FREQUENCY TABLES, GRAPHS, AND DISTRIBUTIONS

You will need to use the following from the previous chapter:

Symbols
Σ: Summation sign

Concepts
Continuous versus discrete scales

Chapter 2

A CONCEPTUAL FOUNDATION

I used to give a diagnostic quiz during the first session of my course in statistics. The quiz consisted of 10 multiple-choice questions requiring simple algebraic manipulations, designed to show whether students had the basic mathematical tools to handle a course in statistics. (I have since stopped giving the quiz because the results were frequently misleading, especially when very bright students panicked and produced deceptively low scores.) Most students were curious to see the results of the quiz and to know how their performance compared to those of their classmates. To show how the class did on the quiz, about the cruelest thing I could have done would have been to put all of the raw scores, in no particular order, on the blackboard. This is the way the data would first appear after I graded the quizzes. For a class of 25 students, the scores would typically look like those in Table 2.1.

Table 2.1

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

You can see that there are a lot of 6s and 8s and not a lot of 10s or scores below 4, but this is not the best way to get a sense of how the class performed. A very simple and logical step makes it easier to understand the scores: Put them in order. A string of scores arranged in numerical order (customarily starting with the highest value) is often called an *array*. Putting the scores from Table 2.1 into an array produces Table 2.2 (read left to right starting with the top row).

Table 2.2

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

Frequency Distributions

The array in Table 2.2 is certainly an improvement, but the table could be made more compact. If the class contained 100 students, an array would be quite difficult to look at. A more informative way to display these data is in a *simple frequency distribution*, which is a table consisting of two columns. The first column lists all of the possible scores, beginning with the highest score in the array and going down to the lowest score. The second column lists the *frequency* of each score—that is, how many times that score is

repeated in the array. You don't have to actually write out the array before constructing a simple frequency distribution, but doing so makes the task easier. Table 2.3 is a simple frequency distribution of the data in Table 2.2. *X* stands for any score, and *f* stands for the frequency of that score. Notice that the score of 3 is included in the table even though it has a frequency of zero (i.e., there are no 3s in the data array). The rule is to list all the possible scores from the highest to the lowest, whether a particular score actually appears in the data or not. To check whether you have included all your scores in the frequency distribution, add up all of the frequencies (i.e., Σf), and make sure that the total is equal to the number of scores in the array (i.e., check that $\Sigma f = N$).

Table 2.3

X	f
10	2
9	2
8	5
7	3
6	7
5	1
4	4
3	0
2	1
$\Sigma f =$	25

A simple frequency distribution is very helpful when the number of different values listed is not very high (nine in the case of Table 2.3), but imagine 25 scores on a midterm exam graded from 0 to 100. The scores might range from 64 to 98, requiring 35 different scores to be listed, at least 10 of which would have zero frequencies. In that case a simple frequency distribution would not be much more informative than a data array. A better solution would be to group the scores into equal-sized intervals (e.g., 80–84, 85–89, etc.) and construct a *grouped frequency distribution*. Because the mechanics of dealing with such distributions are a bit more complicated, I will save this topic for Section B.

The Mode of a Distribution

The score that occurs most frequently in a distribution is called the *mode* of the distribution. For the preceding distribution, the mode is 6 because that score occurs seven times in the distribution—more often than any other score. Complicating things is the fact that a distribution can have more than one mode (e.g., if there were seven instead of only five 8s in Table 2.3, the distribution would have two modes: 6 and 8). The mode will be discussed further in the next chapter, when I deal with ways for summarizing a distribution with just one number.

The Cumulative Frequency Distribution

To evaluate his or her own performance in a class, a student will frequently ask, "How many students in the class had lower scores than mine?" To answer this question for any particular student you need only sum the frequencies for scores below that student's own score. However, you can perform a procedure that will answer that question for any student in the class: You can construct a *cumulative frequency distribution*. The *X* and *f* columns of such a distribution are the same as in the simple frequency distribution, but each entry in the cumulative frequencies (*cf*) column contains a sum of the frequencies for the corresponding score and all scores below it. Table 2.4 shows the cumulative frequencies for the data in Table 2.3.

Table 2.4

X	f	cf
10	2	25
9	2	23
8	5	21
7	3	16
6	7	13
5	1	6
4	4	5
3	0	1
2	1	1

If a student attained a score of 7 on the quiz, we can look at the entry in the *cf* column for a score of 6 to see that this student performed better than 13 other students. The *cf* entry corresponding to a score of 7 (i.e., 16) answers the question, How many scores are either lower than or tied with a score of 7?

The mechanics of creating the *cf* column are easy enough. The *cf* entry for the lowest score is just the same as the frequency of that score. The *cf* for the next highest score is the frequency of that score plus the frequency of the score below. Each *cf* entry equals the frequency of that score plus the *cf* for the score below. For example, the *cf* for a score of 7 is the frequency

of 7, which is 3, plus the *cf* for 6, which is 13: *cf* for 7 = 3 + 13 = 16. The entry at the top of the *cf* column should equal the total number of scores, *N*, which also equals Σ*f*.

The Relative Frequency and Cumulative Relative Frequency Distributions

Although it may be satisfying to know that you scored better than many other students, what usually matters in terms of getting good grades is what *fraction* of the class scored below you. Outscoring 15 students in a class of 25 is pretty good because you beat 3/5 of the class. Having 15 students below you in a class of 100 is not very good because in that case you have outperformed only 3/20, or .15, of the class. The kind of table that can tell you what fraction of the scores are lower than yours is called a *cumulative relative frequency distribution*. There are two different ways to arrive at this distribution.

As a first step, you can create a relative frequency distribution by dividing each entry in the *f* column of a simple frequency distribution by *N*. The resulting fraction is the relative frequency (*rf*), and it tells you what proportion of the group attained each score. Notice that in Table 2.5, each *rf* entry was created by dividing the corresponding *f* by 25. The cumulative relative frequencies (*crf*) are then found by accumulating the *rf*'s starting from the bottom, just as we did with the *f* column to obtain the *cf* entries. Alternatively, you can convert each entry in the *cf* column into a proportion by dividing it by *N*. (For example, the *crf* of .64 for a score of 7 can be found either by dividing 16 by 25 or by adding .12 to the *crf* of .52 for the score below.) Either way you get the *crf* column, as shown in Table 2.5. Note that the *crf* for the top score in the table must be 1.0—if it isn't, you made some kind of mistake (perhaps too much rounding off in lower entries).

Table 2.5

X	f	cf	rf	crf
10	2	25	.08	1.00
9	2	23	.08	.92
8	5	21	.20	.84
7	3	16	.12	.64
6	7	13	.28	.52
5	1	6	.04	.24
4	4	5	.16	.20
3	0	1	0	.04
2	1	1	.04	.04

The Cumulative Percentage Distribution

Let us again focus on a quiz score of 7. I pointed out earlier that by looking at the *cf* entry for 6 you can see that 13 students scored below 7. Now we can look at the *crf* entry for 6 to see that a score of 7 beats .52, or a little more than half, of the class $\left(\frac{13}{25} = .52\right)$. A score of 6, however, beats only .24 (the *crf* entry for 5), or about one fourth, of the class. Sometimes people find it more convenient to work with percentages. If you want a *cumulative percentage frequency* (*cpf*) column, you need only multiply each *crf* entry by 100. A score of 7 is better than the scores of 52% of the class; a 6 beats only 24% of the scores. Because the *cpf* column is especially useful for describing scores in a group, let's look at Table 2.6 and focus only on that column. The entries in the *cpf* column have a special name: They are called *percentile*

Table 2.6

X	f	cpf
10	2	100%
9	2	92
8	5	84
7	3	64
6	7	52
5	1	24
4	4	20
3	0	4
2	1	4

ranks (PR). By convention, a percentile rank is defined as the percentage of the group that is at or below a given score. To find the PR of a particular score we look straight across at the *cpf* entry, rather than looking at the score below. Thus, the PR of a score of 7 is 64; 64% of the group scored 7 or below. Similarly, the PR for 6 is 52. The way percentile ranks are found changes a bit when dealing with a continuous scale or when dealing with grouped frequency distributions, but the concept is the same, as you will see in Section B.

Percentiles

Instead of being concerned with the percentage at or below a particular score, sometimes you may want to focus on a particular percentage and find the score that has that percentile rank. For instance, before seeing the results of the diagnostic quiz, a professor might decide that the bottom 20% of the class must receive some remedial training on algebra, regardless of their actual scores on the quiz. That is, whether the whole class does well or poorly, whoever is in the bottom 20% will have to get help. In this case, we want to find the score in the distribution that has a PR of 20. You can see from Table 2.6 that a score of 4 has a PR of 20, so that is the score we are interested in. This score is called the 20th *percentile*. Anyone with this score or a lower score will have to get algebra help.

A percentile can be defined as a score that has a given PR—the 25th percentile is a score whose PR is 25. In other words, a percentile is the score at or below which a given percentage of the group falls. The most interesting percentiles are either *quartiles* (i.e., 25%, 50%, or 75%) or *deciles* (i.e., 10%, 20%, etc.). Unfortunately, these convenient percentiles rarely appear as entries in a *cpf* column. In Table 2.6, the only convenient percentile is the 20th. The score of 6 comes close to the 50th percentile (52%), and the score of 5 is a good approximation for the 25th percentile. On the other hand, the 75th percentile is almost exactly midway between 7 and 8. Later in this section, I will show how you can use a graph to more precisely estimate percentiles (and PRs) that do not appear as entries in a table.

Graphs

The information in a frequency distribution table can usually be presented more clearly and dramatically in the form of a graph. A typical graph is made with two perpendicular lines, one horizontal and the other vertical. The values of some variable (X) are marked off along the horizontal line, which is also called the *horizontal axis* (or X axis). A second variable, labeled Y, is marked off along the *vertical axis* (or Y axis). When graphing a frequency distribution, the variable of interest (e.g., quiz scores) is placed along the X axis, and distance (i.e., height) along the Y axis represents the frequency count for each variable.

The Bar Graph

Probably the simplest type of graph is the *bar graph*, in which a rectangle, or bar, is erected above each value of X. The higher the frequency of that value, the greater the height of the bar. The bar graph is appropriate when the values of X come from a discrete rather than a continuous scale (as defined in Chapter 1). A good example of a variable that always produces discrete values is family size. Whereas quiz scores can sometimes be measured with fractions, family size is *always* a whole number. The appropriate way to display a frequency distribution of family size is with a bar graph. Imagine

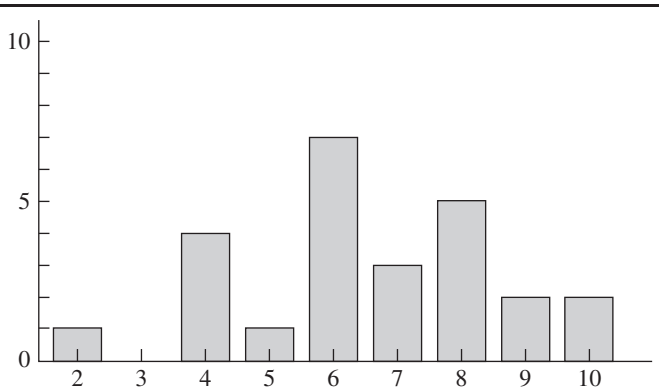


Figure 2.1

Bar Graph

that the *X* values in Table 2.3 are not quiz scores but the sizes (number of parents plus number of children) of 25 randomly selected families in which the parents have been taking fertility drugs. The bar graph for these data is shown in Figure 2.1. Notice that the bars do not touch; we wouldn't want to give the impression that the values come from a continuous scale—that a family can be, for instance, between 3 and 4 in size.

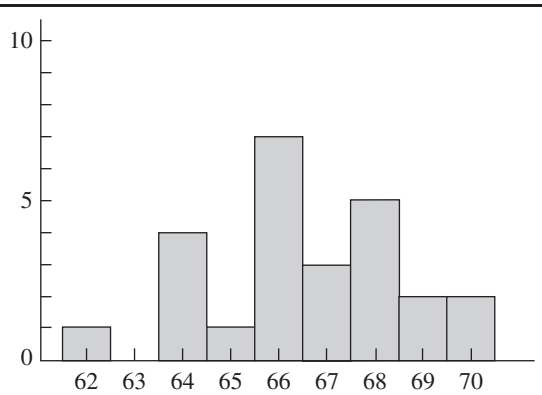
The advantage of a bar graph as compared to a table should be clear from Figure 2.1; the bar graph shows at a glance how the family sizes are distributed among the 25 families. Bar graphs are also appropriate when the variable in question has been measured on a nominal or ordinal scale. For instance, if the 25 members of a statistics class were sorted according to eye color, the values along the *X* axis would be blue, brown, green, and so forth, and the heights of the bars would indicate how many students had each eye color.

The Histogram

A slightly different type of graph is more appropriate if the variable is measured on a continuous scale. A good example of a variable that is almost always measured on a continuous scale is height. Unlike family size, height varies continuously, and it is often represented in terms of fractional values. By convention, however, in the United States height is most commonly reported to the nearest inch. If you ask someone how tall she is, she might say, for example, 5 feet 5 inches, but you know she is rounding off a bit. It is not likely that she is *exactly* 5 feet 5 inches tall. You know that her height could be anywhere between 5 feet $4\frac{1}{2}$ inches and 5 feet $5\frac{1}{2}$ inches. Because height is being measured on a continuous scale, a value like 5 feet 5 inches generally stands for an interval that goes from 5 feet $4\frac{1}{2}$ inches (the lower *real limit*) to 5 feet $5\frac{1}{2}$ inches (the upper real limit). When constructing a bar graph that involves a continuous scale, the bar for each value is drawn wide enough so that it goes from the lower real limit to the upper real limit. Therefore, adjacent bars touch each other. A bar graph based on a continuous scale, in which the bars touch, is called a *frequency histogram*. The data from Table 2.3 can be displayed in a histogram if we assume that the *X* values represent inches above 5 feet for a group of 25 women whose heights have been measured. (That is, a value of 2 represents 5 feet 2 inches, or 62 inches; 3 is 5 feet 3 inches, or 63 inches; etc.) The histogram is shown in Figure 2.2. As with the bar graph, the heights of the bars represent the

Figure 2.2

Frequency Histogram



frequency count for each value. A glance at this figure shows you how the women are distributed in terms of height.

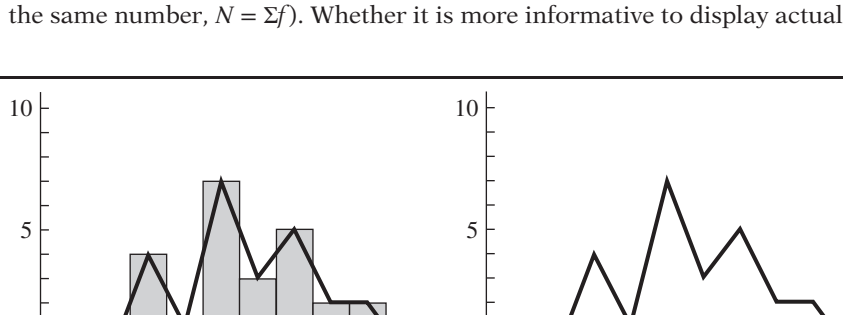
The Frequency Polygon

For some purposes, researchers find the bars of a histogram to be distracting and prefer an alternative format, the *frequency polygon*. An easy way to think of a frequency polygon is to imagine placing a dot in the middle of the top of each bar in a histogram and connecting the dots with straight lines (and then getting rid of the bars), as shown in Figure 2.3a. Of course, normally a frequency polygon is drawn without first constructing the histogram, as shown in Figure 2.3b. Notice that the frequency polygon is connected to the horizontal axis at the high end by a straight line from the bar representing the frequency count of the highest value, and is similarly connected at the low end. Thus, the area enclosed by the polygon is clearly defined and can be used in ways to be described later. A frequency polygon is particularly useful when comparing two overlapping distributions on the same graph. The bars of a histogram would only get in the way in that case.

Just as a simple frequency distribution can be displayed as a histogram or as a polygon, so too can the other distributions we have discussed: the relative frequency distribution, the cumulative frequency distribution, and so forth. It should be obvious, however, that a histogram or polygon based on a relative frequency distribution will have exactly the same shape as the corresponding graph of a simple frequency distribution—only the scale of the Y axis will change (because all of the frequency counts are divided by the same number, $N = \Sigma f$). Whether it is more informative to display actual

Figure 2.3

Frequency Polygon



frequencies or relative frequencies depends on the situation. If the group from which the data have been taken is very large, relative frequencies will probably make more sense.

Whether your polygon is based on simple or relative frequencies, it is easy to find the mode of your distribution (defined earlier in this section) from looking at the polygon. The mode is the score on the *X* axis that is directly under the highest point of the polygon. Because the height of the polygon at each point represents the frequency of the score below it, the score at which the polygon is highest is the most popular score in the distribution, and therefore it is the mode. However, as mentioned before, there can be more than one mode in a distribution (e.g., the polygon can look a bit like a camel with two humps). Even if one mode is actually a bit higher than the other (in which case, technically, there is really only one mode), if the polygon rises to one distinct peak, decreases, and then rises again to another distinct peak, it is common to say that the distribution has two modes. The role of the mode in describing distributions will be discussed further in the next chapter.

The Cumulative Frequency Polygon

A *cumulative frequency polygon* (also called an *ogive*) has a very different shape than a simple frequency polygon does. For one thing, the *cf* polygon never slopes downward as you move to the right in the graph, as you can see in Figure 2.4 (which was drawn using the same data as in all the examples above). That is because the cumulative frequency can never decrease. It can stay the same, if the next value has a zero frequency, but there are no negative frequency counts, so a cumulative frequency can never go down as the number of values increases. This is a case for which the polygon is definitely easier to look at and interpret than the corresponding histogram. Notice that in the cumulative frequency polygon the dots of the graph are not centered above the values being counted, but rather are above the *upper real limit* of each value (e.g., 5 feet $4\frac{1}{2}$ inches, instead of 5 feet 4 inches). The rationale is that to make sure you have accumulated, for instance, all of the heights labeled 5 feet 4 inches, you have to include all measurements up to 5 feet $4\frac{1}{2}$ inches.

The ogive can be quite useful when the percentile, or PR, in which you are interested falls between two of the entries in a table. In these common

Figure 2.4

Cumulative Frequency Polygon (Ogive)

cases, expressing the *Y* axis of the ogive in terms of cumulative percentages can help you to estimate the intermediate value that you are seeking. In the case of Figure 2.4, you would need only to multiply the frequencies on the *Y* axis by 4 (i.e., 100/*N*) to create a cumulative percentage polygon. Then, to find the percentile rank of any score you first find the score on the *X* axis of the polygon, draw a vertical line from that score up to intersect the cumulative polygon, and finally draw a horizontal line from the point of intersection to the *Y* axis. The percentage at the point where the horizontal line intersects the *Y* axis is the PR of the score in question. For example, if you start with a score of 6.0 on the horizontal axis of Figure 2.4, move up until you hit the curve, and then move to the left, you will hit the vertical axis near the frequency of 10, which corresponds to 40%. So the PR of a score of 6.0 is about 40.

Naturally, the procedure for finding percentiles is exactly the opposite of the one just described. For instance, to find the score at the 70th percentile, start at this percentage on the *Y* axis of Figure 2.4 (midway between the frequencies of 15 and 20, which correspond to 60% and 80%, respectively), and move to the right on a horizontal line until you hit the ogive. From the point of intersection, go straight down to the horizontal axis, and you should hit a score of about 7.8 on the *X* axis. So the 70th percentile is about 7.8.

Of course, the accuracy of these graphical procedures depends on how carefully the lines are drawn. Drawing the graph to a larger scale tends to increase accuracy. Also, note that the cumulative percentage polygon consists of straight *lines*; therefore, these approximations are a form of *linear interpolation*. The procedure can be made more accurate by fitting a curve to the points of the cumulative polygon, but how the curve is drawn depends inevitably on assumptions about how the distribution would look if it were smooth (i.e., if you had infinitely precise measurements of the variable on an infinitely large population). These days you are usually better off just letting your computer draw the graphs and/or find the percentiles and PRs in which you are interested (see Section C).

I will discuss the preceding graphs again when I apply these graphing techniques to grouped frequency distributions in Section B. For now, I will compare the relative frequency polygon with the concept of a *theoretical frequency distribution*.

Real Versus Theoretical Distributions

Frequency polygons make it easy to see the distribution of values in your data set. For instance, if you measured the anxiety level of each new student entering a particular college and made a frequency polygon out of the distribution, you could see which levels of anxiety were common and which were not. If you are a dean at the college and you see a high and wide peak over some pretty high anxiety levels, you would be concerned about the students and would consider various interventions to make the students more at ease. If new anxiety measurements were taken after some interventions, you would hope to see the frequency polygon change so that the line is high over the low anxiety values and gets quite low for high anxiety levels.

Unfortunately, the frequency polygon is harder to look at and interpret simply when it is based on a small number of scores. For instance, the frequency polygon in Figure 2.3b is based on only 25 height measurements (rounded to the nearest inch), and therefore it is not at all smooth; it consists of straight lines and sharp angles, which at no point resemble a curve. However, if height were measured to the nearest tenth of an inch and many more people were included in the distribution, the polygon would

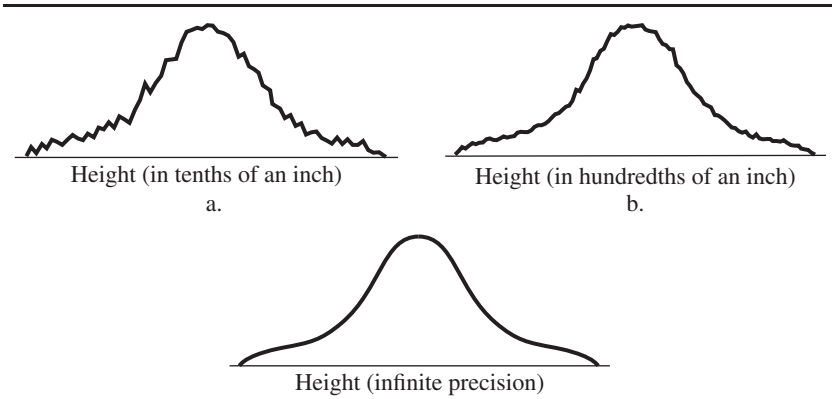


Figure 2.5

Frequency Polygons With Precise Measurements

consist of many more lines, which would be shorter, and many more angles, which would be less sharp (see Figure 2.5a). If height were measured to the nearest hundredth of an inch on a large population, the frequency polygon would consist of very many tiny lines, and it would begin to look fairly smooth (see Figure 2.5b). If we kept increasing the precision of the height measurements and the number of people measured, eventually the frequency polygon will not be distinguishable from a smooth curve (see Figure 2.5c). Smooth curves are easier to summarize and to interpret in a simple way.

The frequency polygons that psychological researchers create from their own data are usually far from smooth due to relatively few measurements and, often, an imprecise scale (that is one reason why psychologists are not likely to publish such displays, using them instead as tools for inspecting their data). On the other hand, a mathematical (or theoretical) distribution is determined by an equation and usually appears as a perfectly smooth curve. The best-known mathematical distribution is the *normal distribution*, which looks something like a bell viewed from the side (as in Figure 2.5c). With a precise scale, and enough people of one gender in a distribution of height (the distribution gets more complicated when the heights of both genders are included, as you will see in the next chapter), the frequency polygon for height will look a lot like the normal curve (except that the true normal curve actually never ends, extending to infinity in both directions before touching the horizontal axis). This resemblance is important because many advanced statistical procedures become quite easy if you assume that the variable of interest follows a normal distribution. I will have much more to say about the normal distribution in the next few chapters, and about other theoretical distributions in later chapters.

SUMMARY

1. Often the first step in understanding a group of scores is to put them in order, thus forming an *array*.
2. If the number of different values in the group is not too large, a *simple frequency distribution* may make it easy to see where the various scores lie. To create a simple frequency distribution, write down all the possible scores in a column with the highest score in the group on top and the lowest on the bottom, even if some of the intermediate possible scores do not appear in the group. Add a second column in which you record the frequency of occurrence in the group for each value in the first column. The score with the highest frequency is the *mode* of the distribution.

3. It is easy to create a cumulative frequency (*cf*) distribution from a simple frequency distribution: The *cf* entry for each score is equal to the frequency for that score plus the frequencies for all lower scores. (This is the same as saying that the *cf* for a given score is the frequency for that score, plus the *cf* for the next lower score.) The *cf* entry for the highest score must equal $\Sigma f = N$ (the total number of scores in the group).
4. To convert a simple or cumulative frequency distribution to a relative or cumulative relative distribution, divide each entry by *N*. The relative distribution tells you the proportion of scores at each value, and the cumulative relative distribution tells you what proportion of the scores is at or below each value.
5. Multiplying each entry of a cumulative relative frequency distribution by 100 gives a cumulative percentage distribution. The entries of the latter distribution are *percentile ranks* (PRs); each entry tells you the percentage of the distribution that is at or below the corresponding score. A percentile, on the other hand, is the score corresponding to a particular cumulative percentage. For example, the 40th percentile is the score that has a PR of 40. If the percentile or PR of interest does not appear in the table, it can be estimated with the appropriate graph (see point 9).
6. If the scores in a distribution represent a discrete variable (e.g., number of children in a family), and you want to display the frequency distribution as a graph, a *bar graph* should be used. In a bar graph, the heights of the bars represent the frequency counts, and adjacent bars do not touch. A bar graph is also appropriate for distributions involving nominal or ordinal scales (e.g., the frequency of different eye colors in the population).
7. When dealing with a continuous scale (e.g., height measured in inches), the distribution can be graphed as a *histogram*, which is a bar graph in which adjacent bars *do* touch. In a histogram, the width of the bar that represents a particular value goes from the *lower* to the *upper real limit* of that value.
8. An alternative to the histogram is the frequency polygon, in which a point is drawn above each value. The height of the point above the value on the *X* axis represents the frequency of that value. These points are then connected by straight lines, and the polygon is connected to the *X* axis at either end to form a closed figure. It is usually easier to compare two polygons on the same graph (e.g., separate distributions for males and females) than two histograms.
9. A cumulative frequency distribution can be graphed as a cumulative frequency polygon, called an *ogive*, in the same manner as the ordinary frequency polygon—just place the dot representing the *cf* over the upper real limit of each corresponding score. If you convert the cumulative frequencies to cumulative percentages, the ogive can be used to estimate percentiles and PRs not in your original table. Move straight up from a score until you hit the curve and then horizontally to the left until you hit the *Y* axis to find the PR of the score. Reversing this procedure allows you to estimate percentiles.
10. A frequency polygon can let you see at a glance which scores are popular in a distribution and which are not. As the number of people in the distribution and the precision of the measurements increase, the polygon begins to look fairly smooth. Ideally, the frequency polygon can somewhat resemble a perfectly smooth mathematical distribution, such as the normal curve.

EXERCISES

*1. A psychotherapist has rated all 20 of her patients in terms of their progress in therapy, using a 7-point scale. The results are shown in the following table:

	f
Greatly improved	5
Moderately improved	4
Slightly improved	6
Unchanged	2
Slightly worse	2
Moderately worse	1
Greatly worse	0

   a. Draw a bar graph to represent the above results. To answer the following questions, create relative frequency (*rf*), cumulative frequency (*cf*), cumulative relative frequency (*crf*), and cumulative percentage frequency (*cpf*) columns for the table.
   b. What proportion of the patients was greatly improved?
   c. How many patients did not improve (i.e., were unchanged or became worse)? What proportion of the patients did not improve?
   d. What is the percentile rank of a patient who improved slightly? Of a patient who became slightly worse?
   e. Which category of improvement corresponds to the third quartile (i.e., 75th percentile)? To the first quartile?

*2. A cognitive psychologist is training volunteers to use efficient strategies for memorizing lists of words. After the training period, 25 participants are each tested on the same list of 30 words. The numbers of words correctly recalled by the participants are as follows: 25, 23, 26, 24, 19, 25, 24, 28, 26, 21, 24, 24, 29, 23, 19, 24, 23, 24, 25, 23, 24, 25, 26, 28, 25. Create a simple frequency distribution to display these data, and then add columns for *rf*, *cf*, *crf*, and *cpf*.
   a. What proportion of the participants recalled exactly 24 words?
   b. How many participants recalled no more than 23 words? What proportion of the total does this represent?
   c. What is the percentile rank of a participant who scored 25? Of a participant who scored 27?
   d. Which score is close to being at the first quartile? The third quartile?
   e. Draw a histogram to represent the data.

3. A boot camp sergeant recorded the number of attempts each of 20 soldiers required to complete an obstacle course. The results were 2, 5, 3, 1, 2, 7, 1, 4, 2, 4, 8, 1, 3, 2, 6, 5, 2, 4, 3, 1. Create a simple frequency table to display these data. Add columns for *cf*, *rf*, *crf*, and *cpf*. (*Note*: Because lower numbers reflect better scores, you may want to put the lowest number on top of the table.)
   a. What proportion of the soldiers could complete the course on the first attempt?
   b. What proportion of them needed four or more attempts?
   c. What is the percentile rank of someone who needed five attempts?
   d. What score is closest to being the third quartile?
   e. Draw a frequency polygon to represent the data.

4. An ethnographer surveyed 25 homes to determine the number of people per household. She found the following household sizes: 2, 1, 3, 5, 1, 4, 3, 2, 2, 6, 3, 4, 5, 1, 2, 4, 2, 7, 4, 6, 5, 5, 6, 6, 5. Construct a simple frequency table to display these results. Add columns for *cf*, *rf*, *crf*, and *cpf*.
   a. What percentage of households have three or fewer members?
   b. What household size corresponds to the 80th percentile?
   c. How many households have only one member? To what proportion does that correspond?
   d. What proportion of households have five or more members?
   e. Draw a bar graph to represent the data.

*5. A physics professor gave a quiz with 10 questions to a class of 20 students. The scores were 10, 3, 8, 7, 1, 6, 5, 9, 8, 4, 2, 7, 7, 10, 9, 6, 8, 3, 8, 5. Create a simple frequency table to display these results. Add columns for *cf*, *rf*, *crf*, and *cpf*.
   a. How many students obtained a perfect score? What proportion does that represent?

b. What score is closest to the 50th percentile?
c. What is the percentile rank of a student who scored a 5? Of a student who scored a 9?
d. What proportion of the students scored 9 or more?
e. Draw a frequency polygon to represent the data.

6. Draw a cumulative percentage polygon (ogive) to represent the data in Exercise 3. Use your graph to answer the following questions (approximate your answer to the nearest tenth of a point):
   a. What score is at the 30th percentile?
   b. What score is at the 50th percentile?
   c. What is the percentile rank that corresponds to a score of 3.5?
   d. What is the percentile rank that corresponds to a score of 6.5?

*7. Draw a cumulative percentage polygon (ogive) to represent the data in Exercise 5. Use your graph to answer the following questions (approximate your answer to the nearest tenth of a point):
   a. What score is at the 50th percentile?
   b. What score is at the 75th percentile?
   c. What is the percentile rank that corresponds to a score of 4?
   d. What is the percentile rank that corresponds to a score of 7?

8. The following data represent the scores of 50 students on a difficult 20-question quiz: 17, 12, 6, 13, 9, 15, 11, 16, 4, 15, 12, 13, 10, 13, 2, 11, 13, 10, 20, 14, 12, 17, 10, 15, 12, 17, 9, 14, 11, 15, 11, 16, 9, 13, 18, 10, 13, 0, 11, 16, 9, 8, 12, 13, 12, 17, 8, 16, 12, 15. Create a simple frequency table for these data, add columns for *cf* and *cpf*, and then graph the cumulative percentage polygon in order to answer the following questions.
   a. Find the (approximate) values for the three quartiles of this distribution.
   b. Find the (approximate) values for the first and ninth deciles of this distribution.
   c. What is the (approximate) percentile rank of a student who scored an 8 on the quiz?
   d. What is the (approximate) percentile rank of a student who scored an 18 on the quiz?

B BASIC STATISTICAL PROCEDURES

Grouped Frequency Distributions

Constructing a simple frequency distribution is, as the name implies, simple. Unfortunately, measurements on an interval/ratio scale usually result in too many different values for a simple frequency distribution to be helpful. The example of quiz scores was particularly convenient because there were only eight different values. However, suppose the example involved 25 scores on a midterm exam graded from 0 to 100. Hypothetical scores are listed in the form of an array in Table 2.7, as defined in Section A.

Table 2.7

98	96	93	92	92	89	89	88	86	86	86	85	85
84	83	81	81	81	81	79	75	75	72	68	64	

To put these scores in a simple frequency distribution, we would have to include all of the values from 98 down to 64, which means that many potential scores would have a frequency of zero (e.g., 97, 95, 94).

The simple frequency distribution obviously would not be very helpful in this case. In fact, it seems little better than merely placing the scores in order in an array. The problem, of course, is that the simple frequency distribution has too many different values. The solution is to group the possible score values into equal-sized ranges, called *class intervals*. A table that shows the frequency for each class interval is called a *grouped frequency distribution*. The data from Table 2.7 were used to form the grouped frequency distribution in Table 2.8. Notice how much more informative the frequency distribution becomes when scores are grouped in this way.

Table 2.8

Class Interval *X*	*f*	Class Interval *X*	*f*
95–99	2	75–79	3
90–94	3	70–74	1
85–89	8	65–69	1
80–84	6	60–64	1

Apparent Versus Real Limits

To describe the construction of a grouped frequency distribution, I will begin by focusing on just one class interval from Table 2.8—for example, 80–84. The interval is defined by its *apparent limits*. A score of 80 is the *lower apparent limit* of this class interval, and 84 is the *upper apparent limit*. If the variable is thought of as continuous, however, the apparent limits are not the *real* limits of the class interval. For instance, if the score values from 64 to 98 represented the heights of 1-year-old infants in centimeters, any fractional value would be possible. In particular, any height above 79.5 cm would be rounded to 80 cm and included in the interval 80–84. Similarly, any height below 84.5 cm would be rounded to 84 and also included in the interval 80–84. Therefore, the *real limits* of the class interval are 79.5 (lower real limit) and 84.5 (upper real limit).

In general, the real limits are just half a unit above or below the apparent limits—whatever the unit of measurement happens to be. In the example of infant heights, the unit is centimeters. If, however, you were measuring the lengths of people's index fingers to the nearest tenth of an inch, you might have an interval (in inches) from 2.0 to 2.4, in which case the real limits would be 1.95 to 2.45. In this case, half a unit of measurement is half of a tenth of an inch, which is one twentieth of an inch, or .05. To find the width of a class interval (usually symbolized by i), we use the real limits rather than the apparent limits. The width of the interval from 2.0 to 2.4 inches would be $2.45 - 1.95 = .5$ inch. In the case of the 80–84 interval we have been discussing, the width is $84.5 - 79.5 = 5$ cm (if the values are thought of as the heights of infants), not the 4cm that the apparent limits would suggest. If the values are thought of as midterm grades, they will not include any fractional values (exams graded from 0 to 100 rarely involve fractions). Nevertheless, the ability being measured by the midterm is viewed as a continuous variable.

Constructing Class Intervals

Notice that the different class intervals in Table 2.8 do not overlap. Consider, for example, the interval 80–84 and the next highest one, 85–89. It is impossible for a score to be in both intervals simultaneously. This is important because it would become very confusing if a single score contributed to the frequency count in more than one interval. It is also important that there is no gap between the two intervals; otherwise, a score could fall between the cracks and not get counted at all. Bear in mind that even though there appears to be a gap when you look at the apparent limits (80–84, 85–89), the gap disappears when you look at the real limits (79.5–84.5, 84.5–89.5) and yet there is still no overlap. Perhaps you are wondering what happens if a score turns out to be exactly 84.5. First, when dealing with a continuous scale, the probability of any particular *exact* value (e.g., 84.500) arising is considered to be too small to worry about. In reality, however, measurements are not so precise, and such values do arise. In that case, a simple rule can be adopted, such as any value ending in exactly .5 should be placed in the higher interval if the number before the .5 is even.

Choosing the Class Interval Width

Before you can create a grouped frequency distribution, you must first decide how wide to make the class intervals. This is an important decision. If you make the class interval too large, there will be too few intervals to give you much detail about the distribution. For instance, suppose we chose to put the data from Table 2.7 into a grouped frequency distribution in

which i (the interval width) equals 10. The result would be as shown in Table 2.9. Such a grouping could be useful if these class intervals actually corresponded with some external criterion; for instance, the class intervals could correspond to the letter grades A, B, C, and D. However, in the absence of some external criterion for grouping, it is preferable to have at least 10 class intervals to get a detailed picture of the distribution. On the other hand, if you make the class intervals too narrow, you may have so many intervals that you are not much better off than with the simple frequency distribution. In general, more than 20 intervals is considered too many to get a good picture of the distribution.

Table 2.9

Class Interval X	f
90–99	5
80–89	14
70–79	4
60–69	2

You may have noticed that Table 2.8, with only eight intervals, does not follow the recommendation of 10 to 20 intervals. There is, however, at least one other guideline to consider in selecting a class interval width: multiples of 5 are particularly easy to work with. To have a number of class intervals between 10 and 20, the data from Table 2.7 would have to be grouped into intervals with $i = 3$ or $i = 2$. The distribution with $i = 2$ is too similar to the simple frequency distribution (i.e., $i = 1$) to be of much value, but the distribution with $i = 3$ is informative, as shown in Table 2.10.

Table 2.10

Class Interval X	f	Class Interval X	f
96–98	2	78–80	1
93–95	1	75–77	2
90–92	2	72–74	1
87–89	3	69–71	0
84–86	6	66–68	1
81–83	5	63–65	1

Finally, note that it is a good idea to make all of the intervals the same size. Although there can be reasons to vary the size of the intervals within the same distribution, it is rarely done, and this text will not discuss such cases.

Finding the Number of Intervals Corresponding to a Particular Class Width

Whether Table 2.10 is really an improvement over Table 2.8 depends on your purposes and preferences. In trying to decide which size class interval to use, you can use a quick way to determine how many intervals you will wind up with for a particular interval width. First, find the *range* of your scores by taking the highest score in the array and subtracting the lowest score. (Actually, you have to start with the *upper real limit* of the highest score and subtract the *lower real limit* of the lowest score. If you prefer, instead of dealing with real limits, you can usually just subtract the lowest from the highest score and add 1.) For the midterm scores, the range is $98.5 - 63.5 = 35$. Second, divide the range by a convenient interval width, and round up if there is any fraction at all. This gives you the number of intervals. For example, using $i = 3$ with the midterm scores, we get $35/3 = 11.67$, which rounds up to 12, which is the number of intervals in Table 2.10. Note that if the range of your values is less than 20 to start with, it is reasonable to stick with the simple frequency distribution, although you may want to use $i = 2$ if the number of scores in your array is small (which would result in many zero frequencies). To avoid having too many intervals with low or zero frequency, it has been suggested that the number of classes not be much more than the square root of the sample size (e.g., if $N = 25$, this rule suggests the use of $\sqrt{25} = 5$ classes; this rule would argue in favor of Table 2.9, but Table 2.8 would still be considered a reasonable choice).

Choosing the Limits of the Lowest Interval

Having chosen the width of your class interval, you must decide on the apparent limits of the lowest interval; the rest of the intervals will then be determined. Naturally, the lowest class interval must contain the lowest score in the array, but that still leaves room for some choice. A useful guideline is to make sure that either the lower apparent limit or the upper apparent limit of the lowest interval is a multiple of i. (If the lower limit

of one interval is a multiple of *i*, all the lower limits will be multiples of *i*.) This is true in Table 2.10: The lower limit of the lowest interval (63) is a multiple of *i*, which is 3. It also would have been reasonable to start with 64–66 as the lowest interval because then the upper limit (66) would have been a multiple of *i*. Choosing the limits of the lowest interval is a matter of convenience, and a judgment can be made after seeing the alternatives.

Relative and Cumulative Frequency Distributions

Once a grouped frequency distribution has been constructed, it is easy to add columns for cumulative, relative, and cumulative relative frequencies, as described in Section A. These columns have been added to the grouped frequency distribution in Table 2.8 to create Table 2.11.

Table 2.11

Interval	*f*	*cf*	*rf*	*crf*
95–99	2	25	.08	1.00
90–94	3	23	.12	.92
85–89	8	20	.32	.80
80–84	6	12	.24	.48
75–79	3	6	.12	.24
70–74	1	3	.04	.12
65–69	1	2	.04	.08
60–64	1	1	.04	.04

Cumulative Percentage Distribution

Perhaps the most useful table of all is one that shows cumulative percent frequencies because (as noted in Section A) such a table allows you to find percentile ranks (PRs) and percentiles. The cumulative percent frequencies for the midterm scores are shown in Table 2.12. It is important to note that the cumulative percentage entry (as with any cumulative entry) for a particular interval corresponds to the *upper real limit* of that interval. For example, across from the interval 85–89 is the *cpf%* entry of 80. This means that a score of 89.5 is the 80th percentile (that is why the table includes a separate column for the upper real limit, labeled *url*). To score better than 80% of those in the class, a student must have a score that beats not only all the scores below the 85–89 interval but all the scores *in* the 85–89 interval. And the only way a student can be sure of beating all the scores in the 85–89 interval is to score at the top of that interval: 89.5.

On the other hand, if you wanted to know what your percentile rank would be if you scored 79.5 on the midterm, you would look at the cumulative percent frequency entry for the 75–79 interval, which tells you that your PR is 24 (i.e., you beat 24% of the group). If you wanted to know the PR for a score

Table 2.12

Interval	*f*	*pf%*	*url*	*cf*	*cpf%*
95–99	2	8	99.5	25	100
90–94	3	12	94.5	23	92
85–89	8	32	89.5	20	80
80–84	6	24	84.5	12	48
75–79	3	12	79.5	6	24
70–74	1	4	74.5	3	12
65–69	1	4	69.5	2	8
60–64	1	4	64.5	1	4

of 67 or 81, or you wanted to know what score was at the 40th percentile, you could not find that information directly in Table 2.12. However, you could use a graph to help you estimate these answers, as demonstrated in Section A, or you could use linear interpolation more directly, as described next.

Estimating Percentiles and Percentile Ranks by Linear Interpolation

If you are dealing with a grouped distribution, and therefore know how many scores are in each interval but not where within each interval those scores lie (i.e., I am assuming that you don't have access to the raw data from which the frequency table was constructed), you can use *linear interpolation* to estimate both percentiles and percentile ranks. The key assumption behind linear interpolation is that the scores are spread evenly (i.e., linearly) throughout the interval.

Estimating Percentile Ranks

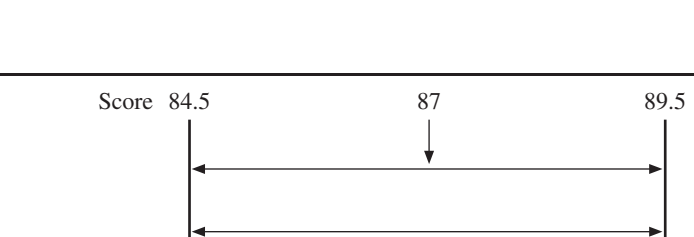
Consider the interval 85–89 in Table 2.12, for which the frequency is 8. We assume that the eight scores are spread evenly from 84.5 to 89.5 so that, for instance, four of the scores are between 84.5 and 87 (the *midpoint* of the interval), and the other four are between 87 and 89.5. This reasoning also applies to the percentages. The cumulative percentage at 84.5 is 48 (the *cpf* entry for 80–84), and at 89.5 it is 80 (the *cpf* entry for 85–89), as shown in Figure 2.6. On the basis of our assumption of linearity, we can say that the midpoint of the interval, 87, should correspond to a cumulative percentage midway between 48 and 80, which is 64% [(48 + 80)/2 = 128/2 = 64]. Thus, the PR for a score of 87 is 64.

A more complicated question to ask is: What is the PR for a score of 86 in Table 2.12? Because 86 is not right in the middle of an interval, we need to know how far across the interval it is. Then we can use linear interpolation to find the cumulative percentage that corresponds to that score. To go from the lower real limit, 84.5, to 86 we have to go 1.5 score points. To go across the entire interval requires five points (the width of the interval). So 86 is 1.5 out of 5 points across the interval; 1.5 out of 5 = 1.5/5 = .3, or 30%. A score of 86 is 30% of the way across the interval. That means that to find the cumulative percentage for 86, we must go 30% of the way from 48 to 80, as shown in Figure 2.7. From 48 to 80 there are 32 percentage points, and 30% of 32 is (.3)(32) = 9.6. So we have to add 9.6 percentage points to 48 to get 57.6, which is the PR for a score of 86. In sum, 86 is 30% of the way from 84.5 to 89.5, so we go 30% of the way from 48 to 80, which is 57.6.

Bear in mind that it is not terribly important to be exact about estimating a percentile rank from a grouped frequency distribution. First, the estimate

Figure 2.6

Finding a Percentile Rank by Linear Interpolation (Midpoint)



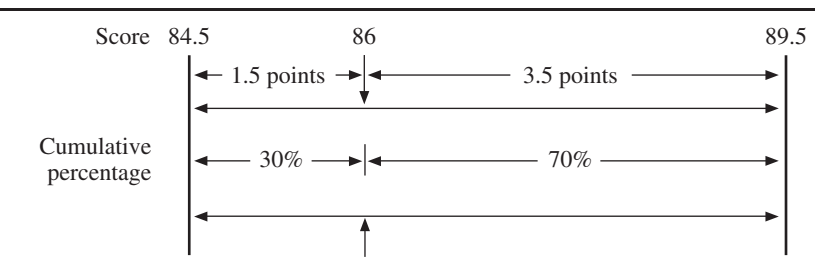


Figure 2.7

Finding Any Percentile Rank by Linear Interpolation

is based on the assumption that the scores are spread evenly throughout the interval, which may not be true. Second, the estimate may be considerably different if the class interval width or the starting score of the lowest interval changes, Now that I have described how to estimate a PR corresponding to any score in a grouped distribution, it will be easy to describe the reverse process of estimating the score that corresponds to a given percentile rank.

Estimating Percentiles

Suppose you want to find the sixtieth percentile (i.e., the score for which the PR is 60) for the midterm exam scores. First, you can see from Table 2.12 that 60% lands between the entries 48% (corresponding to 84.5) and 80% (corresponding to 89.5). Because 60 is somewhat closer to 48 than it is to 80, you know that the 60th percentile should be somewhat closer to 84.5 than to 89.5—that is, in the neighborhood of 86. More exactly, the proportion of the way from 48 to 80 you have to go to get to 60 (which is the same proportion you will have to go from 84.5 to 89.5) is $(60 - 48)/32 = 12/32 = .375$. Adding .375 of 5 (the width of the class interval) to 84.5 yields $84.5 + (.375) \cdot 5 = 84.5 + 1.875 = 86.375$. It would be reasonable to round off in this case, and say that the 60th percentile is 86.4.

Graphing a Grouped Frequency Distribution

A grouped frequency distribution can be displayed as a histogram, like the one used to represent the simple frequency distribution in Section A (see Figure 2.2). In a graph of a grouped distribution, however, the width of each bar extends from the lower real limit to the upper real limit of the class interval that the bar represents. As before, the height of the bar indicates the frequency of the interval. (This is only true when all the class intervals have the same width, but because this is the simplest and most common arrangement, we will consider only this case.) A histogram for a grouped frequency distribution is shown in Figure 2.8, which is a graph of the data in Table 2.8.

If you prefer to use a frequency polygon, place a dot at the top of each bar of the histogram at the midpoint of the class interval. (A quick way to calculate the midpoint is to add the upper and lower apparent limits and divide by 2—this also works with the real limits.) Place dots on the horizontal axis (to represent zero frequency) on either side of the distribution—that is, at the midpoint of the next interval below the lowest and above the highest, as shown in Figure 2.9. Connecting the polygon to these additional dots on either side closes the polygon, with the horizontal axis serving as one of the sides. Thus, the frequency polygon encloses a particular amount of area,

Figure 2.8

Frequency Histogram for a Grouped Distribution

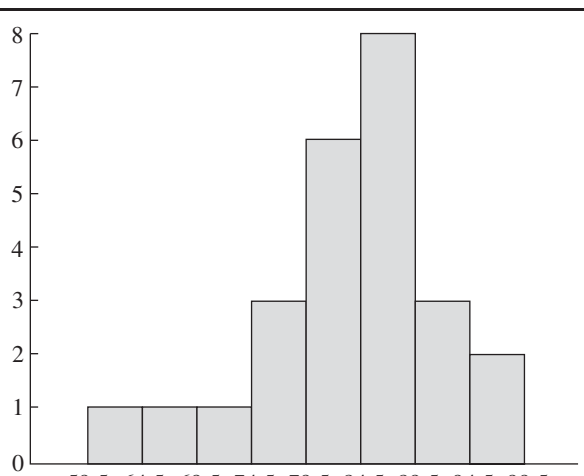
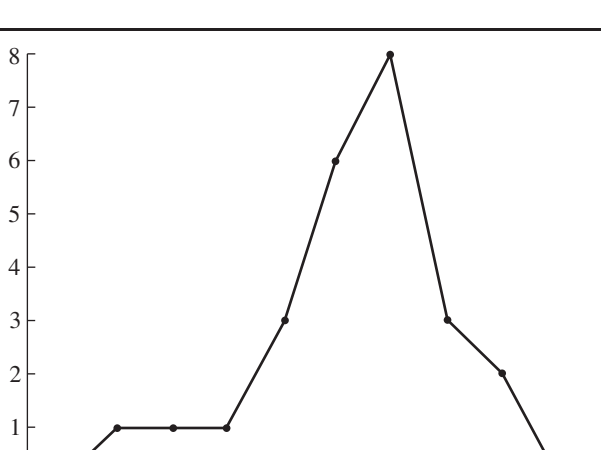


Figure 2.9

Frequency Polygon for a Grouped Distribution



which represents the total number of scores in the distribution. A third of that area, for example, would represent a third of the scores. I will have a lot more to say about the areas enclosed by frequency polygons and smooth distributions in Chapter 4. Of course, you can also create a cumulative frequency or percentage polygon (an ogive) as described in Section A. Just place the dot representing the cumulative frequency or percentage over the upper real limit of the interval to which it corresponds. Then, you can use the ogive you plotted to find percentiles and PRs, also as described in Section A.

Guidelines for Drawing Graphs of Frequency Distributions

Graphs of frequency distributions are not often published in psychological journals, but there are general guidelines for creating any kind of line graph that should be followed to make the graphs easier to interpret. (These guidelines appear in most statistics texts; here, I adapt them for use with frequency distributions.) The first guideline is that you should make the *X*

axis longer than the *Y* axis by about 50% so that the height of the graph is only about two thirds of the width. (Some researchers suggest that the height be closer to three quarters of the width. The exact ratio is not critical, but a proportion in this vicinity is considered easiest to interpret visually.) The second guideline is that the scores or measurement values should be placed along the horizontal axis and the frequency for each value indicated on the vertical axis. This creates a profile, like the skyline of a big city in the distance, that is easy to grasp. A third guideline is obvious: The units should be equally spaced on both axes (e.g., a single frequency count should be represented by the same distance anywhere along the *Y* axis). The fourth guideline is that the intersection of the *X* and *Y* axes should be the zero point for both axes, with numbers getting larger (i.e., more positive) as you move up or to the right. The fifth guideline is that you should choose a measurement unit and a scale (i.e., how much distance on the graph equals one unit) so that the histogram or polygon fills up nearly all of the graph without at any point going beyond the axes of the graph.

Sometimes it is difficult to satisfy the last two guidelines simultaneously. Suppose you want to graph a distribution of normal body temperatures (measured to the nearest tenth of a degree Fahrenheit) for a large group of people. You would like to mark off the *X* axis in units of .1 degree, but if you

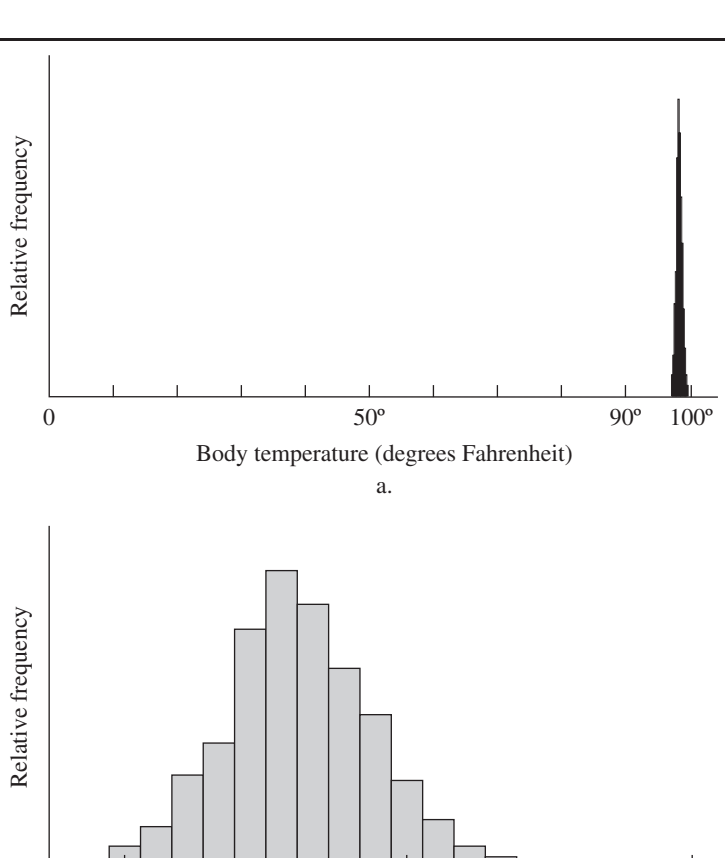


Figure 2.10

Frequency Histograms: Continuous Scale and Broken Scale

start with zero on the left and mark off equal intervals, each representing .1 degree, you will have to mark off 1,000 intervals to get to 100 degrees. Assuming that your distribution extends from about 97° F to about 100° F (98.6° being average body temperature), you will be using only a tiny portion of the *X* axis, as indicated in Figure 2.10a. The solution to this dilemma is to increase the scale so that .1 degree takes more distance along the *X* axis and not to mark off units continuously from 0 to 97°. Instead, you can indicate a break along the *X* axis, as shown in Figure 2.10b so that the zero point can still be included but the distribution can fill the graph. Similarly, a break can be used on the *Y* axis if all the frequency counts are high but do not differ greatly.

The sixth and last guideline is that both axes should be clearly labeled. In the case of a frequency distribution the *X* axis should be labeled with the name of the variable and the unit in which it was measured.

B SUMMARY

1. When a distribution contains too many possible values to fit conveniently in a regular frequency distribution, class intervals may be created (usually all of which are the same size) such that no two intervals overlap and that there are no gaps between intervals. If the variable is measured on a continuous scale, the upper and lower real limits of the interval are half of a measurement unit above and below the upper and lower apparent limits, respectively.
2. One way to help you decide on a class width to use is to first find the range of scores by subtracting the lowest score in your distribution from the highest and adding 1. Then divide the range by a convenient class width (a multiple of 5, or a number less than 5, if appropriate), and round up if there is any fraction, to find the number of intervals that would result. If the number of intervals is between 10 and 20, the width is probably reasonable; otherwise, you can try another convenient value for *i*. However, you may want to use fewer intervals if there are fewer than 100 scores in your distribution.
3. The lowest class interval must contain the lowest score in the distribution. In addition, it is highly desirable for the lower or upper limit to be a multiple of the chosen interval width.
4. In a grouped cumulative percentage distribution, the entry corresponding to a particular class interval is the percentile rank of the upper real limit of that interval. To find the PR of a score that is not at one of the upper real limits in your table, you can use linear interpolation. If the score is *X%* of the interval width above the lower limit of some interval, look at the PR for the upper and lower limits of that interval, and add *X%* of the difference of the two PRs to the lower one.
5. To find a percentile that does not appear as an entry in your table, first locate the two table entries for cumulative percentage that it is between—that will determine the interval that the percentile is in. You can then interpolate within that interval to estimate the percentile.
6. In a histogram for a grouped frequency distribution, the bars for each class interval extend from its lower to its upper real limit, and therefore neighboring bars touch each other. To create a polygon for a grouped distribution, place the dot over the midpoint of the class interval, and for a cumulative polygon (ogive), the dot is placed over the upper real limit of the interval.
7. The guidelines for graphs of frequency distributions that follow apply, for the most part, to other types of line graphs published in psychological journals.

a. The *Y* axis should be only about two-thirds as long as the *X* axis.
b. For frequency distributions, the variable of interest is placed along the *X* axis and the frequency counts (or relative frequency) are represented along the *Y* axis.
c. The measurement units are equally spaced along the entire length of both axes.
d. The intersection of the *X* and *Y* axes is the zero point for both dimensions.
e. Choose a scale to represent the measurement units on the graph so that the histogram or polygon fills the space of the graph as much as possible. Indicating a break in the scale on one or both axes may be necessary to achieve this goal.
f. Both axes should be clearly labeled, and the *X* axis should include the name of the variable and the unit of measurement.

EXERCISES

*1. The following are the IQ scores for the 50 sixth-grade students in Happy Valley Elementary school: 104, 111, 98, 132, 128, 106, 126, 99, 111, 120, 125, 106, 99, 112, 145, 136, 124, 130, 129, 114, 103, 121, 109, 101, 117, 119, 122, 115, 103, 130, 120, 115, 108, 113, 116, 109, 135, 121, 114, 118, 110, 136, 112, 105, 119, 111, 123, 115, 113, 117.
   a. Construct the appropriate grouped frequency distribution, and add *crf* and *cpf* columns (treat IQ as a continuous scale).
   b. Draw a frequency histogram to represent the above data.
   c. Estimate the first and third quartiles.
   d. Estimate the 40th and 60th percentiles.
   e. Estimate the percentile rank of a student whose IQ is 125.
   f. Estimate the percentile rank of a student whose IQ is 108.

*2. An industrial psychologist has devised an aptitude test for selecting employees to work as cashiers using a new computerized cash register. The aptitude test, on which scores can range from 0 to 100, has been given to 60 new applicants, whose scores were as follows: 83, 76, 80, 81, 74, 68, 92, 64, 95, 96, 55, 70, 78, 86, 85, 94, 76, 77, 82, 85, 81, 71, 72, 99, 63, 75, 76, 83, 92, 79, 82, 69, 91, 84, 87, 90, 80, 65, 84, 87, 97, 61, 73, 75, 77, 86, 89, 92, 79, 80, 85, 87, 82, 94, 90, 89, 85, 84, 86, 56.
   a. Construct a grouped frequency distribution table for the above data.
   b. Draw a frequency polygon to display the distribution of these applicants.
   c. Suppose the psychologist is willing to hire only those applicants who scored at the 80th percentile or higher (i.e., the top 20%). Estimate the appropriate cutoff score.
   d. Estimate the 75th and 60th percentiles.
   e. If the psychologist wants to use a score of 88 as the cutoff for hiring, what percentage of the new applicants will qualify?
   f. Estimate the percentile rank for a score of 81.

3. A telephone company is interested in the number of long-distance calls its customers make. Company statisticians randomly selected 40 customers and recorded the number of long-distance calls they made the previous month. They found the following results: 17, 0, 52, 35, 2, 8, 12, 28, 9, 43, 53, 39, 4, 21, 17, 47, 19, 13, 7, 32, 6, 2, 0, 45, 4, 29, 5, 10, 8, 57, 9, 41, 22, 1, 31, 6, 30, 12, 11, 20.
   a. Construct a grouped frequency distribution for the data.
   b. Draw a cumulative percentage polygon for these data.
   c. What percentage of customers made fewer than 10 long-distance calls?
   d. What is the percentile rank of a customer who made 50 calls?
   e. What percentage of customers made 30 or more calls?

*4. A state trooper, interested in finding out the proportion of drivers exceeding the posted speed limit of 55 mph, measured the speed of 25 cars in an hour. Their speeds in miles per hour were as follows: 65, 57, 49, 75, 82, 60, 52, 63, 49, 75, 58, 66, 54, 59, 72, 63, 85, 69, 74, 48, 79, 55, 45, 58, 51.

a. Create a grouped frequency distribution table for these data. Add columns for *cf* and *cpf*.
b. Approximately what percentage of the drivers were exceeding the speed limit?
c. Suppose the state trooper only gave tickets to those exceeding the speed limit by 10 mph or more. Approximately what proportion of these drivers would have received a ticket?
d. Estimate the 40th percentile.
e. Estimate the first and third quartiles.
f. What is the percentile rank of a driver going 62 mph?

*5. A psychologist is interested in the number of dreams people remember. She asked 40 participants to write down the number of dreams they remember over the course of a month and found the following results: 21, 15, 36, 24, 18, 4, 13, 31, 26, 28, 16, 12, 38, 26, 0, 13, 8, 37, 22, 32, 23, 0, 11, 33, 19, 11, 1, 24, 38, 27, 7, 14, 0, 13, 23, 20, 25, 3, 23, 26.
a. Create a grouped frequency distribution for these data with a class interval width of 5. Add columns for *cf* and *cpf*. (*Note*: Treat the number of dreams as a continuous variable.)
b. Draw a frequency histogram to display the distribution of the number of dreams remembered.
c. Suppose that the psychologist would like to select participants who remembered 30 or more dreams for further study. How many participants would she select? What proportion does this represent? What percentile rank does that correspond to?
d. Approximately what number of dreams corresponds to the 90th percentile?
e. What is the percentile rank of a participant who recalled 10 dreams?
f. What is the percentile rank of a participant who recalled 20 dreams?

6. Estimate all three quartiles for the data in the following table. (*Hint*: Each value for X can be assumed to represent a class that ranges from a half unit below to a half unit above the value shown; for example, $X = 16$ represents the range from 15.5 to 16.5.)

X	f	X	f
18	1	9	1
17	0	8	3
16	2	7	5
15	0	6	5
14	1	5	7
13	0	4	5
12	0	3	4
11	1	2	2
10	2	1	1

*7. Construct a grouped frequency distribution (width = 2) for the data in Exercise 2A8.
a. Add a *cpf* column and graph the cumulative percentage polygon.
b. Find the (approximate) values for all three quartiles.
c. Find the (approximate) values for the first and ninth deciles.
d. What is the (approximate) PR of a student who scored an 8 on the quiz?
e. What is the (approximate) PR of a student who scored an 18 on the quiz?

8. Redo Exercise 5 using a class interval width of 3. Discuss the similarities and differences between your answers to this exercise and your answers to Exercise 5. Describe the relative advantages and disadvantages of using a class interval of 3 for these data as compared to a width of 5.

Note: Some chapters will refer to exercises from previous sections of the chapter or from earlier chapters for purposes of comparison. A shorthand notation, consisting of the chapter number and section letter followed by the problem number, will be used to refer to exercises. For example, Exercise 3B2a refers to Chapter 3, Section B, Exercise 2, part a.

C

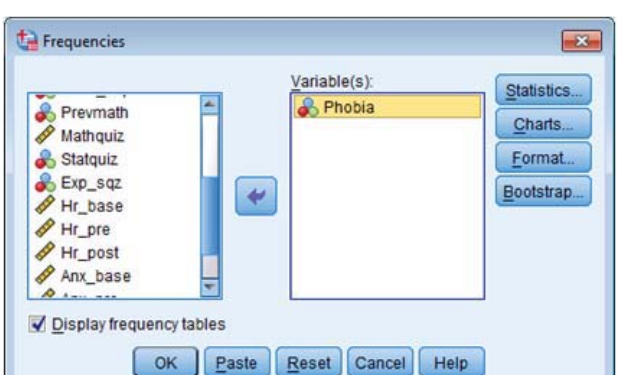
ANALYSIS BY SPSS

Creating Frequency Distributions

To create a frequency distribution, follow these six steps:

1. Select **Descriptive Statistics** from the **ANALYZE** menu, and click on **Frequencies**. . .
2. Move the variables for which you want to see frequency distributions into the *Variable(s)*: space (see Figure 2.11).

Figure 2.11



3. Click the **Statistics** button if you want to request percentiles or other summary statistics.
4. Click the **Charts** button if you want to request a bar chart, pie chart, or histogram.
5. Uncheck the little box labeled "Display frequency tables" if you selected a chart, and do not want to see a frequency table.
6. Click **OK** from the main **Frequencies** dialog box.

If you did not uncheck the little box labeled "Display frequency tables," then for each of the variables you moved into the *Variable(s)* space, you will get a table with five columns, the first of which contains every different score that was obtained in your data for that variable (not all possible scores). That is, SPSS gives you a regular frequency distribution, and does not create a grouped frequency distribution no matter how many different scores you have. The second column, Frequency, contains the number of times each of the different scores occurs (scores that have a frequency of zero just won't appear in this table at all). In the third column, Percent, the entry for Frequency is divided by the total number of cases (i.e., rows) in your spreadsheet, and then multiplied by 100. If there are no missing data for that variable, the column labeled Valid Percent will be identical to the one for Percent. The Cumulative Percent column is based on adding entries from the Valid Percent column, and its entry always tells you the percentage of cases in your spreadsheet that have a value less than or equal to the corresponding score in the leftmost column. Table 2.13 is the Frequency

Table 2.13

Phobia

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	12	12.0	12.0	12.0
	1	15	15.0	15.0	27.0
	2	12	12.0	12.0	39.0
	3	16	16.0	16.0	55.0
	4	21	21.0	21.0	76.0
	5	11	11.0	11.0	87.0
	6	1	1.0	1.0	88.0
	7	4	4.0	4.0	92.0
	8	4	4.0	4.0	96.0
	9	1	1.0	1.0	97.0
	10	3	3.0	3.0	100.0
	Total	100	100.0	100.0	

table created by SPSS for the variable Phobia from Ihno's data (note that Phobia was the selected variable in Figure 2.11).

Percentile Ranks and Missing Values

For instance, you can see from the table that the percentile rank (i.e., cumulative percentage) for a phobia score of 4 is 76. Note that had there been any missing data for Phobia the bottom row, Total, would have been followed by two more rows. Table 2.14 displays only the bottom few rows for the *mathquiz* variable, to illustrate how missing values are handled.

Table 2.14

mathquiz

		Frequency	Percent	Valid Percent	Cumulative Percent
	49	1	1.0	1.2	100.0
	Total	85	85.0	100.0	
Missing	System	15	15.0		
Total		100	100.0		

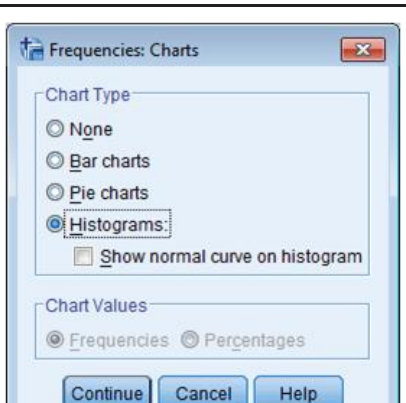
The first row labeled Total has a Frequency entry of 85, because the sum of the entries in the Frequency column will be 85—that's how many students had *mathquiz* scores. The next row indicates how many cases had missing data for that variable, and the last row tells you the total number of cases in your spreadsheet. You can see from the table that only one student scored a 49 on the quiz, which represents exactly 1 percent of the total cases (1/100 * 100), but the Valid Percent is 1/85 = .0118 * 100, which rounds off to 1.2, because one student represents about 1.2% of the students who actually received scores on the *mathquiz*. If the Cumulative Percent column were based on the Percent entries, instead of the Valid Percents, the student with the highest score would have a PR of only 85, rather than a 100, which would be misleading.

Graphing Your Distribution

You can uncheck the Display frequency tables box only if you select at least one option after clicking on either the Statistics or Charts buttons (otherwise SPSS will warn you that there will be no output). I will discuss one useful function of the Statistics button later in this section. For now, let's consider your choices, if you click on the **Charts** button.

The two Charts choices that are relevant to this chapter are *Bar charts* and *Histograms* (see Figure 2.12). If you select Bar charts, SPSS will create a graph based on a regular frequency distribution of your variable; class intervals will not be created, no matter how many different score values your data contain. Moreover, a Bar chart will not only treat your variable as discrete (inserting slim spaces between adjacent bars), but as though it were measured on a nominal or ordinal scale. For instance, no place is held for a value within your variable's range that has zero frequency (e.g., if three students each took one, two, and four prior math courses, but no student took three math courses, you would see three equally high and equally spaced bars, with no extra gap to represent the zero frequency for three prior math courses taken). Selecting Bar charts gives you two choices with respect to the scaling of the vertical axis: frequencies (the default choice), and percentages. The relative heights of the bars will look the same, but if you choose percentages the Y axis will be marked off to correspond with the

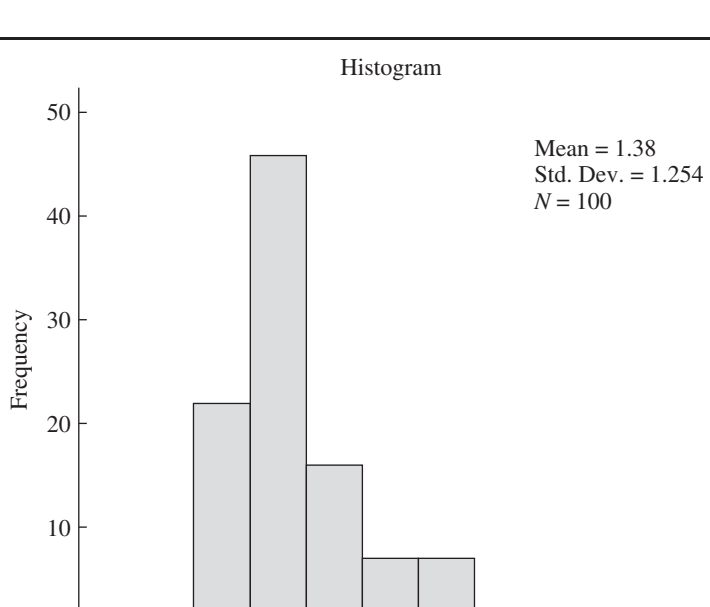
Figure 2.12



fact that the frequencies are being divided by the valid (i.e., *non*missing) *N* and multiplied by 100.

If your variable has been measured on a scale that can be considered quantitative (interval or ratio, and in some cases, ordinal), you will most likely want to choose Histograms, instead of Bar charts. If you choose Histograms for your selected variables, each variable will be treated as though measured on an interval/ratio scale: adjacent bars will touch, and if there are many different values, they will be grouped into convenient class intervals (a full bar-width will be left for each empty class interval within your range of scores). However, the bars of the histogram will be labeled in terms of the midpoints of the intervals; the real limits of the intervals are not shown. A histogram for the *prevmath* variable is shown in Figure 2.13.

Figure 2.13



Obtaining Percentiles

When you click on the **Statistics** button, the upper-left quadrant of the **Frequencies: Statistics** box that opens (see Figure 2.14) presents three choices for obtaining percentiles. The topmost choice, *Quartiles*, will give you, of course, the 25th, 50th, and 75th percentiles. The second choice, *Cut points for . . .*, will give you the deciles, if you use the default number of 10. If you change that number to 5, for instance, you will obtain the 20th, 40th, 60th, and 80th percentiles. The third choice, *Percentile(s):*, allows you to specify any number of particular percentiles that you would like to see—just click "Add" after typing in each one. Click Continue to return to the main Frequencies dialog box, and then click OK. You will get a table of all of the percentiles you requested in numerical order (e.g., if you requested quartiles, as well as the particular percentiles 35 and 60, the table will list the scores corresponding to the 25th, 35th, 50th, 60th, and 75th percentiles, in that order). I will discuss the other choices in the Statistics box in the next chapter. At the end of this section, I consider an interesting alternative to the histogram for observing the shape of your distribution.

The Split File Function

It is not uncommon to want to look at the distribution of a variable separately for important subgroups in your data. For instance, you may want to look at the (math) phobia distribution separately for the male and female students. A general way to perform any SPSS analysis separately for subgroups that are identified by a variable in your data set is to use the **Split File** function. You can open the Split File dialog box by first clicking on **Data**, and then selecting *Split File . . .* from the drop-down menu (third from the bottom). When you first open the Split File box, the topmost of the three choices (see Figure 2.15)—*Analyze all cases, do not create groups*—will already be checked. Either of the other two choices will turn on Split File; later, you can turn off the Split File function by checking the top choice. (Note that it is easy to forget that Split File is on, because it is indicated only in a small area below the lower-right corner of the spreadsheet.)

Figure 2.14

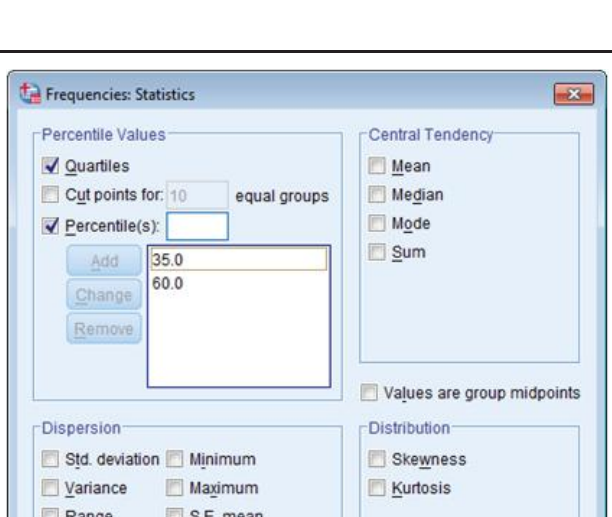
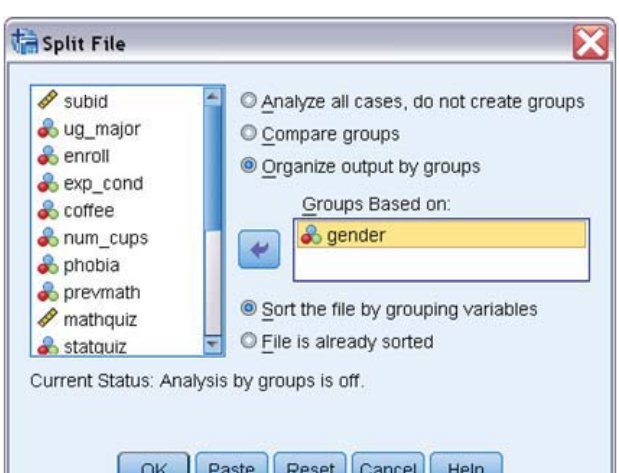


Figure 2.15



Before you can select one of your spreadsheet variables as the basis for splitting your file, you have to make a choice about how your output will be displayed. If you check *Organize output by groups*, SPSS will create a full set of results for the first level of your grouping variable, and then repeat that entire set of results for the second level, and so on. If you check *Compare groups* instead, SPSS will repeat each portion of the results for all levels of your grouping variable before presenting the next portion of the results. Obviously, this choice won't make any difference if your output contains only one box of results. Only after you select one of those choices can you move the grouping variable of interest (e.g., gender) to the *Groups Based on . . .* space (see Figure 2.15).

Stem-and-Leaf Plots

J. W. Tukey (1977) is well known for urging psychologists to engage in a more extensive inspection of their data than they usually do, before moving on to inferential statistics. He called this detailed inspection exploratory data analysis (EDA), and he provided a number of straightforward and useful methods with which to perform EDA. One of those methods serves as a reasonable alternative to the histogram; it is called the **stem-and-leaf display**, or *stemplot*, for short. I did not discuss stemplots earlier in this chapter, so I will first show you an example of one before telling you how I obtained it from SPSS. Figure 2.16 contains the 85 scores from the *mathquiz* variable.

To construct a stemplot by hand, you would first write down the *stems* in a vertical column. For two-digit numbers, it makes sense to use the first digits as the stems, and the second digits as the *leaves*. Because the *mathquiz* scores range from 9 (first digit is considered to be zero) to 49, the stems would be the digits from 0 to 4. However, as in the case of a grouped frequency distribution, having only five intervals (i.e., stems) does not give much detail, so it is desirable to double the number of stems. For instance, in Figure 2.16 you will see that there are two stems labeled by the number 2. The first of these is used to hold the "leaves" ranging from 0 to 4, and the second one, 5 to 9.

You can see at a glance that the big advantage of the stemplot over the histogram is that all of the original data are still visible. For example, by

Figure 2.16

```
mathquiz Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1.00        0 .  9
     7.00        1 .  0111344
     5.00        1 .  55588
    11.00        2 .  00111122234
    17.00        2 .  55666666688889999
    20.00        3 .  00000111222222333344
    13.00        3 .  5557777888999
     7.00        4 .  0112334
     4.00        4 .  5679

 Stem width:        10
 Each leaf:       1 case(s)
```

looking at the first stem labeled 2 and its leaves, you can see that the data contain the following scores: 20, 20, 21, 21, 21, 21, 22, 22, 22, 23, and 24. The column labeled frequency tells you that that stem has 11 scores, but that column is not a necessary part of the stemplot; you can get that information by counting the leaves. The stemplot also provides the main advantage of the histogram by displaying the shape of the distribution, though you may want to rotate the stemplot to look more like the typical histogram. In the preceding figure, you can see that the distribution is bell-shaped with its peak near the middle, though with a slight negative skew. Depending on the range of your scores, how many digits they each contain, and how many scores there are in total, there are different schemes to make the stemplot easy to interpret at a glance. If you want SPSS to make the decisions and create the stemplot for you, you will have to use an alternative to the **Frequencies** subprogram called **Explore**.

To create stem-and-leaf displays:

1. Select **Descriptive Statistics** from the **ANALYZE** menu, and click on **Explore** . . .
2. Move the variables for which you want to see stemplots into the space labeled *Dependent List*. If you do *not* want to see descriptive statistics for those variables, select *Plots* rather than *Both* in the section labeled "Display."
3. Click the **Plots** button.
4. In the upper-right section (labeled "Descriptive") of the **Explore: Plots** box make sure that *Stem-and-leaf* has already been selected (it is one of the defaults). Select *None* in the upper-left section (labeled "Boxplots"), if you do not want this (default) option (explained in the next chapter), and then click **Continue**.
5. Click **OK** from the main **Explore** dialog box.

Note that you can create separate stem-and-leaf displays for each level of a categorical variable (e.g., male and female) by moving the categorical variable (e.g., *gender*) into the *Factor List*, which is just under the *Dependent List* in the **Explore** dialog box. This is a convenient alternative to using the **Split File** function for this particular procedure. The **Explore** dialog box has a number of other useful functions, especially for evaluating the shape of your sample's distribution, which we explore in subsequent chapters.

EXERCISES

1. Request a frequency distribution and a bar chart for the Undergraduate Major variable for Ihno's students.
2. Repeat Exercise 1 for the variables *prevmath* and *phobia*. Would it make sense to request a histogram instead of a bar chart for *phobia*? Discuss.
3. Request a frequency distribution and a histogram for the variable *statquiz*. Describe the shape of this distribution.
4. Request a frequency distribution and a histogram for the variables baseline anxiety (*anx_base*) and baseline heart rate (*hr_base*). Comment on SPSS's choice of class intervals for each histogram.
5. Request stem-and-leaf displays for the variables *anx_base* and *hr_base*.
6. Request stem-and-leaf plots and histograms for the variables *anx_base* and *hr_base* divided by *gender*.
7. Request the deciles for the variable *statquiz*.
8. Request the quartiles for the variables *anx_base* and *anx_pre*.
9. Request the deciles and quartiles for the *phobia* variable.
10. Request the following percentiles for the variables *hr_base* and *hr_pre*: 15, 30, 42.5, 81, and 96.