CHAPTER 3

Frequency Distributions

3.1 Stemplots

The **stem-and-leaf plot (stemplot)** is a graphical technique that organizes data into a histogram-like display. It is an excellent way to begin an analysis and is a good way to learn several important statistical principals.

To construct a stemplot, begin by dividing each data point into a stem component and a leaf component. Considering this small sample of $n = 10$:

21 42 05 11 30 50 28 27 24 52

For these data points, the "tens place" will become stem values and the "ones place" will become leaf values. For example, the data point 21 has a stem value of 2 and leaf value of 1.

A stem-like axis is drawn. Because data range from 05 to 52, stem values will range from 0 to 5. Stem values are listed in ascending (or descending) order at regularly spaced intervals to form a number line. A vertical line may be drawn next to the stem to separate it from where the leaves will be placed.

```
0|
1|
2|
3|
4|
5|
×10
```

An **axis multiplier** (×10) is included below the stem to show that a stem value of 5 represents 50 (and not say 5 or 500). Leaf values are placed adjacent to their associated stem values. For example, "21" is:

```
0|
1|
2|1
3|
4|
5|
×10
```

The remaining leaves are plotted:

```
0|5
1|1
2|1874
3|0
4|2
5|02
×10
```

Leaves are then rearranged to appear in rank order:

```
0|5
1|1
2|1478
3|0
4|2
5|02
×10
```

The stemplot now resembles a histogram on its side. Rotate the plot 90 degrees to display the distribution in the more familiar horizontal orientation.

```
    8
    7
    4     2
5 1 1 0 2 0
-----------
0 1 2 3 4 5 (×10)
-----------
```

Three aspects of the distribution are now visible. These are its:

1. **Shape**
2. **Location**
3. **Spread**

Shape

Shape refers to the configuration of data points as they appear on the graph. This is seen as a "skyline silhouette":

```
    X
    X
    X     X
X X X X X X
-----------
0 1 2 3 4 5 (×10)
-----------
```

It is difficult to make statements about shape when the data set is this small (a few more data points landing just so can entirely change our impression of its shape), so let us look at larger data set.

Figure 3.1 is a histogram of about a thousand intelligence quotient scores. Overlaying the histogram is a fitted curve. Although the fit of the curve is imperfect, the curve still provides a convenient way to discuss the shape of the distribution.

A distribution's shape can be discussed in terms of its symmetry, modality, and kurtosis.

- **Symmetry** refers to the degree to which the shape reflects a mirror image of itself around its center.
- **Modality** refers to the number of peaks on the distribution.
- **Kurtosis** refers to the steepness of the mound.

Figure 3.2 illustrates these characteristics.

- Distributions (a)–(c) are *symmetrical*
- Distributions (d)–(f) are *asymmetrical*
- Distribution (d) is **bimodal**; the rest of the distributions are **unimodal**.

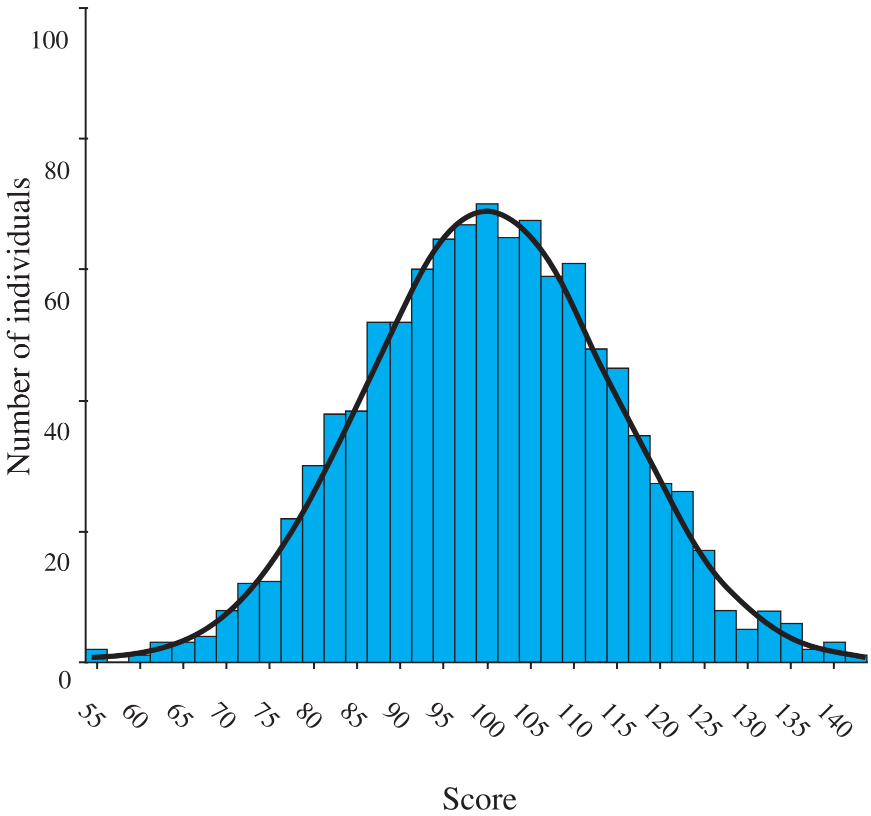

FIGURE 3.1 Histogram with overlying curve showing distribution's shape.

- Distribution (b) is flat with broad tails. This is a **platykuric** distribution (like a platypus). A tall curve with long skinny tails (not shown) is said to be **leptokurtic**. A curve with medium kurtosis is *mesokurtic*.[a]
- Distributions (e) and (f) are **skewed**. Figure (e) has a **positive skew** (tail toward larger numbers on a number line). Figure (f) has a **negative skew** (tail toward smaller numbers).

[a]Beware that it is often difficult to assess the degree of kurtosis visually in applied situations.

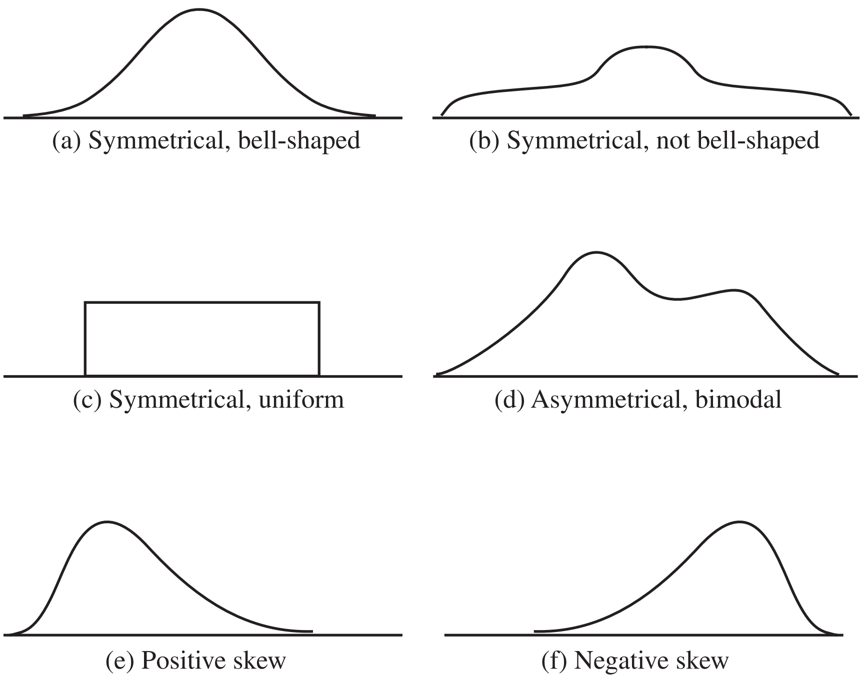

FIGURE 3.2 Examples of distributional shapes.

An **outlier** is a striking deviation from the overall pattern or shape of the distribution. As an example, the value of 50 on this stemplot is an outlier:

```
0|689
1|0124667
2|
3|
4|
5|0
×10
```

Location

We summarize the **location** of a distribution in terms of its center. **Figure 3.3** shows distributions with different locations. Although the two distributions overlap, distribution 2 has higher values on average, as portrayed by its shift toward the right.

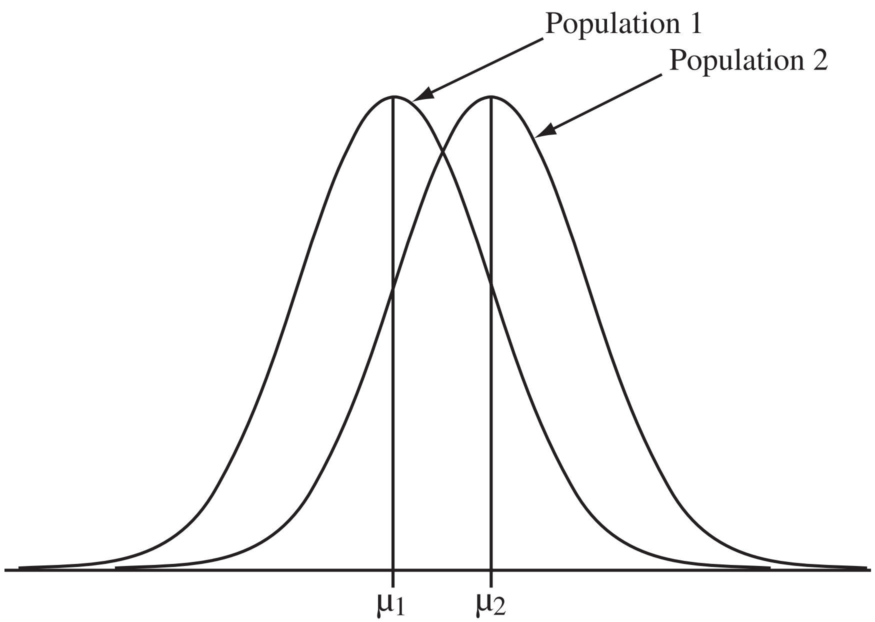

FIGURE 3.3 Distributions with different locations.

The term **average** refers to the center of a distribution.[b] There are different ways to identify a distribution's average, the two most common being the arithmetic average and the median.

The **arithmetic average** is a distribution's gravitational center. This is where the distribution would balance if placed on a scale. The balancing point for the stemplot here is somewhere between 20 and 30:

```
  8
  7
  4     2
5 1 1 0 2 0
-----------
0 1 2 3 4 5 (×10)
-----------
     ^
Gravitational
  Center
```

[b]Sometimes the term average is used restrictively to refer only to the arithmetic mean of a data set.

Of course this is only the *approximate* balancing point. The exact balancing point (arithmetic average) is determined by adding up all the values in the data set and dividing by the sample size. In this case, the arithmetic average = (21 + 42 + 5 + 11 + 30 + 50 + 28 + 27 + 24 + 52) ÷ 10 = 29. Our "eyeball estimate" was pretty good.

The **median** is the point that divides the data set into a top half and bottom half; it is halfway up (or down) the ordered list.

> **The depth of a data point corresponds to its rank from either the top or bottom of the ordered list of values.**

It is a little easier to determine the median if we stretch out the data to form an **ordered array**. The ordered array and median for the current data is:

```
05  11  21  24  27  28  30  42  50  52
                   ^
                 median
```

More formally, the median has a depth of $\frac{n\;1\;1}{2}$, where n is the sample size. When n is even, the median will fall between two values, in which case you simply average these values to get the median. For example, the median in our illustrative data set has a depth of $\frac{10\;1\;1}{2}\;5\;\;5.5$, placing it between the 5th (27) and 6th (28) ordered value. The average of these two points (27.5) is the median.

> **The arithmetic average is a distribution's balancing point.**
> **The median is its middle value.**

Spread

The term **spread** is an informal way to refer to the dispersion or *variability* of data points. **Figure 3.4** shows distributions with different variability. Population 1 and population 2 have the same central locations, but population 2 has greater spread (variability).

It is best to quantify spread with a statistic based on typical distance around the center of the distribution.

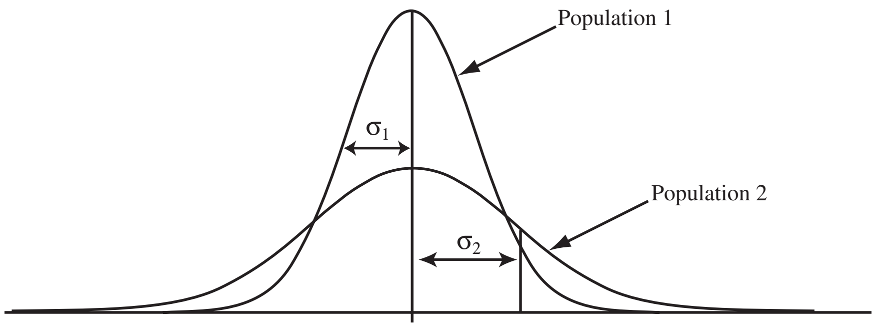

FIGURE 3.4 Distributions with different spreads.

There are several ways to measure spread, and we will learn several of these methods in Chapter 4. For now, let us simply describe spread in terms of the range of values, lowest to highest.

Additional Illustrations of Stemplots

The next couple of illustrations show how to draw a stemplot for data that might not immediately lend itself to plotting.

Illustrative Example: *Truncating leaf values*. Consider these eight data points:

1.47 2.06 2.36 3.43 3.74 3.78 3.94 4.42

Data have three significant digits although only two are needed for plotting. Our rule will be to "prune" the leaves by **truncating** extra digits before plotting. For example, the value 1.47 is truncated to 1.4, the value 2.06 is truncated to 2.0, and so on. We also drop the decimal point before plotting. For example, "1.47" appears as "1|4".

Here is the stemplot:

```
1|4
2|03
3|4779
4|4
×1
```

continues

continues

How do we interpret this plot? As always, consider its shape, location, and spread.

- Shape: A mound shape with no apparent outliers. (With such a small data set, little else can be said about shape.)
- Location: Because there are $n = 8$ data points, the median has a depth of $\frac{8+1}{2} = 4.5$. Count to a depth of 4½ to see that the median falls between 3.4 and 3.7 (underlined in the stemplot). Average these values; the median is about 3.55.
- Spread: Data spread from about 1.4 to 4.4. ■

Illustrative Example: *Irish health care web sites.* The Irish Department of Health recommends a reading level of 12 to 14 years of age for health information leaflets aimed at the public. **Table 3.1** lists reading levels for $n = 46$ Irish health care Web sites.

Table 3.1. Reading levels for Irish healthcare web sites ($n = 46$).

08	10	11	11	12	13	13	13	13	14
14	15	15	15	15	15	15	15	16	16
16	14	17	17	17	17	17	17	17	17
17	17	17	17	17	17	17	17	17	17
17	17	17	17	17	17				

Source: O'Mahoney, B. (1999). Irish Health Care Web Sites: A Review. *Irish Medical Journal, 92*(4), 334–336. Data are stored online in the file IRISHWEB.*.

continues

continues

If we imagine that each data point has an invisible ".0", plotting these zeros makes this revealing plot:

```
08|0
09|
10|0
11|00
12|0
13|0000
14|000
15|0000000
16|0
17|00000000000000000000000
×1
```

This distribution has a negative skew and a low outlier of 8.0 (shape). The median is 17 (location).[c] Data range from 8 to 17 (spread). ■

Splitting Stem Values

Sometimes a stemplot will be too squished to reveal its shape. In such circumstance, we can use **split stem values** to stretch out the stem. As an example, consider this plot.

```
1|4789
2|223466789
3|000123445678
×1
```

This plot is too squashed to reveal its shape, so we will split each stem value in two, listing two "1s" where there had been one, two "2s" and so on. Think of each stem value as a "bin." The first "1" will be a bin to hold values between 1.0 and 1.4. The second "1" will hold values between 1.5 and 1.9 (and so on). Here's the plot with split stem values:

---

[c] The median has a depth of $(n + 1)/2 = (46 + 1)/2 = 23.5$. This position is underlined revealing a median of 17.

```
1|4
1|789
2|2234
2|66789
3|00012344
3|5678
×1
```

This plot does a better job showing the shape of the distribution, revealing its negative skew.

When needed, we can also split stem values into five subunits. The following codes can be used to tag stem values:

*	for leaves of zero and one
T	for leaves of two and three
F	for leaves of four and five
S	for leaves of six and seven
·	for leaves of eight and nine

Consider these nine values:

3.5 8.1 7.4 4.0 0.7 4.9 8.4 7.0 5.5

A stemplot with quintuple-split stem values makes a nice picture:

```
0*|0
 T|3
 F|445
 S|77
 .|88
×10
```

How Many Stem Values?

When creating stemplots, you must choose how to scale the stem. Again, think of stem values as "bins" for collecting leaves. You can start with between 3 and 12 "bins" and make adjustments from there. If the plot is too squished, split the stem values. If it is too spread out, use a larger stem multiplier. Finding the most revealing plot may entail trial and error.

Illustrative Example: *Health insurance coverage.* A U.S. Census Bureau report looked at health insurance coverage in the United States for the period 2002 to 2004. **Table 3.2** lists the average percentage of people without health insurance coverage by state for this period.

Table 3.2 Percentage of residents without health insurance by state, U.S., 2004, $n = 51$.

State	%	State	%	State	%
Alabama	13.5	Kentucky	13.9	N. Dakota	11.0
Alaska	18.2	Louisiana	18.8	Ohio	11.8
Arizona	17.0	Maine	10.6	Oklahoma	19.2
Arkansas	16.7	Maryland	14.0	Oregon	16.1
California	18.4	Massachusetts	10.8	Pennsylvania	11.5
Colorado	16.8	Michigan	11.4	Rhode Is.	10.5
Connecticut	10.9	Minnesota	08.5	S. Carolina	13.8
Delaware	11.8	Mississippi	17.2	S. Dakota	11.9
Dist. Col.	13.5	Missouri	11.7	Tennessee	12.7
Florida	18.5	Montana	17.9	Texas	25.1
Georgia	16.6	Nebraska	11.0	Utah	13.4
Hawaii	09.9	Nevada	19.1	Vermont	10.5
Idaho	17.3	N. Hamp.	10.6	Virginia	13.6
Illinois	14.2	N. Jersey	14.4	Washington	14.2
Indiana	13.7	N. Mexico	21.4	W. Virginia	15.9
Iowa	10.1	New York	15.0	Wisconsin	10.4
Kansas	10.8	N. Carolina	16.6	Wyoming	15.9

Source: DeNavas-Walt, C., Proctor, B. D., & Lee, C. H. (2005). *Income, Poverty, and Health Insurance Coverage in the United States: 2004* (No. P60-229). Washington, D.C.: U.S. Government Printing Office. Data are stored online in the file INC-POV-HLTHINS.* as the variable NOINS.

The stemplot with single stem values looks like this:

```
0|89
1|000000000111111123333333444455566666777788899
2|15
×10
```

This plot is too squished, so we split the stem values to come up with the following plot:

continues

continues

```
0|89
1|0000000001111111233333334444
1|5556666677778888899
2|1
2|5
×10
```

This plot is improved, but still seems too compressed. Let's try a quintuple split of stem values:

```
0.|89
1*|0000000001111111
 T|23333333
 F|4444555
 S|666667777
 .|888899
2*|1
 T|
 F|5
×10
```

This plot reveals a positive skew and high outlier.

Illustrative Example: *Student weights.* Table 3.3 lists body weights of 53 students.

Table 3.3 Body weight (pounds) of students in a class, $n = 53$.

192	110	195	180	170	215
152	120	170	130	130	125
135	185	120	155	101	194
110	165	185	220	180	
128	212	175	140	187	
180	119	203	157	148	
260	165	185	150	106	
170	210	123	172	180	
165	186	139	175	127	
150	100	106	133	124	

Data are stored online in the file BODY-WEIGHT.*.

continues

continues

How would we plot this data in a way that is most revealing? First notice that values range from 100 to 260 pounds. Using a multiplier of ×100 would result in only two stem values (100–199 and 200–299). Splitting stem values with the ×100 in two would help, but would still result in only four stem values: 100–149, 150–199, 200–249, and 250–299. Using quintuple-split stem values produces this nice plot:

```
1*|0000111
1T|22222233333
1F|4455555
1S|666777777
1.|88888888999
2*|0111
2T|2
2F|
2S|6
×100
```

This plot has a positive skew and high outlier. The location of its median is underlined (median = 160), and data spread from 100 to 260. ■

Back-to-Back Stemplots

We can compare two distributions with **back-to-back stemplots**. To create this type of plot, draw a stem in a central gutter and place leaves from groups on either side of this central stem. Back-to-back plots make it easy to compare group shapes, locations, and spreads.

Illustrative Example: *Back-to-back stemplots.* Table 3.4 lists fasting cholesterol values (mg/dl) for two groups of men.

Table 3.4 Fasting cholesterol values (mg/dl) in two groups of men. Group 1 men were classified as type A personalities.

Group 1:									
233	291	312	250	246	197	268	224	239	239
254	276	234	181	248	252	202	218	212	325
Group 2:									
344	185	263	246	224	212	188	250	148	169
226	175	242	252	153	183	137	202	194	213

Data stored online in the file WCGS.*.

Here is the back-to-back stemplot of the data using quintuple-split stem values:

```
Group 1|   |Group 2
--------------
       |1t|3
       |1f|45
       |1s|67
     98|1.|8889
    110|2*|011
  33332|2t|22
  55544|2f|4455
     76|2s|6
      9|2.|
      1|3*|
      2|3t|
       |3f|4
       (×100)
```

Notice that the distribution of group 1 is shifted down the axis toward the higher values on the stem showing it to have higher values on average. ■

Exercises

3.1 ***Poverty in eastern states, 2000.*** **Table 3.5** lists the percentage of people living below the poverty line in the 26 states east of the Mississippi River for the year 2000. Make a stemplot of these values. After creating the plot, describe the distribution's shape, location, and spread. Are there any outliers? Which states straddle the median?

Table 3.5 Percentage of people living below the poverty line in each of the 26 States east of the Mississippi River for the year 2000.

Alabama	14.6	Maryland	07.3	Pennsylvania	09.9
Connecticut	07.6	Massachusetts	10.2	Rhode Is.	10.0
Delaware	09.8	Michigan	10.2	S. Carolina	11.9
Florida	12.1	Mississippi	15.5	Tennessee	13.3
Georgia	12.6	New Hamp.	07.4	Vermont	10.1
Illinois	10.5	New Jersey	08.1	Virginia	08.1
Indiana	08.2	New York	14.7	West Virginia	15.8
Kentucky	12.5	N. Carolina	13.2	Wisconsin	08.8
Maine	09.8	Ohio	11.1		

Source: Delaker, J. (2001). *Poverty in the United States, 2000* (No. P60-214). Washington, D.C.: U.S. Census Bureau. Table D, p. 11. Data are stored in the file POV-EAST-2000.*.

3.2 ***Hospitalization.*** **Table 3.6** list lengths of stays (days) for a sample 25 patients.

Table 3.6 Duration of hospitalization (days), $n = 25$.

5	10	6	11	5	14	30	11	17	3
9	3	8	8	5	5	7	4	3	7
9	11	11	9	4					

Data are stored online in the file HDUR.*as the variable DUR.

(a) Create a stemplot with single stem values for these data. (Use an axis multiplier of $\times 10$).
(b) Create a stemplot with split stem values.
(c) Which of the stemplots do you prefer?
(d) Describe in plain language the distribution's shape, location, and spread.

3.3 ***Leaves on a common stem.*** For each of the following comparisons, plot the data as back-to-back stemplots on a common stem. Then, compare group locations and spreads.

(a) Comparison A

Group 1:	90	70	50	30	10
Group 2:	70	60	50	40	30

(b) Comparison B

Group 1:	90	80	70	60	50
Group 2:	70	60	50	40	30

(c) Comparison C

Group 1:	90	70	50	30	10
Group 2:	90	80	70	60	50

3.4 ***Cholesterol comparison.*** **Table 3.7** lists plasma cholesterol levels ($mmol/m^3$) in two independent groups. Plot these data on a common stem. Then, compare group locations and spreads.

Table 3.7 Plasma cholesterol levels ($mmol/m^3$) in two independent groups.

Group 1 (mildly hypercholesterolemic)										
6.0	6.4	7.0	5.8	6.0	5.8	5.9	6.7	6.1	6.5	6.3
5.8										
Group 2 (controls)										
6.4	5.4	5.6	5.0	4.0	4.5	6.0				

Source: Data for Group 1 are from Rassias, G., Kestin, M., & Nestel, P. J. (1991). Linoleic acid lowers LDL cholesterol without a proportionate displacement of saturated fatty acid. *European Journal of Clinical Nutrition, 45*(6), 315–320. Data for Group 2 were generated with a Normal random variable number generator. Data for both groups are stored online in the file CHOLESTEROL.SAV.

3.2 Frequency Tables

Frequency Counts from Stemplots

Frequency means the number of times something occurs. **Figure 3.5** is a stemplot with frequency counts shown to the left of the stem. **Figure 3.6** displays a similar technique for a larger data set. Below the second plot, it states: Each leaf: 2 case(s):[d] This plot uses 327 leaves to represent the 654 observations. Having counts next to the stem provides a convenient frequency listing.

[d]SPSS uses the term "cases" to refer to "observations."

Frequency	Stem and Leaf	
1.00	0 .	5
1.00	1 .	1
4.00	2 .	1478
1.00	3 .	0
1.00	4 .	2
2.00	5 .	02

Stem width: 10

Each leaf: 1 case(s)

FIGURE 3.5 Frequency counts from stemplot, SPSS output.

Frequency	Stem and Leaf	
2.00	3 .	0
9.00	4 .	0000
28.00	5 .	0000000000000000
37.00	6 .	00000000000000000000
54.00	7 .	00000000000000000000000000000
85.00	8 .	00000000000000000000000000000000000000000000
94.00	9 .	0000000000000000000000000000000000000000000000000
81.00	10 .	000000000000000000000000000000000000000000
90.00	11 .	00000000000000000000000000000000000000000000000
57.00	12 .	000000000000000000000000000
43.00	13 .	00000000000000000000
25.00	14 .	000000000000
19.00	15 .	000000000
13.00	16.	000000
8.00	17.	0000
9.00	Extremes	(>=18)

Stem width: 1

Each leaf: 2 case(s)

FIGURE 3.6 Ages of participants in a childhood and adolescent health survey. Plot produced with SPSS for Windows, Rel. 11.0.1. 2001. Chicago: SPSS Inc.

Frequency Tables

Frequency tables are a traditional way to describe the distribution of counts in a dataset. **Table 3.8** shows a typical frequency table. Three types of frequencies are listed.

- The **frequency** column contains counts.
- The **relative frequency (%)** column contains frequency counts divided by the total with values expressed as a percentage[e], for example:

  2 ÷ 654 = .003, or 0.3% are 3 years of age
  9 ÷ 654 = .014, or 1.4% are 4 years of age
  28 ÷ 654 = .043, or 4.3% are 5 years of age
  And so on.

Relative frequencies are important because their values do not depend on the size of the data set.

Table 3.8 Frequency table, ages of subjects in a respiratory health study.

Age (years)	Frequency	Relative frequency (%)	Cumulative frequency (%)
3	2	0.3%	0.3%
4	9	1.4	1.7
5	28	4.3	6.0
6	37	5.7	11.6
7	54	8.3	19.9
8	85	13.0	32.9
9	94	14.4	47.2
10	81	12.4	59.6
11	90	13.8	73.4
12	57	8.7	82.1
13	43	6.6	88.7
14	25	3.8	92.5
15	19	2.9	95.4
16	13	2.0	97.4
17	8	1.2	98.6
18	6	.9	99.5
19	3	.5	100.0%
Total	654	100.0%	—

[e]"Per-cent" literally means "per hundred." To turn a proportion into a percentage, simply multiply by 100%. This does *not* change the value, since multiplying 100% is the same as multiplying by 1.

- The **cumulative frequency (%)** column contains percents that falls within or below a given level. To determine cumulative percents, add the current relative frequency to the prior cumulative relative frequency. Start with the fact that 0.3% of individuals are 3 years old. Then determine

  0.3% + 1.4% = 1.7% are 4 years of age or younger
  1.7% + 4.3% = 6.0% are 5 years of age or younger
  And so on.

> **Frequency** ≡ count
> **Relative frequency** ≡ proportion
> **Cumulative relative frequency** ≡ proportion that falls in or below a certain level

Class-Interval Frequency Tables (Grouped Data)

It is often necessary to group data into **class intervals** before tallying frequencies. This process is analogous to deciding on the number of stem values to use when creating a stemplot. Again, begin by grouping data into 3 to 12 class intervals. Then, by trial and error, find a grouping that best suits your needs.

Class intervals can be set up with equal spacing or unequal spacing; either way, **endpoint conventions** are needed to ensure that each observation falls into only one interval. Class intervals may either: (a) include the left boundary and exclude the right boundary; or (b) include the right boundary and exclude the left boundary. Here are 10-year age intervals that include the left boundary and exclude the right boundary:

$0.0 \le \text{years} < 10$
$10.0 \le \text{years} < 20$
.
.
.
$50.0 \le \text{years} < 60$

To construct a frequency table with class intervals:

1. List class intervals in ascending or descending order.
2. Tally frequencies that fall within each interval.
3. Sum frequencies to determine the total sample size: $n = \sum f_i$, where f_i represents the frequency in class *i*.

4. Determine relative frequencies (proportions) that fall within each interval: $p_i = f_i \div n$, where p_i represents the proportion in class interval i. (It is often useful to report proportions as percents by moving the decimal point over two places to the right and adding a "%" sign.)
5. Determine the cumulative proportions by summing proportions from prior levels to the current level: $c_i = p_i + c_{i-1}$, where c_i represents the cumulative relative frequency in class i.

Here is a listing of the 10 data points that began the chapter:

5 11 21 24 27 28 30 42 50 52

Here are the data as a frequency table with uniform 15-year class intervals:

Class interval (i)	Age range (years)	Talley	Frequency	Age range frequency (%)	Cum. relative frequency (%)
1	0 – 14	//	2	20%	20%
2	15 – 29	////	4	40	60
3	30 – 44	//	2	20	80
4	45 – 59	//	2	20	100%
Totals	All	—	10	100%	—

Some situations call for **nonuniform class intervals**. Here are the data from Figure 3.6 with data grouped in classes corresponding to school-age populations.

Age range	Frequency	Relative frequency	Cum. relative frequency
Pre-school (3 – 4 years)	11	1.7%	1.7%
Elementary (5 – 11 years)	469	71.7	74.4
Junior High (12 – 13 years)	100	15.3	88.7
Senior High (14 – 19 years)	74	11.3	100.0%
Total	654	100.0%	—

3.3 Additional Frequency Charts

Stemplots display frequencies with leaf counts. **Bar charts** display frequencies with bars that correspond in height to frequencies or relative frequencies. **Histograms** are bar charts with contiguous bars. **Figure 3.7** is a histogram of the data from Figure 3.6.

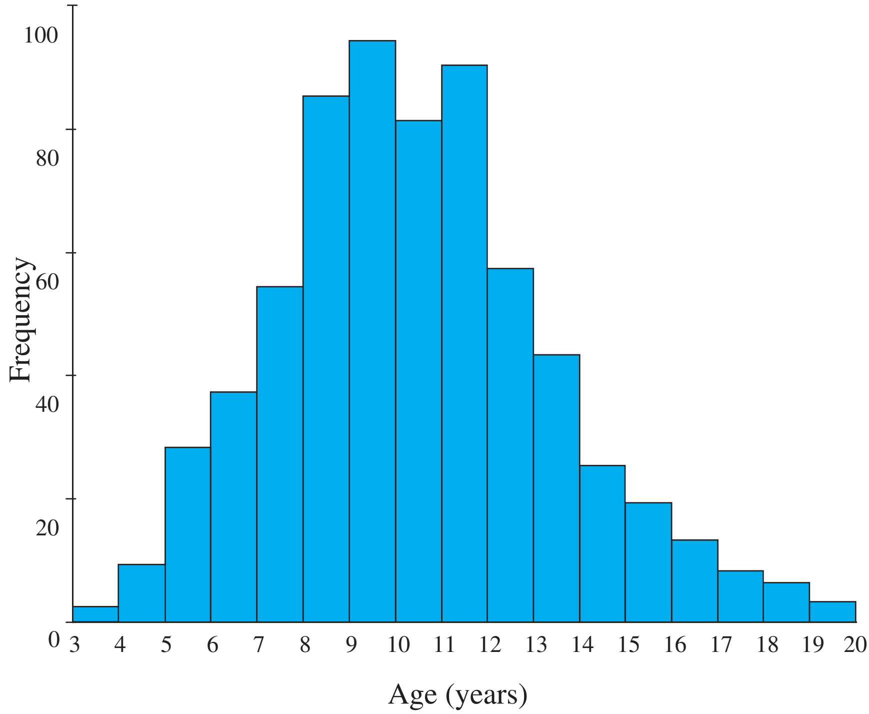

FIGURE 3.7 Histogram bars touch. They are best suited for continuous quantitative data.

Frequency polygons replace histogram bars with a line connecting frequency levels. **Figure 3.8** is a frequency polygon of the same data plotted in **Figure 3.7**.

Histograms and frequency polygons are reserved for use with quantitative variables. Frequencies for categorical variables should be displayed via **bar charts** with noncontiguous bars (e.g., **Figure 3.9**) or **pie charts** (e.g., **Figure 3.10**).

For a remarkably interesting book on good graphical practices in statistics, see Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphic Press.

Exercise

3.5 ***Hospital stay duration.*** In Exercise 3.2, you created a stemplot of lengths of hospital stays for 25 patients. Table 3.6 lists the data.

(a) Construct a frequency table for these data using 5-day class intervals. Include columns for the frequency counts, relative frequencies, and cumulative frequencies.

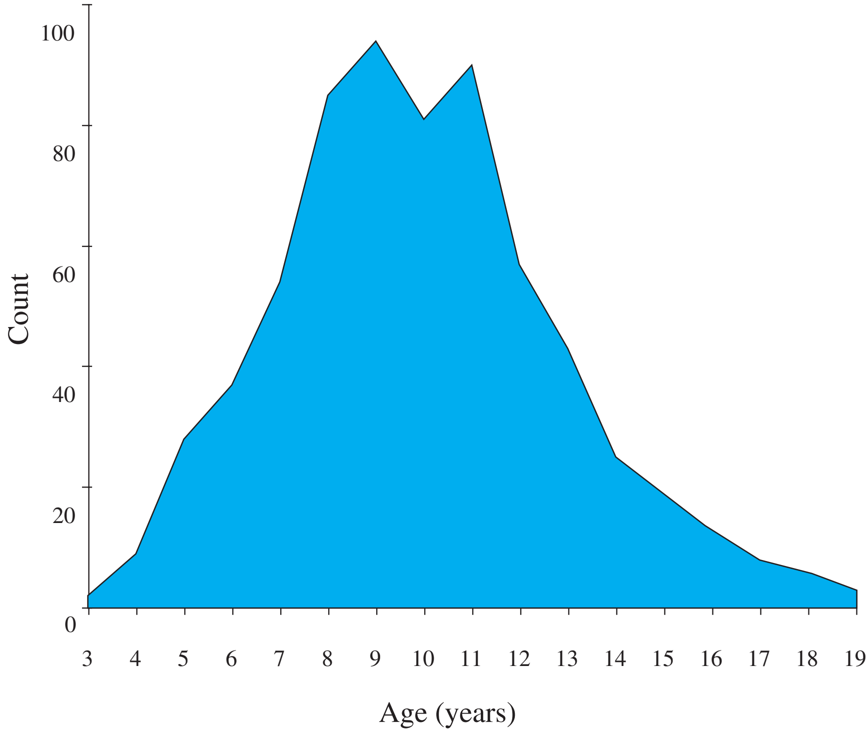

FIGURE 3.8 Frequency polygon.

(b) What percentage of hospital stays were less than 5 days?
(c) What percentage were less than 15 days?
(d) What percentage of hospital stays were at least 15 days in length?

Vocabulary

Arithmetic average
Average
Axis multiplier
Back-to-back stemplot
Bar charts
Bimodal
Class intervals
Cumulative frequency
Depth
Endpoint conventions
Frequency
Frequency polygons
Histograms
Kurtosis
Leaf
Leptokurtic
Location
Median
Modality
Negative skew

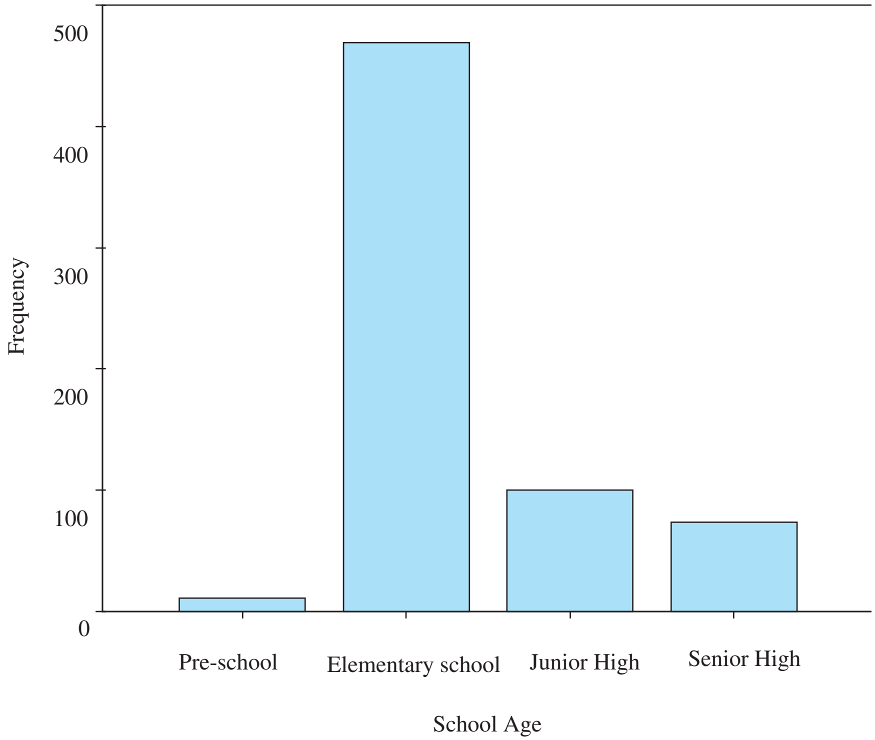

FIGURE 3.9 Bar charts with noncontinuous bars are better suited for nonuniform class intervals and categorical data.

- Nonuniform class intervals
- Ordered array
- Outlier
- Pie charts
- Platykurtic
- Positive skew
- Relative frequency
- Shape
- Split stem values
- Spread
- Stem-and-leaf plot (Stemplot)
- Symmetry
- Truncate

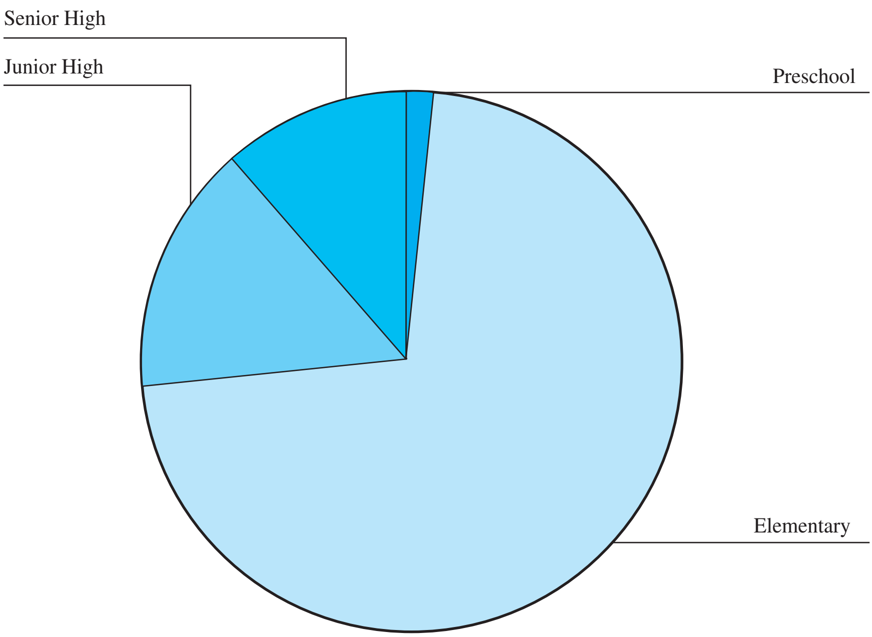

FIGURE 3.10 Pie chart are also well suited for displaying frequencies.

Exercises

3.6 ***Outpatient wait time.*** Waiting times (minutes) for 25 patients at a public health clinic are[f]:

35	22	63	6	49	19	16	31	24	29
23	32	72	13	51	45	77	16	33	55
10	42	28	72	13					

(a) Create a stemplot of these data. Describe the distribution's shape, location, spread.

(b) From your stemplot, create a frequency table with counts, relative frequencies, and cumulative relative frequencies. What percentage of wait times were less than 20 minutes? What percentage were at least 20 minutes?

3.7 ***Body weight expressed as a percentage of ideal.*** Body weights of 18 diabetics expressed as a percentage of ideal (defined as body weight ÷

[f]Data fictitious but realistic. Stored online in the file WAITTIME.*.

ideal body weight × 100) are shown here.[g] Construct a stem-and-leaf plot of these data and interpret your findings.

107	119	99	114	120	104	88	114	124	116
101	121	152	100	125	114	95	117		

3.8 ***Docs' kids.*** The numbers of children of 24 physicians who work at a particular clinic are shown here.[h] Create a stemplot with these data. Consider its shape, location, and spread. What percentage of physicians at this clinic have less than three children?

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

3.9 ***Seizures following bacterial meningitis.*** A study examined the induction time between bacterial meningitis and the onset of seizures in 13 cases (months). Data are shown here.[i] Construct a stemplot of these data and describe what you see. [*Suggestion*: Use a stem multiplier of ×10 so that the value 0.10 is truncated to 00 and is plotted as "0|0".]

0.10	0.25	0.5	4	12	12	24	24	31	36
42	55	96							

3.10 ***Surgical times.*** Durations of surgeries (hours) for 15 patients receiving artificial hearts are shown here.[j] Create a stem plot of these data. Describe the distribution. Are there any outliers?

7.0	6.5	3.5	3.1	2.8	2.5	2.6	2.4	2.1
1.8	2.3	3.1	3.0	2.5				

3.11 ***U.S. Hispanic population.*** **Table 3.9** lists the percent of residents in the 50 states who identified themselves in the 2000 census as Spanish, Hispanic, or Latino. Create a stemplot of these data using single stem

[g]Saudek, C. D., Selam, J. L., Pitt, H. A., Waxman, K., Rubio, M., Jeandidier, N., et al. (1989). A preliminary trial of the programmable implantable medication system for insulin delivery. *New England Journal of Medicine*, *321*(9), 574–579. Data are stored online in the file %IDEAL.*.

[h]Data are fictitious and are stored online in the file DOCKIDS.*.

[i]Source: Unknown. Data stored online in the file SEIZURE.*.

[j]Source: Unknown. Data stored online in the file SURG-TIME.*.

Table 3.9 Data for Exercise 3.11. Percent of residents who identified themselves as Spanish, Hispanic, or Latino by state, U.S., 2000.

State	Percent	State	Percent	State	Percent
Alabama	1.5	Louisiana	2.4	Ohio	1.9
Alaska	4.1	Maine	.7	Oklahoma	5.2
Arizona	25.3	Maryland	4.3	Oregon	8.0
Arkansas	2.8	Massachusetts	6.8	Pennsylvania	3.2
California	32.4	Michigan	3.3	Rhode Island	8.7
Colorado	17.1	Minnesota	2.9	South Carolina	2.4
Connecticut	9.4	Mississippi	1.3	South Dakota	1.4
Delaware	4.8	Missouri	2.1	Tennessee	2.0
Florida	16.8	Montana	2.0	Texas	32.0
Georgia	5.3	Nebraska	5.5	Utah	9.0
Hawaii	7.2	Nevada	19.7	Vermont	0.9
Idaho	7.9	New Hampshire	1.7	Virginia	4.7
Illinois	10.7	New Jersey	13.3	Washington	7.2
Indiana	3.5	New Mexico	42.1	West Virginia	0.7
Iowa	2.8	New York	15.1	Wisconsin	3.6
Kansas	7.0	North Carolina	4.7	Wyoming	6.4
Kentucky	1.5	North Dakota	1.2		

Source: www.census.gov. Data stored online in the file PER-HISP.*.

values and an axis multiplier of ×10. Then create a stemplot using double-split stem values. Which plot do you prefer?

3.12 ***Low-birth weight rates.*** A birth weight of less than 2500 grams (about 5.5 pounds) qualifies as "low-birth weight" according to international standards. **Table 3.10** lists low-birth weight rates (per 100 births) by country for the year 1991 in 109 countries. Explore these data with a stemplot. Where does the United States rank in this listing?

3.13 ***Air samples.*** An environmental study looked at suspended particulate matter in air samples (μgms/m^3) at two different sites. Data are listed here.[k] Construct side-by-side stemplots to compare the two sites.

Site 1:	68	22	36	32	42	24	28	38
Site 2:	36	38	39	40	36	34	33	32

[k]Source: Unknown. Data stored online in the file AIRSAMPLES.*.

Table 3.10 Data for Exercise 3.12. Low-birth weights per 100 births for 109 countries, 1991.

Spain	1	UK	7	Cote d'Ivoire	14
Finland	4	USA	7	Guatemala	14
Ireland	4	Yugoslavia	7	Indonesia	14
Norway	4	Benin	8	Tanzania	14
Sweden	4	Botswana	8	Zambia	14
Belgium	5	Brazil	8	Central African Rep.	15
Egypt	5	Colombia	8	El Salvador	15
France	5	Cuba	8	Kenya	15
Hong Kong	5	Jamaica	8	Mexico	15
Iran	5	Panama	8	Nicaragua	15
Japan	5	Poland	8	Niger	15
Jordan	5	Tunisia	8	Zimbabwe	15
New Zealand	5	Uruguay	8	Congo	16
Portugal	5	Algeria	9	Dominican Rep.	16
Switzerland	5	Burundi	9	Myanmar	16
Australia	6	China	9	Angola	17
Austria	6	Iraq	9	Ghana	17
Bulgaria	6	Korea, Rep.	9	Haiti	17
Canada	6	Mauritius	9	Mali	17
Czechoslovakia	6	Peru	9	Rwanda	17
Denmark	6	Venezuela	9	Sierra Leone	17
Germany, Fed.	6	Costa Rica	10	Philippines	18
Germany, Rep.	6	Hungary	10	Viet Nam	18
Greece	6	Lebanon	10	Afghanistan	20
Romania	6	Madagascar	10	Honduras	20
Saudi Arabia	6	Malaysia	10	Malawi	20
USSR	6	Mongolia	10	Mozambique	20
Albania	7	Ecuador	11	Nigeria	20
Chile	7	Lesotho	11	Togo	20
Israel	7	Mauritania	11	Pakistan	25
Italy	7	Senegal	11	Papua New Guinea	25
Kuwait	7	Syria	11	Bangladesh	28
Oman	7	Bolivia	12	Sri Lanka	28
Paraguay	7	South Africa	12	India	30
Singapore	7	Thailand	12	Laos	39
Turkey	7	Cameroon	13		
United Arab Emirates	7	Zaire	13		

Source: Grant, J. P. (1992). *The State of the World's Children; Vol. 1991 (United Nations Children's Fund)*. New York: Oxford. Data stored online in the file UNICEF.*.