 5 8

(a)

Stems	Leaves
3	6 8 9
4	7
5	0 1 1 3 5 5 6 7
6	0 2 3 4 4 5 6 7 8 9
7	0 0 0 1 5 8 9
8	0 1 3 5 6 7 9
9	5 8 9 9

(b)

Example:

Cholesterol Levels According to the *National Health and Nutrition Examination Survey*, published by the Centers for Disease Control, the average cholesterol level for children between 4 and 19 years of age is 165 mg/dL. A pediatrician tested the cholesterol levels of several young patients and was alarmed to find that many had levels higher than 200 mg/dL. Table 2.13 presents the readings of 20 patients with high levels. Construct a stem-and-leaf diagram for these data by using

a. one line per stem. **b.** two lines per stem.

TABLE 2.13
Cholesterol levels for 20 high-level patients

210	209	212	208
217	207	210	203
208	210	210	199
215	221	213	218
202	218	200	214

Solution:

Because these observations are three-digit numbers, we use the first two digits of each number as the stem and the third digit as the leaf.

a. Using one line per stem and applying Procedure 2.7, we obtain the stem-and-leaf diagram displayed in Fig. 2.9(a).

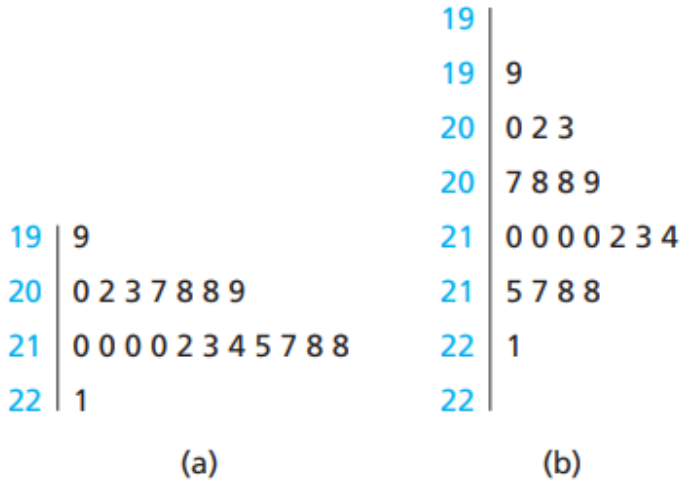


FIGURE 2.9
Stem-and-leaf diagram for cholesterol levels: (a) one line per stem; (b) two lines per stem.

b. The stem-and-leaf diagram in Fig. 2.9(a) is only moderately helpful because there are so few stems. Figure 2.9(b) is a better stem-and-leaf diagram for these data. It uses two lines for each stem, with the first line for the leaf digits 0–4 and the second line for the leaf digits 5–9.

Example:

To illustrate the construction of a stem-and-leaf plot, consider the data of Table 1.4, which specifies the "life" of 40 similar car batteries recorded to the nearest tenth of a year. The batteries are guaranteed to last 3 years. First, split each observation into two parts consisting of a stem and a leaf such that the stem represents the digit preceding the decimal and the leaf corresponds to the decimal part of the number. In other words, for the number 3.7, the digit 3 is designated the stem and the digit 7 is the leaf. The four stems 1, 2, 3, and 4 for our data are listed vertically on the left side in Table 1.5; the leaves are recorded on the right side opposite the appropriate stem value. Thus, the leaf 6 of the number 1.6 is recorded opposite the stem 1; the leaf 5 of the number 2.5 is recorded opposite the stem 2; and so forth. The number of leaves recorded opposite each stem is summarized under the frequency column.

Table 1.4: Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Table 1.5: Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1	69	2
2	25669	5
3	00111122233344455677889 9	25
4	11234577	8

The stem-and-leaf plot of Table 1.5 contains only four stems and consequently does not provide an adequate picture of the distribution. To remedy this problem, we need to increase the number of stems in our plot. One simple way to accomplish this is to write each stem value twice and then record the leaves 0, 1, 2, 3, and 4 opposite the appropriate stem value where it appears for the first time, and the leaves 5, 6, 7, 8, and 9 opposite this same stem value where it appears for the second time. This modified double-stem-and-leaf plot is illustrated in Table 1.6, where the stems corresponding to leaves 0 through 4 have been coded by the symbol $\star$ and the stems corresponding to leaves 5 through 9 by the symbol $\cdot$.

Table 1.6: Double-Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1$\cdot$	69	2
2$\star$	2	1
2$\cdot$	5669	4
3$\star$	001111222333444	15
3$\cdot$	5567778899	10
4$\star$	11234	5
4$\cdot$	577	3

Distribution Shapes:

Distribution of a Data Set

The **distribution of a data set** is a table, graph, or formula that provides the values of the observations and how often they occur.

Up to now, we have portrayed distributions of data sets by frequency distributions, relative-frequency distributions, frequency histograms, relative-frequency histograms, dot plots, stem-and-leaf diagrams, pie charts, and bar charts. An important aspect of the distribution of a quantitative data set is its shape. Indeed, as we demonstrate in later chapters, the shape of a distribution frequently plays a role in determining the appropriate method of statistical analysis. To identify the shape of a distribution, the best approach usually is to use a smooth curve that approximates the overall shape.

Example:

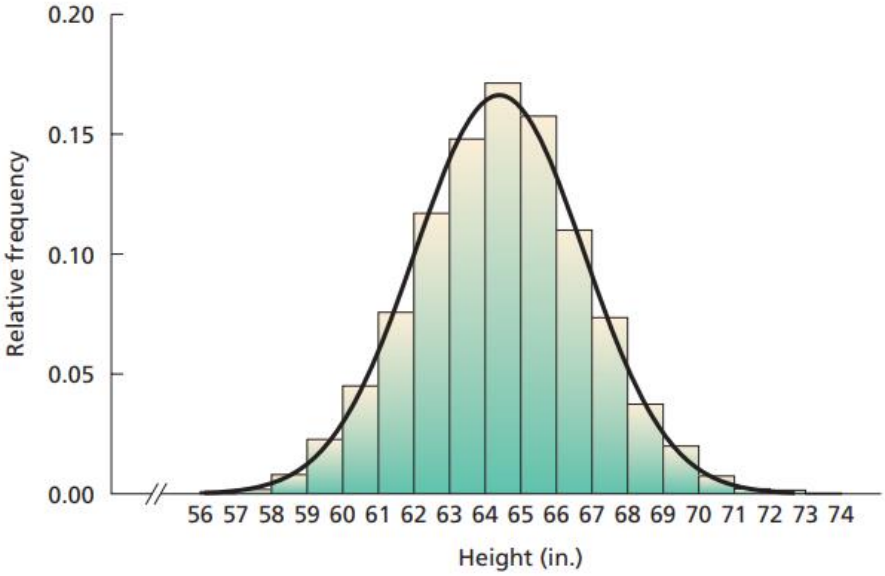


FIGURE 2.10

Relative-frequency histogram and approximating smooth curve for the distribution of heights

Above figure displays a relative-frequency histogram for the heights of the 3264 female students who attend a mid-western college. Also included in Fig histogram and the smooth curve show that this distribution of heights is bell shaped (or mound shaped), but the smooth curve makes seeing the shape a little easier. Another advantage of using smooth curves to identify distribution shapes is that we need not worry about minor differences in shape. Instead we can concentrate on overall patterns, which, in turn, allows us to classify most distributions by designating relatively few shapes.

Distribution Shapes:

Figure 2.11 displays some common distribution shapes: **bell shaped, triangular, uniform, reverse J shaped, J shaped, right skewed, left skewed, bimodal,** and **multimodal.** A distribution doesn't have to have one of these exact shapes in order to take the name: it need only approximate the shape, especially if the data set is small. So, for instance, we describe the distribution of heights in Fig. 2.10 as bell shaped, even though the histogram does not form a perfect bell.

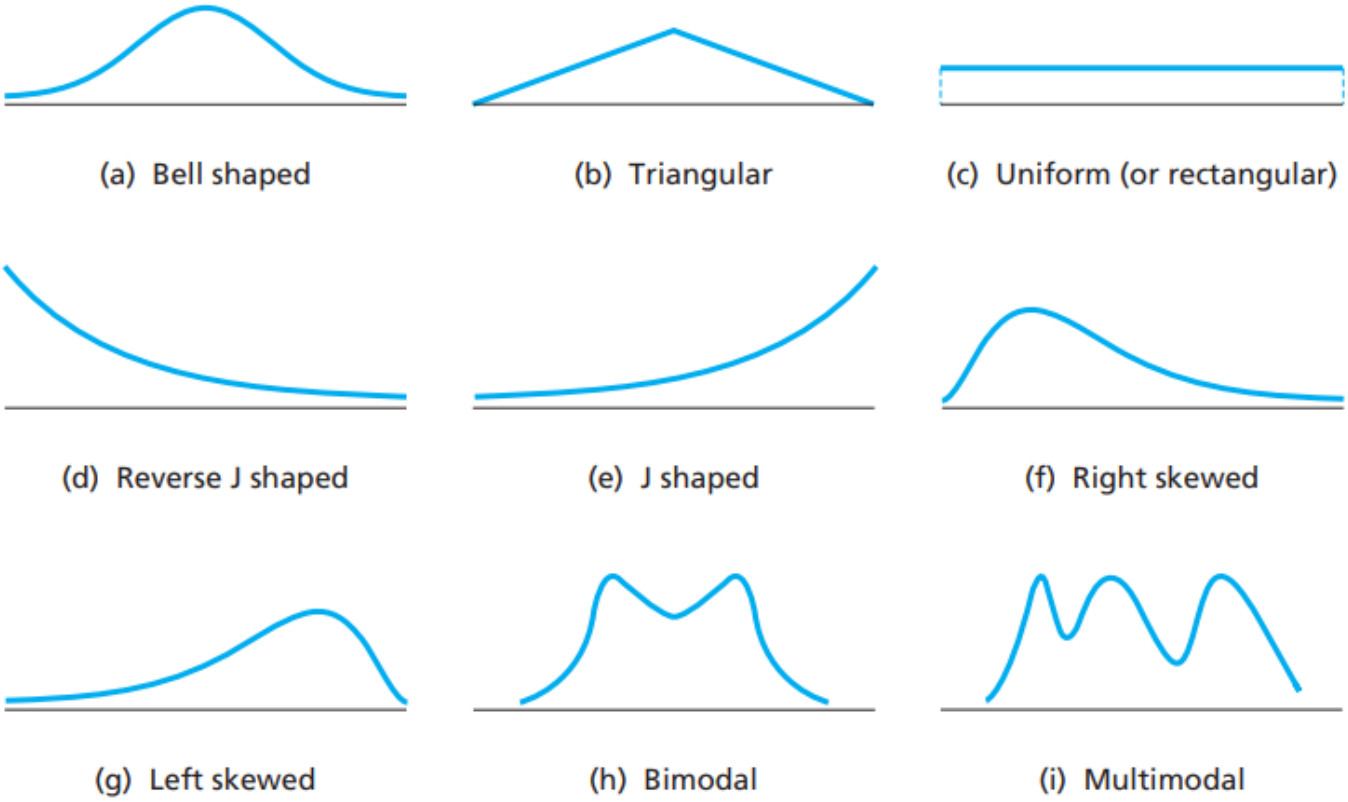


FIGURE 2.11
Common distribution shapes

Modality:

When considering the shape of a distribution, you should observe its number of peaks (highest points). A distribution is **unimodal** if it has one peak; **bimodal** if it has two peaks; and **multimodal** if it has three or more peaks.

The distribution of heights in Fig. 2.10 is unimodal. More generally, we see from Fig. 2.11 that bell-shaped, triangular, reverse J-shaped, J-shaped, right-skewed, and left-skewed distributions are unimodal. Representations of bimodal and multimodal distributions are displayed in Figs. 2.11(h) and (i), respectively.‡

Symmetry and Skewness:

Each of the three distributions in Figs. 2.11(a)–(c) can be divided into two pieces that are mirror images of one another. A distribution with that property is called **symmetric.** Therefore bell-shaped, triangular, and uniform distributions are symmetric. The bimodal distribution pictured in Fig. 2.11(h) also happens to be symmetric, but it is not always true that bimodal or multimodal distributions are symmetric. Figure 2.11(i) shows an asymmetric multimodal distribution.

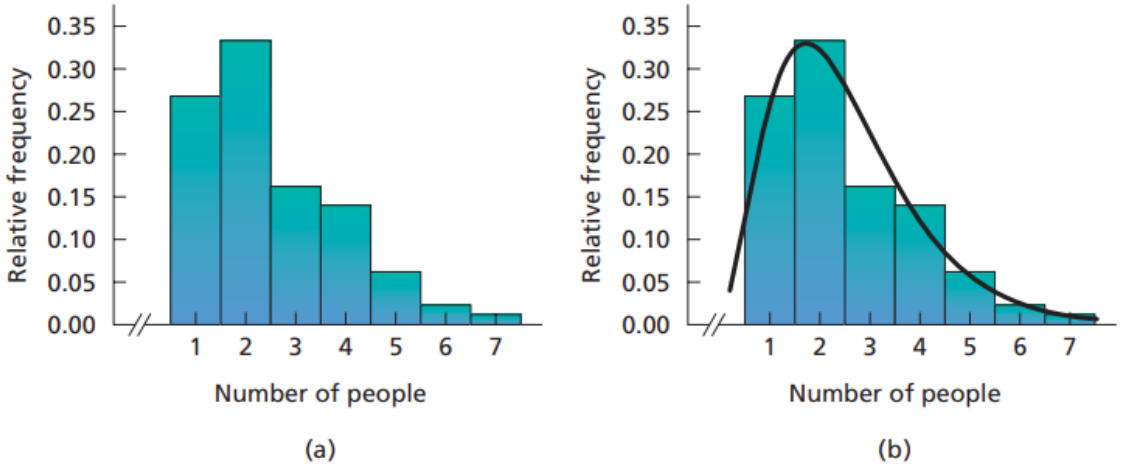
Again, when classifying distributions, we must be flexible. Thus, exact symmetry is not required to classify a distribution as symmetric. For example, the distribution of heights in Fig. 2.10 is considered symmetric.

A unimodal distribution that is not symmetric is either right skewed, as in Fig. 2.11(f), or left skewed, as in Fig. 2.11(g). A right-skewed distribution rises to its peak rapidly and comes back toward the horizontal axis more slowly—its "right tail" is longer than its "left tail." A left-skewed distribution rises to its peak slowly and comes back toward the horizontal axis more rapidly—its "left tail" is longer than its "right tail." Note that reverse J-shaped distributions [Fig. 2.11(d)] and J-shaped distributions [Fig. 2.11(e)] are special types of right-skewed and left-skewed distributions, respectively.

Example: Identifying Distribution Shapes

Household Size The relative-frequency histogram for household size in the United States shown in Fig. 2.12(a) is based on data contained in *Current Population Reports*, a publication of the U.S. Census Bureau.† Identify the distribution shape for sizes of U.S. households.

FIGURE 2.12 Relative-frequency histogram for household size



Solution First, we draw a smooth curve through the histogram shown in Fig. 2.12(a) to get Fig. 2.12(b). Then, by referring to Fig. 2.11, we find that the distribution of household sizes is right skewed.

Frequency Polygon:

A polygon is another device that can be used to present quantitative data in graphic form. To draw a frequency polygon, we first mark a dot above the midpoint of each class at a height equal to the frequency of that class. This is the same as marking the midpoint at the top of each bar in a histogram. Next we include two more classes, one at each end, and mark their midpoints. Note that these two classes have zero frequencies. In the last step, we join the adjacent dots with straight lines. The resulting line graph is called a frequency polygon or simply a polygon. A polygon with relative frequencies marked on the vertical axis is called a relative frequency polygon. Similarly, a polygon with percentages marked on the vertical axis is called a percentage polygon.

Polygon A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines is called a *polygon*.

Figure 2.6 shows the frequency polygon for the frequency distribution of Table 2.8.

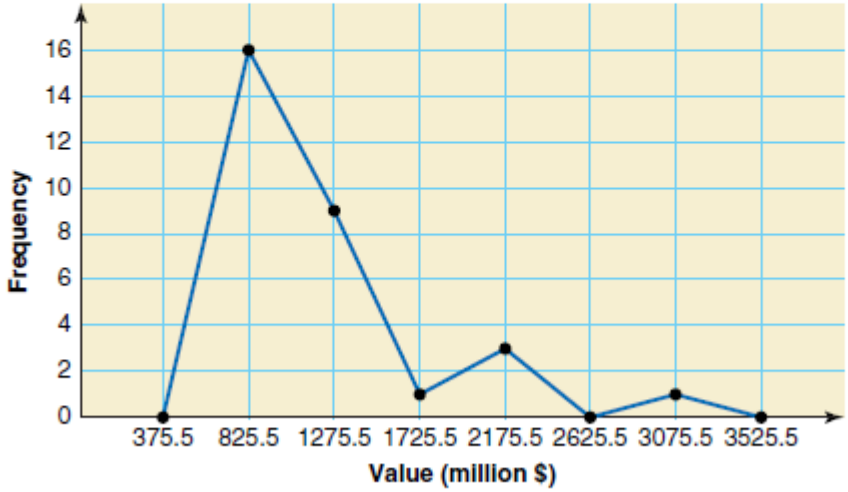


Figure 2.6 Frequency polygon for Table 2.8.

For a very large data set, as the number of classes is increased (and the width of classes is decreased), the frequency polygon eventually becomes a smooth curve. Such a curve is called a frequency distribution curve or simply a frequency curve. Figure 2.7 shows the frequency curve for a large data set with a large number of classes.

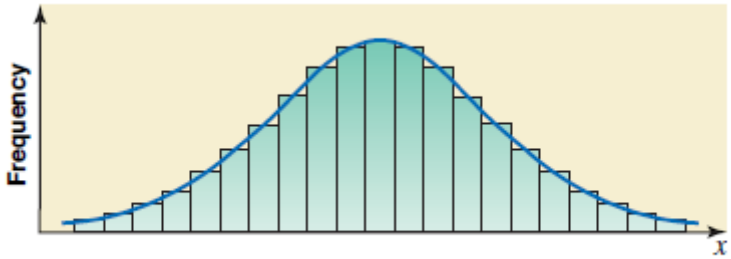


Figure 2.7 Frequency distribution curve.