MEASURES OF CENTRAL TENDENCY AND VARIABILITY

3 Chapter

You will need to use the following from previous chapters:

Symbols
Σ: Summation sign

Concepts
Scales of measurement
Frequency histograms and polygons

Procedures
Rules for using the summation sign

A CONCEPTUAL FOUNDATION

In Chapter 2, I began with an example in which I wanted to tell a class of 25 students how well the class had performed on a diagnostic quiz and make it possible for each student to evaluate how his or her score compared to the rest of the class. As I demonstrated, a simple frequency distribution, especially when graphed as a histogram or a polygon, displays the scores at a glance, and a cumulative percentage distribution makes it easy to find the percentile rank for each possible quiz score. However, the first question a student is likely to ask about the class performance is, "What is the average for the class?" And it is certainly a question worth asking. Although the techniques described in Chapter 2 provide much more information than does a simple average for a set of scores, the average is usually a good summary of that information. In trying to find the average for a group of scores, we are looking for one spot that seems to be the center of the distribution. Thus, we say that we are seeking the *central tendency* of the distribution. But, as the expression "central tendency" implies, there may not be a single spot that is clearly and precisely at the center of the distribution. In fact, there are several procedures for finding the central tendency of a group of scores, and which procedure is optimal can depend on the shape of the distribution involved. I will begin this chapter by describing the common ways that central tendency can be measured and the reasons for choosing one measure over another. Then, I will explain how central tendency measures can be used as a basis from which to quantify the variability of a distribution. Finally, I will consider some more advanced measures for assessing the shape of a distribution.

Measures of Central Tendency

The Arithmetic Mean

When most students ask about the average on an exam, they have in mind the value that is obtained when all of the scores are added and then divided by the total number of scores. Statisticians call this value the *arithmetic mean*, and it is symbolized by the Greek letter μ (mu, pronounced "myoo") when it refers to the mean of a population. Later in this chapter, we will also be interested in the mean for a sample, in which case the mean is symbolized either by a bar over the letter representing the variable (e.g., $\overline{X}$, called "*X* bar") or by the capital letter *M*, for mean. There are other types of means, such as the harmonic mean (which will be introduced in

Chapter 8) and the geometric mean, but the arithmetic mean is by far the most commonly used. Therefore, when I use the terms *mean* or *average* without further specification, it is the arithmetic mean to which I am referring. The arithmetic mean, when applicable, is undoubtedly the most useful measure of central tendency. However, before we consider the many statistical properties of the mean, we need to consider two lesser known, but nonetheless useful, measures of central tendency.

The Mode

Often the main purpose in trying to find a measure of central tendency is to characterize a large group of scores by one value that could be considered the most typical of the group. If you want to know how smart a class of students is (perhaps because you have to prepare to teach them), you would like to know how smart the typical student in that class is. If you want to know how rich a country is, you might want to know the annual income of a typical family. The simplest and crudest way to define the most typical score in a group is in terms of which score occurs with the highest frequency. That score is called the *mode* of the distribution.

The mode is easy to find once you have constructed a frequency distribution; it is the score that has the highest frequency. It is perhaps even easier to identify the mode when a frequency distribution has been displayed as a histogram or a graph. Simply look for the highest bar in the histogram or the highest point in the polygon—the score that is directly below that highest bar or point is the mode. The mode is defined in the same way for a grouped distribution as for a simple distribution, except with a grouped distribution the mode is the most frequently occurring *interval* (or the midpoint of that interval) rather than a single score. One potential drawback of using the mode with grouped distributions is that the mode depends a good deal on the way the scores are grouped (i.e., on your choice for the lowest interval and your choice for the width of the interval). However, even with a simple frequency distribution the mode has its problems, as I will show next.

Table 3.1

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

Advantages and Disadvantages of the Mode A major disadvantage of the mode is that it is not a very reliable measure of central tendency. Consider the simple frequency distribution on the left side of Table 3.1. The mode of that distribution is 7 because that score has the highest frequency (6). However, if just one of the students who scored a 7 was later found really to have scored a 4, that frequency distribution would change to the one on the right side of Table 3.1, and the mode would consequently move from a score of 7 to a score of 4. Naturally, we would like to see more stability in our measures of central tendency. Moreover, if the score with a frequency of 6 in either distribution in Table 3.1 had a frequency of 5, there would be a whole range of scores at the mode, which would make the mode a rather imprecise measure.

Imagine that in either of the distributions in Table 3.1, the scores of 4 and 7 *both* have a frequency of 6. Now the distribution has more than one mode. If a distribution has many modes, finding these modes is not likely to be useful. However, a distribution that contains two distinct subgroups (e.g., men and women measured on the amount of weight they can lift over their heads) may have two meaningful modes (one for each subgroup), as shown in Figure 3.1. Such a distribution is described as *bimodal*. If a distribution has two or three distinct modes (it is hard to imagine a realistic situation with more modes), finding these modes can be useful indeed, and the

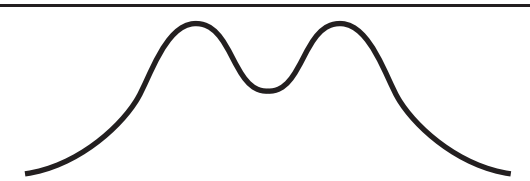


Figure 3.1

A Bimodal Distribution

modes would provide information not available from the more commonly used mean or median. The most common shape for a smooth, or nearly smooth, distribution involves having only one mode. Such distributions are described as *unimodal* and are the only types of distributions that we will be dealing with in this text.

When dealing with interval/ratio scales, it seems that the main advantage of the mode as a measure of central tendency is in terms of distinguishing multimodal from unimodal distributions. The ease with which the mode can be found used to be its main advantage, but in the age of high-speed computers, this is no longer a significant factor. However, the mode has the unique advantage that it can be found for any kind of measurement scale. In fact, when dealing with nominal scales, other measures of central tendency (such as the mean) cannot be calculated; the mode is the *only* measure of central tendency in this case. For instance, suppose you are in charge of a psychiatric emergency room and you want to know the most typical diagnosis of a patient coming for emergency treatment. You cannot take the average of 20 schizophrenics, 15 depressives, and so forth. All you can do to assess central tendency is to find the most frequent diagnosis (e.g., schizophrenia may be the *modal* diagnosis in the psychiatric emergency room).

The Median

If you are looking for one score that is in the middle of a distribution, a logical score to focus on is the score that is at the 50th percentile (i.e., a score whose PR is 50). This score is called the *median*. The median is a very useful measure of central tendency, as you will see, and it is very easy to find. If the scores in a distribution are arranged in an array (i.e., in numerical order), and there are an *odd* number of scores, the median is literally the score in the middle. If there are an *even* number of scores, as in the distribution on the left side of Table 3.1 $\left(N = \sum f = 30\right)$, the median is the average of the two middle scores (as though the scores were measured on an interval/ratio scale). For the left distribution in Table 3.1, the median is the average of 5 and 6, which equals 5.5.

The Median for Ordinal Data Unlike the mode, the median cannot be found for a nominal scale because the values (e.g., different psychiatric diagnoses) do not have any inherent order (e.g., we cannot say which diagnoses are "above" bipolar disorder and which "below"). However, if the values can be placed in a meaningful order, you are then dealing with an ordinal scale, and the median *can* be found for ordinal scales. For example, suppose that the coach of a debating team has rated the effectiveness of the 25 members of the team on a scale from 1 to 10. The data in Table 3.2 (reproduced here) could represent those ratings.

Once the scores have been placed in order, the median is the middle score. (Unfortunately, if there are two middle scores and you are dealing

Table 3.2

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

with ordinal data, it is not proper to average the two scores, although this is often done anyway as an approximation.) Even though the ratings from 1 to 10 cannot be considered equally spaced, we can assume, for example, that the debaters rated between 1 and 5 are all considered less effective than one who is rated 6. Thus, we can find a ranking or rating such that half the group is below it and half above, except for those who are tied with the middle score (or one of the two middle scores). The median is more informative if there are not many ties. In general, having many tied scores diminishes the usefulness of an ordinal scale.

Dealing With Undeterminable Scores and Open-Ended Categories One situation that is particularly appropriate for the use of the median occurs when the scores for some subjects cannot be determined exactly, but we know on which end of the scale those scores fall. For instance, in a typical study involving reaction time (RT), an experimenter will not wait forever for a subject to respond. Usually some arbitrary limit is imposed on the high end—for example, if the subject does not respond after 10 seconds, record 10 s as the RT and go on to the next trial. Calculating the mean would be misleading, however, if any of these 10-second responses were included. First, these 10-second responses are really *undeterminable scores*—the researcher doesn't know how long it would have taken for the subject to respond. Second, averaging in a few 10-second responses with the rest of the responses, which may be less than 1 second, can produce a mean that misrepresents the results. On the other hand, the median will not change if the response is recorded as 10 or 100 s (assuming that the median is less than 10 s to begin with). Thus, when some of the scores are undeterminable, the median has a strong advantage over the mean as a descriptive statistic.

Sometimes when data are collected for a study, some of the categories are deliberately left *open ended*. For instance, in a study of AIDS awareness, subjects might be asked how many sexual partners they have had in the past 6 months, with the highest category being 10 or more. Once a subject has had at least 10 different partners in the given period, it may be considered relatively unimportant to the study to determine exactly how many more than 10 partners were involved. (Perhaps the researchers fear that the accuracy of numbers greater than 10 could be questioned.) However, this presents the same problem for calculating the mean as an undeterminable score. It would be misleading to average in the number 10 when the subject reported having 10 *or more* partners. Again, this is not a problem for finding the median; all of the subjects reporting 10 or more partners would simply be tied for the highest position in the distribution. (A problem in determining the median would arise only if as many as half the subjects reported 10 or more partners.)

The Median and the Area of a Distribution As mentioned, the mode is particularly easy to find from a frequency polygon—it is the score that corresponds to the highest point. The median also bears a simple relationship to the frequency polygon. If a vertical line is drawn at the median on a frequency polygon so that it extends from the horizontal axis until it meets the top of the frequency polygon, the area of the polygon will be divided in half. This is because the median divides the total number of scores in half, and the area of the polygon is proportional to the number of scores.

To better understand the relation between the frequency of scores and the area of a frequency polygon, take another look at a frequency histogram

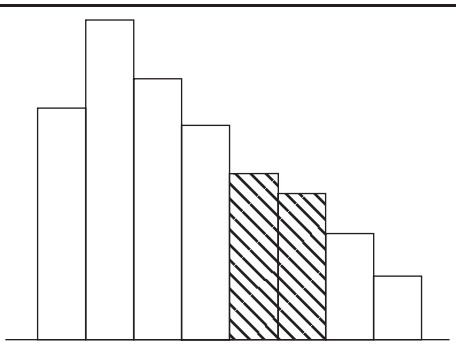


Figure 3.2

Area of a Frequency Histogram

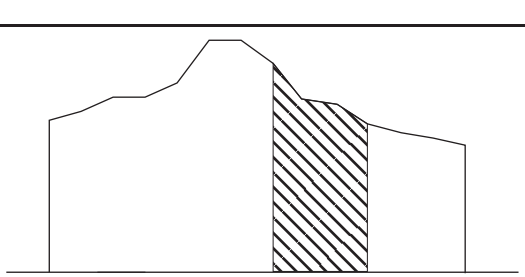


Figure 3.3

Area of a Frequency Polygon

(see Figure 3.2). The height of each bar in the histogram is proportional to the frequency of the score or interval that the bar represents. (This is true for the simplest type of histogram, which is the only type we will consider.) Because the bars all have the same width, the area of each bar is also proportional to the frequency. You can imagine that each bar is a building and that the taller the building, the more people live in it. The entire histogram can be thought of as the skyline of a city; you can see at a glance where (in terms of scores on the *X* axis) the bulk of the people live. All the bars together contain all the scores in the distribution. If two of the bars, for instance, take up an area that is 20% of the total, you know that 20% of the scores fall in the intervals represented by those two bars.

A relationship similar to the one between scores and areas of the histogram bars can be observed in a frequency polygon. The polygon encloses an area that represents the total number of scores. If you draw two vertical lines within the polygon, at two different values on the *X* axis, you enclose a smaller area, as shown in Figure 3.3. Whatever proportion of the total area is enclosed between the two values (.20 in Figure 3.3) is the proportion of the scores in the distribution that fall between those two values. We will use this principle to solve problems in the next chapter. At this point I just wanted to give you a feeling for why a vertical line drawn at the median divides the distribution into two equal areas.

Measures of Variability

Finding the right measure of central tendency for a distribution is certainly important, and I will have more to say about this process with respect to the shape of the distribution, but there is another very important aspect of describing a set of data that I do not want to postpone any longer.

The following hypothetical situation will highlight the importance of this other dimension.

Suppose you're an eighth-grade English teacher entering a new school, and the principal is giving you a choice of teaching either class A or class B. Having read this chapter thus far, you inquire about the mean reading level of each class. (To simplify matters you can assume that the distributions of both classes are unimodal.) The principal tells you that class A has a mean reading level of 8.0, whereas class B has a mean of 8.2. All else being equal, you are inclined to take the slightly more advanced class. But all is not equal. Look at the two distributions in Figure 3.4.

What the principal neglected to mention is that reading levels in class B are much more spread out. It should be obvious that class A would be easier to teach. If you geared your lessons toward the 8.0 reader, no one in class A is so much below that level that he or she would be lost, nor is anyone so far above that level that he or she would be completely bored. On the other hand, teaching class B at the 8.2 level could leave many students either lost or bored.

The fact is that no measure of central tendency is very representative of the scores, if the distribution contains a great deal of variability. The principal could have shown you both distributions to help you make your decision; the difference in variability (also called the *dispersion*) is so obvious that if you had seen the distributions you could have made your decision instantly. For less obvious cases, and for the purposes of advanced statistical techniques, it would be useful to measure the width of each distribution. However, there is more than one way to measure the spread of a distribution. The rest of this section is mainly about the different ways of measuring variability.

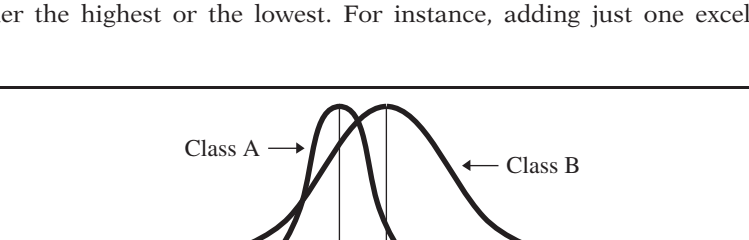
The Range

The simplest and most obvious way to measure the width of a distribution is to subtract the lowest score from the highest score. The resulting number is called the *range* of the distribution. For instance, judging Figure 3.4 by eye, in class A the lowest reading score appears to be about 7.6 and the highest about 8.4. Subtracting these two scores we obtain 8.4 − 7.6 = .8. However, if these scores are considered to be measured on a continuous scale, we should subtract the lower real limit of 7.6 (i.e., 7.55) from the upper real limit of 8.4 (i.e., 8.45) to obtain 8.45 − 7.55 = .9. For class B, the lowest and highest scores appear to be 7.1 and 9.3, respectively, so the range would be 9.35 − 7.05 = 2.3—considerably larger than the range for class A.

The major drawback to the range as a measure of variability is that, like the mode, it can be quite unreliable. The range can be changed drastically by moving only one score in the distribution, if that score happens to be either the highest or the lowest. For instance, adding just one excellent

Figure 3.4

Mean Reading Levels in Two Eighth-Grade Classes



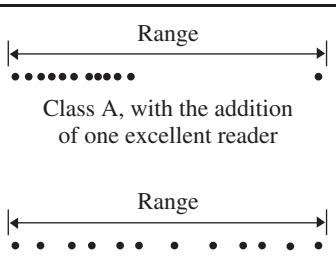


Figure 3.5

The Ranges of Two Different Distributions

reader to class A can make the range of class A as large as the range of class B. But the range of class A would then be very misleading as a descriptor of the variability of the bulk of the distribution (see Figure 3.5). In general, the range will tend to be misleading whenever a distribution includes a few extreme scores (such scores are usually referred to as *outliers*). Another drawback to the range is that it cannot be determined for a distribution that contains undeterminable scores at one end or the other.

On the positive side, the range not only is the easiest measure of variability to find, it also has the advantage of capturing the entire distribution without exception. For instance, in designing handcuffs for use by police departments, a manufacturer would want to know the entire range of wrist sizes in the adult population so that the handcuffs could be made to adjust over this range. It would be important to make the handcuffs large enough so that no wrist would be too large to fit but able to become small enough so that no adult could wriggle out and get free.

The Semi-Interquartile Range

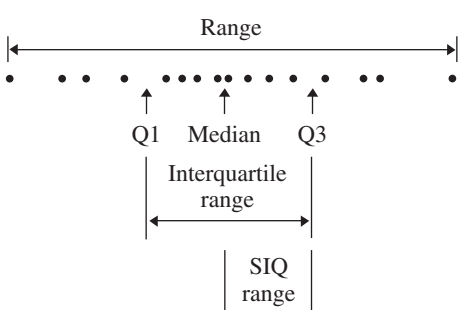
There is one measure of variability that can be used with open-ended distributions and is virtually unaffected by extreme scores because, like the median, it is based on percentiles. It is called the *interquartile* (*IQ*) *range*, and it is found by subtracting the 25th percentile from the 75th percentile. The 25th percentile is often called the first quartile and symbolized as Q1; similarly, the 75th percentile is known as the third quartile (Q3). Thus the interquartile range (IQ) can be symbolized as Q3 − Q1. The IQ range gives the width of the middle half of the distribution, therefore avoiding any problems caused by outliers or undeterminable scores at either end of the distribution. A more popular variation of the IQ range is the *semi-interquartile* (*SIQ*) *range*, which is simply half of the interquartile range, as shown in Formula 3.1:

$$\text{SIQ range} = \frac{\text{Q3} - \text{Q1}}{2} \qquad \textbf{Formula 3.1}$$

The SIQ range is preferred because it gives the distance of a typical score from the median; that is, roughly half the scores in the distribution will be closer to the median than the length of the SIQ range, and about half will be further away. The SIQ range is often used in the same situations for which the median is preferred to the mean as a measure of central tendency, and it can be very useful for descriptive purposes. However, the SIQ range's chief advantage—its unresponsiveness to extreme scores—can also be its chief disadvantage. Quite a few scores on both ends of a distribution can be moved much further from the center without affecting the SIQ range.

Figure 3.6

The Interquartile and Semi-Interquartile Ranges



Thus the SIQ range does not always give an accurate indication of the width of the entire distribution (see Figure 3.6). Moreover, the SIQ range shares with the median the disadvantage of not fitting easily into advanced statistical procedures.

The Mean Deviation

The SIQ range can be said to indicate the typical distance of a score from the median. This is a very useful way to describe the variability of a distribution. For instance, if you were teaching an English class and were aiming your lessons at the middle of the distribution, it would be helpful to know how far off your teaching level would be, on the average. However, the SIQ range does not take into account the distances of *all* the scores from the center. A more straightforward approach would be to find the distance of every score from the middle of the distribution and then average those distances. Let us look at the mathematics involved in creating such a measure of variability.

First, we have to decide on a measure of central tendency from which to calculate the distance of each score. The median would be a reasonable choice, but because we are developing a measure to use in advanced statistical procedures, the mean is preferable. The distance of any score from the mean $(X_i - \mu)$ is called a *deviation score*. (A deviation score is sometimes symbolized by a lowercase x; but in my opinion that notation is too confusing, so it will not be used in this text.) The average of these deviation scores would be given by $\sum(X_i - \mu)/N$. Unfortunately, there is a problem with using this expression. According to one of the properties of the mean (these properties will be explained more fully in Section B), $\sum(X_i - \mu)$ will always equal zero, which means that the average of the deviation scores will also always equal zero (about half the deviations will be above the mean and about half will be below). This problem disappears when you realize that it is the distances we want to average, regardless of their direction (i.e., sign). What we really want to do is take the *absolute values* of the deviation scores before averaging to find the typical amount by which scores deviate from the mean. (Taking the absolute values turns the minus signs into plus signs and leaves the plus signs alone; in symbols, $|X|$ means take the absolute value of X.) This measure is called the *mean deviation*, or more accurately, the mean absolute deviation (MAD), and it is found using Formula 3.2:

$$\text{Mean deviation} = \frac{\sum |X_i - \mu|}{N} \qquad \textbf{Formula 3.2}$$

To clarify the use of Formula 3.2, I will find the mean deviation of the following three numbers: 1, 3, 8. The mean of these numbers is 4. Applying Formula 3.2 yields:

$$\frac{|1-4|+|3-4|+|8-4|}{3}=\frac{|-3|+|-1|+|+4|}{3}=\frac{3+1+4}{3}=\frac{8}{3}=2.67$$

The mean deviation makes a lot of sense, and it should be easy to understand; it is literally the average amount by which scores deviate from the mean. It is too bad that the mean deviation does not fit in well with more advanced statistical procedures. Fortunately, there is a measure that is closely related to the mean deviation that does fit well with the statistical procedures that are commonly used. I will get to this measure soon. First, another intermediate statistic must be described.

The Variance

If you square all the deviations from the mean, instead of taking the absolute values, and sum all of these squared deviations together, you get a quantity called the *sum of squares* (*SS*), which is less for deviations around the mean than for deviations around any other point in the distribution. (Note that the squaring eliminates all the minus signs, just as taking the absolute values did.) Formula 3.3 for *SS* is:

$$SS=\sum(X_i-\mu)^2 \qquad \textbf{Formula 3.3}$$

If you divide *SS* by the total number of scores (*N*), you are finding the mean of the squared deviations, which can be used as a measure of variability. The mean of the squared deviations is most often called the *population variance*, and it is symbolized by the lowercase Greek letter sigma squared (σ^2; the uppercase sigma, $\sum$, is used as the summation sign). Formula 3.4A for the variance is as follows:

$$\sigma^2=\frac{\sum(X_i-\mu)^2}{N} \qquad \textbf{Formula 3.4A}$$

Because the variance is literally the mean of the squared deviations from the mean, it is sometimes referred to as a mean square, or *MS* for short. This notation is commonly used in the context of the analysis of variance procedure, as you will see in Part IV of this text. Recall that the numerator of the variance formula is often referred to as *SS*; the relationship between *MS* and *SS* is expressed in Formula 3.5A:

$$\sigma^2=MS=\frac{SS}{N} \qquad \textbf{Formula 3.5A}$$

It is certainly worth the effort to understand the variance because this measure plays an important role in advanced statistical procedures, especially those included in this text. However, it is easy to see that the variance does not provide a good descriptive measure of the spread of a

distribution. As an example, consider the variance of the numbers 1, 3, and 8:

$$\sigma^2 = \frac{(1-4)^2 + (3-4)^2 + (8-4)^2}{3}$$
$$= \frac{3^2 + 1^2 + 4^2}{3} = \frac{9+1+16}{3} + \frac{26}{3} = 8.67$$

The variance (8.67) is larger than the range of the numbers. This is because the variance is based on *squared* deviations. The obvious remedy to this problem is to take the square root of the variance, which leads to our final measure of dispersion.

The Standard Deviation

Taking the square root of the variance produces a measure that provides a good description of the variability of a distribution and one that plays a role in advanced statistical procedures as well. The square root of the population variance is called the *population standard deviation* (*SD*), and it is symbolized by the lowercase Greek letter sigma (σ). (Notice that the symbol is *not* squared—squaring the standard deviation gives the variance.) The basic definitional formula for the standard deviation is:

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} \qquad \textbf{Formula 3.4B}$$

An alternative way to express this relationship is:

$$\sigma = \sqrt{MS} = \sqrt{\frac{SS}{N}} \qquad \textbf{Formula 3.5B}$$

To remind you that each formula for the standard deviation will be the square root of a variance formula, I will use the same number for both formulas, adding "A" for the variance formula and "B" for the corresponding *SD* formula. Because σ is the square root of *MS*, it is sometimes referred to as the *root-mean-square* (RMS) of the deviations from the mean.

At this point, you may be wondering why you would bother squaring all the deviations if after averaging you plan to take the square root. First, we need to make it clear that squaring, averaging, and then taking the square root of the deviations is not the same as just averaging the absolute values of the deviations. If the two procedures were equivalent, the standard deviation would always equal the mean deviation. An example will show that this is not the case. The standard deviation of the numbers 1, 3, and 8 is equal to the square root of their variance, which was found earlier to be 8.67. So, $\sigma = \sqrt{8.67} = 2.94$, which is clearly larger than the mean deviation (2.67) for the same set of numbers.

The process of squaring and averaging gives extra weight to large scores, which is not removed by taking the square root. Thus, the standard deviation is never smaller than the mean deviation, although the two measures can be equal. In fact, the standard deviation will be equal to the mean deviation whenever there are only two numbers in the set. In this case, both measures of variability will equal half the distance between the two numbers. I mentioned previously that the standard deviation gives more weight to large scores than does the mean deviation. This is true because

squaring a large deviation has a great effect on the variance. This sensitivity to large scores can be a problem if there are a few very extreme scores in a distribution, which result in a misleadingly large standard deviation. If you are dealing with a distribution that contains a few extreme scores (whether low, high, or some of each), you may want to consider an alternative to the standard deviation, such as the mean deviation, which is less affected by extreme scores, or the semi-interquartile range, which may not be affected at all. On the other hand, you could consider a method for eliminating outliers or transforming the data, such as those outlined in Section B.

The Variance of a Sample

Thus far the discussion of the variance and standard deviation has been confined to the situation in which you are describing the variability of an entire population of scores (i.e., your interests do not extend beyond describing the set of scores at hand). Later chapters, however, will consider the case in which you have only a sample of scores from a larger population, and you want to use your description of the sample to extrapolate to that population. Anticipating that need, I will now consider the case in which you want to describe the variability of a sample.

To find the variance of a sample, you can use the procedure expressed in Formula 3.4A, but it will be appropriate to change some of the notation. First, I will use s^2 to symbolize the sample variance, according to the custom of using Roman letters for sample statistics. Along these lines, the mean subtracted from each score will be symbolized as $\overline{X}$ instead of μ, because it is the mean of a sample. Thus Formula 3.4A becomes:

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{N}$$

The Biased and Unbiased Sample Variances The preceding formula represents a perfectly reasonable way to describe the variability in a sample, but a problem arises when the variance thus calculated is used to estimate the variance of the larger population. The problem is that the variance of the sample tends to underestimate the variance of the population. Of course, the variance of every sample will be a little different, even if all of the samples are the same size and they are from the same population. Some sample variances will be a little larger than the population variance and some a little smaller, but unfortunately the average of infinitely many sample variances (when calculated by the formula above) will be *less* than the population variance. This tendency of a sample statistic to consistently underestimate (or overestimate) a population parameter is called *bias*. The sample variance as defined by the (unnumbered) formula above is therefore called a *biased estimator*.

Fortunately, the underestimation just described is so well understood that it can be corrected easily by making a slight change in the formula for calculating the sample variance. To calculate an *unbiased sample variance*, you can use Formula 3.6A:

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{n - 1} \qquad \textbf{Formula 3.6A}$$

If infinitely many sample variances are calculated with Formula 3.6A, the average of these sample variances *will* equal the population variance σ^2.

Note that I used a lowercase n in the preceding formula to remind you that this formula is designed to be used on a sample and not on a population. If a formula is intended for a population, or could just as easily apply to a population as a sample, I'll use an uppercase N.

Notation for the Variance and the Standard Deviation

You've seen that there are two different versions of the variance of a sample: biased and unbiased. Some texts use different symbols to indicate the two types of sample variances, such as an uppercase S for biased and a lowercase s for unbiased, or a plain s for biased and $\hat{s}$ (pronounced "s hat") for unbiased. I will adopt the simplest notation by assuming that the variance of a sample will always be calculated using Formula 3.6A (or its algebraic equivalent). Therefore, the symbol s^2 for the sample variance will always (in my text, at least) refer to the *unbiased* sample variance. Whenever the biased formula is used (i.e., the formula with N or n rather than $n-1$ in the denominator), you can assume that the set of numbers at hand is being treated like a population, and therefore the variance will be identified by σ^2. When you are finding the variance of a population, you are never interested in extrapolating to a larger group, so there would be no reason to calculate an unbiased variance. Thus when you see σ^2, you know that it was obtained by Formula 3.4A (or its equivalent), and when you see s^2, you know that Formula 3.6A (or its equivalent) was used.

As you might guess from the preceding discussion, using Formula 3.4B to find the standard deviation of a sample produces a biased estimate of the population standard deviation. The solution to this problem would seem to be to use the square root of the unbiased sample variance whenever you are finding the standard deviation of a sample. This produces a new formula for the standard deviation:

$$s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}} \qquad \textbf{Formula 3.6B}$$

Surprisingly, this formula does not entirely correct the bias in the standard deviation, but fortunately the bias that remains is small enough to be ignored (at least that is what researchers in psychology do). Therefore, I will refer to s (defined by Formula 3.6B) as the *unbiased sample standard deviation*, and I will use σ (defined by Formula 3.4B) as the symbol for the standard deviation of a population.

Degrees of Freedom

The adjustment in the variance formula that made the sample variance an unbiased estimator of the population variance was quite simple: $n-1$ was substituted for N in the denominator. Explaining why this simple adjustment corrects the bias described previously is not so simple, but I can give you some feeling for why $n-1$ makes sense in the formula. Return to the example of finding the variance of the numbers 1, 3, and 8. As you saw before, the three deviations from the mean are −3, −1, and 4, which add up to zero (as will always be the case). The fact that these three deviations must add up to zero implies that knowing only two of the deviations automatically tells you what the third deviation will be. That is, if you know that two of the deviations are −1 and −3, you know that the third deviation must be +4 so that the deviations will sum to zero. Thus, only two of the three

deviations are free to vary (i.e., $n-1$) from the mean of the three numbers; once two deviations have been fixed, the third is determined. The number of deviations that are free to vary is called the number of *degrees of freedom* (df). Generally, when there are n scores in a sample, df = $n-1$.

Another way to think about degrees of freedom is as the number of separate pieces of information that you have about variability. If you are trying to find out about the body temperatures of a newly discovered race of humans native to Antarctica and you sample just one person, you have one piece of information about the population mean, but no ($n - 1 = 1 - 1 = 0$) information about variability. If you sample two people, you have just one piece of information about variability ($2 - 1 = 1$)—the difference between the two people. Note, however, that the number of pieces of information about variability would be n rather than $n-1$ if you knew the population mean before doing any sampling. If you knew that the Antarcticans must have 98.6 degrees Fahrenheit as their population mean for body temperature, but that they could have more or less variability than other people, a single Antarctican would give you one piece of information about variability. If that one Antarctican had a normal body temperature of 96, more variability for Antarcticans would be suggested than if he or she had a temperature of 98.2. It is when you do not know the population mean that variability must be calculated from the mean of your sample, and that entails losing one degree of freedom.

Once the deviation scores have been squared and summed (i.e., *SS*) for a sample, dividing by the number of degrees of freedom is necessary to produce an unbiased estimate of the population variance. This new notation can be used to create shorthand formulas for the sample variance and standard deviation, as follows:

$$s^2 = \frac{SS}{n-1} = \frac{SS}{\text{df}} \qquad \textbf{Formula 3.7A}$$

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{\text{df}}} \qquad \textbf{Formula 3.7B}$$

In applying these formulas to the sample of three numbers (1, 3, 8), you do not have to recalculate *SS*, which was the numerator when we found σ^2 by Formula 3.5A. Given that $SS = 26$, Formula 3.7A tells you that $s^2 = SS/(n-1) = 26/2 = 13$, which is considerably larger than σ^2 (8.67). The increase from σ^2 to s^2 is necessary to correct the underestimation created by Formula 3.5A when estimating the true variance of the larger population. Formula 3.7B shows that $\sigma = \sqrt{13} = 3.61$, which, of course, is considerably larger than σ (2.94). The large differences between the biased and unbiased versions of the variance and standard deviation are caused by our unusually tiny sample ($n = 3$). As n becomes larger, the difference between n and $n-1$ diminishes, as does the difference between σ^2 and s^2 (or σ and s). When n is very large (e.g., over 100), the distinction between the biased and unbiased formulas is so small that for some purposes, it can be ignored.

Skewed Distributions

There are many ways in which the shapes of two unimodal distributions can differ, but one aspect of shape that is particularly relevant to psychological variables and plays an important role in choosing measures of central tendency and variability is *skewness*. A distribution is *skewed* if the bulk of the scores are concentrated on one side of the scale, with relatively few scores on the other side. When graphed as a frequency polygon, a skewed

Figure 3.7

Skewed Distributions

a. Negatively skewed distribution b. Positively skewed distribution

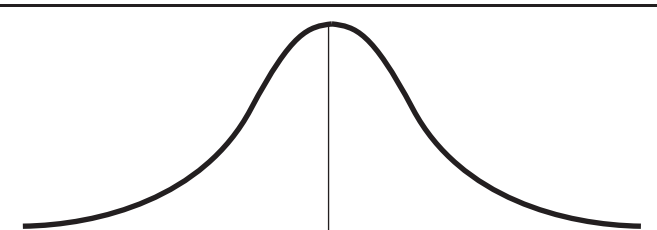
distribution will look something like those in Figure 3.7. The distribution in Figure 3.7a is said to be *negatively skewed*, whereas the one in Figure 3.7b is called *positively skewed*. To remember which shape involves a negative skew and which a positive skew, think of the *tail of the distribution* as a long, thin skewer. If the skewer points to the left (in the direction in which the numbers eventually become negative), the distribution is negatively "skewered" (i.e., negatively skewed); if the skewer points to the right (the direction in which the numbers become positive), the distribution is positively skewed.

Recalling the description of the relation between the area of a polygon and the proportion of scores can help you understand the skewed distribution. A section of the tail with a particular width (i.e., range along the horizontal axis) will have a relatively small area (and therefore relatively few scores) as compared to a section with the same width in the thick part of the distribution (the "hump"). The latter section will have a lot more area and thus a lot more scores.

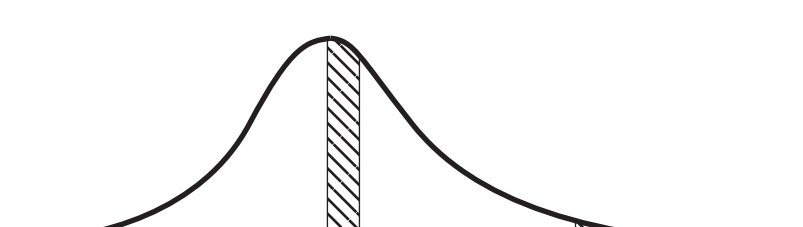
The Central Tendency of a Skewed Distribution

When a unimodal distribution is strongly skewed, it can be difficult to decide whether to use the median or the mean to represent the central tendency of the distribution (the mode would never be the best of the three in this situation). On the other hand, for a symmetrical unimodal distribution, as depicted in Figure 3.8a, the mean and the median are both exactly in the center, right at the mode. Because the distribution is symmetrical, there is the same amount of area on each side of the mode. Now let's see what happens when we turn this distribution into a positively skewed distribution by adding a few high scores, as shown in Figure 3.8b. Adding a small number of scores on the right increases the area on the right slightly. To have the same area on both sides, the median must move to the right a bit. Notice, however, that the median does not have to move very far along the *X* axis. Because the median is in the thick part of the distribution, moving only slightly to the right shifts enough area to compensate for the few high scores that were added. (See how the shaded area on the right end of the graph in Figure 3.8b equals the shaded area between the median and the mode.) Thus, the median is not strongly affected by the skewing of a distribution, and that can be an advantage in describing the central tendency of a distribution.

In fact, once you have found the median of a distribution, you can take a score on one side of the distribution and move it much further away from the median. As long as the score stays on the same side of the median, you can move it out as far as you want—the median will not change its location. This is *not* true for the mean. The mean is affected by the numerical value of every score in the distribution. Consequently the mean will be pulled in the direction of the skew, sometimes quite a bit, as illustrated in Figure 3.9. When the distribution is negatively skewed (Figure 3.9a), the mean will be to the left of (i.e., more negative than) the median, whereas the reverse will be true for a positively skewed distribution (Figure 3.9b). Conversely,



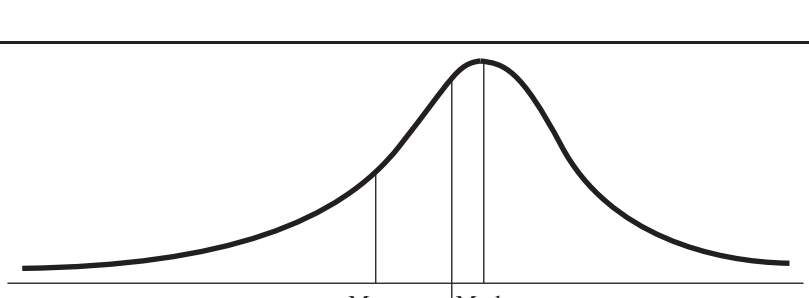
a. Symmetrical distribution



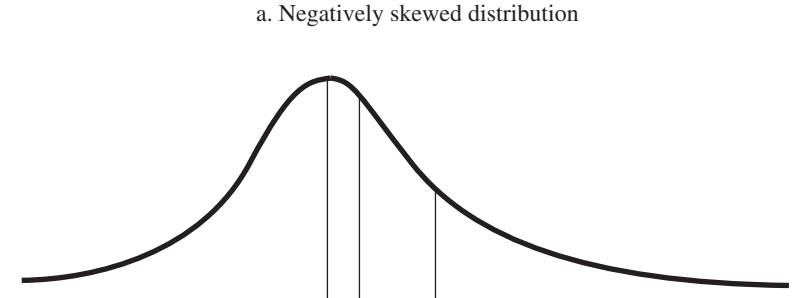
b. Positively skewed distribution

Figure 3.8

Median of a Skewed Distribution



a. Negatively skewed distribution



b. Positively skewed distribution

Figure 3.9

Mean of a Skewed Distribution

if you find both the mean and the median for a distribution, and the median is higher (i.e., more positive), the distribution has a negative skew; if the mean is higher, the skew is positive. In a positively skewed distribution, more than half of the scores will be below the mean, whereas the opposite is true when dealing with a negative skew. If the mean and median are the same, the distribution is probably symmetric around its center.

Choosing Between the Mean and Median

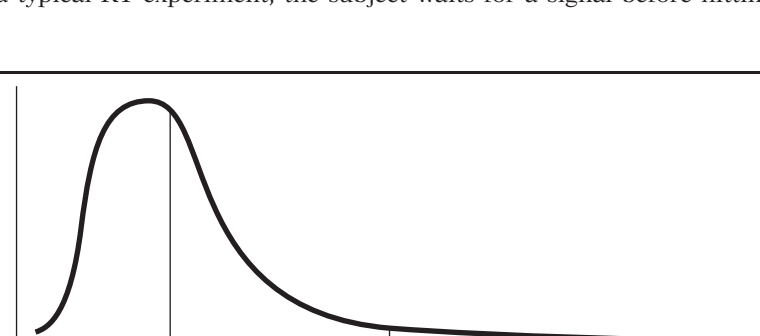
Let us consider an example of a skewed distribution for which choosing a measure of central tendency has practical consequences. There has been much publicity in recent years about the astronomical salaries paid to a few superstar athletes. However, the bulk of the professional athletes in any particular sport are paid a more reasonable salary. For example, the distribution of salaries for major league baseball players in the United States is positively skewed, as shown in Figure 3.10. When the Players' Association is negotiating with management, guess which measure of central tendency each side prefers to use? Of course, management points out that the average (i.e., mean) salary is already quite high (a bit over $3 million as of this writing). The players can point out, however, that the high mean is caused by the salaries of relatively few superstars, and that the mean salary is not very representative of the majority of players (the distribution in Figure 3.10 does not end on the right until it reaches nearly $30 million!). The argument of the Players' Association would be that the median provides a better representation of the salaries of the majority of players. In this case, it seems that the players have a good point (though a median salary of about $1 million is not all that bad). However, the mean has some very useful mathematical properties, which will be explored in detail in Section B.

Floor and Ceiling Effects

Positively skewed distributions are likely whenever there is a limit on values of the variable at the low end but not the high end, or when the bulk of the values are clustered near the lower limit rather than the upper limit. This kind of one-sided limitation is called a *floor effect*. One of the most common examples in psychological research is reaction time (RT). In a typical RT experiment, the subject waits for a signal before hitting a

Figure 3.10

Annual Salaries of Major League Baseball Players



response button; the time between the onset of the signal and the depression of the button is recorded as the reaction time. There is a physiological limit to how quickly a subject can respond to a stimulus, although this limit is somewhat longer if the subject must make some complex choice before responding. After the subjects have had some practice, most of their responses will cluster just above an approximate lower limit, with relatively few responses taking considerably longer. The occasional long RTs may reflect momentary fatigue or inattention, and they create the positive skew (the RT distribution would have a shape similar to the distributions shown in Figures 3.7b and 3.10). Another example of a floor effect involves measurements of clinical depression in a large random group of college students. A third example is scores on a test that is too difficult for the group being tested; many scores would be near zero and there would be only a few high scores.

The opposite of a floor effect is, not surprisingly, a *ceiling effect*, which occurs when the scores in a distribution approach an upper limit but are not near any lower limit. A rather easy exam will show a ceiling effect, with most scores near the maximum and relatively few scores (e.g., only those of students who didn't study or didn't do their homework) near the low end. For example, certain tests are given to patients with brain damage or chronic schizophrenia to assess their orientation to the environment, knowledge of current events, and so forth. Giving such a test to a random group of adults will produce a negatively skewed distribution (such as the one shown in Figure 3.7a). For descriptive purposes, the median is often preferred to the mean whenever either a floor or a ceiling effect is exerting a strong influence on the distribution.

Variability of Skewed Distributions

The standard deviation (*SD*) is a very useful measure of variability, but if you take a score at one end of your distribution and move it much further away from the center, it will have a considerable effect on the *SD*, even though the spread of the bulk of your scores has not changed at all. The mean deviation (*MD*) is somewhat less affected because the extreme score is not being squared, but like the *SD*, the *MD* also becomes misleading when your distribution is very skewed. Of course, the ordinary range is even more misleading in such cases. The only well-known measure of variability that is not affected by extreme scores, and therefore gives a good description of the spread of the main part of your distribution, even if it is very skewed, is the SIQ range.

SUMMARY

1. The mode of a distribution, the most frequent score, is the only descriptive statistic that must correspond to an actual score in the distribution. It is also the only statistic that can be used with all four measurement scales and the only statistic that can take on more than one value in the same distribution (this can be useful, for instance, when a distribution is distinctly bimodal). Unfortunately, the mode is too unreliable for many statistical purposes.
2. The median of a distribution, the 50th percentile, is a particularly good descriptive statistic when the distribution is strongly skewed. Also, it is the point that minimizes the magnitude (i.e., absolute value) of the sum of the (unsquared) deviations. However, the median lacks many of the convenient properties of the mean.

3. The arithmetic mean, the simple average of all the scores, is the most convenient measure of central tendency for use with inferential statistics.
4. The simplest measure of the variability (or *dispersion*) of a distribution is the *range*, the difference between the highest and lowest scores in the distribution. The range is the only measure of variability that tells you the total extent of the distribution, but unfortunately, it tends to be too unreliable for most statistical purposes.
5. The *semi-interquartile range*, half the distance between the first and third quartiles, is a particularly good descriptive measure when dealing with strongly skewed distributions and outliers, but it does not play a role in inferential statistical procedures.
6. The *mean deviation* (*MD*), the average distance of the scores from the mean, is a good description of the variability in a distribution and is easy to understand conceptually, but is rarely used in inferential statistics.
7. The *variance*, the average of the *squared* deviations from the mean, plays an important role in inferential statistics, but it does not provide a convenient description of the spread of a distribution.
8. The *standard deviation*, the square root of the variance, serves as a good description of the variability in a distribution (except when there are very extreme scores), and it also lends itself to use in inferential statistics.
9. Some additional properties of the measures discussed in this section are as follows: the mode, median, range, and SIQ range all require a minimal amount of calculation, and all can be used with ordinal scales; the mode, median, and SIQ range can be used even when there are undeterminable or open-ended scores, and they are virtually unaffected by outliers; the mean, mean deviation, variance, and standard deviation can be used only with an interval or ratio scale, and each of these measures is based on (and is affected by) all of the scores in a distribution.
10. The population variance formula, when applied to data from a sample, tends to underestimate the variance of the population. To correct this *bias*, the sample variance (s^2) is calculated by dividing the sum of squared deviations (*SS*) by $n-1$, instead of by n. The symbol σ^2 will be reserved for any calculation of variance in which N or n, rather than $n-1$, is used in the denominator.
11. The denominator of the formula for the unbiased sample variance, $n-1$, is known as the *degrees of freedom* (df) associated with the variance, because once you know the mean, df is the number of deviations from the mean that are free to vary. Although the sample standard deviation ($\sqrt{s^2} = s$) is not a perfectly unbiased estimation of the standard deviation of the population, the bias is so small that s is referred to as the unbiased sample standard deviation.
12. A *floor effect* occurs when the scores in a distribution come up against a lower limit but are not near any upper limit. This often results in a positively skewed distribution, such that the scores are mostly bunched up on the left side of the distribution with relatively few scores that form a *tail* of the distribution pointing to the right. On the other hand, a *ceiling effect* occurs when scores come close to an upper limit, in which case a negatively skewed distribution (tail pointing to the left) is likely.
13. In a positively skewed distribution, the mean will be pulled toward the right more (and therefore be larger) than the median. The reverse will occur for a negatively skewed distribution.

14. In a strongly skewed distribution, the median is usually the better descriptive measure of central tendency because it is closer to the bulk of the scores than the mean. The mean deviation is less affected by the skewing than the standard deviation, but the SIQ range is less affected still, making it the best descriptive measure of the spread of the bulk of the scores.

EXERCISES

*1. Select the measure of central tendency (mean, median, or mode) that would be most appropriate for describing each of the following hypothetical sets of data:
   a. Religious preferences of delegates to the United Nations
   b. Heart rates for a group of women before they start their first aerobics class
   c. Types of phobias exhibited by patients attending a phobia clinic
   d. Amounts of time participants spend solving a classic cognitive problem, with some of the participants unable to solve it
   e. Height in inches for a group of boys in the first grade

2. Describe a realistic situation in which you would expect to obtain each of the following:
   a. A negatively skewed distribution
   b. A positively skewed distribution
   c. A bimodal distribution

*3. A midterm exam was given in a large introductory psychology class. The median score was 85, the mean was 81, and the mode was 87. What kind of distribution would you expect from these exam scores?

4. A veterinarian is interested in the life span of golden retrievers. She recorded the age at death (in years) of the retrievers treated in her clinic. The ages were 12, 9, 11, 10, 8, 14, 12, 1, 9, 12.
   a. Calculate the mean, median, and mode for age at death.
   b. After examining her records, the veterinarian determined that the dog that had died at 1 year was killed by a car. Recalculate the mean, median, and mode without that dog's data.
   c. Which measure of central tendency in part b changed the most, compared to the values originally calculated in part a?

5. Which of the three most popular measures of variability (range, SIQ range, standard deviation) would you choose in each of the following situations?
   a. The distribution is badly skewed with a few extreme outliers in one direction.
   b. You are planning to perform advanced statistical procedures (e.g., draw inferences about population parameters).
   c. You need to know the maximum width taken up by the distribution.
   d. You need a statistic that takes into account every score in the population.
   e. The highest score in the distribution is "more than 10."

*6. a. Calculate the mean, *SS*, and variance (i.e., σ^2) for the following set of scores: 11, 17, 14, 10, 13, 8, 7, 14.
   b. Calculate the mean deviation and the standard deviation (i.e., σ) for the set of scores in part a.

*7. How many degrees of freedom are contained in the set of scores in Exercise 6? Calculate the unbiased sample variance (i.e., s^2) and standard deviation (i.e., s) for that set of scores. Compare your answers to σ^2 and σ, which you found in Exercise 6.

8. Eliminate the score of 17 from the data in Exercise 6, and recalculate both *MD* and σ. Compared to the values calculated in Exercise 6b, which of these two statistics changed more? What does this tell you about these two statistical measures?

*9. Calculate the mean, mode, median, range, SIQ range, mean deviation, and standard deviation (s) for the following set of scores: 17, 19, 22, 23, 26, 26, 26, 27, 28, 28, 29, 30, 32, 35, 35, 36.

10. a. Calculate the range, SIQ range, mean deviation, and standard deviation (s) for the following set of scores: 3, 8, 13, 23, 26, 26, 26, 27, 28, 28, 29, 30, 32, 41, 49, 56.
   b. How would you describe the relationship between the set of data above and the set of data in Exercise 9?
   c. Compared to the values calculated in Exercise 9, which measures of variability have changed the most, which the least, and which not at all?

B

BASIC STATISTICAL PROCEDURES

Formulas for the Mean

In Section A the arithmetic mean was defined informally as the sum of all of the scores divided by the number of scores added. It is more useful to express the mean as a formula in terms of the summation notation that was presented in the first chapter. The formula for the *population mean* is:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

which tells you to sum all the X's from X_1 to X_N before dividing by N. If you simplify the summation notation by leaving off the indexes (as I promised I would in Chapter 1), you end up with Formula 3.8:

$$\mu = \frac{\sum X}{N} \qquad \textbf{Formula 3.8}$$

The procedure for finding the mean of a sample is exactly the same as the procedure for finding the mean of a population, as shown by Formula 3.9 for the sample mean (note again the use of a lowercase n for the size of a sample):

$$\overline{X} = \frac{\sum X}{n} \qquad \textbf{Formula 3.9}$$

(Recall that the symbol for the sample mean, $\overline{X}$, is pronounced "X bar" when said aloud.) Suppose that the following set of scores represents measurements of clinical depression in seven normal college students: 0, 3, 5, 6, 8, 8, 9. I will use Formula 3.8 to find the mean: $\mu = 39/7 = 5.57$. (If I had considered this set of scores a sample, I would have used Formula 3.9 and of course obtained the same answer, which would have been referred to as $\overline{X}$). To appreciate the sensitivity of the mean to extreme scores, imagine that all of the students have been measured again and all have attained the same rating as before, except for the student who had scored 9. This student has become clinically depressed and therefore receives a new rating of 40. Thus, the new set of scores is 0, 3, 5, 6, 8, 8, 40. The new mean is $\mu = 70/7 = 10$. Note that although the mean has changed a good deal, the median is 6, in both cases.

The Weighted Mean

The statistical procedure for finding the *weighted mean*, better known as the *weighted average*, has many applications in statistics as well as in real life. I will begin this explanation with the simplest possible example. Suppose a professor who is teaching two sections of statistics has given a diagnostic quiz at the first meeting of each section. One class has 30 students who score an average of 7 on the quiz, whereas the other class has only 20 students who average an 8. The professor wants to know the average quiz score for all of the students taking statistics (i.e., both sections combined). The naive approach would be to take the average of the two section means (i.e., 7.5), but as you have probably guessed, this would give you the wrong answer. The correct thing to do is to take the *weighted* average of the two section means. It's not fair to count the class of 30 equally with the class of 20

(imagine giving equal weights to a class of 10 and a class of 100). Instead, the larger class should be given more *weight* in finding the average of the two classes. The amount of weight should depend on the class size, as it does in Formula 3.10. Note that Formula 3.10 could be used to average together any number of class sections or other groups, where n_i is the number of scores in one of the groups and $\overline{X}_i$ is the mean of that group. The formula uses the symbol for the sample mean because weighted averages are often applied to samples to make better guesses about populations.

$$\overline{X}_w = \frac{\sum n_i X_i}{\sum n_i} = \frac{n_1 X_1 + n_2 X_2 + \cdots}{n_1 + n_2 + \cdots} \qquad \textbf{Formula 3.10}$$

We can apply Formula 3.10 to the means of the two statistics sections:

$$\overline{X}_w = \frac{(30)(7) + (20)(8)}{30 + 20} = \frac{210 + 160}{50} = \frac{370}{50} = 7.4$$

Notice that the weighted mean (7.4) is a little closer to the mean of the larger class (7) than to the mean of the smaller class. The weighted average of two groups will always be between the two group means and closer to the mean of the larger group. For more than two groups, the weighted average will be somewhere between the smallest and the largest of the group means.

Let us look more closely at how the weighted average formula works. In the case of the two sections of the statistics course, the weighted average indicates what the mean would be if the two sections were combined into one large class of 50 students. To find the mean of the combined class directly, you would need to know the sum of scores for the combined class and then to divide it by 50. To find $\sum X$ for the combined class, you would need to know the sum for each section. You already know the mean and n for each section, so it is easy to find the sum for each section. First, take another look at Formula 3.9 for the sample mean:

$$\overline{X} = \frac{\sum X}{n}$$

If you multiply both sides of the equation by n, you get $\sum X = n\overline{X}$. (Note that it is also true that $\sum X = n\mu$; we will use this equation in the next subsection.) You can use this new equation to find the sum for each statistics section. For the first class, $\sum X_1 = (30)(7) = 210$, and for the second class, $\sum X_2 = (20)(8) = 160$. Thus, the total for the combined class is $210 + 160 = 370$, which divided by 50 is 7.4. Of course, this is the same answer we obtained with the weighted average formula. What the weighted average formula is actually doing is finding the sum for each group, adding all the group sums to find the total sum, and then dividing by the total number of scores from all the groups.

Computational Formulas for the Variance and Standard Deviation

The statistical procedure for finding *SS*, which in turn forms the basis for calculating the variance and standard deviation, can be tedious, particularly if you are using the definitional Formula 3.3, as reproduced below:

$$SS = \sum (X_i - \mu)^2$$

Formula 3.3 is also called the *deviational formula* because it is based directly on deviation scores. The reason that using this formula is tedious is that each score must be subtracted from the mean, usually resulting in fractions even when all the scores are integers, and then each of these differences must be squared. Compare this process to the *computational formula* for *SS*:

$$SS = \sum X^2 - N\mu^2 \qquad \textbf{Formula 3.11}$$

Note that according to this formula all the X^2 values must be summed, and then the term $N\mu^2$ is subtracted only once, after $\sum X^2$ has been found. (I am using an uppercase N, because this formula might apply either to a population or a sample). It may seem unlikely to you that Formula 3.11 yields exactly the same value as the more tedious Formula 3.3—except that the latter is likely to produce more error due to rounding off at intermediate stages—but it takes just a few steps of algebra to transform one formula into the other.

Some statisticians might point out that if you want a "raw-score" formula for *SS*, Formula 3.11 does not qualify because it requires that the mean be computed first. I think that anyone would want to find the mean before assessing variability—but if you want to find *SS* more directly from the data, you can use Formula 3.12:

$$SS = \sum X^2 - \frac{\left(\sum X\right)^2}{N} \qquad \textbf{Formula 3.12}$$

As I pointed out in Chapter 1, ΣX^2 and $(\Sigma X)^2$ are very different values; the parentheses in the latter term instruct you to add up all the X values *before* squaring (i.e., you square only once at the end), whereas in the former term you square each X before adding.

All you need to do to create a computational formula for the population variance (σ^2) is to divide any formula for *SS* by N. For example, if you divide Formula 3.11 by N, you get Formula 3.13A for the population variance:

$$\sigma^2 = \frac{\sum X^2}{N} - \mu^2 \qquad \textbf{Formula 3.13A}$$

There is an easy way to remember this formula. The term $\sum X^2/N$ is the mean of the squared scores, whereas the term μ^2 is the square of the mean score. So the variance, which is the mean of the squared deviation scores, is equal to the mean of the squared scores minus the square of the mean score. A raw-score formula for the population variance, which does not require you to compute μ first, is found by dividing Formula 3.12 by N, as follows:

$$\sigma^2 = \frac{1}{N}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right] \qquad \textbf{Formula 3.14A}$$

The formula above may look a bit awkward, but it lends itself to an easy comparison with a similar formula for the unbiased sample variance, which I will present shortly.

As usual, formulas for the population standard deviation (σ) are created simply by taking the square root of the variance formulas. (Note that I am

continuing to use "A" for the variance formula and "B" for the standard deviation.)

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \mu^2} \qquad \textbf{Formula 3.13B}$$

$$\sigma = \sqrt{\frac{1}{N}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right]} \qquad \textbf{Formula 3.14B}$$

To illustrate the use of these computational formulas I will find *SS*, using Formula 3.11, for the three numbers (1, 3, 8) that I used as an example in Section A. The first step is to find that the mean of the three numbers is 4. Next, $\sum X^2 = 1^2 + 3^2 + 8^2 = 1 + 9 + 64 = 74$. Then, $N\mu^2 = 3 \times 4^2 = 3 \times 16 = 48$. Finally, $SS = \sum X^2 - N\mu^2 = 74 - 48 = 26$. Of course, all you have to do is divide 26 by *N*, which is 3 in this case, to get the population variance, but because it is common to use one of the variance formulas directly, without stopping to calculate *SS* first, I will next illustrate the use of Formulas 3.13A and 3.14A for the numbers 1, 3, 8:

$$\sigma^2 = \frac{\sum X^2}{N} - \mu^2 = \frac{74}{3} - 4^2 = 24.67 - 16 = 8.67$$

$$\sigma^2 = \frac{1}{N}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right] = \frac{1}{3}\left[74 - \frac{12^2}{3}\right] = \frac{1}{3}(74 - 48) = \frac{1}{3}(26)$$
$$= 8.67$$

Finding the population standard deviation entails nothing more than taking the square root of the population variance, so I will not bother to illustrate the use of the standard deviation formulas at this point.

Unbiased Computational Formulas

When calculating the variance of a set of numbers that is considered a sample of a larger population, it is usually desirable to use a variance formula that yields an unbiased estimate of the population variance. An unbiased sample variance (s^2) can be calculated by dividing *SS* by $n-1$, instead of by *N*. A computational formula for s^2 can therefore be derived by taking any computational formula for *SS* and dividing by $n-1$. For instance, dividing Formula 3.12 by $n-1$ produces Formula 3.15A:

$$s^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right] \qquad \textbf{Formula 3.15A}$$

You should recognize the portion of the above formula in brackets as Formula 3.12. Also, note the similarity between Formulas 3.14A and 3.15A—the latter being the unbiased version of the former.

The square root of the unbiased sample variance is used as an unbiased estimate of the population standard deviation, even though, as I pointed out

before, it is not strictly unbiased. Taking the square root of Formula 3.15A yields Formula 3.15B for the standard deviation of a sample (s):

$$s = \sqrt{\frac{1}{n-1}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]} \qquad \textbf{Formula 3.15B}$$

Obtaining the Standard Deviation Directly From Your Calculator

Fortunately, scientific calculators that provide standard deviation as a built-in function have become very common and very inexpensive. These calculators have a statistics mode; once the calculator is in that mode, there is a special key that must be pressed after each score in your data set to enter that number. When all your numbers have been entered, a variety of statistics are available by pressing the appropriate keys. Usually the key for the biased standard deviation is labeled σ_N; the subscript N or n is used to remind you that N or n rather than $n-1$ is being used to calculate this standard deviation. Unfortunately, the symbol for the *unbiased* standard deviation is often σ_{N-1}, which is not consistent with my use of s and n for the sample statistic, but at least the $N-1$ or sometimes $n-1$ is there to remind you that this standard deviation is calculated with the unbiased formula. To get either type of variance on most of these calculators, you must square the corresponding standard deviation, and to get SS, you must multiply the variance by n or $n-1$, depending on which standard deviation you started with.

Converting Biased to Unbiased Variance and Vice Versa

If your calculator has only the biased or unbiased standard deviation built in (but not both), it is easy to obtain the other one with only a little additional calculation. The procedure I'm about to describe could also be used if you see one type of standard deviation published in an article and would like to determine the other one. If you are starting with the biased standard deviation, square it and then multiply it by n to find SS. Then, to obtain s you divide the SS you just found by $n-1$ and take its square root. Fortunately, there is an even shorter way to do this, as shown in Formula 3.16A:

$$s = \sigma\sqrt{\frac{n}{n-1}} \qquad \textbf{Formula 3.16A}$$

For the numbers 1, 3, and 8, I have already calculated the biased variance (8.67), and therefore the biased standard deviation is $\sqrt{8.67} = 2.94$. To find the unbiased standard deviation, you can use Formula 3.16A:

$$s = 2.94\sqrt{\frac{3}{2}} = 2.94(1.225) = 3.60$$

This result agrees, within rounding error, with the unbiased standard deviation I found for these numbers more directly at the end of Section A.

If you are starting out with the unbiased standard deviation, you can use Formula 3.16A with n and $n-1$ reversed, and changed to uppercase, as follows:

$$\sigma = s\sqrt{\frac{N-1}{N}} \qquad \textbf{Formula 3.16B}$$

If you are dealing with variances instead of standard deviations, you can use the preceding formulas by removing the square root signs and squaring both s and σ.

Properties of the Mean

The mean and standard deviation are often used together to describe a set of numbers. Both of these measures have a number of mathematical properties that make them desirable not only for descriptive purposes but also for various inferential purposes, many of which will be discussed in later chapters. I will describe some of the most important and useful properties for both of these measures beginning with the mean:

1. *If a constant is added* (*or subtracted*) *to every score in a distribution, the mean is increased* (*or decreased*) *by that constant*. For instance, if the mean of a midterm exam is only 70, and the professor decides to add 10 points to every student's score, the new mean will be $70 + 10 = 80$ (i.e., $\mu_{new} = \mu_{old} + C$). The rules of summation presented in Chapter 1 prove that if you find the mean after adding a constant to every score (i.e., $\sum(X + C)/N$), the new mean will equal $\mu + C$. First, note that $\sum(X + C) = \sum X + \sum C$ (according to Summation Rule 1A). Next, note that $\sum C = NC$ (according to Summation Rule 3). So,

$$\frac{\sum(X+C)}{N} = \frac{\sum X + \sum C}{N} = \frac{\sum X + NC}{N} = \frac{\sum X}{N} + \frac{NC}{N} = \frac{\sum X}{N} + C = \mu + C$$

   (A separate proof for subtracting a constant is not necessary; the constant being added could be negative without changing the proof.)
2. *If every score is multiplied* (*or divided*) *by a constant, the mean will be multiplied* (*or divided*) *by that constant*. For instance, suppose that the average for a statistics quiz is 7.4 (out of 10), but later the professor wants the quiz to count as one of the exams in the course. To put the scores on a scale from 0 to 100, the professor multiplies each student's quiz score by 10. The mean is also multiplied by 10, so the mean of the new exam scores is $7.4 \times 10 = 74$.

   We can prove that this property holds for any constant. The mean of the scores after multiplication by a constant is $\left(\sum CX\right)/N$. By Summation Rule 2A, you know that $\sum CX = C\sum X$, so

$$\frac{\sum CX}{N} = \frac{C\sum X}{N} = C\frac{\sum X}{N} = C\mu$$

   There is no need to prove that this property also holds for dividing by a constant because the constant in the above proof could be less than 1.0 without changing the proof.
3. *The sum of the deviations from the mean will always equal zero*. To make this idea concrete, imagine that a group of waiters has agreed to share all of their tips. At the end of the evening, each waiter puts his or her tips in a big bowl; the money is counted and then divided equally among the

waiters. Because the sum of all the tips is being divided by the number of waiters, each waiter is actually getting the mean amount of the tips. Any waiter who had pulled in more than the average tip would lose something in this deal, whereas any waiter whose tips for the evening were initially below the average would gain. These gains or losses can be expressed symbolically as deviations from the mean, $X_i - \mu$, where X_i is the amount of tips collected by the ith waiter and μ is the mean of the tips for all the waiters. The property above can be stated in symbols as $\sum(X_i - \mu) = 0$.

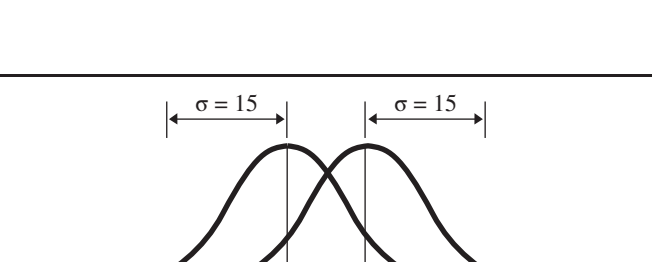
In terms of the waiters, this property says that the sum of the gains must equal the sum of the losses. This makes sense—the gains of the waiters who come out ahead in this system come entirely from the losses of the waiters who come out behind. Note, however, that the *number* of gains does not have to equal the *number* of losses. For instance, suppose that 10 waiters decided to share tips and that 9 waiters receive \$10 each and a 10th waiter gets \$100. The sum will be $(9 \times 10) + 100 = \$90 + \$100 = \$190$, so the mean is $\$190/10 = \$19$. The nine waiters who pulled in \$10 each will each gain \$9, and one waiter will lose \$81 (\$100 – \$19). But although there are nine gains and one loss, the total amount of gain ($9 \times \$9 = \$81$) equals the total amount of loss ($1 \times \$81 = \$81$). Note that in this kind of distribution, the majority of scores can be below the mean (the distribution is positively skewed, as described in the previous section).

The property above can also be proven to be generally true. First, note that $\sum(X_i - \mu) = \sum X_i - \sum \mu$, according to Summation Rule 1B. Because μ is a constant, $\sum \mu = N\mu$ (Summation Rule 3), so $\sum(X_i - \mu) = \sum X_i - N\mu$. Multiplying both sides of the equation for the mean by N, we get: $\sum X_i = N\mu, so \sum(X_i - \mu) = N\mu - N\mu = 0$.

4. *The sum of the squared deviations from the mean will be less than the sum of squared deviations around any other point in the distribution*. To make matters simple, I will use my usual example: 1, 3, 8. The mean of these numbers is 4. The deviations from the mean (i.e., $X_i - 4$) are −3, −1, and +4. (Note that these sum to zero, as required by Property 3 above.) The squared deviations are 9, 1, and 16, the sum of which is 26. If you take any number other than the mean (4), the sum of the squared deviations from that number will be more than 26. For example, the deviations from 3 (which happens to be the median) are −2, 0, +5; note that these do *not* sum to zero. The squared deviations are 4, 0, and 25, which sum to more than 26. Also note, however, that the absolute values of the deviations from the median add up to 7, which is less than the sum of the absolute deviations from the mean (8). It is the median that minimizes the sum of absolute deviations, whereas the mean minimizes the sum of *squared* deviations. Proving that the latter property is always true is a bit tricky,

Figure 3.11

Adding a Constant to a Distribution



but the interested reader can find such a proof in some advanced texts (e.g., Hays, 1994). This property, often called the *least-squares property*, is a very important one and will be mentioned in the context of several statistical procedures later in this text.

Properties of the Standard Deviation

Note: These properties apply equally to the biased and unbiased formulas.

1. *If a constant is added* (*or subtracted*) *from every score in a distribution, the standard deviation will not be affected*. To illustrate a property of the mean, I used the example of an exam on which the mean score was 70. The professor decided to add 10 points to each student's score, which caused the mean to rise from 70 to 80. Had the standard deviation been 15 points for the original exam scores, the standard deviation would still be 15 points after 10 points were added to each student's exam score. Because the mean moves with the scores, and the scores stay in the same relative positions with respect to each other, shifting the location of the distribution (by adding or subtracting a constant) does not alter its spread (see Figure 3.11). This can be shown to be true in general by using simple algebra. The standard deviation of a set of scores after a constant has been added to each one is:

$$\sigma_{new} = \sqrt{\frac{\sum [(X + C) - \mu_{new}]^2}{N}}$$

According to the first property of the mean just described, $\mu_{new} = \mu_{old} + C$. Therefore,

$$\sigma_{new} = \sqrt{\frac{\sum [(X + C) - (\mu_{old} + C)]^2}{N}} = \sqrt{\frac{\sum (X + C - \mu_{old} - C)^2}{N}}$$

Rearranging the order of terms gives the following expression:

$$\sigma_{new} = \sqrt{\frac{\sum (X - \mu_{old} + C - C)^2}{N}} = \sqrt{\frac{\sum (X - \mu_{old})^2}{N}} = \sigma_{old}$$

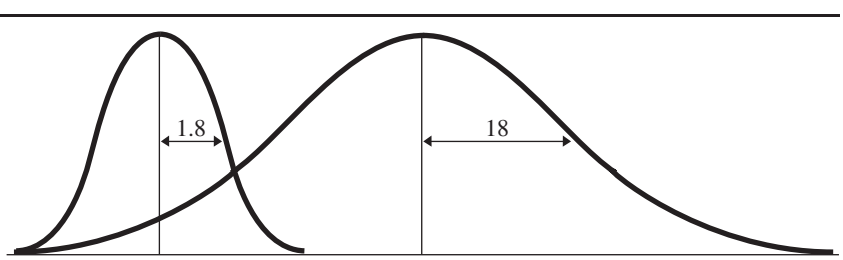
The above proof works the same way if you are subtracting, rather than adding, a constant.

2. *If every score is multiplied* (*or divided*) *by a constant, the standard deviation will be multiplied* (*or divided*) *by that constant*. In describing a corresponding property of the mean, I used an example of a quiz with a mean of 7.4; each student's score was multiplied by 10, resulting in an exam with a mean of 74. Had the standard deviation of the quiz been 1.8, the standard deviation after scores were multiplied by 10 would have been 18. Whereas adding a constant does not increase the spread of the distribution, multiplying by a constant does (see Figure 3.12). For example, quiz scores of 4 and 7 are spread by only 3 points, but after they are multiplied by 10 the scores are 40 and 70, which are 30 points apart. Once again we can show that this property is true in general by using some algebra and the rules of summation. The standard deviation of a set of scores after multiplication by a constant is:

$$\sigma_{new} = \sqrt{\frac{\sum (CX_i - \mu_{new})^2}{N}}$$

Figure 3.12

Multiplying a Distribution by a Constant



According to the second property of the mean described above, $\mu_{new} = C\mu_{old}$. Therefore:

$$\sigma_{new} = \sqrt{\frac{\sum (CX_i - C\mu_{old})^2}{N}} = \sqrt{\frac{\sum [C(X_i - \mu_{old})]^2}{N}} = \sqrt{\frac{\sum C^2(X_i - \mu_{old})^2}{N}}$$

The term C^2 is a constant, so according to Summation Rule 2, we can move this term in front of the summation sign. Then a little bit of algebraic manipulation proves the preceding property:

$$\sigma_{new} = \sqrt{\frac{C^2 \sum (X_i - \mu_{old})^2}{N}} = \sqrt{C^2}\sqrt{\frac{\sum (X_i - \mu_{old})^2}{N}} = C\sigma_{old}$$

3. *The standard deviation from the mean will be smaller than the standard deviation from any other point in the distribution*. This property follows from property 4 of the mean, as described previously. If *SS* is minimized by taking deviations from the mean rather than from any other location, it makes sense that σ, which is $\sqrt{SS/N}$, will also be minimized. Proving this requires some algebra and the rules of summation; the proof can be found in some advanced texts (e.g., Hays, 1994, p. 188).

Measuring Skewness

Skewness can be detected informally by inspecting a graph of the distribution in your sample in the form of, for example, a frequency polygon, or a stem-and-leaf plot. However, quantifying skewness can be useful in deciding when the skewing is so extreme that you ought to take steps to modify your distribution in your sample or use different types of statistics. For this reason most statistical packages provide a measure of skewness when a full set of descriptive statistics is requested. Whereas the variance is based on the average of squared deviations from the mean, skewness is based on the average of *cubed* deviations from the mean:

$$\text{Average cubed deviation} = \frac{\sum (X_i - \mu)^3}{N}$$

Recall that when you square a number, the result will be positive whether the original number was negative or positive. However, the cube (or third power) of a number has the same sign as the original number. If the number is negative, the cube will be negative ($-2^3 = -2 \times -2 \times -2 = -8$), and if the number is positive, the cube will be positive (e.g., $+2^3 = +2 \times +2 \times +2 = +8$). Deviations below the mean will still be negative after being

cubed, and positive deviations will remain positive after being cubed. Thus skewness will be the average of a mixture of positive and negative numbers, which will balance out to zero *only* if the distribution is symmetric. (Note that the deviations from the mean will always average to zero before being cubed, but after being cubed they need not.) Any negative skew will cause the skewness measure to be negative, and any positive skew will produce a positive skewness measure. Unfortunately, like the variance, this measure of skewness does not provide a good description of a distribution—in this case, because it is in cubed units. Rather than taking the cube root of the preceding formula, you can derive a more useful measure of the skewness of a population distribution by dividing that formula by σ^3 (the cube of the standard deviation calculated as for a population) to produce Formula 3.17:

$$\text{Skewness} = \frac{\sum (X_i - \mu)^3}{N\sigma^3} \qquad \textbf{Formula 3.17}$$

Formula 3.17 has the very useful property of being dimensionless (cubed units are being divided, and thus canceled out, by cubed units); it is a pure measure of the shape of the distribution. Not only is this measure of skewness unaffected by adding or subtracting constants (as is the variance), it is also unaffected by multiplying or dividing by constants. For instance, if you take a large group of people and measure each person's weight in pounds, the distribution is likely to have a positive skew that will be reflected in the measure obtained from Formula 3.17. Then, if you convert each person's weight to kilograms, the *shape* of the distribution will remain the same (although the variance will be multiplied by a constant), and fortunately the skewness measure will also remain the same. The only drawback to Formula 3.17 is that if you use it to measure the skewness of a sample, the result will be a biased estimate of the population skewness. This is only a problem if you plan to test your measure of skewness with inferential methods, but this is rarely done. However, even a descriptive measure of skewness can be very useful for comparing one distribution to another.

To illustrate the use of Formula 3.17, I will calculate the skewness of four numbers: 2, 3, 5, 10. First, using Formula 3.4B you can verify that $\sigma = 3.082$, so $\sigma^3 = 29.28$. Next, $\Sigma(X-\mu)^3 = (2-5)^3 + (3-5)^3 + (5-5)^3 + (10-5)^3 = -3^3 + -2^3 + 0^3 + 5^3 = -27 + (-8) + 0 + 125 = 90$. (Note how important it is to keep track of the sign of each number.) Now we can plug these values into Formula 3.17:

$$\text{Skewness} = \frac{90}{4(29.28)} = \frac{90}{117.1} = .768$$

As you can see, the skewness is positive (recall that the skewness of the normal distribution is zero). Although the total amount of deviation below the mean is the same as the amount of deviation above, one larger deviation (i.e., +5) counts more than two smaller ones (i.e., −2 and −3).

Measuring Kurtosis

It is important to note that two distributions can both be symmetric (i.e., skewness equals zero), unimodal, and bell-shaped and yet not be identical in shape. (Bell-shaped is a crude designation—many variations are possible.) Moreover, the two distributions just mentioned can even have the same mean and variance and still differ fundamentally in shape. The simplest way that two such distributions can differ is in the degree of flatness

that characterizes the curve. If a distribution tends to have relatively thick or heavy tails and then bends sharply so as to have a relatively greater concentration near its center (more "peakedness"), that distribution is called *leptokurtic*. Compared to the normal distribution, a leptokurtic distribution lacks scores in the "shoulders" of the distribution (the areas on each side of the distribution that are between the tails and the middle of the distribution). On the other hand, a distribution that tends to be flat (i.e., it has no shortage of scores in the shoulder area and therefore does not bend sharply in that area), with relatively thin tails and less peakedness, is called *platykurtic*. (The Greek prefixes platy- and lepto- describe the middle portion of the distribution, lepto-, meaning "slim," and platy-, meaning "wide.") These two different shapes are illustrated in Figure 3.13, along with a distribution that is midway between in its degree of *kurtosis*—a *mesokurtic distribution*. The normal distribution is used as the basis for comparison in determining kurtosis, so it is mesokurtic, by definition. (Because a distribution can be leptokurtic due to very heavy tails *or* extreme peakedness, there are debates about the relative importance of these two factors in determining kurtosis. This debate goes well beyond the scope of this text, but you can read more about it in an article by DeCarlo, 1997.)

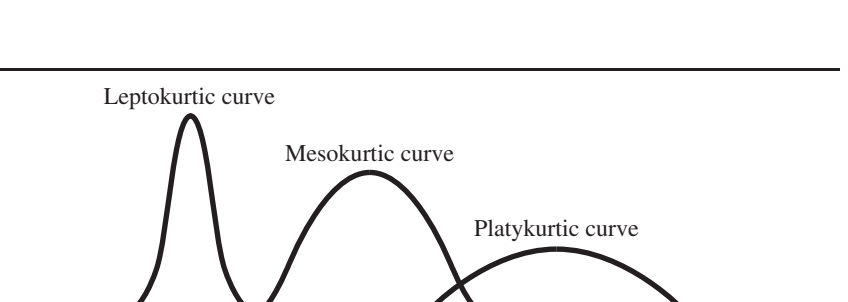
Just as the measure of skewness is based on cubed deviations from the mean, the measure of kurtosis is based on deviations from the mean raised to the fourth power. This measure of kurtosis must then be divided by the standard deviation raised to the fourth power, to create a dimensionless measure of distribution shape that will not change if you add, subtract, multiply, or divide all the data by a constant. Formula 3.18 for the kurtosis of a population is as follows:

$$\text{Kurtosis} = \frac{\sum (X_i - \mu)^4}{N\sigma^4} - 3 \qquad \textbf{Formula 3.18}$$

Notice that Formula 3.18 parallels Formula 3.17, except for the change from the third to the fourth power, and the subtraction of 3 in the kurtosis formula. The subtraction appears in most kurtosis formulas to facilitate comparison with the normal distribution. Subtracting 3 ensures that the kurtosis of the normal distribution will come out to zero. Thus a distribution that has relatively fatter tails than the normal distribution (and greater peakedness) will have a positive kurtosis (i.e., it will be leptokurtic), whereas a relatively thin-tailed, less peaked distribution will have a negative kurtosis (it will be platykurtic). In fact, unless you subtract 3, the population kurtosis will never be less than +1. After you subtract 3, kurtosis can range from −2 to positive infinity.

Figure 3.13

Degrees of Kurtosis



I will illustrate the calculation of kurtosis for the following four numbers: 1, 5, 7, 11. First, we find that the mean of these numbers is 6 and the population variance 13. To find σ^4, we need only square the biased variance: $13^2 = 169$. (Note that in general, $(x^2)^2 = x^4$.) Next, we find $\Sigma(X-\mu)^4 = (1-6)^4 + (5-6)^4 + (7-6)^4 + (11-6)^4 = -5^4 + (-1)^4 + 1^4 + 5^4 = 625 + 1 + 1 + 625 = 1252$. Now we are ready to use Formula 3.18:

$$\text{Kurtosis} = \frac{1252}{4(169)} - 3 = 1.85 - 3 = -1.15$$

The calculated value of −1.15 suggests that the population from which the four numbers were drawn has negative kurtosis (i.e., somewhat lighter tails than the normal distribution). In practice, however, one would never draw any conclusions about kurtosis when dealing with only four numbers.

The most common reason for calculating the skewness and kurtosis of a set of data is to help you decide whether your sample comes from a population that is normally distributed. All of the statistical procedures in Parts II through VI of this text are based on the assumption that the variable being measured has a normal distribution in the population. The next few chapters offer additional techniques for comparing your data to a normal distribution, and dealing with data that contains extreme scores on one or both ends of the sample distribution.

B

SUMMARY

1. If several groups are to be combined into a larger group, the mean of the larger group will be the weighted average of the means of the smaller groups, where the weights are the sizes of the groups. Finding the weighted average, in this case, can be accomplished by finding the sum of each group (which equals its size times its mean), adding all the sums together, and then dividing by the size of the combined group (i.e., the sum of the sizes of the groups being combined).
2. Convenient computational formulas for the variance and *SD* can be created by starting with a computational formula for *SS* (e.g., the sum of the squared scores minus *N* times the square of the mean) and then dividing by *N* for the biased variance, or $n-1$ for the unbiased variance. The computational formula for the biased or unbiased *SD* is just the square root of the corresponding variance formula.
3. The standard deviation can be found directly using virtually any inexpensive scientific or statistical calculator (the calculator must be in statistics mode, and, usually, a special key must be used to enter each score in your data set). The variance is then found by squaring the *SD*, and the *SS* can be found by multiplying the variance by *N* (if the variance is biased), or $n-1$ (if the variance is unbiased).
4. A biased *SD* can be converted to an unbiased *SD* by multiplying it by the square root of the ratio of n over $n-1$, a factor that is very slightly larger than 1.0 for large samples. To convert from unbiased to biased, the ratio is flipped over.
5. Properties of the Mean
   a. If a constant is added (or subtracted) to every score in a distribution, the mean of the distribution will be increased (or decreased) by that constant (i.e., $\mu_{new} = \mu_{old} \pm C$).
   b. If every score in a distribution is multiplied (or divided) by a constant, the mean of the distribution will be multiplied (or divided) by that constant (i.e., $\mu_{new} = C\mu_{old}$).
   c. The sum of the deviations from the mean will always equal zero (i.e., $\sum(X_i - \mu) = 0$).

d. The sum of the squared deviations from the mean will be less than the sum of squared deviations from any other point in the distribution (i.e., $\sum (X_i - \mu)^2 < \sum (X_i - C)^2$, where C represents some location in the distribution other than the mean).

6. Properties of the Standard Deviation
   a. If a constant is added (or subtracted) from every score in a distribution, the standard deviation will remain the same (i.e., $\sigma_{new} = \sigma_{old}$).
   b. If every score is multiplied (or divided) by a constant, the standard deviation will be multiplied (or divided) by that constant (i.e., $\sigma_{new} = C\sigma_{old}$).
   c. The standard deviation around the mean will be smaller than it would be around any other point in the distribution.
7. *Skewness* can be measured by cubing (i.e., raising to the third power) the deviations of scores from the mean of a distribution, taking their average, and then dividing by the cube of the population standard deviation. The measure of skewness will be a negative number for a negatively skewed distribution, a positive number for a positively skewed distribution, and zero if the distribution is perfectly symmetric around its mean.
8. *Kurtosis* can be measured by raising deviations from the mean to the fourth power, taking their average, and then dividing by the square of the population variance. If the kurtosis measure is set to zero for the normal distribution (by subtracting 3 in the just-described formula), positive kurtosis indicates relatively fat tails and more peakedness in the middle of the distribution (a leptokurtic distribution), whereas negative kurtosis indicates relatively thin tails and a lesser peakedness in the middle (a platykurtic distribution).

EXERCISES

*1. There are three fourth-grade classes at Happy Valley Elementary School. The mean IQ for the 10 pupils in the gifted class is 119. For the 20 pupils in the regular class, the mean IQ is 106. Finally, the five pupils in the special class have a mean IQ of 88. Calculate the mean IQ for all 35 fourth-grade pupils.

2. A student has earned 64 credits so far, of which 12 credits are As, 36 credits are Bs, and 16 credits are Cs. If A = 4, B = 3, and C = 2, what is this student's grade point average?

*3. A fifth-grade teacher calculated the mean of the spelling tests for his 12 students; it was 8. Unfortunately, now that the teacher is ready to record the grades, one test seems to be missing. The 11 available scores are 10, 7, 10, 10, 6, 5, 9, 10, 8, 6, 9. Find the missing score. (*Hint*: You can use property 3 of the mean.)

4. A psychology teacher has given an exam on which the highest possible score is 200 points. The mean score for the 30 students who took the exam was 156, and the standard deviation was 24. Because there was one question that every student answered incorrectly, the teacher decides to give each student 10 extra points and then divide each score by 2, so the total possible score is 100. What will the mean and standard deviation of the scores be after this transformation?

5. The IQ scores for 10 sixth-graders are 111, 103, 100, 107, 114, 101, 107, 102, 112, 109.
   a. Calculate σ for the IQ scores using the definitional formula.
   b. Calculate σ for the IQ scores using the computational formula.
   c. Describe one condition under which it is easier to use the definitional than the computational formula.
   d. How could you transform the scores above to make it easier to use the computational formula?

*6. Use the appropriate computational formulas to calculate both the biased and

unbiased standard deviations for the following set of numbers: 21, 21, 24, 24, 27, 30, 33, 39.

*7. a. Calculate s for the following set of numbers: 7, 7, 10, 10, 13, 16, 19, 25. (*Note*: This set of numbers was created by subtracting 14 from each of the numbers in the previous exercise.) Compare your answer to this exercise with your answer to Exercise 6. What general principle is being illustrated?
b. Calculate s for the following set of numbers: 7, 7, 8, 8, 9, 10, 11, 13. (*Note*: This set of numbers was created by dividing each of the numbers in Exercise 6 by 3.) Compare your answer to this exercise with your answer to Exercise 6. What general principle is being illustrated?

8. a. For the data in Exercise 6 use the definitional formula to calculate s around the *median* instead of the mean.
b. What happens to s? What general principle is being illustrated?

*9. If σ for a set of data equals 4.5, what is the corresponding value for s
a. When $n = 5$?
b. When $n = 20$?
c. When $n = 100$?

10. If s for a set of data equals 12.2, what is the corresponding value for σ
a. When $N = 10$?
b. When $N = 200$?

*11. a. Calculate the population standard deviation and skewness for the following set of data: 2, 4, 4, 10, 10, 12, 14, 16, 36.
b. Calculate the population standard deviation and skewness for the following set of data: 1, 2, 2, 5, 5, 6, 7, 8, 18. (This set was formed by halving each number in part a.)
c. How does each value calculated in part a compare to its counterpart calculated in part b? What general principles are being illustrated?

12. a. Calculate the population standard deviation and skewness for the following set of data: 1, 2, 2, 5, 5, 6, 7, 8. (This set was formed by dropping the highest number from the set in Exercise 11 part b.)
b. Comparing your answer to part a with your answer to Exercise 11 part b, what can you say about the effect of one extreme score on variability and skewness?

*13. Take the square root of each of the scores in Exercise 11 part a, and recalculate σ and the skewness. What effect does this transformation have on these measures?

14. Calculate the kurtosis for the following set of data: 3, 9, 10, 11, 12, 13, 19.

*15. a. Calculate the kurtosis for the following set of data: 9, 10, 11, 12, 13.
b. Compare your answer to your answer for Exercise 14. What is the effect on kurtosis when you remove extreme scores from both sides of a distribution?

C ANALYSIS BY SPSS

Summary Statistics

The three measures of central tendency discussed in this chapter, as well as several measures of variability, can be obtained from SPSS by opening the **Frequencies: Statistics** box, described in Chapter 2, for obtaining percentiles.

To obtain basic summary statistics for a variable, follow these five steps:

1. Select **Descriptive Statistics** from the **ANALYZE** menu, and click on **Frequencies** . . .
2. Move the variables for which you want to see summary statistics into the *Variable(s)*: space.
3. Click the **Statistics** button, and then select the Central Tendency, Dispersion, and Distribution statistics you want to see (see Figure 3.14). Click **Continue** to return to the main dialog box.
4. Uncheck the little box labeled "Display frequency tables," if you do not want to see any frequency tables.
5. Click **OK** from the main **Frequencies** dialog box.

Figure 3.14

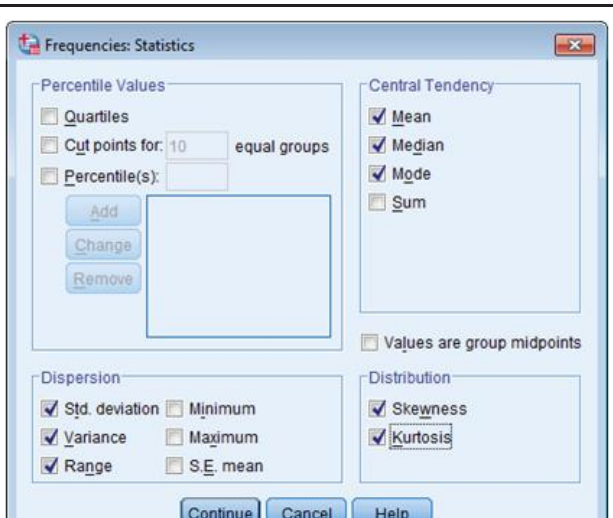


Table 3.3

Statistics

Prevmath

N	Valid	100
	Missing	0
Mean		1.38
Median		1.00
Mode		1
Std. Deviation		1.254
Variance		1.571
Skewness		1.283
Std. Error of Skewness		.241
Kurtosis		1.638
Std. Error of Kurtosis		.478
Range		6

The section of the **Frequencies: Statistics** box labeled "Central Tendency" allows you to select any or all of the following choices: Mean, Median, Mode, and Sum (although the latter is not a measure of central tendency, the creators of SPSS obviously found it convenient to include the Sum of the scores under this heading). The region of this box labeled "Dispersion" provides three of the measures of variability described in this chapter: the standard deviation, variance, and range. Finally, the "Distribution" portion of the box lets you obtain measures of Skewness and/or Kurtosis. The statistics selected in the previous figure were applied to the number of previous math courses taken in order to obtain the results shown in Table 3.3.

The skewness measure indicates a good deal of positive skewing, which is consistent with the mean being considerably larger than the median, and with the distribution shape that you can see in the histogram of the *prevmath* variable, shown in the previous chapter. Note that SPSS computes only the "unbiased" versions of the variance and standard deviation. Because they are rarely used in social science research, and probably to reduce confusion, SPSS does not offer the biased versions of these measures from any of its menus.

If you are not interested in looking at the whole distribution of your variable, but just want some summary statistics, you can open the **Descriptives** dialog box by clicking on **ANALYZE**, **Descriptive Statistics**, and then **Descriptives** . . . , and then after moving the variable(s) of interest to the *Variable(s)*: space, click on the **Options** button. The choices in this Options box are the same as in the **Frequencies: Statistics** box, except that neither the median nor the mode is available, because the **Descriptives** subprogram was designed to be used for interval/ratio data, and not for ordinal or nominal data.

Using Explore to Obtain Additional Statistics

The **Explore** dialog box, used in the previous chapter to obtain stemplots, is useful for exploring your sample data in a variety of ways, as its name implies. For example, if you click on the **Statistics** button in the **Explore** dialog box and then select *Descriptives*, you will get, for measures of

variability, not only the standard deviation, variance, and range, but the interquartile range, as well. If you want the SIQ range, just divide the latter measure by two. Next, we will use **Explore** to take a more detailed look at the distribution of data within a sample.

Boxplots

Box-and-whisker plots (boxplots, for short), like stemplots, represent one of the EDA techniques developed by Tukey (1977) to aid researchers in understanding their data distributions before they apply the methods of inferential statistics. I have not covered boxplots yet in this chapter, so I will begin by showing you a boxplot of the 100 phobia ratings in Ihno's data set (see Figure 3.15).

Let's begin by looking at the "box" in the boxplot. The top side of the box is always placed at the 75th percentile, which you can see in this case corresponds to a phobia rating of 4. The bottom part of the box, which is always drawn at the 25th percentile, corresponds to a rating of 1. (Technically, the top and bottom sides of the box are called the *hinges*, and the way Tukey defined them does not perfectly align with the 25th and 75th percentiles, but there is so little difference that it is not worth getting into further details here.) The horizontal line within the box is always located at the median (i.e., 50th percentile), which for these data is 3. The fact that the median is closer to the 75th than the 25th percentile tells us that the distribution has a positive skew. This becomes even more obvious by looking at the "whiskers."

The height of the box (essentially the same as the interquartile range) is 3 in this example, and the whiskers can extend, at most, a distance of 1.5 times that height in each direction (these whisker limits are called the *inner fences* of the plot). Thus, the upper whisker could have extended to a score of $4 + 1.5 \times 3 = 8.5$ (the *upper* inner fence), except that the upper

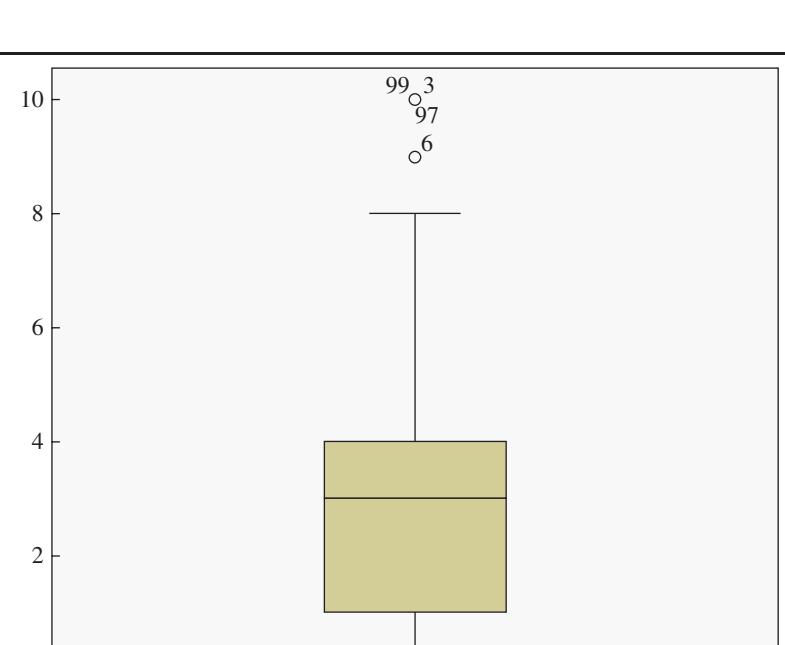


Figure 3.15

whisker must stop at an actual score, called an *adjacent value*, which cannot be higher than the upper inner fence. So, for this example, the upper whisker ends at a rating of 8, which is the highest score that actually appears in the data *and* is not higher than the upper inner fence. Any scores higher than the end of the upper whisker are defined as *outliers*. In this example, there are a total of four outliers in the positive direction—one 9 and three 10's—and SPSS labels them by their case (i.e., row) numbers, as you can see in Figure 3.15. Outlying scores are always good candidates for closer inspection, as they may be the result of transcription errors, participant errors, or even accurate measurements of unusual participants. However, given the upper limit of the scale in this example, the outliers seem unlikely to be errors or strange cases.

Of course, all of the rules I just described for the upper whisker and so on apply equally to the lower end of the box. However, practical constraints on the data can place their own limitations on the boxplot. In this example, ratings cannot be lower than zero, so the lower whisker must end at zero, making it impossible to have outliers on the low end. This is the well-known "floor" effect. Comparing the lengths of the two whiskers makes it even clearer that you are dealing with a sample distribution that is positively skewed. Now that you know what a basic boxplot looks like, I will explain how to obtain one from SPSS.

To create Boxplots:

1. Select **Descriptive Statistics** from the **ANALYZE** menu, and click on **Explore** . . .
2. Move the variables for which you want to see boxplots 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" (see Figure 3.16).
3. Click the **Plots** button.
4. In the upper-left section (labeled "Boxplots") of the **Explore: Plots** box make sure that *Factor levels together* has already been selected (it is one of the defaults). Unselect *Stem-and-leaf* in the upper-right section, if you do not want this (default) option (explained in the previous chapter), and then click **Continue**.
5. Click **OK** from the main **Explore** dialog box.

If all you want is a simple boxplot for one of your variables, it doesn't matter whether you select *Factor levels together* (the default) or *Dependents together* in the **Explore: Plots** box. However, if you were to add a second variable to the Dependent List (see Figure 3.16), then checking *Dependents together* would create a pair of boxplots side-by-side on the same graph. This is only desirable, of course, if the two variables have been measured on the same scale. Checking *Factor levels together* instead would result in two separate boxplots, one after the other.

Another option is to have one variable in the Dependent List and one in the Factor List, say *hr_base* and *gender*, respectively. Again, it doesn't matter whether you check *Factor levels together* or *Dependents together*; in either case, you will get a boxplot for each level of the variable in the Factor list, all in the same graph, as shown in Figure 3.17.

These side-by-side boxplots show clearly that females have the higher median heart rate, and also that females have more of a negative skew (their median is closer to the bottom of the box), whereas the males have an outlier

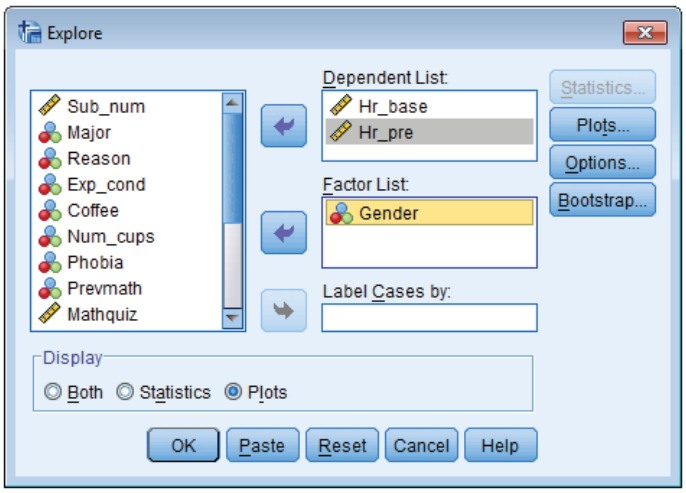


Figure 3.16

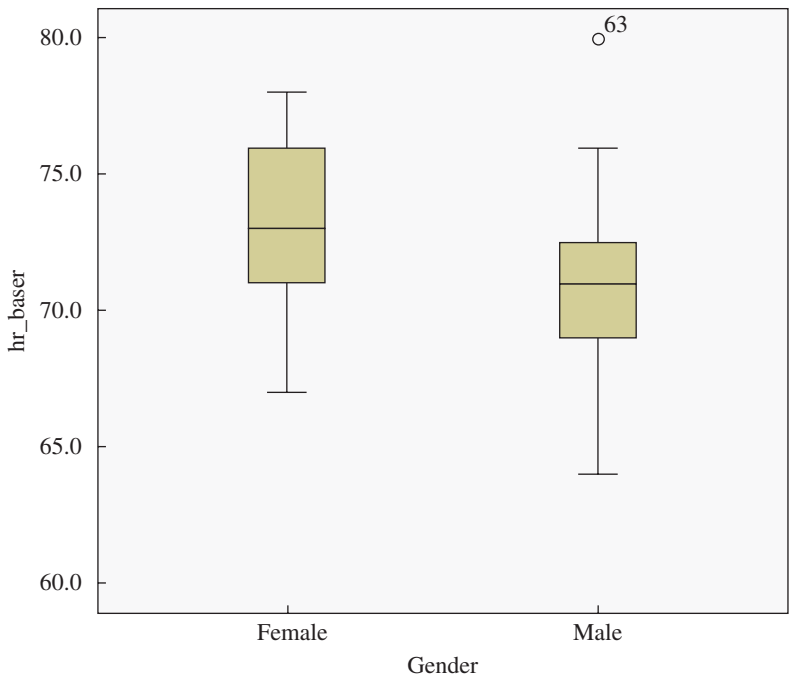
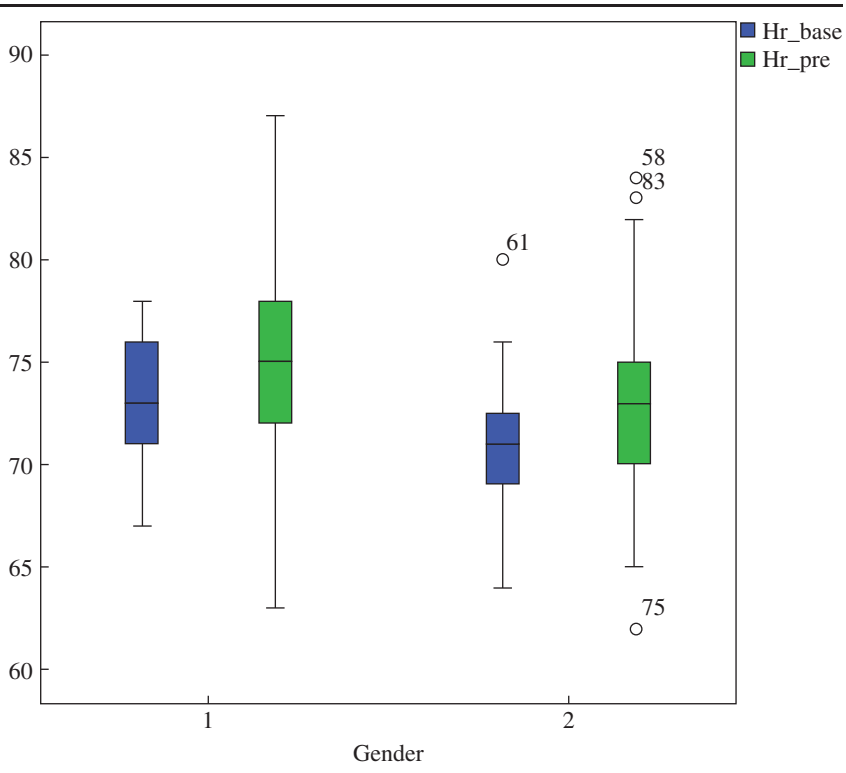


Figure 3.17

in the positive direction. If you were to include two dependent variables (e.g., *hr_base* and *hr_pre*), as well as one factor (e.g., *gender*), as in the **Explore** box I showed earlier, it certainly would make a difference whether you checked *Factor levels together* or *Dependents together*. Selecting *Factor levels together* would create separate graphs for each DV, each containing side-by-side boxplots for the two genders. Selecting *Dependents together* would create a single graph with four boxplots in this case, as shown in Figure 3.18. You can explore more complex combinations of boxplots by combining several dependent variables with several multilevel factors.

Figure 3.18



Selecting Cases

Having identified a few outliers in your data that are not based on mistakes, you may want to run your analyses with and without the outliers to see just how much difference that makes. For example, you may want SPSS to compute the mean and *SD* of the phobia variable without the four students who rated their phobia as a 9 or 10. To filter out these outliers follow these steps:

1. Choose **Select Cases** . . . from the Data menu.
2. Check the second choice (*If condition is satisfied*), and then click on the **If** . . . button.
3. In the dialog box that opens, move "phobia" from the variable list on the left to the workspace.
4. Type "< 9" just to the right of "phobia" (you can separate symbols with spaces or not), and then click **Continue**.
5. Finally, click **OK** in the original **Select Cases** box.

The following will occur after you click OK: a new variable, named "filter_$," will appear in the rightmost column of your spreadsheet, with 1's for selected cases, and 0's for excluded cases; slashes will appear through the row numbers (leftmost column of the spreadsheet) of the cases that are being filtered out; the words "Filter On" will appear in the lower-right corner of the window in which your spreadsheet appears. The filtering will stay on until you turn it off either by deleting the filter_$ variable, or going back to the Select Cases box and checking the first choice, *All cases*. Note that in the Output section of this box you have the option of deleting filtered

cases permanently from your spreadsheet, which would make sense only if you thought those cases had such serious mistakes that they could not be salvaged. Normally, you will want to go with the default choice: *Filter out unselected cases*.

It is important to keep in mind that the expression you type in the **Select Cases: If** box determines the cases that will be *included* (i.e., selected)—for example, cases with phobia ratings less than 9—rather than the cases which will be filtered out. Suppose you wanted to eliminate only cases with phobia ratings of exactly 9; in that case, you would type "phobia ~= 9." The tilde (~) followed by the equals sign means *not equal*, so the entire expression says: Include a case if its value for phobia does *not* equal 9 (see Figure 3.19).

Select Cases can be used as an alternative to Split File if you want to analyze only one major subgroup of your data, but not the others. For example, using the expression "gender = 1" as your **Select Cases: If** condition means that only the female students will be included in the following analyses (until you turn Select Cases off). You can become even more "selective" in selecting cases by setting multiple conditions that must be satisfied for a case to be included. If you want to perform an analysis on just the male psychology majors, you could do that by using the expression "*major* = 1 & *gender* = 2". The ampersand (&) implies that *both* conditions must be met for a case to be included. If you wanted to include only psychology and sociology majors, you would type: *major* = 1 | *major* = 4. Note that the vertical line in the preceding expression means *or*; it may appear as a "broken" vertical line on your keyboard, and it is sometimes referred to as the "pipe." The pipe character implies that a case will be included if it satisfies *either or both* of the conditions. Unfortunately, you cannot abbreviate the preceding expression like this: "*major* = 1 | 4"; the syntax rules of SPSS require that the variable name be repeated for each value.

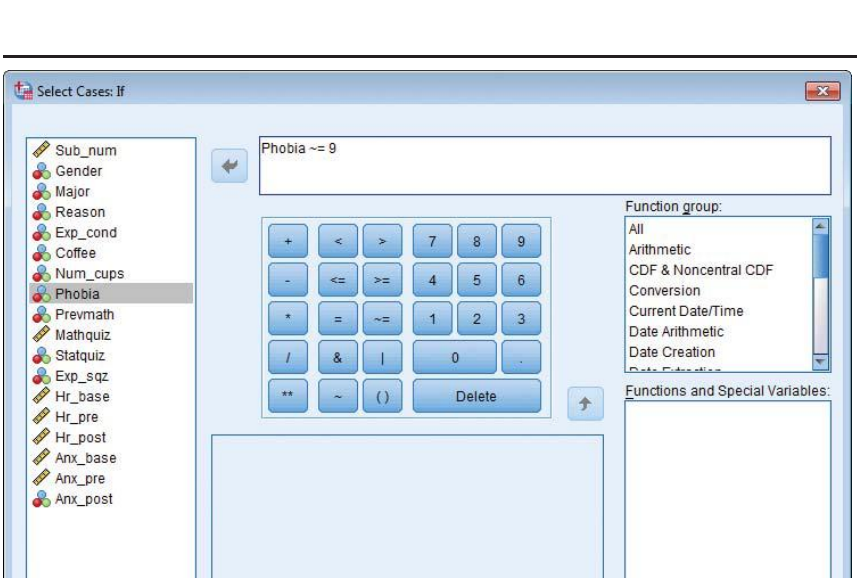


Figure 3.19

EXERCISES

1. Find the mode, median, and mean for each of the quantitative variables in Ihno's data set.
2. Find the mode for the undergraduate major variable.
3. Find the range, semi-interquartile range, unbiased variance, and unbiased standard deviation for each of the quantitative variables in Ihno's data set.
4. a. Create a boxplot for the *statquiz* variable. Then, use Split File to create a separate boxplot for the *statquiz* variable for each level of the *major* variable.
   b. Create boxplots for the statquiz variable for each level of the *major* variable so that all of the boxplots appear on the same graph.
   c. Use Select Cases to create a boxplot for the *statquiz* variable for just the female Biology majors.
   d. Use Select Cases to create a single boxplot for the *statquiz* variable that contains only the female Psychology majors and female Biology majors.
5. Create boxplots for both baseline and pre-quiz anxiety, so that they appear side-by-side on the same graph.
6. Use both Select Cases and Split File to find the mean and standard deviation for each of the quantitative variables separately for the male and female econ majors.

KEY FORMULAS

The semi-interquartile range after the 25th (Q1) and 75th (Q3) percentiles have been determined:

$$\text{SIQ range} = \frac{Q3 - Q1}{2} \qquad \textbf{Formula 3.1}$$

The mean deviation (after the mean of the distribution has been found):

$$\text{Mean deviation} = \frac{\sum |X_i - \mu|}{N} \qquad \textbf{Formula 3.2}$$

The sum of squares, definitional formula (requires that the mean of the distribution be found first):

$$SS = \sum (X_i - \mu)^2 \qquad \textbf{Formula 3.3}$$

The population variance, definitional formula (requires that the mean of the distribution be found first):

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \qquad \textbf{Formula 3.4A}$$

The population standard deviation, definitional formula (requires that the mean of the distribution be found first):

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} \qquad \textbf{Formula 3.4B}$$

The population variance (after *SS* has already been calculated):

$$\sigma^2 = MS = \frac{SS}{N} \qquad \textbf{Formula 3.5A}$$

The population standard deviation (after *SS* has been calculated):

$$\sigma = \sqrt{MS} = \sqrt{\frac{SS}{N}} \qquad \textbf{Formula 3.5B}$$

The unbiased sample variance, definitional formula:

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} \qquad \textbf{Formula 3.6A}$$

The unbiased sample standard deviation, definitional formula:

$$s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}} \qquad \textbf{Formula 3.6B}$$

The unbiased sample variance (after *SS* has been calculated):

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df} \qquad \textbf{Formula 3.7A}$$

The unbiased sample standard deviation (after *SS* has been calculated):

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}} \qquad \textbf{Formula 3.7B}$$

The arithmetic mean of a population:

$$\mu = \frac{\sum X}{N} \qquad \textbf{Formula 3.8}$$

The arithmetic mean of a sample:

$$\overline{X} = \frac{\sum X}{n} \qquad \textbf{Formula 3.9}$$

The weighted mean of two or more samples:

$$\overline{X}_w = \frac{\sum n_i \overline{X}_i}{\sum n_i} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \cdots}{n_1 + n_2 + \cdots} \qquad \textbf{Formula 3.10}$$

The sum of squares, computational formula (requires that the mean has been calculated):

$$SS = \sum X^2 - N\mu^2 \qquad \textbf{Formula 3.11}$$

The sum of squares, computational formula (direct from raw data):

$$SS = \sum X^2 - \frac{\left(\sum X\right)^2}{N} \qquad \textbf{Formula 3.12}$$

The population variance, computational formula (requires that the mean has been calculated):

$$\sigma^2 = \frac{\sum X^2}{N} - \mu^2 \qquad \textbf{Formula 3.13A}$$

The population standard deviation, computational formula (requires that the mean has been calculated):

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \mu^2} \qquad \textbf{Formula 3.13B}$$

The population variance, computational formula (direct from raw data):

$$\sigma^2 = \frac{1}{N}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right] \qquad \textbf{Formula 3.14A}$$

The population standard deviation, computational formula (direct from raw data):

$$\sigma = \sqrt{\frac{1}{N}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right]} \qquad \textbf{Formula 3.14B}$$

The unbiased sample variance, computational formula (direct from raw data):

$$s^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right] \qquad \textbf{Formula 3.15A}$$

The unbiased sample standard deviation, computational formula (direct from raw data):

$$s = \sqrt{\frac{1}{n-1}\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]} \qquad \textbf{Formula 3.15B}$$

The unbiased standard deviation (if the biased formula has already been used):

$$s = \sigma\sqrt{\frac{n}{n-1}} \qquad \textbf{Formula 3.16A}$$

The biased standard deviation (if the unbiased formula has already been used):

$$\sigma = s\sqrt{\frac{N-1}{N}} \qquad \textbf{Formula 3.16B}$$

Skewness of a population in dimensionless units:

$$Skewness = \frac{\sum (X_i - \mu)^3}{N\sigma^3} \qquad \textbf{Formula 3.17}$$

Kurtosis of a population in dimensionless units, adjusted so that the normal distribution has zero kurtosis:

$$Kurtosis = \frac{\sum (X_i - \mu)^4}{N\sigma^4} - 3 \qquad \textbf{Formula 3.18}$$