CHAPTER

2

Frequency Distributions

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Proportions (math review, Appendix A)
  - Fractions
  - Decimals
  - Percentages
- Scales of measurement (Chapter 1): Nominal, ordinal, interval, and ratio
- Continuous and discrete variables (Chapter 1)
- Real limits (Chapter 1)

Preview

If at first you don't succeed, you are probably not related to the boss.

Did we make you chuckle or, at least, smile a little? The use of humor is a common technique to capture attention and to communicate ideas. Advertisers, for example, often try to make a commercial funny so that people notice it and, perhaps, remember the product. After-dinner speakers always put a few jokes into the speech in an effort to maintain the audience's interest. Although humor seems to capture our attention, does it actually affect our memory?

In an attempt to answer this question, Stephen Schmidt (1994) conducted a series of experiments examining the effects of humor on memory for sentences. Humorous sentences were collected from a variety of sources and then a nonhumorous version was constructed for each sentence. For example, the nonhumorous version of our opening sentence was:

People who are related to the boss often succeed the very first time.

Participants were then presented with a list containing half humorous and half nonhumorous sentences. Later, each person was asked to recall as many sentences as possible. The researcher measured the number of humorous sentences and the number of nonhumorous sentences recalled by each participant. Data similar to the results obtained by Schmidt are shown in Table 2.1.

TABLE 2.1

Memory scores for a sample of 16 participants. The scores represent the number of sentences recalled from each category.

Humorous Sentences				Nonhumorous Sentences			
4	5	2	4	5	2	4	2
6	7	6	6	2	3	1	6
2	5	4	3	3	2	3	3
1	3	5	5	4	1	5	3

The Problem: It is difficult to see any clear pattern simply by looking at the list of numbers. Can you tell whether the memory scores for one type of sentence are generally higher than those for the other type?

The Solution: A frequency distribution provides an overview of the entire group of scores making it easy to see the general level of performance for each type of sentence. For example, the same memory scores that are shown in Table 2.1 have been organized in a frequency distribution graph in Figure 2.1. In the figure, each individual is represented by a block that is placed above that individual's score. The resulting pile of blocks shows a picture of how the individual scores are distributed. For this example, it is now easy to see that the scores for the humorous sentences are generally higher than the scores for the nonhumorous sentences; on average, participants recalled around 5 humorous sentences but only about 3 of the nonhumorous sentences.

In this chapter we present techniques for organizing data into tables and graphs so that an entire set of scores can be presented in a relatively simple display or illustration.

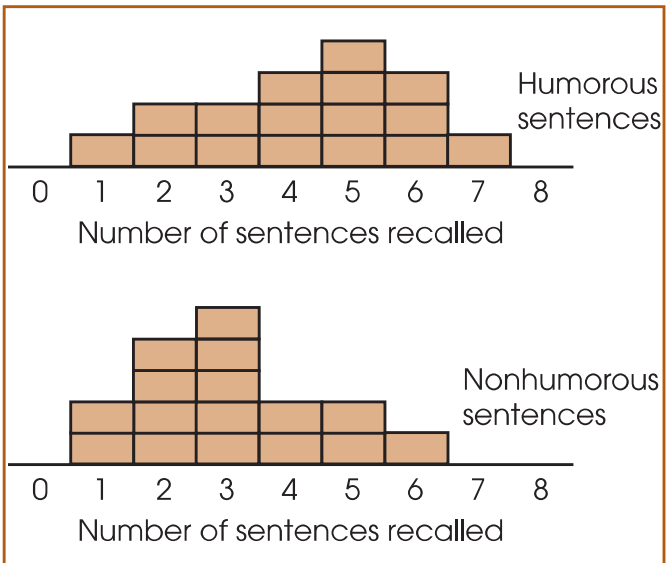


FIGURE 2.1

Hypothetical data showing the number of humorous sentences and the number of nonhumorous sentences recalled by participants in a memory experiment.

2.1 INTRODUCTION TO FREQUENCY DISTRIBUTIONS

The results from a research study usually consist of pages of numbers corresponding to the measurements, or scores, collected during the study. The immediate problem for the researcher is to organize the scores into some comprehensible form so that any patterns in the data can be seen easily and communicated to others. This is the job of descriptive statistics: to simplify the organization and presentation of data. One of the most common procedures for organizing a set of data is to place the scores in a *frequency distribution.*

DEFINITION

A **frequency distribution** is an organized tabulation of the number of individuals located in each category on the scale of measurement.

A frequency distribution takes a disorganized set of scores and places them in order from highest to lowest, grouping together individuals who all have the same score. If the highest score is $X = 10$, for example, the frequency distribution groups together all the 10s, then all the 9s, then the 8s, and so on. Thus, a frequency distribution allows the researcher to see "at a glance" the entire set of scores. It shows whether the scores are generally high or low, whether they are concentrated in one area or spread out across the entire scale, and generally provides an organized picture of the data. In addition to providing a picture of the entire set of scores, a frequency distribution allows you to see the location of any individual score relative to all of the other scores in the set.

A frequency distribution can be structured either as a table or as a graph, but in either case, the distribution presents the same two elements:

1. The set of categories that make up the original measurement scale.
2. A record of the frequency, or number of individuals in each category.

Thus, a frequency distribution presents a picture of how the individual scores are distributed on the measurement scale—hence the name *frequency distribution*.

2.2 FREQUENCY DISTRIBUTION TABLES

It is customary to list categories from highest to lowest, but this is an arbitrary arrangement. Many computer programs list categories from lowest to highest.

The simplest frequency distribution table presents the measurement scale by listing the different measurement categories (X values) in a column from highest to lowest. Beside each X value, we indicate the frequency, or the number of times that particular measurement occurred in the data. It is customary to use an X as the column heading for the scores and an f as the column heading for the frequencies. An example of a frequency distribution table follows.

EXAMPLE 2.1

The following set of $N = 20$ scores was obtained from a 10-point statistics quiz. We organize these scores by constructing a frequency distribution table. Scores:

8, 9, 8, 7, 10, 9, 6, 4, 9, 8,
7, 8, 10, 9, 8, 6, 9, 7, 8, 8

1. The highest score is $X = 10$, and the lowest score is $X = 4$. Therefore, the first column of the table lists the categories that make up the scale of measurement

X	f
10	2
9	5
8	7
7	3
6	2
5	0
4	1

(X values) from 10 down to 4. Notice that all of the possible values are listed in the table. For example, no one had a score of $X = 5$, but this value is included. With an ordinal, interval, or ratio scale, the categories are listed in order (usually highest to lowest). For a nominal scale, the categories can be listed in any order.

2. The frequency associated with each score is recorded in the second column. For example, two people had scores of $X = 10$, so there is a 2 in the f column beside $X = 10$.

Because the table organizes the scores, it is possible to see the general quiz results very quickly. For example, there were only two perfect scores, but most of the class had high grades (8s and 9s). With one exception (the score of $X = 4$), it appears that the class has learned the material fairly well.

Notice that the X values in a frequency distribution table represent the scale of measurement, *not* the actual set of scores. For example, the X column lists the value 10 only one time, but the frequency column indicates that there are actually two values of $X = 10$. Also, the X column lists a value of $X = 5$, but the frequency column indicates that no one actually had a score of $X = 5$.

You also should notice that the frequencies can be used to find the total number of scores in the distribution. By adding up the frequencies, you obtain the total number of individuals:

$$\Sigma f = N$$

OBTAINING ΣX FROM A FREQUENCY DISTRIBUTION TABLE

There may be times when you need to compute the sum of the scores, ΣX, or perform other computations for a set of scores that has been organized into a frequency distribution table. To complete these calculations correctly, you must use all the information presented in the table. That is, it is essential to use the information in the f column as well as the X column to obtain the full set of scores.

When it is necessary to perform calculations for scores that have been organized into a frequency distribution table, the safest procedure is to take the individual scores out of the table before you begin any computations. This process is demonstrated in the following example.

EXAMPLE 2.2

X	f
5	1
4	2
3	3
2	3
1	1

Consider the frequency distribution table shown in the margin. The table shows that the distribution has one 5, two 4s, three 3s, three 2s, and one 1, for a total of 10 scores. If you simply list all 10 scores, you can safely proceed with calculations such as finding ΣX or ΣX^2. For example, to compute ΣX you must add all 10 scores:

$$\Sigma X = 5 + 4 + 4 + 3 + 3 + 3 + 2 + 2 + 2 + 1$$

For the distribution in this table, you should obtain $\Sigma X = 29$. Try it yourself. Similarly, to compute ΣX^2 you square each of the 10 scores and then add the squared values.

$$\Sigma X^2 = 5^2 + 4^2 + 4^2 + 3^2 + 3^2 + 3^2 + 2^2 + 2^2 + 2^2 + 1^2$$

This time you should obtain $\Sigma X^2 = 97$.

An alternative way to get ΣX from a frequency distribution table is to multiply each X value by its frequency and then add these products. This sum may be

expressed in symbols as ΣfX. The computation is summarized as follows for the data in Example 2.2:

Caution: Doing calculations within the table works well for ΣX but can lead to errors for more complex formulas.

X	f	fX	
5	1	5	(the one 5 totals 5)
4	2	8	(the two 4s total 8)
3	3	9	(the three 3s total 9)
2	3	6	(the three 2s total 6)
1	1	1	(the one 1 totals 1)
	$\Sigma X = 29$		

No matter which method you use to find ΣX, the important point is that you must use the information given in the frequency column in addition to the information in the X column.

PROPORTIONS AND PERCENTAGES

In addition to the two basic columns of a frequency distribution, there are other measures that describe the distribution of scores and can be incorporated into the table. The two most common are proportion and percentage.

Proportion measures the fraction of the total group that is associated with each score. In Example 2.2, there were two individuals with $X = 4$. Thus, 2 out of 10 people had $X = 4$, so the proportion would be $\frac{2}{10} = 0.20$. In general, the proportion associated with each score is

$$\text{proportion} = p = \frac{f}{N}$$

Because proportions describe the frequency (f) in relation to the total number (N), they often are called *relative frequencies*. Although proportions can be expressed as fractions (for example, $\frac{2}{10}$), they more commonly appear as decimals. A column of proportions, headed with a p, can be added to the basic frequency distribution table (see Example 2.3).

In addition to using frequencies (f) and proportions (p), researchers often describe a distribution of scores with percentages. For example, an instructor might describe the results of an exam by saying that 15% of the class earned *A*s, 23% earned *B*s, and so on. To compute the percentage associated with each score, you first find the proportion (p) and then multiply by 100:

$$\text{percentage} = p(100) = \frac{f}{N}(100)$$

Percentages can be included in a frequency distribution table by adding a column headed with % (see Example 2.3).

EXAMPLE 2.3

The frequency distribution table from Example 2.2 is repeated here. This time we have added columns showing the proportion (p) and the percentage (%) associated with each score.

X	f	$p = f/N$	% = $p(100)$
5	1	1/10 = 0.10	10%
4	2	2/10 = 0.20	20%
3	3	3/10 = 0.30	30%
2	3	3/10 = 0.30	30%
1	1	1/10 = 0.10	10%

LEARNING CHECK

1. Construct a frequency distribution table for the following set of scores.

Scores: 3, 2, 3, 2, 4, 1, 3, 3, 5

2. Find each of the following values for the sample in the following frequency distribution table.

a. n
b. ΣX
c. ΣX^2

X	f
5	1
4	2
3	2
2	4
1	1

ANSWERS

1.

X	f
5	1
4	1
3	4
2	2
1	1

2. a. $n = 10$ **b.** $\Sigma X = 28$ **c.** $\Sigma X^2 = 92$ (square then add all 10 scores)

GROUPED FREQUENCY DISTRIBUTION TABLES

When a set of data covers a wide range of values, it is unreasonable to list all the individual scores in a frequency distribution table. Consider, for example, a set of exam scores that range from a low of $X = 41$ to a high of $X = 96$. These scores cover a *range* of more than 50 points.

If we were to list all of the individual scores from $X = 96$ down to $X = 41$, it would take 56 rows to complete the frequency distribution table. Although this would organize the data, the table would be long and cumbersome. Remember: The purpose for constructing a table is to obtain a relatively simple, organized picture of the data. This can be accomplished by grouping the scores into intervals and then listing the intervals in the table instead of listing each individual score. For example, we could construct a table showing the number of students who had scores in the 90s, the number with scores in the 80s, and so on. The result is called a *grouped frequency distribution table* because we are presenting groups of scores rather than individual values. The groups, or intervals, are called *class intervals*.

When the scores are whole numbers, the total number of rows for a regular table can be obtained by finding the difference between the highest and the lowest scores and adding 1:

rows = highest – lowest + 1

There are several guidelines that help guide you in the construction of a grouped frequency distribution table. Note that these are simply guidelines, rather than absolute requirements, but they do help produce a simple, well-organized, and easily understood table.

GUIDELINE 1

The grouped frequency distribution table should have about 10 class intervals. If a table has many more than 10 intervals, it becomes cumbersome and defeats the purpose of a frequency distribution table. On the other hand, if you have too few intervals, you begin to lose information about the distribution of the scores. At the extreme, with only one interval, the table would not tell you anything about how the scores are distributed. Remember that the purpose of a frequency distribution is to help a researcher see the data. With too few or too many intervals, the table will not provide a clear picture. You

should note that 10 intervals is a general guide. If you are constructing a table on a blackboard, for example, you probably want only 5 or 6 intervals. If the table is to be printed in a scientific report, you may want 12 or 15 intervals. In each case, your goal is to present a table that is relatively easy to see and understand.

GUIDELINE 2 The width of each interval should be a relatively simple number. For example, 2, 5, 10, or 20 would be a good choice for the interval width. Notice that it is easy to count by 5s or 10s. These numbers are easy to understand and make it possible for someone to see quickly how you have divided the range of scores.

GUIDELINE 3 The bottom score in each class interval should be a multiple of the width. If you are using a width of 10 points, for example, the intervals should start with 10, 20, 30, 40, and so on. Again, this makes it easier for someone to understand how the table has been constructed.

GUIDELINE 4 All intervals should be the same width. They should cover the range of scores completely with no gaps and no overlaps, so that any particular score belongs in exactly one interval.

The application of these rules is demonstrated in Example 2.4.

EXAMPLE 2.4

An instructor has obtained the set of $N = 25$ exam scores shown here. To help organize these scores, we will place them in a frequency distribution table. The scores are:

82, 75, 88, 93, 53, 84, 87, 58, 72, 94, 69, 84, 61,
91, 64, 87, 84, 70, 76, 89, 75, 80, 73, 78, 60

Remember, when the scores are whole numbers, the number of rows is determined by

highest – lowest + 1

The first step is to determine the range of scores. For these data, the smallest score is $X = 53$ and the largest score is $X = 94$, so a total of 42 rows would be needed for a table that lists each individual score. Because 42 rows would not provide a simple table, we have to group the scores into class intervals.

The best method for finding a good interval width is a systematic trial-and-error approach that uses guidelines 1 and 2 simultaneously. Specifically, we want about 10 intervals and we want the interval width to be a simple number. For this example, the scores cover a range of 42 points, so we will try several different interval widths to see how many intervals are needed to cover this range. For example, if each interval is 2 points wide, it would take 21 intervals to cover a range of 42 points. This is too many, so we move on to an interval width of 5 or 10 points. The following table shows how many intervals would be needed for these possible widths:

Because the bottom interval usually extends below the lowest score and the top interval extends beyond the highest score, you often need slightly more than the computed number of intervals.

Width	Number of Intervals Needed to Cover a Range of 42 Points	
2	21	(too many)
5	9	(OK)
10	5	(too few)

Notice that an interval width of 5 will result in about 10 intervals, which is exactly what we want.

The next step is to actually identify the intervals. The lowest score for these data is $X = 53$, so the lowest interval should contain this value. Because the interval should have a multiple of 5 as its bottom score, the interval should begin at 50. The

interval has a width of 5, so it should contain 5 values: 50, 51, 52, 53, and 54. Thus, the bottom interval is 50–54. The next interval would start at 55 and go to 59. Note that this interval also has a bottom score that is a multiple of 5, and contains exactly 5 scores (55, 56, 57, 58, and 59). The complete frequency distribution table showing all of the class intervals is presented in Table 2.2.

Once the class intervals are listed, you complete the table by adding a column of frequencies. The values in the frequency column indicate the number of individuals who have scores located in that class interval. For this example, there were three students with scores in the 60–64 interval, so the frequency for this class interval is $f = 3$ (see Table 2.2). The basic table can be extended by adding columns showing the proportion and percentage associated with each class interval.

Finally, you should note that after the scores have been placed in a grouped table, you lose information about the specific value for any individual score. For example, Table 2.2 shows that one person had a score between 65 and 69, but the table does not identify the exact value for the score. In general, the wider the class intervals are, the more information is lost. In Table 2.2 the interval width is 5 points, and the table shows that there are three people with scores in the lower 60s and one person with a score in the upper 60s. This information would be lost if the interval width were increased to 10 points. With an interval width of 10, all of the 60s would be grouped together into one interval labeled 60–69. The table would show a frequency of four people in the 60–69 interval, but it would not tell whether the scores were in the upper 60s or the lower 60s.

REAL LIMITS AND FREQUENCY DISTRIBUTIONS

Recall from Chapter 1 that a continuous variable has an infinite number of possible values and can be represented by a number line that is continuous and contains an infinite number of points. However, when a continuous variable is measured, the resulting measurements correspond to *intervals* on the number line rather than single points. If you are measuring time in seconds, for example, a score of $X = 8$ seconds actually represents an interval bounded by the real limits 7.5 seconds and 8.5 seconds. Thus, a frequency distribution table showing a frequency of $f = 3$ individuals all assigned a score of $X = 8$ does not mean that all three individuals had exactly the same measurement. Instead, you should realize that the three measurements are simply located in the same interval between 7.5 and 8.5.

The concept of real limits also applies to the class intervals of a grouped frequency distribution table. For example, a class interval of 40–49 contains scores from $X = 40$ to $X = 49$. These values are called the *apparent limits* of the interval because it appears

TABLE 2.2

This grouped frequency distribution table shows the data from Example 2.4. The original scores range from a high of $X = 94$ to a low of $X = 53$. This range has been divided into 9 intervals with each interval exactly 5 points wide. The frequency column (f) lists the number of individuals with scores in each of the class intervals.

X	f
90–94	3
85–89	4
80–84	5
75–79	4
70–74	3
65–69	1
60–64	3
55–59	1
50–54	1

that they form the upper and lower boundaries for the class interval. If you are measuring a continuous variable, however, a score of $X = 40$ is actually an interval from 39.5 to 40.5. Similarly, $X = 49$ is an interval from 48.5 to 49.5. Therefore, the real limits of the interval are 39.5 (the lower real limit) and 49.5 (the upper real limit). Notice that the next higher class interval is 50–59, which has a lower real limit of 49.5. Thus, the two intervals meet at the real limit 49.5, so there are no gaps in the scale. You also should notice that the width of each class interval becomes easier to understand when you consider the real limits of an interval. For example, the interval 50–59 has real limits of 49.5 and 59.5. The distance between these two real limits (10 points) is the width of the interval.

LEARNING CHECK

1. For each of the following situations, determine what interval width is most appropriate for a grouped frequency distribution and identify the apparent limits of the bottom interval.
   - **a.** Scores range from $X = 7$ to $X = 21$.
   - **b.** Scores range from $X = 52$ to $X = 98$.
   - **c.** Scores range from $X = 16$ to $X = 93$.
2. Using only the frequency distribution table presented in Table 2.2, how many individuals had a score of $X = 73$?

ANSWERS

1. **a.** A width of 2 points would require 8 intervals. Bottom interval is 6–7.
   b. A width of 5 points would require 10 intervals. Bottom interval is 50–54.
   c. A width of 10 points would require 9 intervals. Bottom interval is 10–19.
3. After a set of scores has been summarized in a grouped table, you cannot determine the frequency for any specific score. There is no way to determine how many individuals had $X = 73$ from the table alone. (You can say that *at most* three people had $X = 73$.)

2.3 FREQUENCY DISTRIBUTION GRAPHS

A frequency distribution graph is basically a picture of the information available in a frequency distribution table. We consider several different types of graphs, but all start with two perpendicular lines called *axes*. The horizontal line is the *X*-axis, or the abscissa (ab-SIS-uh). The vertical line is the *Y*-axis, or the ordinate. The measurement scale (set of *X* values) is listed along the *X*-axis with values increasing from left to right. The frequencies are listed on the *Y*-axis with values increasing from bottom to top. As a general rule, the point where the two axes intersect should have a value of zero for both the scores and the frequencies. A final general rule is that the graph should be constructed so that its height (*Y*-axis) is approximately two-thirds to three-quarters of its length (*X*-axis). Violating these guidelines can result in graphs that give a misleading picture of the data (see Box 2.1).

GRAPHS FOR INTERVAL OR RATIO DATA

When the data consist of numerical scores that have been measured on an interval or ratio scale, there are two options for constructing a frequency distribution graph. The two types of graphs are called *histograms* and *polygons*.

Histograms To construct a histogram, you first list the numerical scores (the categories of measurement) along the *X*-axis. Then you draw a bar above each *X* value so that

a. The height of the bar corresponds to the frequency for that category.

b. For continuous variables, the width of the bar extends to the real limits of the category. For discrete variables, each bar extends exactly half the distance to the adjacent category on each side.

For both continuous and discrete variables, each bar in a histogram extends to the midpoint between adjacent categories. As a result, adjacent bars touch and there are no spaces or gaps between bars. An example of a histogram is shown in Figure 2.2.

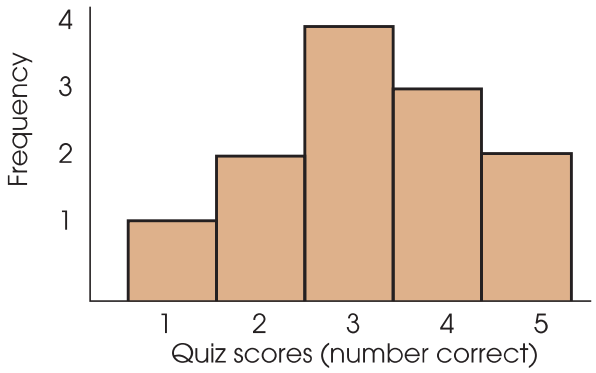
When data have been grouped into class intervals, you can construct a frequency distribution histogram by drawing a bar above each interval so that the width of the bar extends exactly half the distance to the adjacent category on each side. This process is demonstrated in Figure 2.3.

For the two histograms shown in Figures 2.2 and 2.3, notice that the values on both the vertical and horizontal axes are clearly marked and that both axes are labeled. Also note that, whenever possible, the units of measurement are specified; for example, Figure 2.3 shows a distribution of heights measured in inches. Finally, notice that the horizontal axis in Figure 2.3 does not list all of the possible heights starting from zero and going up to 48 inches. Instead, the graph clearly shows a break between zero and 30, indicating that some scores have been omitted.

A modified histogram A slight modification to the traditional histogram produces a very easy to draw and simple to understand sketch of a frequency distribution. Instead

FIGURE 2.2

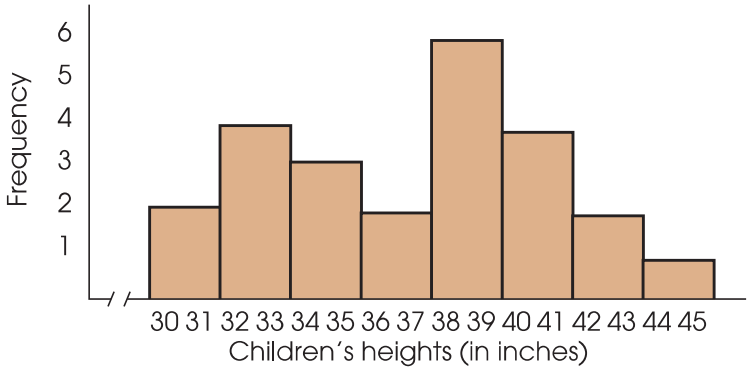
An example of a frequency distribution histogram. The same set of quiz scores is presented in a frequency distribution table and in a histogram.



X	*f*
5	2
4	3
3	4
2	2
1	1

FIGURE 2.3

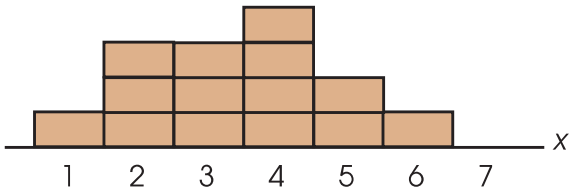
An example of a frequency distribution histogram for grouped data. The same set of children's heights is presented in a frequency distribution table and in a histogram.



X	*f*
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

FIGURE 2.4

A frequency distribution in which each individual is represented by a block placed directly above the individual's score. For example, three people had scores of $X = 2$.



of drawing a bar above each score, the modification consists of drawing a stack of blocks. Each block represents one individual, so the number of blocks above each score corresponds to the frequency for that score. An example is shown in Figure 2.4.

Note that the number of blocks in each stack makes it very easy to see the absolute frequency for each category. In addition, it is easy to see the exact difference in frequency from one category to another. In Figure 2.4, for example, there are exactly two more people with scores of $X = 2$ than with scores of $X = 1$. Because the frequencies are clearly displayed by the number of blocks, this type of display eliminates the need for a vertical line (the *Y*-axis) showing frequencies. In general, this kind of graph provides a simple and concrete picture of the distribution for a sample of scores. Note that we often use this kind of graph to show sample data throughout the rest of the book. You should also note, however, that this kind of display simply provides a sketch of the distribution and is not a substitute for an accurately drawn histogram with two labeled axes.

Polygons The second option for graphing a distribution of numerical scores from an interval or ratio scale of measurement is called a polygon. To construct a polygon, you begin by listing the numerical scores (the categories of measurement) along the *X*-axis. Then,

a. A dot is centered above each score so that the vertical position of the dot corresponds to the frequency for the category.

b. A continuous line is drawn from dot to dot to connect the series of dots.

c. The graph is completed by drawing a line down to the *X*-axis (zero frequency) at each end of the range of scores. The final lines are usually drawn so that they reach the *X*-axis at a point that is one category below the lowest score on the left side and one category above the highest score on the right side. An example of a polygon is shown in Figure 2.5.

A polygon also can be used with data that have been grouped into class intervals. For a grouped distribution, you position each dot directly above the midpoint of the class interval. The midpoint can be found by averaging the highest and the lowest scores in the interval. For example, a class interval that is listed as 20–29 would have a midpoint of 24.5.

$$\text{midpoint} = \frac{20+29}{2} = \frac{49}{2} = 24.5$$

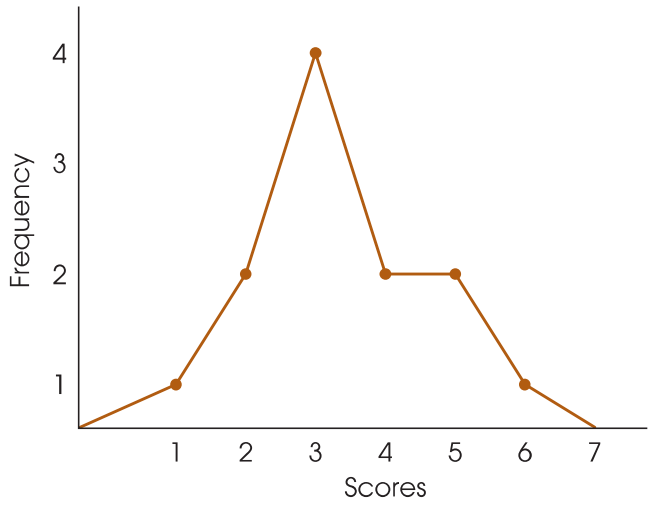
An example of a frequency distribution polygon with grouped data is shown in Figure 2.6.

GRAPHS FOR NOMINAL OR ORDINAL DATA

When the scores are measured on a nominal or ordinal scale (usually non-numerical values), the frequency distribution can be displayed in a *bar graph.*

FIGURE 2.5

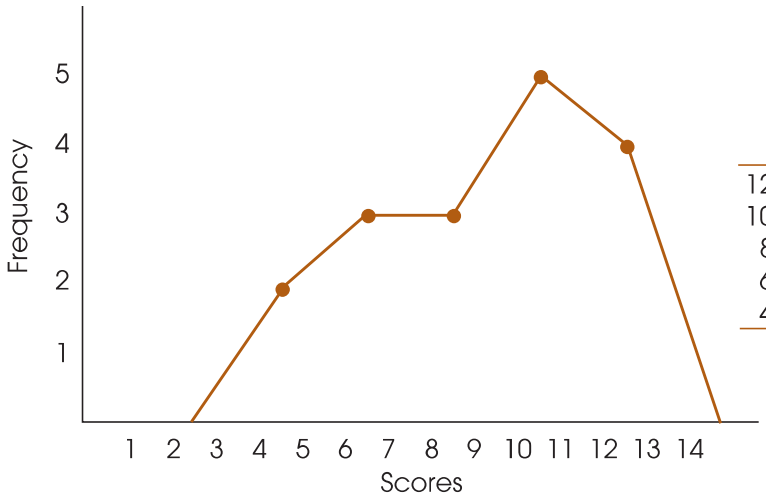
An example of a frequency distribution polygon. The same set of data is presented in a frequency distribution table and in a polygon.



X	f
6	1
5	2
4	2
3	4
2	2
1	1

FIGURE 2.6

An example of a frequency distribution polygon for grouped data. The same set of data is presented in a grouped frequency distribution table and in a polygon.



X	f
12–13	4
10–11	5
8–9	3
6–7	3
4–5	2

Bar graphs A bar graph is essentially the same as a histogram, except that spaces are left between adjacent bars. For a nominal scale, the space between bars emphasizes that the scale consists of separate, distinct categories. For ordinal scales, separate bars are used because you cannot assume that the categories are all the same size.

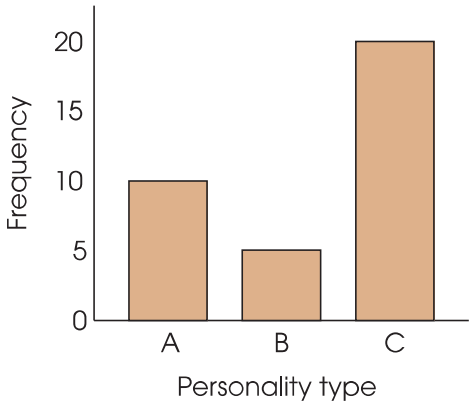
To construct a bar graph, list the categories of measurement along the *X*-axis and then draw a bar above each category so that the height of the bar equals the frequency for the category. An example of a bar graph is shown in Figure 2.7.

GRAPHS FOR POPULATION DISTRIBUTIONS

When you can obtain an exact frequency for each score in a population, you can construct frequency distribution graphs that are exactly the same as the histograms, polygons, and bar graphs that are typically used for samples. For example, if a population is defined as a specific group of $N = 50$ people, we could easily determine how many have IQs of $X = 110$. However, if we are interested in the entire population of adults

FIGURE 2.7

A bar graph showing the distribution of personality types in a sample of college students. Because personality type is a discrete variable measured on a nominal scale, the graph is drawn with space between the bars.



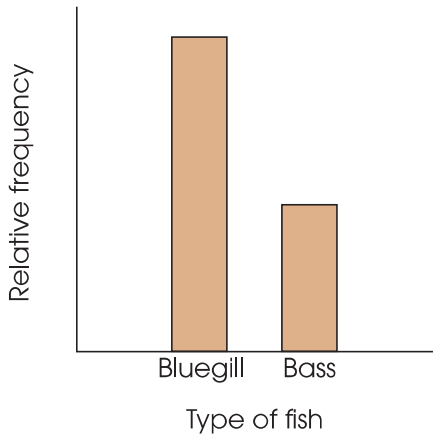
in the United States, it would be impossible to obtain an exact count of the number of people with an IQ of 110. Although it is still possible to construct graphs showing frequency distributions for extremely large populations, the graphs usually involve two special features: relative frequencies and smooth curves.

Relative frequencies Although you usually cannot find the absolute frequency for each score in a population, you very often can obtain *relative frequencies.* For example, you may not know exactly how many fish are in the lake, but after years of fishing you do know that there are twice as many bluegill as there are bass. You can represent these relative frequencies in a bar graph by making the bar above bluegill two times taller than the bar above bass (Figure 2.8). Notice that the graph does not show the absolute number of fish. Instead, it shows the relative number of bluegill and bass.

Smooth curves When a population consists of numerical scores from an interval or a ratio scale, it is customary to draw the distribution with a smooth curve instead of the jagged, step-wise shapes that occur with histograms and polygons. The smooth curve indicates that you are not connecting a series of dots (real frequencies) but instead are showing the relative changes that occur from one score to the next. One commonly occurring population distribution is the normal curve. The word *normal* refers to a specific shape that can be precisely defined by an equation. Less precisely, we can describe

FIGURE 2.8

A frequency distribution showing the relative frequency for two types of fish. Notice that the exact number of fish is not reported; the graph simply says that there are twice as many bluegill as there are bass.



a normal distribution as being symmetrical, with the greatest frequency in the middle and relatively smaller frequencies as you move toward either extreme. A good example of a normal distribution is the population distribution for IQ scores shown in Figure 2.9. Because normal-shaped distributions occur commonly and because this shape is mathematically guaranteed in certain situations, we give it extensive attention throughout this book.

In the future, we will be referring to *distributions of scores*. Whenever the term *distribution* appears, you should conjure up an image of a frequency distribution graph. The graph provides a picture showing exactly where the individual scores are located. To make this concept more concrete, you might find it useful to think of the graph as showing a pile of individuals just like we showed a pile of blocks in Figure 2.4. For the population of IQ scores shown in Figure 2.9, the pile is highest at an IQ score around 100 because most people have average IQs. There are only a few individuals piled up at an IQ of 130; it must be lonely at the top.

2.4 THE SHAPE OF A FREQUENCY DISTRIBUTION

Rather than drawing a complete frequency distribution graph, researchers often simply describe a distribution by listing its characteristics. There are three characteristics that completely describe any distribution: shape, central tendency, and variability. In simple terms, central tendency measures where the center of the distribution is located. Variability tells whether the scores are spread over a wide range or are clustered together. Central tendency and variability will be covered in detail in Chapters 3 and 4. Technically, the shape of a distribution is defined by an equation that prescribes the exact relationship between each X and Y value on the graph. However, we rely on a few less-precise terms that serve to describe the shape of most distributions.

Nearly all distributions can be classified as being either *symmetrical* or *skewed*.

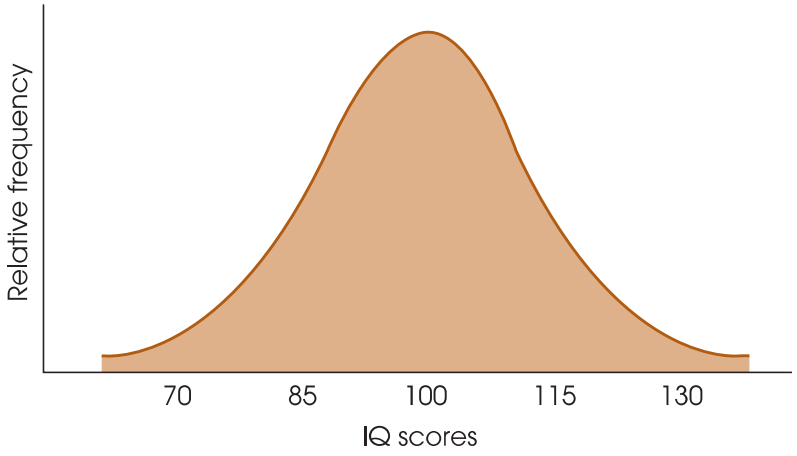
DEFINITIONS

In a **symmetrical distribution**, it is possible to draw a vertical line through the middle so that one side of the distribution is a mirror image of the other (Figure 2.11).

In a **skewed distribution**, the scores tend to pile up toward one end of the scale and taper off gradually at the other end (see Figure 2.11).

FIGURE 2.9

The population distribution of IQ scores: an example of a normal distribution.



BOX 2.1 THE USE AND MISUSE OF GRAPHS

Although graphs are intended to provide an accurate picture of a set of data, they can be used to exaggerate or misrepresent a set of scores. These misrepresentations generally result from failing to follow the basic rules for graph construction. The following example demonstrates how the same set of data can be presented in two entirely different ways by manipulating the structure of a graph.

For the past several years, the city has kept records of the number of homicides. The data are summarized as follows:

Year	Number of Homicides
2007	42
2008	44
2009	47
2010	49

These data are shown in two different graphs in Figure 2.10. In the first graph, we have exaggerated the height and started numbering the *Y*-axis at 40 rather than at zero. As a result, the graph seems to indicate a rapid rise in the number of homicides over the 4-year period. In the second graph, we have stretched out the *X*-axis and used zero as the starting point for the *Y*-axis. The result is a graph that shows little change in the homicide rate over the 4-year period.

Which graph is correct? The answer is that neither one is very good. Remember that the purpose of a graph is to provide an accurate display of the data. The first graph in Figure 2.10 exaggerates the differences between years, and the second graph conceals the differences. Some compromise is needed. Also note that in some cases a graph may not be the best way to display information. For these data, for example, showing the numbers in a table would be better than either graph.

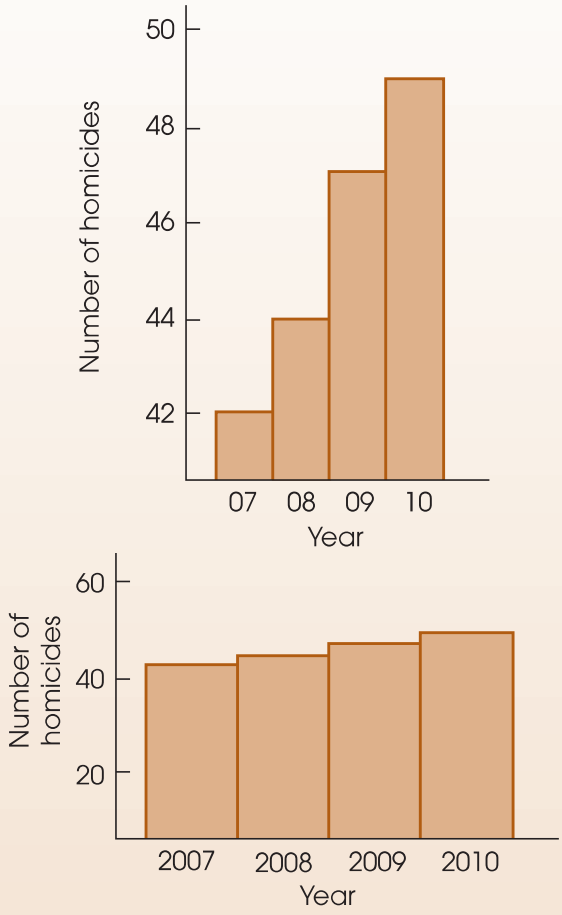


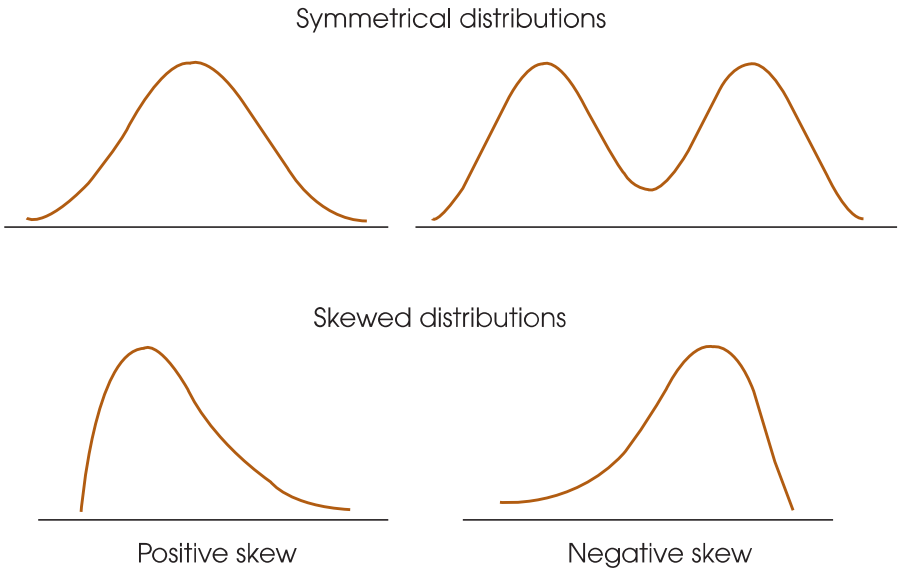
FIGURE 2.10

Two graphs showing the number of homicides in a city over a 4-year period. Both graphs show exactly the same data. However, the first graph gives the appearance that the homicide rate is high and rising rapidly. The second graph gives the impression that homicides rate is low and has not changed over the 4-year period.

The section where the scores taper off toward one end of a distribution is called the **tail** of the distribution.

A skewed distribution with the tail on the right-hand side is **positively skewed** because the tail points toward the positive (above-zero) end of the *X*-axis. If the tail points to the left, the distribution is **negatively skewed** (see Figure 2.11).

FIGURE 2.11
Examples of different shapes for distributions.



For a very difficult exam, most scores tend to be low, with only a few individuals earning high scores. This produces a positively skewed distribution. Similarly, a very easy exam tends to produce a negatively skewed distribution, with most of the students earning high scores and only a few with low values.

LEARNING CHECK

1. Sketch a frequency distribution histogram and a frequency distribution polygon for the data in the following table:

X	*f*
5	4
4	6
3	3
2	1
1	1

2. Describe the shape of the distribution in Exercise 1.

3. A researcher records the gender and academic major for each student at a college basketball game. If the distribution of majors is shown in a frequency distribution graph, what type of graph should be used?

4. If the results from a research study are presented in a frequency distribution histogram, would it also be appropriate to show the same results in a polygon? Explain your answer.

5. A college reports that the youngest registered student is 17 years old, and 20% of the registered students are older than 25. What is the shape of the distribution of ages for registered students?

ANSWERS

1. The graphs are shown in Figure 2.12.
2. The distribution is negatively skewed.
3. A bar graph is used for nominal data.
4. Yes. Histograms and polygons are both used for data from interval or ratio scales.
5. It is positively skewed with most of the distribution around 17–21 and a few scores scattered at 25 and higher.

2.5 PERCENTILES, PERCENTILE RANKS, AND INTERPOLATION

Although the primary purpose of a frequency distribution is to provide a description of an entire set of scores, it also can be used to describe the position of an individual within the set. Individual scores, or X values, are called *raw scores*. By themselves, raw scores do not provide much information. For example, if you are told that your score on an exam is $X = 43$, you cannot tell how well you did relative to other students in the class. To evaluate your score, you need more information, such as the average score or the number of people who had scores above and below you. With this additional information, you would be able to determine your relative position in the class. Because raw scores do not provide much information, it is desirable to transform them into a more meaningful form. One transformation that we consider changes raw scores into *percentiles*.

DEFINITIONS

The **rank** or **percentile rank** of a particular score is defined as the percentage of individuals in the distribution with scores equal to or less than the particular value.

When a score is identified by its percentile rank, the score is called a **percentile**.

Suppose, for example, that you have a score of $X = 43$ on an exam and that you know that exactly 60% of the class had scores of 43 or lower. Then your score $X = 43$ has a percentile rank of 60%, and your score would be called the 60th percentile. Notice that *percentile rank* refers to a percentage and that *percentile* refers to a score. Also notice that your rank or percentile describes your exact position within the distribution.

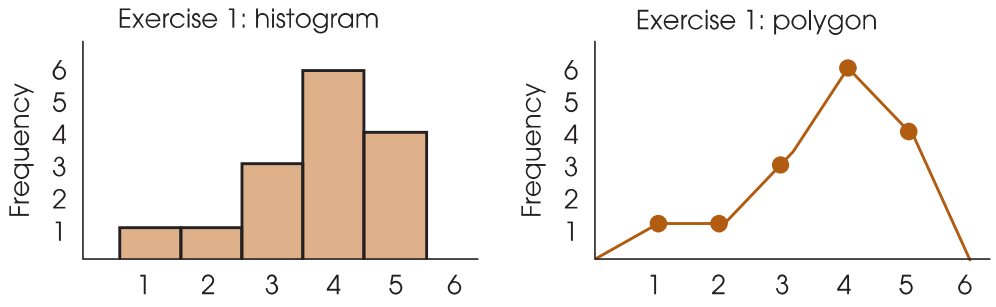


FIGURE 2.12
Answer to the Learning Check Exercise 1.

CUMULATIVE FREQUENCY AND CUMULATIVE PERCENTAGE

To determine percentiles or percentile ranks, the first step is to find the number of individuals who are located at or below each point in the distribution. This can be done most easily with a frequency distribution table by simply counting the number who are in or below each category on the scale. The resulting values are called *cumulative frequencies* because they represent the accumulation of individuals as you move up the scale.

EXAMPLE 2.5

In the following frequency distribution table, we have included a cumulative frequency column headed by *cf*. For each row, the cumulative frequency value is obtained by adding up the frequencies in and below that category. For example, the score $X = 3$ has a cumulative frequency of 14 because exactly 14 individuals had scores of $X = 3$ or less.

X	*f*	*cf*
5	1	20
4	5	19
3	8	14
2	4	6
1	2	2

The cumulative frequencies show the number of individuals located at or below each score. To find percentiles, we must convert these frequencies into percentages. The resulting values are called *cumulative percentages* because they show the percentage of individuals who are accumulated as you move up the scale.

EXAMPLE 2.6

This time we have added a cumulative percentage column (c%) to the frequency distribution table from Example 2.5. The values in this column represent the percentage of individuals who are located in and below each category. For example, 70% of the individuals (14 out of 20) had scores of $X = 3$ or lower. Cumulative percentages can be computed by

$$c\% = \frac{cf}{N}(100\%)$$

X	*f*	*cf*	*c%*
5	1	20	100%
4	5	19	95%
3	8	14	70%
2	4	6	30%
1	2	2	10%

The cumulative percentages in a frequency distribution table give the percentage of individuals with scores at or below each *X* value. However, you must remember that the *X* values in the table are usually measurements of a continuous variable and, therefore, represent intervals on the scale of measurement (see page 22). A score of $X = 2$,

for example, means that the measurement was somewhere between the real limits of 1.5 and 2.5. Thus, when a table shows that a score of $X = 2$ has a cumulative percentage of 30%, you should interpret this as meaning that 30% of the individuals have been accumulated by the time you reach the top of the interval for $X = 2$. Notice that each cumulative percentage value is associated with the upper real limit of its interval. This point is demonstrated in Figure 2.13, which shows the same data that were used in Example 2.6. Figure 2.13 shows that two people, or 10%, had scores of $X = 1$; that is, two people had scores between 0.5 and 1.5. You cannot be sure that both individuals have been accumulated until you reach 1.5, the upper real limit of the interval. Similarly, a cumulative percentage of 30% is reached at 2.5 on the scale, a percentage of 70% is reached at 3.5, and so on.

INTERPOLATION

It is possible to determine some percentiles and percentile ranks directly from a frequency distribution table, provided that the percentiles are upper real limits and the ranks are percentages that appear in the table. Using the table in Example 2.6, for example, you should be able to answer the following questions:

1. What is the 95th percentile? (Answer: $X = 4.5$.)
2. What is the percentile rank for $X = 3.5$? (Answer: 70%.)

However, there are many values that do not appear directly in the table, and it is impossible to determine these values precisely. Referring to the table in Example 2.6 again,

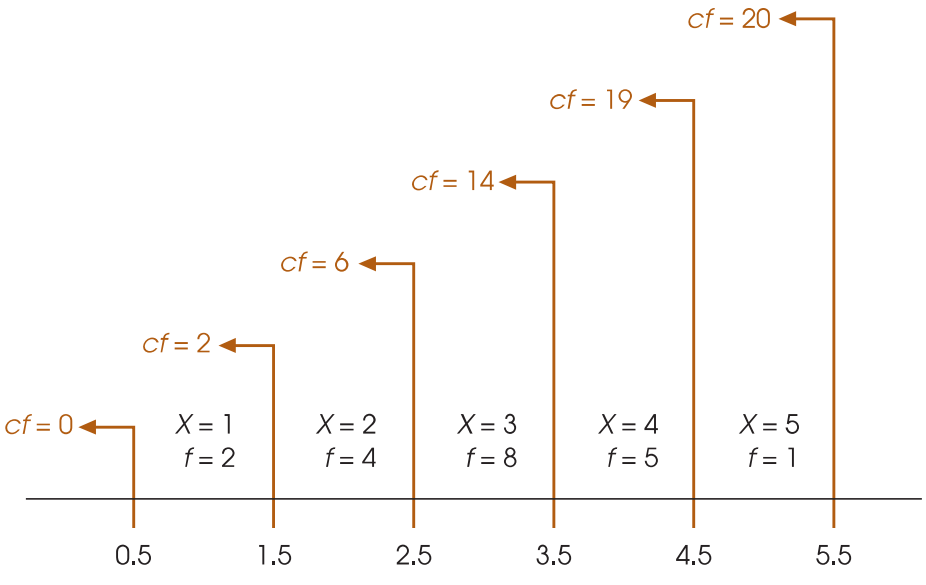
1. What is the 50th percentile?
2. What is the percentile rank for $X = 4$?

Because these values are not specifically reported in the table, you cannot answer the questions. However, it is possible to estimate these intermediate values by using a standard procedure known as *interpolation*.

Before we apply the process of interpolation to percentiles and percentile ranks, we use a simple, commonsense example to introduce this method. Suppose that Bob walks

FIGURE 2.13

The relationship between cumulative frequencies (*cf* values) and upper real limits. Notice that two people have scores of $X = 1$. These two individuals are located between the real limits of 0.5 and 1.5. Although their exact locations are not known, you can be certain that both had scores below the upper limit of 1.



to work each day. The total distance is 2 miles and the trip takes Bob 40 minutes. What is your estimate of how far Bob has walked after 20 minutes? To help, we have created a table showing the time and distance for the start and finish of Bob's trip.

	Time	Distance
Start	0	0
Finish	40	2

If you estimated that Bob walked 1 mile in 20 minutes, you have done interpolation. You probably went through the following logical steps:

1. The total time is 40 minutes.
2. 20 minutes represents half of the total time.
3. Assuming that Bob walks at a steady pace, he should walk half of the total distance in half of the total time.
4. The total distance is 2 miles and half of the total distance is 1.

The process of interpolation is pictured in Figure 2.14. In the figure, the top line shows the time for Bob's walk, from 0 to 40 minutes, and the bottom line shows the time, from 0 to 2 miles. The middle line shows different fractions along the way. Using the figure, try answering the following questions about time and distance.

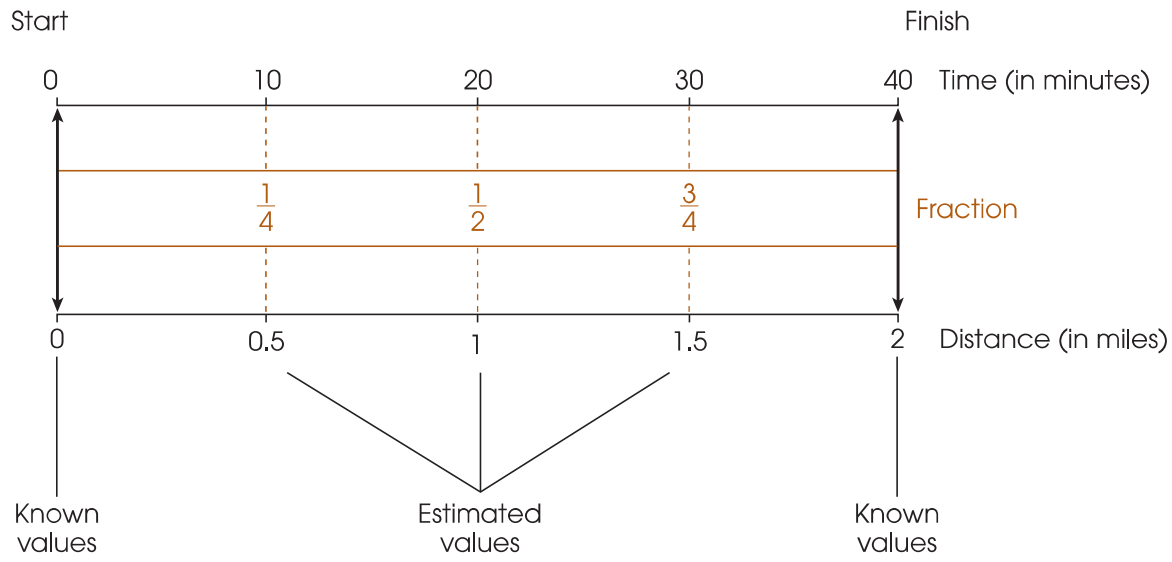


FIGURE 2.14

A graphic representation of the process of interpolation. The same interval is shown on two separate scales, time and distance. Only the endpoints of the scales are known—Bob starts at 0 for both time and distance, and he ends at 40 minutes and 2 miles. Interpolation is used to estimate values within the interval by assuming that fractional portions of one scale correspond to the same fractional portions of the other. For example, it is assumed that halfway through the time scale corresponds to halfway through the distance scale.

1. How much time does it take for Bob to walk 1.5 miles?
2. How far has Bob walked after 10 minutes?

If you got answers of 30 minutes and $\frac{1}{2}$ mile, you have mastered the process of interpolation.

Notice that interpolation provides a method for finding intermediate values—that is, values that are located between two specified numbers. This is exactly the problem we faced with percentiles and percentile ranks. Some values are given in the table, but others are not. Also notice that interpolation only *estimates* the intermediate values. The basic assumption underlying interpolation is that there is a constant rate of change from one end of the interval to the other. In Bob's walking example, we assume that he is walking at a constant rate for the entire trip. Because interpolation is based on this assumption, the values that we calculate are only estimates. The general process of interpolation can be summarized as follows:

1. A single interval is measured on two separate scales (for example, time and distance). The endpoints of the interval are known for each scale.
2. You are given an intermediate value on one of the scales. The problem is to find the corresponding intermediate value on the other scale.
3. The interpolation process requires four steps:
   a. Find the width of the interval on both scales.
   b. Locate the position of the intermediate value in the interval. This position corresponds to a fraction of the whole interval:

$$\text{fraction} = \frac{\text{distance from the top of the interval}}{\text{interval width}}$$

   c. Use the same fraction to determine the corresponding position on the other scale. First, determine the distance from the top of the interval:

$$\text{distance} = (\text{fraction}) \times (\text{width})$$

   d. Use the distance from the top to determine the position on the other scale.

You may notice that in each of these problems we use interpolation working from the *top* of the interval. However, this choice is arbitrary, and you should realize that interpolation can be done just as easily working from the bottom of the interval.

The following examples demonstrate the process of interpolation as it is applied to percentiles and percentile ranks. The key to success in solving these problems is that each cumulative percentage in the table is associated with the upper real limit of its score interval.

EXAMPLE 2.7

Using the following distribution of scores, we will find the percentile rank corresponding to $X = 7.0$:

X	f	cf	$c\%$
10	2	25	100%
9	8	23	92%
8	4	15	60%
7	6	11	44%
6	4	5	20%
5	1	1	4%

Notice that $X = 7.0$ is located in the interval bounded by the real limits of 6.5 and 7.5. The cumulative percentages corresponding to these real limits are 20% and 44%, respectively. These values are shown in the following table:

	Scores (X)	Percentages
Top	7.5	44%
Intermediate value	7.0	?
Bottom	6.5	20%

For interpolation problems, it is always helpful to create a table showing the range on both scales.

STEP 1 For the scores, the width of the interval is 1 point (from 6.5 to 7.5). For the percentages, the width is 24 points (from 20% to 44%).

STEP 2 Our particular score is located 0.5 point from the top of the interval. This is exactly halfway down in the interval.

STEP 3 On the percentage scale, halfway down is

$$\frac{1}{2}\ (24 \text{ points}) = 12 \text{ points}$$

STEP 4 For the percentages, the top of the interval is 44%, so 12 points down would be

$$44\% - 12\% = 32\%$$

This is the answer. A score of $X = 7.0$ corresponds to a percentile rank of 32%

This same interpolation procedure can be used with data that have been grouped into class intervals. Once again, you must remember that the cumulative percentage values are associated with the upper real limits of each interval. The following example demonstrates the calculation of percentiles and percentile ranks using data in a grouped frequency distribution.

EXAMPLE 2.8 Using the following distribution of scores, we can use interpolation to find the 50th percentile:

X	f	cf	$c\%$
20–24	2	20	100%
15–19	3	18	90%
10–14	3	15	75%
5–9	10	12	60%
0–4	2	2	10%

A percentage value of 50% is not given in the table; however, it is located between 10% and 60%, which are given. These two percentage values are associated with the

upper real limits of 4.5 and 9.5, respectively. These values are shown in the following table:

	Scores (X)	Percentages	
Top	9.5	60%	
	?	50%	Intermediate value
Bottom	4.5	10%	

STEP 1 For the scores, the width of the interval is 5 points. For the percentages, the width is 50 points.

STEP 2 The value of 50% is located 10 points from the top of the percentage interval. As a fraction of the whole interval, this is 10 out of 50, or $\frac{1}{5}$ of the total interval.

STEP 3 Using this same fraction for the scores, we obtain a distance of

$$\frac{1}{5}(5 \text{ points}) = 1 \text{ point}$$

The location we want is 1 point down from the top of the score interval.

STEP 4 Because the top of the interval is 9.5, the position we want is

$$9.5 - 1 = 8.5$$

This is the answer. The 50th percentile is $X = 8.5$.

LEARNING CHECK

1. On a statistics exam, would you rather score at the 80th percentile or at the 20th percentile?
2. For the distribution of scores presented in the following table,
   a. Find the 70th percentile.
   b. Find the percentile rank for $X = 9.5$.

X	f	cf	$c\%$
20–24	1	20	100%
15–19	5	19	95%
10–14	8	14	70%
5–9	4	6	20%
0–4	2	2	10%

3. Using the distribution of scores from Exercise 2 and interpolation,
   a. Find the 15th percentile.
   b. Find the percentile rank for $X = 13$.

ANSWERS

1. The 80th percentile is the higher score.
2. a. $X = 14.5$ is the 70th percentile. b. $X = 9.5$ has a rank of 20%.
3. a. Because 15% is between the values of 10% and 20% in the table, you must use interpolation. The score corresponding to a rank of 15% is $X = 7$.
   b. Because $X = 13$ is between the real limits of 9.5 and 14.5, you must use interpolation. The percentile rank for $X = 13$ is 55%.

2.6 STEM AND LEAF DISPLAYS

In 1977, J.W. Tukey presented a technique for organizing data that provides a simple alternative to a grouped frequency distribution table or graph (Tukey, 1977). This technique, called a *stem and leaf display*, requires that each score be separated into two parts: The first digit (or digits) is called the *stem*, and the last digit is called the *leaf*. For example, $X = 85$ would be separated into a stem of 8 and a leaf of 5. Similarly, $X = 42$ would have a stem of 4 and a leaf of 2. To construct a stem and leaf display for a set of data, the first step is to list all the stems in a column. For the data in Table 2.3, for example, the lowest scores are in the 30s and the highest scores are in the 90s, so the list of stems would be

Stems
3
4
5
6
7
8
9

The next step is to go through the data, one score at a time, and write the leaf for each score beside its stem. For the data in Table 2.3, the first score is $X = 83$, so you would write 3 in the leaf column beside the 8 in the column of stems. This process is continued for the entire set of scores. The complete stem and leaf display is shown with the original data in Table 2.3.

COMPARING STEM AND LEAF DISPLAYS WITH FREQUENCY DISTRIBUTIONS

Notice that the stem and leaf display is very similar to a grouped frequency distribution. Each of the stem values corresponds to a class interval. For example, the stem 3 represents all scores in the 30s—that is, all scores in the interval 30–39. The number of leaves in the display shows the frequency associated with each stem. It also should be clear that the stem and leaf display has one important advantage over a traditional grouped frequency distribution. Specifically, the stem and leaf display allows you to identify every individual score in the data. In the display shown in Table 2.3, for example, you know that there were three scores in the 60s and that the specific values were 62, 68, and 63. A frequency distribution would tell you only the frequency, not

TABLE 2.3

A set of $N = 24$ scores presented as raw data and organized in a stem and leaf display.

Data			Stem and Leaf Display	
83	82	63	3	23
62	93	78	4	26
71	68	33	5	6279
76	52	97	6	283
85	42	46	7	1643846
32	57	59	8	3521
56	73	74	9	37
74	81	76		

the specific values. This advantage can be very valuable, especially if you need to do any calculations with the original scores. For example, if you need to add all the scores, you can recover the actual values from the stem and leaf display and compute the total. With a grouped frequency distribution, however, the individual scores are not available.

LEARNING CHECK

1. Use a stem and leaf display to organize the following set of scores:

74, 103, 95, 98, 81, 117, 105, 99, 63, 86, 94, 107

96, 100, 98, 118, 107, 82, 84, 71, 91, 107, 84, 77

2. Explain how a stem and leaf display contains more information than a grouped frequency distribution.

ANSWERS

1. The stem and leaf display for these data is as follows:

6	3
7	417
8	16244
9	5894681
10	357077
11	78

2. A grouped frequency distribution table tells only the number of scores in each interval; it does not identify the exact value for each score. The stem and leaf display identifies the individual scores as well as the number of scores in each interval.

SUMMARY

1. The goal of descriptive statistics is to simplify the organization and presentation of data. One descriptive technique is to place the data in a frequency distribution table or graph that shows exactly how many individuals (or scores) are located in each category on the scale of measurement.

2. A frequency distribution table lists the categories that make up the scale of measurement (the X values) in one column. Beside each X value, in a second column, is the frequency or number of individuals in that category. The table may include a proportion column showing the relative frequency for each category:

$$\text{proportion} = p = \frac{f}{n}$$

The table may include a percentage column showing the percentage associated with each X value:

$$\text{percentage} = p(100) = \frac{f}{n}(100)$$

3. It is recommended that a frequency distribution table have a maximum of 10 to 15 rows to keep it simple. If the scores cover a range that is wider than this suggested maximum, it is customary to divide the range into sections called *class intervals*. These intervals are then listed in the frequency distribution table along with the frequency or number of individuals with scores in each interval. The result is called a *grouped frequency distribution*. The guidelines for constructing a grouped frequency distribution table are as follows:

a. There should be about 10 intervals.

b. The width of each interval should be a simple number (e.g., 2, 5, or 10).

c. The bottom score in each interval should be a multiple of the width.

d. All intervals should be the same width, and they should cover the range of scores with no gaps.

4. A frequency distribution graph lists scores on the horizontal axis and frequencies on the vertical axis. The type of graph used to display a distribution depends on the scale of measurement used. For interval or ratio scales, you should use a histogram or a polygon. For a histogram, a bar is drawn above each score so that the height of the bar corresponds to the frequency. Each bar extends to the real limits of the score, so that adjacent bars touch. For a polygon, a dot is placed above the midpoint of each score or class interval so that the height of the dot corresponds to the frequency; then lines are drawn to connect the dots. Bar graphs are used with nominal or ordinal scales. Bar graphs are similar to histograms except that gaps are left between adjacent bars.
5. Shape is one of the basic characteristics used to describe a distribution of scores. Most distributions can be classified as either symmetrical or skewed. A skewed distribution with the tail on the right is said to be positively skewed. If it has the tail on the left, it is negatively skewed.
6. The cumulative percentage is the percentage of individuals with scores at or below a particular point in the distribution. The cumulative percentage values are associated with the upper real limits of the corresponding scores or intervals.
7. Percentiles and percentile ranks are used to describe the position of individual scores within a distribution. Percentile rank gives the cumulative percentage associated with a particular score. A score that is identified by its rank is called a *percentile*.
8. When a desired percentile or percentile rank is located between two known values, it is possible to estimate the desired value using the process of interpolation. Interpolation assumes a regular linear change between the two known values.
9. A stem and leaf display is an alternative procedure for organizing data. Each score is separated into a stem (the first digit or digits) and a leaf (the last digit or digits). The display consists of the stems listed in a column with the leaf for each score written beside its stem. A stem and leaf display combines the characteristics of a table and a graph and produces a concise, well-organized picture of the data.

KEY TERMS

frequency distribution (39)
range (42)
grouped frequency distribution (42)
class interval (42)
apparent limits (44)
histogram (46)
polygon (47)
bar graph (48)
relative frequency (49)
symmetrical distribution (50)
tail(s) of a distribution (51)
positively skewed distribution (51)
negatively skewed distribution (51)
percentile rank (53)
percentile (53)
cumulative frequency (*cf*) (54)
cumulative percentage (*c*%) (54)
interpolation (55)
stem and leaf display (60)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 2 on the book companion website.

Improve your understanding of statistics with Aplia's auto-graded problem sets and immediate, detailed explanations for every question. To learn more, visit **www.aplia.com/statistics**.

CENGAGEbrain.com

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to produce **Frequency Distribution Tables or Graphs**.

Frequency Distribution Tables

Data Entry

1. Enter all the scores in one column of the data editor, probably VAR00001.

Data Analysis

1. Click **Analyze** on the tool bar, select **Descriptive Statistics,** and click on **Frequencies.**
2. Highlight the column label for the set of scores (VAR00001) in the left box and click the arrow to move it into the **Variable** box.
3. Be sure that the option to **Display Frequency Table** is selected.
4. Click **OK.**

SPSS Output

The frequency distribution table lists the score values in a column from smallest to largest, with the percentage and cumulative percentage also listed for each score. Score values that do not occur (zero frequencies) are not included in the table, and the program does not group scores into class intervals (all values are listed).

Frequency Distribution Histograms or Bar Graphs

Data Entry

1. Enter all the scores in one column of the data editor, probably VAR00001.

Data Analysis

1. Click **Analyze** on the tool bar, select **Descriptive Statistics,** and click on **Frequencies.**
2. Highlight the column label for the set of scores (VAR00001) in the left box and click the arrow to move it into the **Variable** box.
3. Click **Charts.**
4. Select either **Bar Graphs** or **Histogram.**
5. Click **Continue.**
6. Click **OK**.

SPSS Output

After a brief delay, SPSS displays a frequency distribution table and a graph. Note that SPSS often produces a histogram that groups the scores in unpredictable intervals. A bar graph usually produces a clearer picture of the actual frequency associated with each score.

FOCUS ON PROBLEM SOLVING

1. The reason for constructing frequency distributions is to put a disorganized set of raw data into a comprehensible, organized format. Because several different types of frequency distribution tables and graphs are available, one problem is deciding which type to use. Tables have the advantage of being easier to construct, but graphs generally give a better picture of the data and are easier to understand.

   To help you decide which type of frequency distribution is best, consider the following points:

   a. What is the range of scores? With a wide range, you need to group the scores into class intervals.

   b. What is the scale of measurement? With an interval or a ratio scale, you can use a polygon or a histogram. With a nominal or an ordinal scale, you must use a bar graph.

2. When using a grouped frequency distribution table, a common mistake is to calculate the interval width by using the highest and lowest values that define each interval. For example, some students are tricked into thinking that an interval identified as 20–24 is only 4 points wide. To determine the correct interval width, you can:

   a. Count the individual scores in the interval. For this example, the scores are 20, 21, 22, 23, and 24, for a total of 5 values. Thus, the interval width is 5 points.

   b. Use the real limits to determine the real width of the interval. For example, an interval identified as 20–24 has a lower real limit of 19.5 and an upper real limit of 24.5 (halfway to the next score). Using the real limits, the interval width is

   $$24.5 - 19.5 = 5 \text{ points}$$

3. Percentiles and percentile ranks are intended to identify specific locations within a distribution of scores. When solving percentile problems, especially with interpolation, it is helpful to sketch a frequency distribution graph. Use the graph to make a preliminary estimate of the answer before you begin any calculations. For example, to find the 60th percentile, draw a vertical line through the graph so that slightly more than half (60%) of the distribution is on the left-hand side of the line. Locating this position in your sketch gives you a rough estimate of what the final answer should be. When doing interpolation problems, you should keep several points in mind:

   a. Remember that the cumulative percentage values correspond to the upper real limits of each score or interval.

   b. You should always identify the interval with which you are working. The easiest way to do this is to create a table showing the endpoints on both scales (scores and cumulative percentages). This is illustrated in Example 2.7 on pages 57–58.

   c. The word *interpolation* means *between two poles*. Remember: Your goal is to find an intermediate value between the two ends of the interval. Check your answer to be sure that it is located between the two endpoints. If it is not, then check your calculations.

DEMONSTRATION 2.1

A GROUPED FREQUENCY DISTRIBUTION TABLE

For the following set of $N = 20$ scores, construct a grouped frequency distribution table using an interval width of 5 points. The scores are:

14, 8, 27, 16, 10, 22, 9, 13, 16, 12,
10, 9, 15, 17, 6, 14, 11, 18, 14, 11

STEP 1 Set up the class intervals.

The largest score in this distribution is $X = 27$, and the lowest is $X = 6$. Therefore, a frequency distribution table for these data would have 22 rows and would be too large. A grouped frequency distribution table would be better. We have asked specifically for an interval width of 5 points, and the resulting table has five rows.

X
25–29
20–24
15–19
10–14
5–9

Remember that the interval width is determined by the real limits of the interval. For example, the class interval 25–29 has an upper real limit of 29.5 and a lower real limit of 24.5. The difference between these two values is the width of the interval—namely, 5.

STEP 2 Determine the frequencies for each interval.

Examine the scores, and count how many fall into the class interval of 25–29. Cross out each score that you have already counted. Record the frequency for this class interval. Now repeat this process for the remaining intervals. The result is the following table:

X	f	
25–29	1	(the score $X = 27$)
20–24	1	($X = 22$)
15–19	5	(the scores $X = 16$, 16, 15, 17, and 18)
10–14	9	($X = 14$, 10, 13, 12, 10, 14, 11, 14, and 11)
5–9	4	($X = 8$, 9, 9, and 6)

DEMONSTRATION 2.2

USING INTERPOLATION TO FIND PERCENTILES AND PERCENTILE RANKS

Find the 50th percentile for the set of scores in the grouped frequency distribution table that was constructed in Demonstration 2.1.

STEP 1 Find the cumulative frequency (*cf*) and cumulative percentage values, and add these values to the basic frequency distribution table.

Cumulative frequencies indicate the number of individuals located in or below each category (class interval). To find these frequencies, begin with the bottom interval, and then accumulate the frequencies as you move up the scale.

Cumulative percentages are determined from the cumulative frequencies by the relationship

$$c\% = \left(\frac{cf}{N}\right)100\%$$

For example, the *cf* column shows that 4 individuals (out of the total set of $N = 20$) have scores in or below the 5–9 interval. The corresponding cumulative percentage is

$$c\% = \left(\frac{4}{20}\right)100\% = \left(\frac{1}{5}\right)100\% = 20\%$$

The complete set of cumulative frequencies and cumulative percentages is shown in the following table:

X	*f*	*cf*	*c*%
25–29	1	20	100%
20–24	1	19	95%
15–19	5	18	90%
10–14	9	13	65%
5–9	4	4	20%

STEP 2 Locate the interval that contains the value that you want to calculate.

We are looking for the 50th percentile, which is located between the values of 20% and 65% in the table. The scores (upper real limits) corresponding to these two percentages are 9.5 and 14.5, respectively. The interval, measured in terms of scores and percentages, is shown in the following table:

X	*c*%
14.5	65%
??	50%
9.5	20%

STEP 3 Locate the intermediate value as a fraction of the total interval.

Our intermediate value is 50%, which is located in the interval between 65% and 20%. The total width of the interval is 45 points ($65 - 20 = 45$), and the value of 50% is located 15 points down from the top of the interval. As a fraction, the 50th percentile is located $\frac{15}{45} = \frac{1}{3}$ down from the top of the interval.

STEP 4 Use the fraction to determine the corresponding location on the other scale.

Our intermediate value, 50%, is located $\frac{1}{3}$ of the way down from the top of the interval. Our goal is to find the score, the *X* value, that also is located $\frac{1}{3}$ of the way down from the top of the interval.

On the score (*X*) side of the interval, the top value is 14.5, and the bottom value is 9.5, so the total interval width is 5 points ($14.5 - 9.5 = 5$). The position we are seeking is $\frac{1}{3}$ of the way from the top of the interval. One-third of the total interval is

$$\left(\frac{1}{3}\right)5 = \frac{5}{3} = 1.67 \text{ points}$$

To find this location, begin at the top of the interval, and come down 1.67 points:

$$14.5 - 1.67 = 12.83$$

This is our answer. The 50th percentile is $X = 12.83$.

PROBLEMS

1. Place the following sample of $n = 20$ scores in a frequency distribution table.

6, 9, 9, 10, 8, 9, 4, 7, 10, 9
5, 8, 10, 6, 9, 6, 8, 8, 7, 9

2. Construct a frequency distribution table for the following set of scores. Include columns for proportion and percentage in your table.

Scores: 5, 7, 8, 4, 7, 9, 6, 6, 5, 3
9, 6, 4, 7, 7, 8, 6, 7, 8, 5

3. Find each value requested for the distribution of scores in the following table.

a. n
b. ΣX
c. ΣX^2

X	f
5	2
4	3
3	5
2	1
1	1

4. Find each value requested for the distribution of scores in the following table.

a. n
b. ΣX
c. ΣX^2

X	f
5	1
4	2
3	3
2	5
1	3

5. For the following scores, the smallest value is $X = 8$ and the largest value is $X = 29$. Place the scores in a grouped frequency distribution table

a. using an interval width of 2 points.
b. using an interval width of 5 points.

24, 19, 23, 10, 25, 27, 22, 26
25, 20, 8, 24, 29, 21, 24, 13
23, 27, 24, 16, 22, 18, 26, 25

6. The following scores are the ages for a random sample of $n = 30$ drivers who were issued speeding tickets in New York during 2008. Determine the best interval width and place the scores in a grouped frequency distribution table. From looking at your table, does it appear that tickets are issued equally across age groups?

17, 30, 45, 20, 39, 53, 28, 19,
24, 21, 34, 38, 22, 29, 64,
22, 44, 36, 16, 56, 20, 23, 58,
32, 25, 28, 22, 51, 26, 43

7. For each of the following samples, determine the interval width that is most appropriate for a grouped frequency distribution and identify the approximate number of intervals needed to cover the range of scores.

a. Sample scores range from $X = 24$ to $X = 41$
b. Sample scores range from $X = 46$ to $X = 103$
c. Sample scores range from $X = 46$ to $X = 133$

8. What information can you obtain about the scores in a regular frequency distribution table that is not available from a grouped table?

9. Describe the difference in appearance between a bar graph and a histogram and describe the circumstances in which each type of graph is used.

10. For the following set of quiz scores:

3, 5, 4, 6, 2, 3, 4, 1, 4, 3
7, 7, 3, 4, 5, 8, 2, 4, 7, 10

a. Construct a frequency distribution table to organize the scores.
b. Draw a frequency distribution histogram for these data.

11. Sketch a histogram and a polygon showing the distribution of scores presented in the following table:

X	*f*
7	1
6	1
5	3
4	6
3	4
2	1

12. A survey given to a sample of 200 college students contained questions about the following variables. For each variable, identify the kind of graph that should be used to display the distribution of scores (histogram, polygon, or bar graph).
a. number of pizzas consumed during the previous week
b. size of T-shirt worn (S, M, L, XL)
c. gender (male/female)
d. grade point average for the previous semester
e. college class (freshman, sophomore, junior, senior)

13. Each year the college gives away T-shirts to new students during freshman orientation. The students are allowed to pick the shirt sizes that they want. To determine how many of each size shirt they should order, college officials look at the distribution from last year. The following table shows the distribution of shirt sizes selected last year.

Size	*f*
S	27
M	48
L	136
XL	120
XXL	39

a. What kind of graph would be appropriate for showing this distribution?
b. Sketch the frequency distribution graph.

14. A report from the college dean indicates that for the previous semester, the grade distribution for the Department of Psychology included 135 As, 158 Bs, 140 Cs, 94 Ds, and 53 Fs. Determine what kind of graph would be appropriate for showing this distribution and sketch the frequency distribution graph.

15. For the following set of scores

Scores: 5, 8, 5, 7, 6, 6, 5, 7, 4, 6
6, 9, 5, 5, 4, 6, 7, 5, 7, 5

a. Place the scores in a frequency distribution table.
b. Identify the shape of the distribution.

16. Place the following scores in a frequency distribution table. Based on the frequencies, what is the shape of the distribution?

5, 6, 4, 7, 7, 6, 8, 2, 5, 6
3, 1, 7, 4, 6, 8, 2, 6, 5, 7

17. For the following set of scores:

3, 7, 6, 5, 5, 9, 6, 4, 6, 8
10, 2, 7, 4, 9, 5, 6, 3, 8

a. Construct a frequency distribution table.
b. Sketch a polygon showing the distribution.
c. Describe the distribution using the following characteristics:
(1) What is the shape of the distribution?
(2) What score best identifies the center (average) for the distribution?
(3) Are the scores clustered together, or are they spread out across the scale?

18. Fowler and Christakis (2008) report that personal happiness tends to be associated with having a social network including many other happy friends. To test this claim, a researcher obtains a sample of $n = 16$ adults who claim to be happy people and a similar sample of $n = 16$ adults who describe themselves as neutral or unhappy. Each individual is then asked to identify the number of their close friends whom they consider to be happy people. The scores are as follows:

Happy:
8, 7, 4, 10, 6, 6, 8, 9, 8, 8,
7, 5, 6, 9, 8, 9

Unhappy:
5, 8, 4, 6, 6, 7, 9, 6, 2, 8,
5, 6, 4, 7, 5, 6

Sketch a polygon showing the frequency distribution for the happy people. In the same graph, sketch a polygon for the unhappy people. (Use two different colors, or use a solid line for one polygon and a dashed line for the other.) Does one group seem to have more happy friends?

19. Complete the final two columns in the following frequency distribution table and then find the percentiles and percentile ranks requested.

X	*f*	*cf*	*c%*
7	2		
6	3		
5	6		
4	9		
3	4		
2	1		

a. What is the percentile rank for $X = 2.5$?
b. What is the percentile rank for $X = 6.5$?
c. What is the 20th percentile?
d. What is the 80th percentile?

20. Complete the final two columns in the following frequency distribution table and then find the percentiles and percentile ranks requested.

X	*f*	*cf*	*c%*
50–59	1		
40–49	3		
30–39	6		
20–29	5		
10–19	3		
0–9	2		

a. What is the percentile rank for $X = 9.5$?
b. What is the percentile rank for $X = 39.5$?
c. What is the 25th percentile?
d. What is the 50th percentile?

21. Complete the final two columns in the following frequency distribution table and then use interpolation to find the percentiles and percentile ranks requested.

X	*f*	*cf*	*c%*
10	2		
9	5		
8	8		
7	15		
6	10		
5	6		
4	4		

a. What is the percentile rank for $X = 6$?
b. What is the percentile rank for $X = 9$?
c. What is the 25th percentile?
d. What is the 90th percentile?

22. Find the requested percentiles and percentile ranks for the following distribution of quiz scores for a class of $N = 40$ students.

X	*f*	*cf*	*c%*
20	2	40	100.0
19	4	38	95.0
18	6	34	85.0
17	13	28	70.0
16	6	15	37.5
15	4	9	22.5
14	3	5	12.5
13	2	2	5.0

a. What is the percentile rank for X = 15?
b. What is the percentile rank for X = 18?
c. What is the 15th percentile?
d. What is the 90th percentile?

23. Use interpolation to find the requested percentiles and percentile ranks requested for the following distribution of scores.

X	*f*	*cf*	*c%*
14–15	3	50	100
12–13	6	47	94
10–11	8	41	82
8–9	18	33	66
6–7	10	15	30
4–5	4	5	10
2–3	1	1	2

a. What is the percentile rank for $X = 5$?
b. What is the percentile rank for $X = 12$?
c. What is the 25th percentile?
d. What is the 70th percentile?

24. The following frequency distribution presents a set of exam scores for a class of $N = 20$ students.

X	*f*	*cf*	*c%*
90–99	4	20	100
80–89	7	16	80
70–79	4	9	45
60–69	3	5	25
50–59	2	2	10

a. Find the 30th percentile.
b. Find the 88th percentile.
c. What is the percentile rank for $X = 77$?
d. What is the percentile rank for $X = 90$?

25. Construct a stem and leaf display for the data in problem 6 using one stem for the scores in the 60s, one for scores in the 50s, and so on.

26. A set of scores has been organized into the following stem and leaf display. For this set of scores:
a. How many scores are in the 70s?
b. Identify the individual scores in the 70s.
c. How many scores are in the 40s?
d. Identify the individual scores in the 40s.

3	8
4	60
5	734
6	81469
7	2184
8	247

27. Use a stem and leaf display to organize the following distribution of scores. Use seven stems with each stem corresponding to a 10-point interval.

Scores:

28,	54,	65,	53,	81
45,	44,	51,	72,	34
43,	59,	65,	39,	20
53,	74,	24,	30,	49
36,	58,	60,	27,	47
22,	52,	46,	39,	65