Introduction to Psychological Statistics

Part One
Descriptive Statistics

1
Chapter

A

CONCEPTUAL FOUNDATION

If you have not already read the Preface, please do so now. Many readers have developed the habit of skipping the Preface because it is often used by the author as a soapbox, or as an opportunity to give his or her autobiography and to thank many people the reader has never heard of. The Preface of this text is different and plays a particularly important role. You may have noticed that this book uses a unique form of organization (each chapter is broken into A, B, and C sections). The Preface explains the rationale for this unique format and explains how you can derive the most benefit from it.

What Is (Are) Statistics?

An obvious way to begin a text about statistics is to pose the rhetorical question, "What *is* statistics?" However, it is also proper to pose the question "What *are* statistics?"—because the term *statistics* can be used in at least two different ways. In one sense *statistics* refers to a collection of numerical facts, such as a set of performance measures for a baseball team (e.g., batting averages of the players) or the results of the latest U.S. census (e.g., the average size of households in each state of the United States). So the answer is that statistics are observations organized into numerical form.

In a second sense, *statistics* refers to a branch of mathematics that is concerned with methods for understanding and summarizing collections of numbers. So the answer to "What is statistics?" is that it is a set of methods for dealing with numerical facts. Psychologists, like other scientists, refer to numerical facts as *data*. The word *data* is a plural noun and always takes a plural verb, as in "the data *were* analyzed." (The singular form, datum, is rarely used.) Actually, there is a third meaning for the term *statistics*, which distinguishes a statistic from a parameter. To explain this distinction, I have to contrast samples with populations, which I will do at the end of this section.

As a part of mathematics, statistics has a theoretical side that can get very abstract. This text, however, deals only with *applied statistics*. It describes methods for data analysis that have been worked out by statisticians, but does not show how these methods were derived from more fundamental mathematical principles. For that part of the story, you would need to read a text on *theoretical* or *mathematical statistics* (e.g., Hogg & Craig, 1995).

The title of this text uses the phrase "psychological statistics." This could mean a collection of numerical facts about psychology (e.g., how large a percentage of the population claims to be happy), but as you have probably guessed, it actually refers to those statistical methods that are commonly applied to the analysis of psychological data. Indeed, just about every kind of statistical method has been used at one time or another to analyze some set of psychological data. The methods presented in this text are the ones usually taught in an intermediate (advanced undergraduate or graduate level) statistics course for psychology students, and they have been chosen because they are not only commonly used but are also simple to explain. Unfortunately, some methods that are now used frequently in psychological research (e.g., structural equation modeling) are too complex to be covered adequately at this level.

One part of applied statistics is concerned only with summarizing the set of data that a researcher has collected; this is called *descriptive statistics*. If all sixth graders in the United States take the same standardized exam, and you want a system for describing each student's standing with respect to the others, you need descriptive statistics. However, most psychological research involves relatively small groups of people from which inferences are drawn about the larger population; this branch of statistics is called *inferential statistics*. If you have a random sample of 100 patients who have been taking a new antidepressant drug, and you want to make a general statement about the drug's possible effectiveness in the entire population, you need inferential statistics. This text begins with a presentation of several procedures that are commonly used to create descriptive statistics. Although such methods can be used just to describe data, it is quite common to use these descriptive statistics as the basis for inferential procedures. The bulk of the text is devoted to some of the most common procedures of inferential statistics.

Statistics and Research

The reason a course in statistics is nearly universally required for psychology students is that statistical methods play a critical role in most types of psychological research. However, not all forms of research rely on statistics. For instance, it was once believed that only humans make and use tools. Then chimpanzees were observed stripping leaves from branches before inserting the branches into holes in logs to "fish" for termites to eat (van Lawick-Goodall, 1971). Certainly such an observation has to be replicated by different scientists in different settings before becoming widely accepted as evidence of toolmaking among chimpanzees, but statistical analysis is not necessary.

On the other hand, suppose you want to know whether a glass of warm milk at bedtime will help insomniacs get to sleep faster. In this case, the results are not likely to be obvious. You don't expect the warm milk to knock out any of the subjects, or even to help every one of them. The effect of the milk is likely to be small and noticeable only after averaging the time it takes a number of participants to fall asleep (the sleep latency) and comparing that to the average for a (control) group that does not get the milk. Descriptive statistics is required to demonstrate that there is a difference between the two groups, and inferential statistics is needed to show that if the experiment were repeated, it would be likely that the difference would be in the same direction. (If warm milk really has *no* effect on sleep latency, the next experiment would be just as likely to show that warm milk slightly increases sleep latency as to show that it slightly decreases it.)

Variables and Constants

A key concept in the above example is that the time it takes to fall asleep varies from one insomniac to another and also varies after a person drinks warm milk. Because sleep latency varies, it is called a *variable*. If sleep latency were the same for everyone, it would be a *constant*, and you really wouldn't need statistics to evaluate your research. It would be obvious after testing a few participants whether the milk was having an effect. But, because sleep latency varies from person to person and from night to night, it would not be obvious whether a particular case of shortened sleep latency was due to warm milk or just to the usual variability. Rather than focusing on any one instance of sleep latency, you would probably use statistics to compare a whole set of sleep latencies of people who drank warm milk with another whole set of people who did not.

In the field of physics there are many important constants (e.g., the speed of light, the mass of a proton), but most human characteristics vary a great deal from person to person. The number of chambers in the heart is a constant for humans (four), but resting heart rate is a variable. Many human variables (e.g., beauty, charisma) are easy to observe but hard to measure precisely or reliably. Because the types of statistical procedures that can be used to analyze the data from a research study depend in part on the way the variables involved were measured, we turn to this topic next.

Scales of Measurement

Measurement is a system for assigning numerical values to observations in a consistent and reproducible way. When most people think of measurement, they think first of physical measurement, in which numbers and measurement units (e.g., minutes and seconds for sleep latency) are used in a precise way. However, in a broad sense, measurement need not involve numbers at all. Due in large part to the seminal work of S. S. Stevens, psychologists have become accustomed to thinking in terms of levels of measurement that range from the merely categorical to the numerically precise. The four-scale system devised by Stevens (1946) is presented next. Note that the utility of this system is a matter of considerable controversy (Velleman & Wilkinson, 1993), but it has become much too popular to ignore. I will address the controversy after I describe the scales.

Nominal Scales

Facial expressions can be classified by the emotions they express (e.g., anger, happiness, surprise). The different emotions can be considered values on a *nominal scale*; the term *nominal* refers to the fact that the values are simply named, rather than assigned numbers. (Some emotions can be identified quite reliably, even across diverse cultures and geographical locations; see Ekman, 1982.) If numbers are assigned to the values of a nominal scale, they are assigned arbitrarily and therefore cannot be used for mathematical operations. For example, the *Diagnostic and Statistical Manual* of the American Psychiatric Association (the latest version is *DSM-5*) assigns a number as well as a name to each psychiatric diagnosis (e.g., the number 300.3 designates obsessive-compulsive disorder). However, it makes no sense to use these numbers mathematically; for instance, you cannot average the numerical diagnoses of all the members in a family to find out the average mental illness of the family. Even the order of the assigned numbers is mostly arbitrary; the higher *DSM-5* numbers do not indicate more severe diagnoses.

Many variables that are important to psychology (e.g., gender, type of psychotherapy) can be measured only on a nominal scale, so we will be dealing with this level of measurement throughout the text. Nominal scales are often referred to as *categorical scales* because the different levels of the scale represent distinct categories; each object measured is assigned to one and only one category. A nominal scale is also referred to as a *qualitative* level of measurement because each level has a different quality and therefore cannot be compared with other levels with respect to quantity.

Ordinal Scales

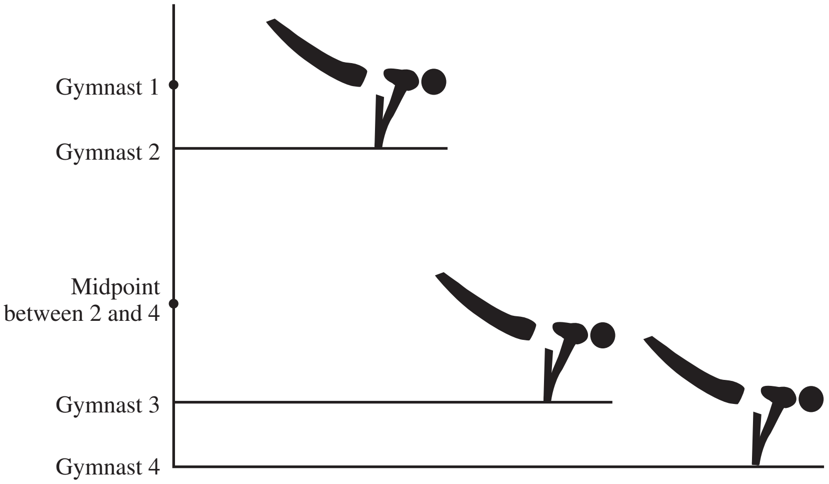
A quantitative level of measurement is being used when the different values of a scale can be placed in order. For instance, an elementary school teacher may rate the handwriting of each student in a class as excellent, good, fair, or poor. Unlike the categories of a nominal scale, these designations have a meaningful order and therefore constitute an *ordinal scale*. One can add the percentage of students rated excellent to the percentage of students rated good, for instance, and then make the statement that a certain percentage of the students have handwriting that is "better than fair."

Often the levels of an ordinal scale are given numbers, as when a coach rank-orders the gymnasts on a team based on ability. These numbers are not arbitrary like the numbers that may be assigned to the categories of a nominal scale; the gymnast ranked number 2 *is* better than the gymnast ranked number 4, and gymnast number 3 is somewhere between. However, the rankings cannot be treated as real numbers; that is, it cannot be assumed that the third-ranked gymnast is midway between the second and the fourth. In fact, it could be the case that the number 2 gymnast is much better than either number 3 or 4, and that number 3 is only slightly better than number 4 (as shown in Figure 1.1). Although the average of the numbers 2 and 4 is 3, the average of the abilities of the number 2 and 4 gymnasts is not equivalent to the abilities of gymnast number 3.

A typical example of the use of an ordinal scale in psychology is when photographs of human faces are rank-ordered for attractiveness. A less obvious example is the measurement of anxiety by means of a self-rated questionnaire (on which subjects indicate the frequency of various anxiety symptoms in their lives using numbers corresponding to never, sometimes, often, etc.). Higher scores can generally be thought of as indicating greater amounts of anxiety, but it is not likely that the anxiety difference between subjects scoring 20 and 30 is going to be exactly the same as the anxiety

Figure 1.1

Ordinal Scale



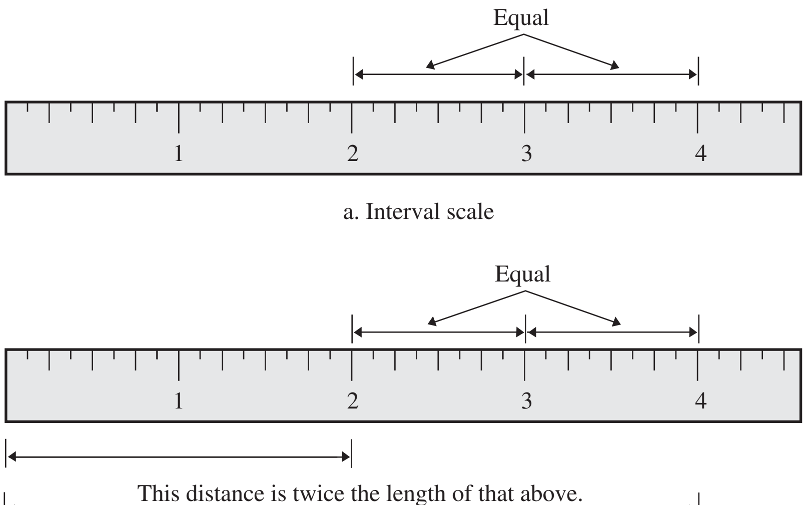
difference between subjects scoring 40 and 50. Nonetheless, scores from anxiety questionnaires and similar psychological measures are usually dealt with mathematically by researchers as though they were certain the scores were equally spaced throughout the scale, and therein lies the main controversy concerning Stevens's breakdown of the four scales.

Those who take Stevens's scale definitions most seriously contend that when dealing with an ordinal scale (when you are sure of the order of the levels but not sure that the levels are equally spaced), you should use statistical procedures that have been devised specifically for use with ordinal data. The descriptive statistics that apply to ordinal data as well as to data measured on the other scales will be discussed in the next two chapters. The use of inferential statistics with ordinal data will not be presented in this text, but will be dealt with in a separate chapter that will be available from the website for this text (see Preface).

Interval and Ratio Scales

In general, physical measurements have a level of precision that goes beyond the ordinal property previously described. We are confident that the inch marks on a ruler are equally spaced; we know that considerable effort goes into making sure of this. Because we know that the space, or interval, between 2 and 3 inches is the same as that between 3 and 4 inches, we can say that this measurement scale possesses the *interval property* (see Figure 1.2a). Such scales are based on *units* of measurement (e.g., the inch); a unit at one part of the scale is always the same size as a unit at any other part of the scale. It is therefore permissible to treat the numbers on this kind of scale as actual numbers and to assume that a measurement of three units is exactly halfway between two and four units.

In addition, most physical measurements possess what is called the *ratio property*. This means that when your measurement scale tells you that you now have twice as many units of the variable as before, you really *do* have twice as much of the variable. Measurements of sleep latency in minutes and seconds have this property. When a subject's sleep latency is



a. Interval scale

b. Ratio scale

Figure 1.2

Interval and Ratio Scales

20 minutes, it has taken that person twice as long to fall asleep as a subject with a sleep latency of 10 minutes. Measuring the lengths of objects with a ruler also involves the ratio property. Scales that have the ratio property in addition to the interval property are called *ratio scales* (see Figure 1.2b).

Whereas all ratio scales have the interval property, there are some scales that have the interval property but not the ratio property. These scales are called *interval scales*. Such scales are relatively rare in the realm of physical measurement; perhaps the best-known examples are the Celsius (also known as centigrade) and Fahrenheit temperature scales. The degrees are equally spaced, according to the interval property, but one cannot say that something that has a temperature of 40 degrees is twice as hot as something that has a temperature of 20 degrees. The reason these two temperature scales lack the ratio property is that the zero point for each is arbitrary. Both scales have different zero points (0 °C = 32 °F), but in neither case does zero indicate a total lack of heat. (Heat comes from the motion of particles within a substance, and as long as there is some motion, there is some heat.) In contrast, the Kelvin scale of temperature is a true ratio scale because its zero point represents *absolute* zero temperature—a total lack of heat. (Theoretically, the motion of internal particles has stopped completely.)

Although interval scales that are not also ratio scales may be rare when dealing with physical measurement, they are not uncommon in psychological research. If we grant that IQ scores have the interval property (which is open to debate), we still would not consider IQ a ratio scale. It doesn't make sense to say that someone who scores a zero on a particular IQ test has no intelligence at all, unless intelligence is defined very narrowly. And does it make sense to say that someone with an IQ of 150 is exactly twice as intelligent as someone who scores 75?

Parametric Versus Nonparametric Statistics

Because nearly all common statistical procedures are just as valid for interval scales as they are for ratio scales (including all of the inferential methods that will be described in Parts II through VI of this text), it is customary to discuss these two types of scales together by referring to their products as *interval/ratio data*. Large amounts of interval/ratio data can usually be arranged into smooth distributions, which will be explained in greater detail in the next few chapters. These empirical data distributions often resemble well-known mathematical distributions, which can be summarized by just a few values called parameters. Statistical procedures based on distributions and their parameters are called *parametric statistics*. With interval/ratio data it is often (but not always) appropriate to use parametric statistics. Conversely, parametric statistics were designed to be used with interval/ratio data. Whether it makes sense to apply parametric statistics to data obtained from ordinal scales will be discussed in the next subsection. The bulk of this text (i.e., Parts II through VI) is devoted to parametric statistics. If all of your variables have been measured on nominal scales, or your interval/ratio data do not even come close to meeting the distributional assumptions of parametric statistics (which will be explained at the appropriate time), you should be using *nonparametric statistics*, as described in Part VII.

For some purposes, it makes sense to describe any scale that measures different amounts of the same variable, so that cases can at least be placed in order with respect to how much of that variable they exhibit, as a *quantitative* scale. Thus, data from ordinal, interval, or ratio scales can be referred to as quantitative data. By contrast, the categories of a nominal scale do *not* differ in the amount of a common variable; the categories differ in a qualitative sense. Therefore, data from a nominal scale are referred to as *qualitative data*. Part VII of this text is devoted to the analysis of qualitative data. Techniques for dealing specifically with ordinal data, which are included under the heading of nonparametric statistics, will be available in a separate chapter, which, as I mentioned earlier, will be available only on the web.

Likert Scales and the Measurement Controversy

One of the most common forms of measurement in psychological research, especially in social psychology, involves participants responding to a statement by indicating their degree of agreement on a Likert scale, named after its creator, Rensis Likert (1932). A typical Likert scale contains the following five ordered choices: strongly disagree; disagree; neither agree nor disagree; agree; strongly agree (a common variation is the 7-point Likert scale). These scales clearly possess the ordinal property, but there is some controversy concerning whether they can be legitimately treated as interval scales. For instance, if the numbers 1 through 5 are assigned to the choices of a 5-point Likert scale, one can ask: “Is it meaningful to average these numbers across a group of individuals responding to the same statement, and compare that average to the average for a different group?”

To take a concrete example, suppose that two psychology majors each choose “agree” in response to the statement “I enjoy reading about statistics,” and two economics majors respond such that one chooses “strongly agree,” and the other chooses the middle response. The choices of the two psychology majors could both be coded as 4, and the choices of the two economics majors could be coded 5 and 3, respectively, so both groups would have an average agreement of 4.0. However, to say that the two groups are expressing an equal amount of enjoyment for reading about statistics requires assuming that the difference in enjoyment between the ratings of “neither agree nor disagree” and “agree” is the same as the difference between the ratings of “agree” and “strongly agree,” which would be required to make this an interval scale. Given that there is no basis for making the interval assumption, it can be argued that Likert scales are no more precise than any other ordinal scales, and, according to Stevens (1951), it is not permissible to perform mathematical operations, like averages, on numbers derived from ordinal scales.

Statisticians have convincingly argued against Stevens’s strict rules about measurement scales and which mathematical operations are permissible for each scale. In summarizing many of these arguments, Velleman and Wilkinson (1993) point out that what matters most in determining which types of statistics can be validly applied to your data is the type of questions you are asking of your data, and what you are trying to accomplish. Norman’s (2010) main argument in favor of applying parametric statistics to ordinal data is that empirical and statistical studies have shown that these procedures are *robust* with respect to the interval scale assumption—that is, a lack of equality of intervals by itself has little impact on the final statistical conclusions.

Note that a single Likert item is rarely used as a major dependent variable. It is much more common to present to participants a series of similar items (e.g., I feel tense; I feel jumpy; I cannot relax), each of which is responded to on the same Likert scale, and then to average the numerically coded responses together to create a single score for, say, experienced anxiety. Some statisticians are more comfortable with attributing the interval property to a sum or average of Likert items than to a single item, but it is common for psychologists to apply parametric statistics, regardless of the number of Likert items contained in the scale. Also note that other rating scales are treated in the same way as the Likert scales I have been describing. For example, ratings of facial attractiveness on a scale from 1 to 10 can be properly characterized as ordinal data, but they are usually averaged together and subjected to parametric statistics as though they possessed the interval property.

Continuous Versus Discrete Variables

One distinction among variables that affects the way they are measured is that some variables vary continuously, whereas others have only a finite (or countable) number of levels with no intermediate values possible. The latter variables are said to be discrete (see Figure 1.3). A simple example of a *continuous variable* is height; no matter how close two people are in height, it is theoretically possible to find someone whose height is somewhere between those two people. (Quantum physics has shown that there are limitations to the precision of measurement, and it may be meaningless to talk of continuous variables at the quantum level, but these concepts have no practical implications for psychological research.)

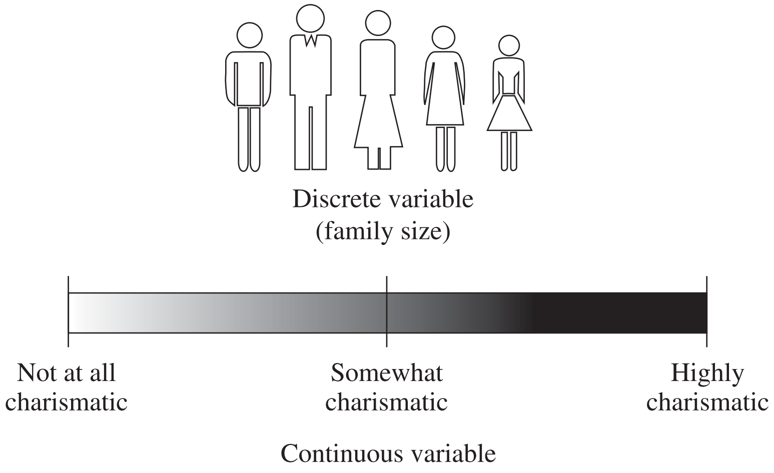
An example of a *discrete variable* is the size of a family. This variable can be measured on a ratio scale by simply counting the family members, but it does not vary continuously—a family can have two or three children, but there is no meaningful value in between. The size of a family will always be a whole number and never involve a fraction (even if Mom is pregnant). The distinction between discrete and continuous variables affects some of the procedures for displaying and describing data, as you will see in the next chapter. Fortunately, however, the inferential statistics discussed in Parts II through VI of this text are *not* affected by whether the variable measured is discrete or continuous, as long as the variable is measured on a quantitative scale.

Scales Versus Variables Versus Underlying Constructs

It is important not to confuse variables with the scales with which they are measured. For instance, the temperature of the air outside can be

Figure 1.3

Discrete and Continuous Variables



measured on an ordinal scale (e.g., the hottest day of the year, the third hottest day), an interval scale (degrees Celsius or Fahrenheit), or a ratio scale (degrees Kelvin); these three scales are measuring the same physical quantity but yield very different measurements. In many cases, a variable that varies continuously, such as charisma, can only be measured crudely, with relatively few levels (e.g., highly charismatic, somewhat charismatic, not at all charismatic). On the other hand, a continuous variable such as generosity can be measured rather precisely by the exact amount of money donated to charity in a year (which is at least one aspect of generosity). Although in an ultimate sense all scales are discrete, scales with very many levels relative to the quantities measured are treated as continuous for display purposes, whereas scales with relatively few levels are usually treated as discrete (see Chapter 2). Of course, the scale used to measure a discrete variable is always treated as discrete.

Choosing a scale is just one part of *operationalizing* a variable, which also includes specifying the method by which an object will be measured. If the variable of interest is the height of human participants in a study, a scale based on inches or centimeters, for instance, can be chosen, and an operation can then be specified: place a measuring tape, marked off by the chosen scale, along the participant's body. Specifying the *operationalization* of the variable helps to ensure that one's measurements can be easily and reliably reproduced by other scientists. In the case of a simple physical measurement such as height, there is little room for confusion or controversy. However, for many important psychological variables, the exact *operationalization* of the variable is critical, as there may be plenty of room for disagreement among researchers studying the same ostensible phenomenon.

Let us reconsider the example of generosity. Unlike height, the term generosity does not refer to some obvious variable that can be measured in an easily agreed-upon way. Rather, it is an *underlying construct* that is understood intuitively, but is hard to define exactly. In some contexts, generosity can be viewed as a *latent variable*, as opposed to a manifest or observed variable. One way to operationalize the measurement of generosity is to record the total amount of charitable deductions on an individual's tax return. This will likely yield a different result, and not necessarily a more accurate one, than asking the individual to report all of his or her charitable donations, including those that might not qualify as a tax deduction. An alternative approach would be to ask a participant in a study to donate some proportion (whatever they are comfortable with) of the amount they were paid for the experiment back to the experimenter so more participants could be run.

So far, all of these operationalized variables involve money, which can have very different meanings to different people. A completely different variable for measuring generosity would involve asking participants to donate their time to helping a charitable cause. However, some people are very generous with their time in helping friends and family, but not strangers. As you can see, whatever variable is chosen as a measure of generosity will capture only an aspect of the underlying construct, and whatever statistical results are based on that variable can only contribute partially and indirectly to the understanding of that construct. This is a humbling reality for many areas of psychological research.

Independent Versus Dependent Variables

Returning to the experiment in which one group of insomniacs gets warm milk before bedtime and the other does not, note that there are actually *two* variables involved in this experiment. One of these, sleep latency, has already been discussed; it is being measured on a ratio scale. The other variable is less obvious; it is group membership. That is, subjects *vary* as to which experimental condition they are in—some receive milk, and some do not. This variable, which in this case has only two levels, is called the *independent variable*. A subject's level on this variable—that is, which group a subject is placed in—is determined at random by the experimenter and is independent of anything that happens during the experiment. The other variable, sleep latency, is called the *dependent variable* because its value depends (it is hoped) at least partially on the value of the independent variable. That is, sleep latency is expected to depend in part on whether the subject drinks milk before bedtime. Notice that the independent variable is measured on a nominal scale (the two categories are "milk" and "no milk"). However, because the dependent variable is being measured on a ratio scale, parametric statistical analysis is appropriate. If neither of the variables were measured on an interval or ratio scale (for example, if sleep latency were categorized as simply less than or greater than 10 minutes), a nonparametric statistical procedure would be needed (see Part VII). If the independent variable were also being measured on an interval/ratio scale (e.g., amount of milk given) you would still use parametric statistics, but of a different type (see Chapter 9). I will discuss different experimental designs as they become relevant to the statistical procedures I am describing. For now, I will simply point out that parametric statistics can be used to analyze the data from an experiment, even if the independent variable is measured on a nominal scale.

Experimental Versus Observational Research

It is important to realize that not all research involves experiments; much of the research in some areas of psychology involves measuring differences between groups that were not created by the researcher. For instance, insomniacs can be compared to normal sleepers on variables such as anxiety. If inferential statistics shows that insomniacs, in general, differ from normal sleepers in daily anxiety, it is interesting, but we still do not know whether the greater anxiety causes the insomnia, the insomnia causes the greater anxiety, or some third variable (e.g., increased muscle tension) causes both. We cannot make causal conclusions because we are not in control of who is an insomniac and who is not. Nonetheless, such *observational* (also called quasi-experimental) studies can produce useful insights and sometimes suggest confirmatory experiments.

To continue this example: If a comparison of insomniacs and normal sleepers reveals a statistically reliable difference in the amount of sugar consumed daily, these results suggest that sugar consumption may be interfering with sleep. In this case, observational research has led to an interesting hypothesis that can be tested more conclusively by means of an experiment. A researcher randomly selects two groups of sugar-eating insomniacs; one group is restricted from eating sugar and the other is not. If the sugar-restricted insomniacs sleep better, that evidence supports the notion that sugar consumption interferes with sleep. If there is no sleep difference between the groups, the causal connection may be in the opposite direction (i.e., lack of sleep may produce an increased craving for sugar), or the insomnia may be due to some as yet unidentified third variable (e.g., maybe anxiety produces both insomnia *and* a craving for sugar). The statistical analysis is generally the same for both experimental and quasi-experimental research; it is the causal conclusions that differ.

Populations Versus Samples

In psychological research, measurements are often performed on some aspect of a person. The psychologist may want to know about people's ability to remember faces, solve anagrams, or experience happiness. The collection of all people who could be measured, or in whom the psychologist is interested, is called the *population*. However, it is not always people who are the subjects of measurement in psychological research. A population can consist of laboratory rats, mental hospitals, married couples, small towns, and so forth. Indeed, as far as theoretical statisticians are concerned, a population is just a set (ideally one that is infinitely large) of numbers. The statistical procedures used to analyze data are the same regardless of where the numbers come from (as long as certain assumptions are met, as subsequent chapters will make clear). In fact, the statistical methods you will be studying in this text were originally devised to solve problems in agriculture, beer manufacturing, human genetics, and other diverse areas.

If you had measurements for an entire population, you would have so many numbers that you would surely want to use descriptive statistics to summarize your results. This would also enable you to compare any individual to the rest of the population, compare two different variables measured on the same population, or even to compare two different populations measured on the same variable. More often, practical limitations will prevent you from gathering all of the measurements that you might want. In such cases you would obtain measurements for some subset of the population. This subset is called a *sample* (see Figure 1.4).

Sampling is something we all do in daily life. If you have tried two or three items from the menu of a nearby restaurant and have not liked any of them, you do not have to try everything on the menu before deciding not to dine at that restaurant anymore. When you are conducting research, you follow a more formal sampling procedure. If you have obtained measurements on a sample, you would probably begin by using descriptive statistics to summarize the data in your sample. But it is not likely that you would stop there. Usually, you would then use the procedures of inferential statistics to draw some conclusions about the entire population from which you obtained your sample. Strictly speaking, these conclusions would be valid only if your sample was a *random sample*. In reality, truly random samples of human beings are virtually impossible to obtain, so most psychology research is conducted on *samples of convenience* (e.g., students in an introductory psychology class who must either "volunteer" for some experiments or complete some alternative assignment). To the extent that one's sample is not truly random, it may be difficult to generalize one's

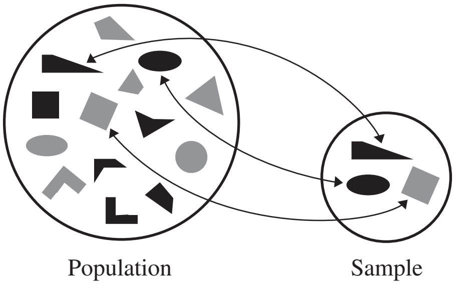


Figure 1.4

A Population and a Sample

results to the larger population. The role of sampling in inferential statistics will be discussed at greater length in Part II.

Now we come to the third definition for the term *statistic*. A *statistic* is a value derived from the data in a sample rather than a population. It could be a value derived from all of the data in the sample, such as the mean, or it could be just one measurement in the sample, such as the maximum value. If the same mathematical operation used to derive a statistic from a sample is performed on the entire population from which you selected the sample, the result is called a population *parameter* rather than a sample statistic. As you will see, sample statistics are often used to make estimates of, or draw inferences about, corresponding population parameters.

Much of the controversy surrounding the use of parametric statistics to evaluate psychological research arises because the distributions of many psychological variables, measured on actual people, do not match the theoretical mathematical distributions on which the common methods are based. Often the researcher has collected so few data points that the empirical distribution (i.e., the distribution of the data collected) gives no clear basis for determining which theoretical distribution would best represent the population. Moreover, using any theoretical distribution to represent a finite population of psychological measurements involves some degree of approximation.

Fortunately, the procedures described in this text are applicable to a wide range of psychological variables, and computer simulation studies have shown that the approximations involved usually do not produce errors large enough to be of practical significance. You can rest assured that I will not have much more to say about the theoretical basis for the applied statistics presented in this text, except to explain, where appropriate, the assumptions underlying the use of inferential statistics to analyze the data from psychological research.

Statistical Formulas

Many descriptive statistics, as well as sample statistics that are used for inference, are found by means of statistical formulas. Often these formulas are applied to all of the measurements that have been collected, so a notational system is needed for referring to many data points at once. It is also frequently necessary to add many measurements together, so a symbol is needed to represent this operation. Throughout the text, Section B will be reserved for a presentation of the nuts and bolts of statistical analysis. The first Section B will present the building blocks of all statistical formulas: subscripted variables and summation signs.

A SUMMARY

1. Descriptive statistics is concerned with summarizing a given set of measurements, whereas inferential statistics is concerned with generalizing beyond the given data to some larger potential set of measurements.
2. The type of descriptive or inferential statistics that can be applied to a set of data depends, in part, on the type of measurement scale that was used to obtain the data.
3. If the different levels of a variable can be named, but not placed in any specific order, a *nominal scale* is being used. The categories in a nominal scale can be numbered, but the numbers cannot be used in any mathematical way—even the ordering of the numbers would be arbitrary.

4. If the levels of a scale can be ordered, but the intervals between adjacent levels are not guaranteed to be the same size, you are dealing with an *ordinal scale*. The levels can be assigned numbers, as when subjects or items are rank-ordered along some dimension, but there is some debate as to whether these numbers can or cannot be used for arithmetical operations, because we cannot be sure that the average of ranks 1 and 3, for instance, equals rank 2.
5. If the intervals corresponding to the units of measurement on a scale are always equal (e.g., the difference between two and three units is the same as between four and five units), the scale has the interval property. Scales that have equal intervals but do not have a true zero point are called *interval scales*.
6. If an interval scale has a true zero point (i.e., zero on the scale indicates a total absence of the variable being measured), the ratio between two measurements will be meaningful (a fish that is 30 inches long is twice as long as one that is 15 inches long). A scale that has both the interval and the ratio properties is called a *ratio scale*.
7. A variable that has countable levels with no values possible between any two adjacent levels is called a *discrete variable*. A variable that can be measured with infinite precision (i.e., intermediate measurements are always possible), at least in theory, is called a *continuous variable*. In practice, most physical measurements are treated as continuous even though they are not infinitely precise.
8. The entire set of measurements about which one is concerned is referred to as a *population*. The measurements that comprise a population can be from individual people, families, animals, hospitals, cities, and so forth. A subset of a population is called a *sample*, especially if the subset is considerably smaller than the population and is chosen at random.
9. Values that are derived from and in some way summarize samples are called *statistics*, whereas values that describe a population are called *parameters*.
10. If at least one of your variables has been measured on an interval or ratio scale, and certain additional assumptions have been met, it may be appropriate to use *parametric statistics* to draw inferences about population parameters from your sample statistics. If all of your variables have been measured on ordinal or nominal scales, or the assumptions of parametric statistics have not been met, it may be necessary to use *nonparametric statistics*.

EXERCISES

1. Give two examples of each of the following:
   a. Nominal scale
   b. Ordinal scale
   c. Interval scale
   d. Ratio scale
   e. Continuous variable
   f. Discrete variable

*2. What type of scale is being used for each of the following measurements?
   a. Number of arithmetic problems correctly solved
   b. Class standing (i.e., one's rank in the graduating class)
   c. Type of phobia
   d. Body temperature (in °F)
   e. Self-esteem, as measured by self-report questionnaire
   f. Annual income in dollars
   g. Theoretical orientation toward psychotherapy
   h. Place in a dog show
   i. Heart rate in beats per minute

*3. Which of the following variables are discrete and which are continuous?
   a. The number of people in one's social network
   b. Intelligence
   c. Size of vocabulary
   d. Blood pressure
   e. Need for achievement

4. a. Give two examples of a population that does not consist of individual people.
   b. For each population described in part a, indicate how you might obtain a sample.

*5. A psychologist records how many words participants recall from a list under three different conditions: large reward for each word recalled, small reward for each word recalled, and no reward.
   a. What is the independent variable?
   b. What is the dependent variable?
   c. What kind of scale is being used to measure the dependent variable?

6. Patients are randomly assigned to one of four types of psychotherapy. The progress of each subject is rated at the end of 6 months.
   a. What is the independent variable?
   b. What is the dependent variable?
   c. What kind of scale is formed by the levels of the independent variable?
   d. Describe one type of scale that might be used to measure the dependent variable.

*7. Which of the following studies are experimental and which are observational?
   a. Comparing pet owners with those who don't own pets on an empathy measure
   b. Comparing men and women with respect to performance on a video game that simulates landing a space shuttle
   c. Comparing participants run by a male experimenter with participants run by a female experimenter with respect to the number of tasks completed in 1 hour
   d. Comparing the solution times of participants given a hint with those not given a hint

8. Which of the following would be called a statistic and which a parameter?
   a. The average income for 100 U.S. citizens selected at random from various telephone books
   b. The average income of citizens in the United States
   c. The highest age among respondents to a sex survey in a popular magazine

Throughout the text, asterisks (*) will precede the exercises that have answers appearing in Appendix B.

B BASIC STATISTICAL PROCEDURES

Variables With Subscripts

Recognizing that statistics is a branch of mathematics, you should not be surprised that its procedures are usually expressed in terms of mathematical notation. For instance, you probably recall from high school math that a variable whose value is unknown is most commonly represented by the letter X. This is also the way a statistician would represent a *random variable*. However, to describe statistical manipulations with samples, we need to refer to collections of random variables. Because this concept is rather abstract, I will use a very concrete example.

When describing the characteristics of a city to people who are considering living there, a realtor typically gives a number of facts such as the average income and the size of the population. Another common statistic is the average temperature in July. The usual way to find the average temperature for the entire month of July is to take the average temperature for each day in July and then average these averages. To express this procedure symbolically it would be helpful to find a way to represent the average temperature for any particular day in July. It should be obvious that it would be awkward to use the same letter, X, for each day of the month if we then want to write a formula that tells us how to combine these 31 different averages into a single average. On the other hand, we certainly cannot use a different letter of the alphabet for each day. The solution is to use subscripts. The average temperature for July 1 can be written X_1, for July 2, X_2, and so on up to X_{31}. We now have a compact way of referring to 31 different variables. If we wanted to indicate a different type of variable, such as high or low temperature for each day, we would need to use a different letter (e.g., Y_1, Y_2, up to Y_{31}). If we want to make some general statement about the average temperature for any day in July without specifying which particular day, we can write X_i. The letter i used as a subscript stands for the word *index* and can take the place of any numerical subscript.

The Summation Sign

To get the average temperature for the month of July, we must add up the average temperatures for each day in July and then divide by 31. Using the subscripts introduced above, the average temperature for July can be expressed as $(X_1 + X_2 + X_3 + \cdots + X_{31})/31$. (Note that because it would take up a lot of room to write out all of the 31 variables, dots are used to indicate that variables have been left out.) Fortunately, there is a neater way of indicating that all the variables from X_1 to X_{31} should be added. The mathematical symbol that indicates that a string of variables is to be added is called the summation sign, and it is symbolized by the uppercase Greek letter sigma (Σ). The summation sign works in conjunction with the subscripts on the variables in the following manner. First, you write $i = 1$ under the summation sign to indicate that the summing should start with the variable that has the subscript 1. (You could write $i = 2$ to indicate that you want to begin with the second variable, but it is rare to start with any subscript other than 1.) On top of the summation sign you indicate the subscript of the last variable to be added. Finally, next to the summation sign you write the letter that stands for the collection of variables to be added, using the subscript i. So the sum of the average temperatures for each day in July can be symbolized as follows:

$$\sum_{i=1}^{31} X_i$$

This expression is a neat, compact way of telling you to perform the following:

1. Take X_i and replace i with the number indicated under the summation sign (in this case, you would write X_1).
2. Put a plus sign to the right of the previous expression (X_1 +).
3. Write X_i again, this time replacing i with the next integer, and add another plus sign ($X_1 + X_2 +$).
4. Continue the above process until i has been replaced by the number on top of the summation sign ($X_1 + X_2 + X_3 + \cdots + X_{31}$).

If you wanted to write a general expression for the sum of the average temperatures on all the days of any month, you could not use the number 31 on top of the summation sign (e.g., June has only 30 days). To be more general, you could use the letter N to stand for the number of days in any month, which leads to the following expression:

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

To find the average temperature for the month in question, we would divide the above sum by N (the number of days in that month). The whole topic of finding averages will be dealt with in detail in Chapter 3. For now we will concentrate on the mathematics of finding sums.

Summation notation can easily be applied to samples from a population, where N represents the sample size. For instance, if N is the number of people who are allowed by law on a particular elevator, and X_i is the weight of any one particular person, the previous expression represents the total weight of the people on some elevator that is full to its legal capacity. When statisticians use summation signs in statistical formulas, $i = 1$ almost always appears under the summation sign and N appears above it. Therefore, most introductory statistics texts leave out these indexes and simply write the summation sign by itself, expecting the reader to assume that the summation goes from $i = 1$ to N. Although mathematical statisticians dislike this lack of precision, I will, for the sake of simplicity, go along with the practice of leaving off the indexes from summation signs, and usually from variables, as well.

The summation sign plays a role in most of the statistical formulas in this text. To understand those formulas fully it is helpful to know several interesting mathematical properties involved with the use of the summation sign. The most important of those properties will be presented in the remainder of this section.

Properties of the Summation Sign

The first property we will discuss concerns the addition of two collections of variables. Returning to our example about the temperature in July, suppose that you are interested in a temperature-humidity index (THI), which is a better indicator of comfort than temperature alone. Assume that the average THI for any day is just equal to the average temperature of that day (X_i) plus the average humidity of that day (Y_i) (although this is not the index that is usually used). Thus we can express the THI for any day as $X_i + Y_i$. If you wanted to add the THI for all the days in the month, you could use the following general expression: $\Sigma(X_i + Y_i)$. This expression produces the same result as adding the Xs and Ys separately. This leads to our first rule for dealing with summation signs.

Summation Rule 1A

$$\sum (X_i + Y_i) = \sum X_i + \sum Y_i$$

The rule works in exactly the same way for subtraction.

Summation Rule 1B

$$\sum (X_i - Y_i) = \sum X_i - \sum Y_i$$

Rule 1A works because if all you're doing is adding, it doesn't matter what order you use. Note that $\Sigma(X_i + Y_i)$ can be written as:

$$(X_1 + Y_1) + (X_2 + Y_2) + (X_3 + Y_3) + \cdots + (X_N + Y_N)$$

If you remove the parentheses and change the order, as follows,

$$X_1 + X_2 + X_3 + \cdots + X_N + Y_1 + Y_2 + Y_3 + \cdots + Y_N$$

you can see that the above expression is equal to $\Sigma X_i + \Sigma Y_i$. The proof for Rule 1B is exactly parallel.

Sometimes the summation sign is applied to a constant: ΣC_i. In this case, we could write $C_1 + C_2 + C_3 + \cdots + C_N$, but all of these terms are just equal to C, the value of the constant. The fact that the number of Cs being added is equal to N leads to the following rule.

Summation Rule 2

$$\sum C = NC$$

In the equation above, the subscript on the letter C was left off because it is unnecessary and is not normally used.

Quite often a variable is multiplied or divided by a constant before the summation sign is applied: ΣCX_i. This expression can be simplified without changing its value by placing the constant in front of the summation sign. This leads to the next summation rule.

Summation Rule 3

$$\sum CX_i = C \sum X_i$$

The advantage of this rule is that it reduces computational effort. Instead of multiplying every value of the variable by the constant before adding, we can first add up all the values and then multiply the sum by the constant. You can see why Rule 3 works by writing out the expression and rearranging the terms:

$$\sum CX_i = CX_1 + CX_2 + CX_3 + \cdots + CX_N$$

The constant C can be factored out of each term, and the rest can be placed in parentheses, as follows: $C(X_1 + X_2 + X_3 + \cdots + X_N)$. The part in parentheses is equal to $\sum X_i$, so the entire expression equals $C \sum X_i$.

The last rule presents a simplification that is *not* allowed. Because $\sum(X_i + Y_i) = \sum X_i + \sum Y_i$, it is tempting to assume that $\sum X_i Y_i$ equals $\left(\sum X_i\right)\left(\sum Y_i\right)$ but unfortunately this is *not* true. In the case of Rule 1A, only addition is involved, so the order of operations does not matter (the same is true with a mixture of subtraction and addition). But when multiplication and addition are mixed together, the order of operations cannot be changed without affecting the value of the expression. This leads to the fourth rule.

Summation Rule 4

$$\sum(X_i Y_i) \neq \left(\sum X_i\right)\left(\sum Y_i\right)$$

This inequality can be demonstrated with a simple numerical example. Assume that:

$$X_1 = 1 \quad X_2 = 2 \quad X_3 = 3 \quad Y_1 = 4 \quad Y_2 = 5 \quad Y_3 = 6$$

$$\sum(X_i Y_i) = 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 = 4 + 10 + 18 = 32$$

$$\left(\sum X_i\right)\left(\sum Y_i\right) = (1 + 2 + 3)(4 + 5 + 6) = (6)(15) = 90$$

As you can see, the two sides of the above inequality do not yield the same numerical value.

An important application of Rule 4 involves the case in which X and Y are equal, so we have $\sum(X_iX_i) \neq \sum(X_i)\sum(X_i)$. Because X_iX_i equals X_i^2 and $\left(\sum X_i\right)\left(\sum X_i\right) = \left(\sum X_i\right)^2$, a consequence of Rule 4 is that:

$$\sum X_i^2 \neq \left(\sum X_i\right)^2$$

This is an important property to remember because both terms play an important role in statistical formulas, and in some cases both terms appear in the same formula. The term on the left, $\sum X_i^2$, says that each X value should be squared *before the values are added*. If $X_1 = 1$, $X_2 = 2$, and $X_3 = 3$, $\sum X_i^2 = 1^2 + 2^2 + 3^2 = 1 + 4 + 9 = 14$. On the other hand, the term on the right $\left(\sum X_i\right)^2$ says that all of the X values should be added *before the total is squared*. Using the same X values as above, $\left(\sum X_i\right)^2 = (1 + 2 + 3)^2 = 6^2 = 36$. Notice that 36 is larger than 14. When all the values are positive, $\left(\sum X_i\right)^2$ will always be larger than $\sum X_i^2$.

In this text, I will use only one summation sign at a time in the main formulas. Summation signs can be doubled or tripled to create more complex formulas, but matters soon become difficult to keep track of, so I will use other notational tricks to avoid such complications.

Rounding Off Numbers

Whereas discrete variables can be measured exactly, the measurement of continuous variables always involves some rounding off. If you are using an interval or ratio scale, the precision of your measurement will depend on the unit you are using. If you are measuring height with a ruler in which the inches are divided into tenths, you must round off to the nearest tenth of an inch. When you report someone's height as 65.3 inches, it really means that the person's height was somewhere between 65.25 inches (half a unit below the reported measurement) and 65.35 inches (half a unit above). You can choose to round off to the nearest inch, of course, but you cannot be more precise than the nearest tenth of an inch.

Rounding off also occurs when calculating statistics, even if the data come from a discrete variable. If three families contain a total of eight people, the average family size is 8/3. To express this fraction in terms of decimals requires rounding off because this is a number with repeating digits past the decimal point (i.e., 2.666 and so on infinitely). When the original data come in the form of whole numbers, it is common to express calculations based on those numbers to two decimal places (i.e., two digits to the right of the decimal point). In the case of 8/3, 2.666 . . . can be rounded off to 2.67. The rule is simple: When rounding to two decimal places, look at the digit in the third decimal place (e.g., 2.666). If this digit is 5 or more, the digit to its left is raised by 1 and the rest of the digits are dropped (e.g., 2.666 becomes 2.67 and 4.5251 is rounded off to 4.53). If the digit in the third decimal place is less than 5, it is just dropped, along with any digits to its right (e.g., 7/3, 2.333 . . . is rounded to 2.33, 4.5209 is rounded to 4.52).

The only exception to this simple rule occurs when the digit in the third decimal place is 5 and the remaining digits are all zero (e.g., 3/8 = .375). In this case, add 1 to the digit in the second decimal place if it is odd, and drop the remaining digits (.375 is rounded to .38); if the digit in the second decimal place is even, simply drop the digits to its right (.425 is rounded to .42). This convention is arbitrary, but it is useful in that about half the numbers will have an odd digit in the second decimal place and will be rounded up and the other half will be rounded down. Of course, these rules can be applied no matter how many digits to the right of the decimal point you want to keep. For instance, if you want to keep five such digits, you look at the sixth one to make your decision.

Extra care must be taken when rounding off numbers that will be used in further calculations (e.g., the mean family size may be used to calculate other statistics, such as a measure of variability). If you are using a calculator, you may want to jot down all the digits that are displayed. When this is not convenient, a good strategy is to hold on to two more decimal places than you want to have in your final answer. If you are using whole numbers and want to express your final answer to two decimal places, your intermediate calculations should be rounded off to not less than four decimal places (e.g., 2.66666 would be rounded to 2.6667).

The amount of *round-off error* that is tolerable depends on what your results are being used for. When it comes to homework exercises or exam questions, your instructor should give you some idea of what he or she considers a tolerable degree of error due to rounding. Fortunately, with the use of computers for statistical analysis, rounding error is rapidly disappearing as a problem in psychological research.

B SUMMARY

1. Summation Rules

Summation Rule 1A

$$\sum(X_i + Y_i) = \sum X_i + \sum Y_i$$

This rule says that when summing the sums of two variables (e.g., each sum is the combined weights of the male and female members of a mixed-doubles tennis team, and you want to sum up the weights of all of these two-person teams), you can get the same answer by summing each variable separately (sum the weights of all of the men and then the weights of all of the women) and then adding these two sums together at the end.

Summation Rule 1B

$$\sum(X_i - Y_i) = \sum X_i - \sum Y_i$$

This rule says that when summing the differences of two variables (e.g., summing the height differences of male-female couples), you can get the same answer by summing each variable separately (sum the heights of all of the men and then the heights of all of the women) and then subtracting the two sums at the end.

Summation Rule 2

$$\sum C = NC$$

For instance, if everyone working for some company earns the same annual salary, C, and there are N of these workers, the total wages paid in a given year, ΣC, is equal to NC.

Summation Rule 3

$$\sum CX_i = C\sum X_i$$

For instance, in a company where the workers earn different annual salaries (X_i), if each worker's salary were multiplied by some constant, C, the total wages paid during a given year (ΣX_i) would be multiplied by the same constant. Because the constant can be some fraction, there is no need to have a separate rule for dividing by a constant.

Summation Rule 4

$$\sum(X_iY_i) \neq \left(\sum X_i\right)\left(\sum Y_i\right)$$

An important corollary of this rule is that $\sum X_i^2 \neq \left(\sum X_i\right)^2$.

2. Rules for Rounding Numbers
If you want to round to N decimal places, look at the digit in the $N + 1$ place.

a. If it is less than 5, do not change the digit in the Nth place.
b. If it is 5 or more, increase the digit in the Nth place by 1.
c. If it is 5 and there are no more digits to the right (or they are all zero), raise the digit in the Nth place by 1 only if it is an odd number. Leave the Nth digit as is if it is an even number.

In all cases, the last step is to drop the digit in the $N + 1$ place and any other digits to its right.

EXERCISES

The first two exercises are based on the following values for two variables: $X_1 = 2$, $X_2 = 4$, $X_3 = 6$, $X_4 = 8$, $X_5 = 10$; $Y_1 = 3$, $Y_2 = 5$, $Y_3 = 7$, $Y_4 = 9$, $Y_5 = 11$.

*1. Find the value of each of the following expressions:

a. $\sum_{i=2}^{5} X_i$ b. $\sum_{i=1}^{4} Y_i$ c. $\sum 5X_i$

d. $\sum 3Y_i$ e. $\sum X_i^2$ f. $\left(\sum 5X_i\right)^2$

g. $\sum Y_i^2$ h. $\left(\sum Y_i\right)^2$

*2. Find the value of each of the following expressions:

a. $\sum(X+Y)$ b. $\sum XY$ c. $\left(\sum X\right)\left(\sum Y\right)$

d. $\sum(X^2+Y^2)$ e. $\sum(X-Y)$ f. $\sum(X+Y)^2$

g. $\sum(X+7)$ h. $\sum(Y-2)$

3. Make up your own set of at least five numbers and demonstrate that $\sum X_i^2 \neq \left(\sum X_i\right)^2$.

*4. Use the appropriate summation rule(s) to simplify each of the following expressions (assume all letters represent variables rather than constants):

a. $\sum(9)$ b. $\sum(A-B)$ c. $\sum(3D)$

d. $\sum(5G+8H)$ e. $\sum(Z^2+4)$

5. Using the appropriate summation rules, show that, as a general rule, $\sum(X_i + C) = \sum X_i + NC$.

*6. Round off the following numbers to two decimal places (assume digits to the right of those shown are zero):

a. 144.0135 b. 67.245 c. 99.707
d. 13.345 e. 7.3451 f. 5.9817
g. 5.997

7. Round off the following numbers to four decimal places (assume digits to the right of those shown are zero):

a. .76995 b. 3.141627 c. 2.7182818
d. 6.89996 e. 1.000819 f. 22.55555

*8. Round off the following numbers to one decimal place (assume digits to the right of those shown are zero):

a. 55.555 b. 267.1919 c. 98.951
d. 99.95 e. 1.444 f. 22.14999

C ANALYSIS BY SPSS

SPSS (originally, the Statistical Package for the Social Sciences) is probably the most commonly used statistical package by psychologists for basic data analysis—that is, the types of analyses that will be described in this text. One consequence of that popularity is that there is a large number of beginner's guides available if you would like more detail than I can provide in the brief C sections for each of these chapters. Also, there are some more advanced guides that can show you how to conduct statistical analyses not included in this text. However, each of these C sections has been written to make it as easy as possible for you to use SPSS to conduct the analyses described in the A and B sections of that chapter, as well as to complete the exercises at the end of these sections. And, I'll show you a few tricks and shortcuts along the way. An equally important goal of these C sections is to help you translate and interpret the statistical output of SPSS in a way that is consistent with the concepts and terminology I will be using in this text.

In recent years, SPSS has been issuing a new version of its software every year, but changes that affect the basic analyses described in these C sections are rare, and not likely to lead to much confusion. The SPSS sections in this text are based on version 21.0, which was released in August 2012, but I will be pointing out any relevant changes, of which I am aware, that have occurred since version 16.0.

Ihno's Data

All of the computer exercises in this text are based on a single set of data that is printed in Appendix C, and is available as an Excel 2007 file on my website: www.psych.nyu.edu/cohen/statstext.html. The data come from a hypothetical study performed by Ihno (pronounced "Eee-know"), an advanced doctoral student, who was the teaching assistant (TA) for several sections of a statistics course. The 100 participants in the data set are the students who were enrolled in Ihno's sections, and voluntarily consented to be in her study, which was approved by the appropriate review board at her hypothetical school. Her data were collected on two different days. On the first day of classes, the students who came to one of Ihno's sections filled in a brief background questionnaire on which they provided contact information, some qualitative data (gender, undergrad major, why they had enrolled in statistics, and whether they have a habit of drinking coffee), and some quantitative data (number of math courses already completed, the score they received on a diagnostic math background quiz they were all required to take before registering for statistics, and a rating of their math phobia on a scale from 0 to 10). (You will see that, due to late registration and other factors, not all of Ihno's students took the diagnostic math background quiz.)

The rest of Ihno's data were collected as part of an experiment that she conducted during her recitation sessions on one day in the middle of the semester. (The one exception is that her students took a regular 10-question quiz the week before her experiment, and she decided to add those scores to her data set.) At the beginning of the experiment, Ihno explained how each student could take his or her own pulse. She then provided a half-minute interval during which they counted the number of beats, and then wrote down twice that number as their (baseline) heart rate in beats per minute (bpm). Then, each student reported how many cups of coffee they had consumed since waking up that morning, and filled out an anxiety questionnaire consisting of 10 items, each rated (0 to 4) on a 5-point Likert scale. Total scores could range from 0 to 40, and provided a measure of baseline anxiety.

Next, Ihno announced a pop quiz. She handed out a page containing 11 multiple-choice statistics questions on material covered during the preceding two weeks, and asked the students to keep this page face down while taking and recording their (prequiz) pulse and filling out a (prequiz) anxiety questionnaire. Then Ihno told the students they had 15 minutes to take the fairly difficult quiz. She also told them that the first 10 questions were worth 1 point each but that the 11th question was worth 3 points of extra credit. Ihno's experimental manipulation consisted of varying the difficulty of the 11th question. Twenty-five quizzes were distributed at each level of difficulty of the final question: easy, moderate, difficult, and impossible to solve. After the quizzes were collected, Ihno asked the students to provide heart rate and anxiety data one more time (i.e., postquiz). Finally, Ihno explained the experiment, adding that the 11th quiz question would not be scored and that, although the students would get back their quizzes with their score for the first 10 items, that score would not influence their grade for the statistics course.

Variable View

In SPSS, data are entered into a spreadsheet, in which the columns represent different variables, and the rows represent the different participants or cases (e.g., the cases could be cities, rather than individual people). This spreadsheet has much in common with an Excel spreadsheet, but there are important differences, as you will see. Data can be viewed and entered when the spreadsheet is in **Data View** mode. Clicking on **Variable View**, just below the lower-left corner of the spreadsheet, switches you to a related spreadsheet in which the same variables are now represented by the rows, and the columns control different aspects related to the appearance and functions of each variable. Of particular relevance to this chapter is the next to last column in the Variable View: **Measure**. Clicking on any cell in this column gives you three measurement choices for the variable in that row: While the terms *Ordinal* and *Nominal* refer to the same measurement scales defined earlier in this chapter, the choice labeled *Scale* refers to what is more often called Interval/Ratio data. Although you are not likely to use any SPSS functions for which these Measure options make a difference, you might as well set them appropriately for each of your variables. In Ihno's data set, the first six variables (Subid through Coffee) are Nominal, and the rest can be designated as Scale, except that it would be reasonable to choose Ordinal for Phobia.

Another column in the Variable view that relates to measurement scales is the second column: Type. There are eight possible variable types that can be set, but only two are commonly used: Numeric and String. If the Type is set to *numeric* for a particular variable, SPSS will not let you enter any symbols other than numbers in that column within the Data View. If you want to enter words as the values for a variable—for example, male or female in each cell of the Gender column—you have to set the Type to *string*, which allows you to enter numbers, letters, and pretty much any other symbols. Note that once the type of a variable has been set to *string*, the value for Measure is set automatically to *nominal*; it can be changed to *ordinal*, but *scale* is not an option for string variables. (*Ordinal* can make sense for a string variable if, for instance, the values are the letter grades A, B, C, etc.)

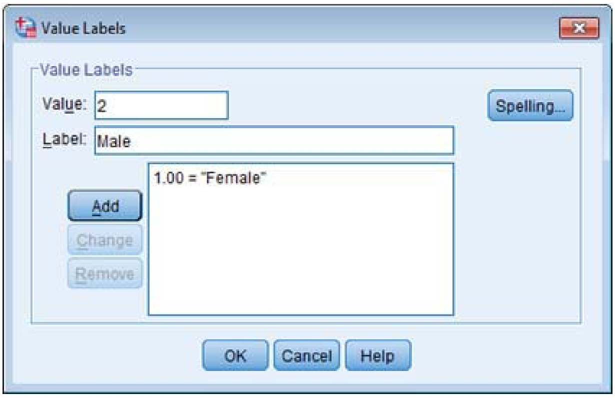


Figure 1.5

Data Coding

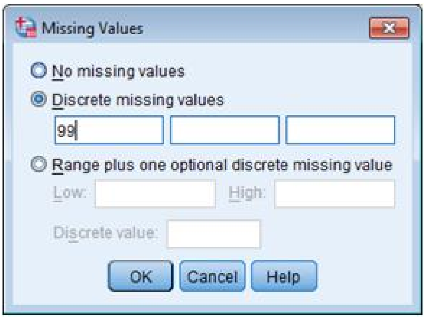
Ihno could have entered the words male and female as the values for gender, but for reasons that will be explained as we go, it was more useful for her to enter all of her data in the form of numbers, even for the nominal variables. She used the number 1 to represent females and 2 for males, which is called *coding* the data, but because this choice is arbitrary it makes sense to put this code into SPSS, rather than relying on our memory for which gender goes with which number. The words that go with the numbers that are entered for a nominal scale are called Value Labels in SPSS, and they are entered using the Values column of the Variable View. For instance, when you click on the right side of the cell corresponding to Gender in the Values column, the Value Labels box appears. To enter Ihno's code for gender, type **1** for Value, and then tab to Value Label and type **Female**, and click the **Add** button. Repeat the preceding steps to enter the label male for the number 2, and then click **OK**. (In Figure 1.5, I am about to click **Add** to enter the value label for the second gender.) The codes for the other nominal variables are shown in Appendix C. Note that when you are in Data View, you can make the value labels appear in the spreadsheet in place of the numerical codes by selecting Value Labels from the View menu.

Don't confuse the Values column with the Label column, which is next to it in Variable View. The Label column allows you to create a more elaborate name for your variable than anything you can enter in the Name column. The *label* you enter in this column will be used on your output for greater clarity. For instance, spaces are not allowed in the Name column, which is why an underscore was used to enter "Exp_cond." However, you could type "Experimental condition" in the Label column. [Tip: It is usually easier to use the Names rather than the Labels when selecting variables from SPSS menus for analysis, so you may want to select **Options** (with the General tab) from the **Edit** menu, and then select "Display names" under *Variable Lists*.]

Missing Values

There is one more column in Variable View that is worth mentioning: the Missing column. SPSS knows that when a cell in Data View is empty it means that the value is *missing*, so it does not, for instance, average in the value for that cell as a zero. However, there may be more than one reason why a value is missing, so you may want to enter arbitrary numbers in the empty

Figure 1.6


Missing Values

○ No missing values

● Discrete missing values

99

○ Range plus one optional discrete missing value

Low: High:

Discrete value:

cells—numbers that could never be real values for that variable (e.g., 888 or 999 for age)—and then associate appropriate labels with those numbers using the Values column as described in the previous section (e.g., 888 is labeled “data were lost,” and 999 is labeled “participant refused to answer”). If you do enter missing value codes in your spreadsheet, giving them value labels is not enough—you must enter those codes using the Missing column in Variable View. For instance, *mathquiz* has missing values in Ihno’s data set. If you were to enter 99 in all of the blank cells (it was not possible to score over 50 on that quiz), you would then have to click in the right side of the cell in the Missing column and the *mathquiz* row to open the Missing Values box. After selecting *Discrete missing values*, you could enter 99 in the first space, and click **OK** (see Figure 1.6). Using Value Labels to attach a label to the value of 99 (the simple label “missing” would make sense if there’s only one missing value code) is optional, but certainly desirable.

Computing New Variables

To create new variables that are based on ones already in your spreadsheet, open the **Transform** menu, and then click on the first choice: **Compute Variable**. In the **Compute Variable** box that opens up, *Target Variable* is a name that you make up (and type into that space) for the new variable; when you have filled in a *Numeric Expression* and then click **OK**, the new variable will automatically appear in the rightmost column of your Data View spreadsheet. Let’s say you want to double all the *mathquiz* scores, so they are based on a maximum of 100 points instead of 50. You would type “2 * mathquiz” or “mathquiz * 2” as the *Numeric Expression*, and perhaps “mathquiz100” as the *Target Variable*. Note that you have the option of using an existing variable as the *Target Variable*. For instance, if you fill in “mathquiz” as the *Target Variable*, and “2 * mathquiz” as the *Numeric Expression*, SPSS will alert you with the question: *Change existing variable?* If you answer by clicking OK, instead of Cancel, the values for *mathquiz* will all be doubled, and no new variable will be created. Usually, it is a good idea to retain your original values when creating new variables from them, but in this case you could always go back to the original values by computing “mathquiz = .5 * mathquiz.”

Reading Excel Files Into SPSS

Fortunately, it is very easy to read an Excel spreadsheet into an SPSS spreadsheet, which is why I have made Ihno’s data set available as an Excel 2007 spreadsheet on my textbook web page: http://psych.nyu.edu/cohen/statstext.html. One particularly convenient option, which I have used for the Excel version of Ihno’s data, is to type in all of your desired SPSS variable names in the first row of your Excel spreadsheet, each name corresponding, of course, to the values in the column beneath it. Just keep in mind that the rules for SPSS variable names are stricter than the Excel rules for column names, so do not include spaces, or other special characters that SPSS forbids, in your variable names. Also remember, when you are trying to open an Excel file in SPSS, you must select *Excel* for the space where it says "**Files of type:**" instead of the default, which is *SPSS Statistics (*.sav)*. Finally, after you have selected your Excel file and clicked on Open, you will see a box, in which the following phrase should be checked (which it usually is by default): "Read variable names from the first row of data." Click OK, and the Excel data should now be in the form of an SPSS Data View spreadsheet.

EXERCISES

1. Read Ihno's data into an SPSS spreadsheet, and then label the values of the categorical (i.e., nominal) variables according to the codes given in Appendix C. Choose the appropriate Measure level for each variable. **Optional:** Fill in missing value codes for the empty cells in *mathquiz*, declare these codes in the Missing column, and give a value label to the missing value code.

*2. Be generous and add 50 points to everyone's *mathquiz*, without creating a new variable. Then, take away the 50 points, so you are back to the original values (just to see that you can do it). Next, add 50 points to *mathquiz* again, but this time create a new variable.

3. Create a new variable by dividing the baseline heart rate by 60; give this new variable a Label to make it clear that it is expressing the baseline heart rate in beats per second (bps). Change the number of decimals to three for this new variable.

*4. a. Create a new variable that adds 2 points to each student's *statquiz* score, and then multiplies it by 10.
b. Create a new variable that multiplies each student's *statquiz* score by 10, and then adds 2 points.

5. a. Create a new variable that equals the *sum* of the three anxiety measures.
b. Create a new variable that equals the *average* of the three heart rate measures.

6. Create a new variable that is equal to *statquiz* minus *exp_sqz*.

Frequency Tables, Graphs, and Distributions

2

Chapter

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

Symbols
Σ: Summation sign

Concepts
Continuous versus discrete scales

A

CONCEPTUAL FOUNDATION

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

Table 2.1

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

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

Table 2.2

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

Frequency Distributions

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

Table 2.3

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

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

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

The Mode of a Distribution

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

The Cumulative Frequency Distribution

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

Table 2.4

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

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

The mechanics of creating the cf column are easy enough. The cf entry for the lowest score is just the same as the frequency of that score. The cf for the next highest score is the frequency of that score plus the frequency of the score below. Each cf entry equals the frequency of that score plus the cf for the score below. For example, the cf for a score of 7 is the frequency of 7, which is 3, plus the *cf* for 6, which is 13: *cf* for 7 = 3 + 13 = 16. The entry at the top of the *cf* column should equal the total number of scores, *N*, which also equals Σ*f*.

The Relative Frequency and Cumulative Relative Frequency Distributions

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

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

Table 2.5

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

The Cumulative Percentage Distribution

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

Table 2.6

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

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

Percentiles

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

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

Graphs

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

The Bar Graph

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

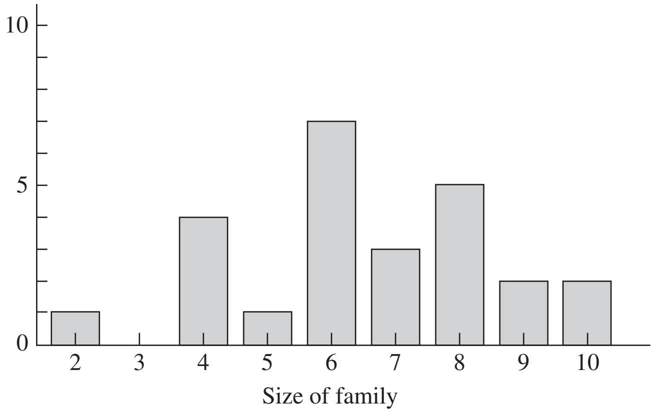


Figure 2.1

Bar Graph

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

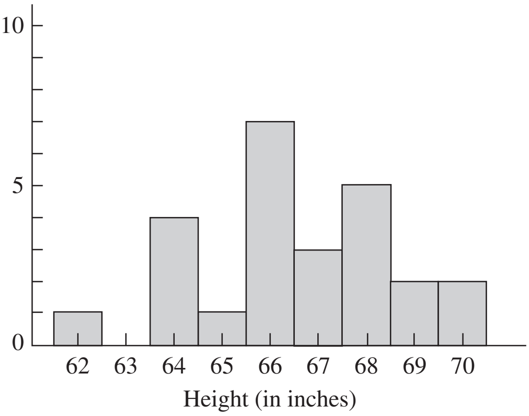
The advantage of a bar graph as compared to a table should be clear from Figure 2.1; the bar graph shows at a glance how the family sizes are distributed among the 25 families. Bar graphs are also appropriate when the variable in question has been measured on a nominal or ordinal scale. For instance,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

The Histogram

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

Figure 2.2

Frequency Histogram



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

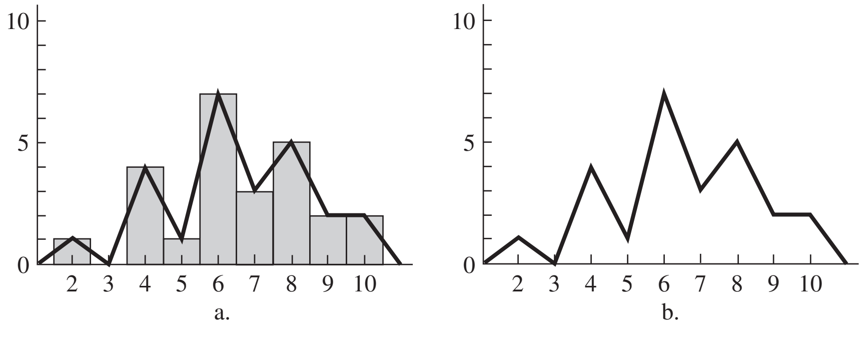
The Frequency Polygon

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

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

Figure 2.3

Frequency Polygon



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

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

The Cumulative Frequency Polygon

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

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

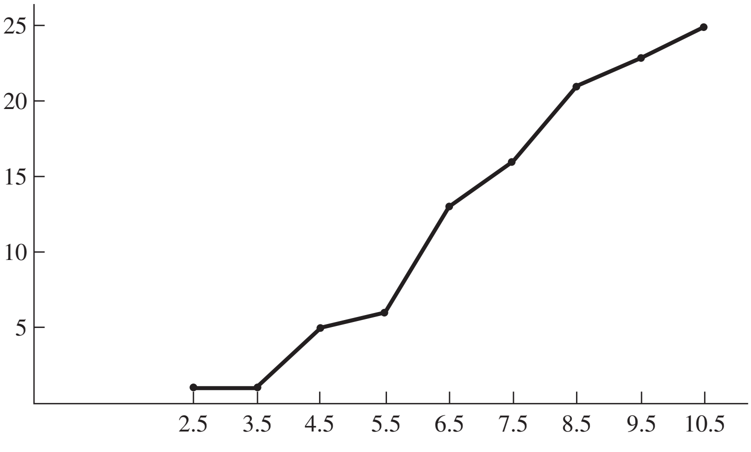


Figure 2.4

Cumulative Frequency Polygon (Ogive)

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

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

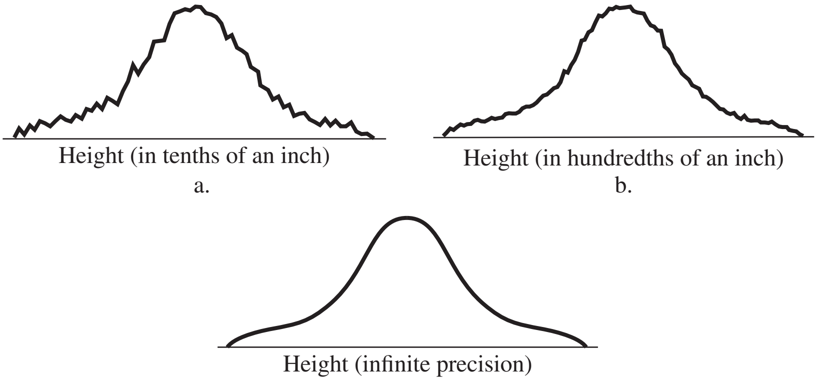
Of course, the accuracy of these graphical procedures depends on how carefully the lines are drawn. Drawing the graph to a larger scale tends to increase accuracy. Also, note that the cumulative percentage polygon consists of straight *lines*; therefore, these approximations are a form of *linear interpolation*.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

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

Real Versus Theoretical Distributions

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

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



Height (in tenths of an inch)
a.

Height (in hundredths of an inch)
b.

Height (infinite precision)
c.

Figure 2.5

Frequency Polygons With Precise Measurements

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

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

A SUMMARY

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

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

EXERCISES

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

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

a. Draw a bar graph to represent the above results. To answer the following questions, create relative frequency (*rf*), cumulative frequency (*cf*), cumulative relative frequency (*crf*), and cumulative percentage frequency (*cpf*) columns for the table.

b. What proportion of the patients was greatly improved?

c. How many patients did not improve (i.e., were unchanged or became worse)? What proportion of the patients did not improve?

d. What is the percentile rank of a patient who improved slightly? Of a patient who became slightly worse?

e. Which category of improvement corresponds to the third quartile (i.e., 75th percentile)? To the first quartile?

*2. A cognitive psychologist is training volunteers to use efficient strategies for memorizing lists of words. After the training period, 25 participants are each tested on the same list of 30 words. The numbers of words correctly recalled by the participants are as follows: 25, 23, 26, 24, 19, 25, 24, 28, 26, 21, 24, 24, 29, 23, 19, 24, 23, 24, 25, 23, 24, 25, 26, 28, 25. Create a simple frequency distribution to display these data, and then add columns for *rf*, *cf*, *crf*, and *cpf*.

a. What proportion of the participants recalled exactly 24 words?

b. How many participants recalled no more than 23 words? What proportion of the total does this represent?

c. What is the percentile rank of a participant who scored 25? Of a participant who scored 27?

d. Which score is close to being at the first quartile? The third quartile?

e. Draw a histogram to represent the data.

3. A boot camp sergeant recorded the number of attempts each of 20 soldiers required to complete an obstacle course. The results were 2, 5, 3, 1, 2, 7, 1, 4, 2, 4, 8, 1, 3, 2, 6, 5, 2, 4, 3, 1. Create a simple frequency table to display these data. Add columns for *cf*, *rf*, *crf*, and *cpf*. (*Note*: Because lower numbers reflect better scores, you may want to put the lowest number on top of the table.)

a. What proportion of the soldiers could complete the course on the first attempt?

b. What proportion of them needed four or more attempts?

c. What is the percentile rank of someone who needed five attempts?

d. What score is closest to being the third quartile?

e. Draw a frequency polygon to represent the data.

4. An ethnographer surveyed 25 homes to determine the number of people per household. She found the following household sizes: 2, 1, 3, 5, 1, 4, 3, 2, 2, 6, 3, 4, 5, 1, 2, 4, 2, 7, 4, 6, 5, 5, 6, 6, 5. Construct a simple frequency table to display these results. Add columns for *cf*, *rf*, *crf*, and *cpf*.

a. What percentage of households have three or fewer members?

b. What household size corresponds to the 80th percentile?

c. How many households have only one member? To what proportion does that correspond?

d. What proportion of households have five or more members?

e. Draw a bar graph to represent the data.

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

a. How many students obtained a perfect score? What proportion does that represent?

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

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

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

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

B BASIC STATISTICAL PROCEDURES

Grouped Frequency Distributions

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

Table 2.7

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

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

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

Table 2.8

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

Apparent Versus Real Limits

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

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

Constructing Class Intervals

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

Choosing the Class Interval Width

Before you can create a grouped frequency distribution, you must first decide how wide to make the class intervals. This is an important decision. If you make the class interval too large, there will be too few intervals to give you much detail about the distribution. For instance, suppose we chose to put the data from Table 2.7 into a grouped frequency distribution in which i (the interval width) equals 10. The result would be as shown in Table 2.9. Such a grouping could be useful if these class intervals actually corresponded with some external criterion; for instance, the class intervals could correspond to the letter grades A, B, C, and D. However, in the absence of some external criterion for grouping, it is preferable to have at least 10 class intervals to get a detailed picture of the distribution. On the other hand, if you make the class intervals too narrow, you may have so many intervals that you are not much better off than with the simple frequency distribution. In general, more than 20 intervals is considered too many to get a good picture of the distribution.

Table 2.9

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

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

Table 2.10

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

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

Finding the Number of Intervals Corresponding to a Particular Class Width

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

Choosing the Limits of the Lowest Interval

Having chosen the width of your class interval, you must decide on the apparent limits of the lowest interval; the rest of the intervals will then be determined. Naturally, the lowest class interval must contain the lowest score in the array, but that still leaves room for some choice. A useful guideline is to make sure that either the lower apparent limit or the upper apparent limit of the lowest interval is a multiple of i. (If the lower limit of one interval is a multiple of *i*, all the lower limits will be multiples of *i*.) This is true in Table 2.10: The lower limit of the lowest interval (63) is a multiple of *i*, which is 3. It also would have been reasonable to start with 64–66 as the lowest interval because then the upper limit (66) would have been a multiple of *i*. Choosing the limits of the lowest interval is a matter of convenience, and a judgment can be made after seeing the alternatives.

Relative and Cumulative Frequency Distributions

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

Table 2.11

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

Cumulative Percentage Distribution

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

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

Table 2.12

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

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

Estimating Percentiles and Percentile Ranks by Linear Interpolation

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

Estimating Percentile Ranks

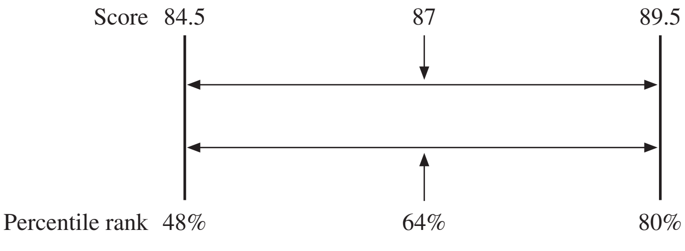
Consider the interval 85–89 in Table 2.12, for which the frequency is 8. We assume that the eight scores are spread evenly from 84.5 to 89.5 so that, for instance, four of the scores are between 84.5 and 87 (the *midpoint* of the interval), and the other four are between 87 and 89.5. This reasoning also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

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

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

Figure 2.6

Finding a Percentile Rank by Linear Interpolation (Midpoint)



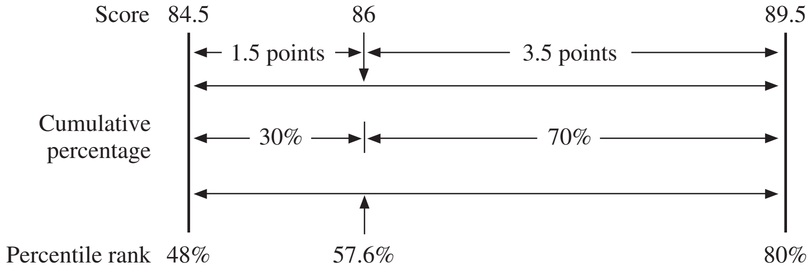


Figure 2.7

Finding Any Percentile Rank by Linear Interpolation

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

Estimating Percentiles

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

Graphing a Grouped Frequency Distribution

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

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

Figure 2.8

Frequency Histogram for a Grouped Distribution

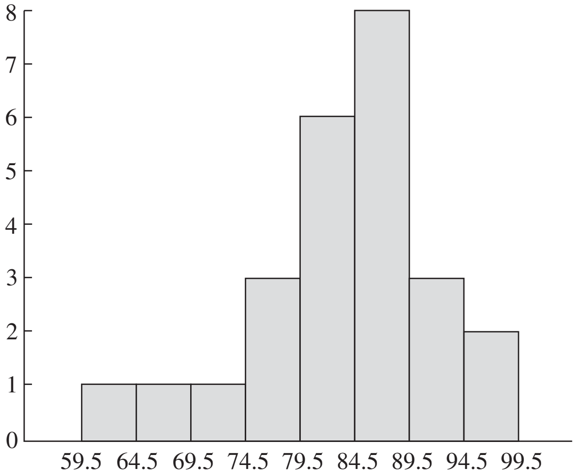
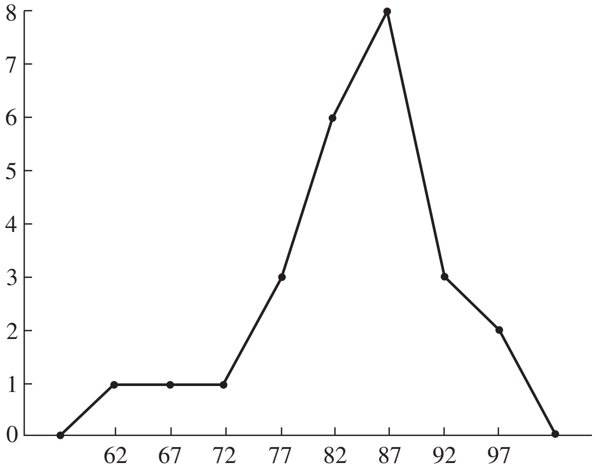


Figure 2.9

Frequency Polygon for a Grouped Distribution



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

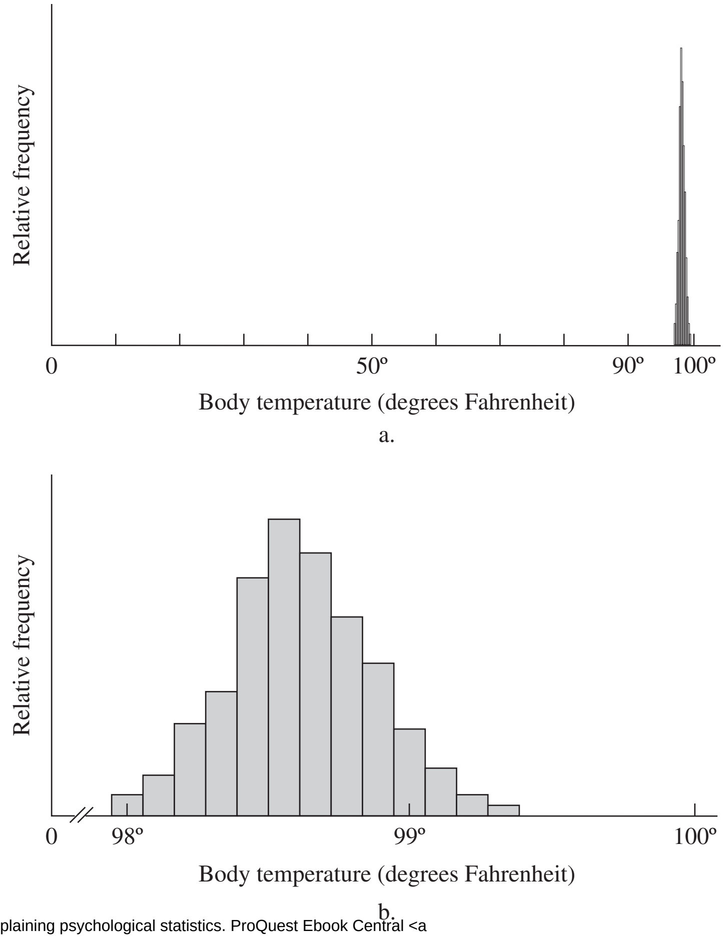
Guidelines for Drawing Graphs of Frequency Distributions

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

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

Figure 2.10

Frequency Histograms: Continuous Scale and Broken Scale



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

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

B SUMMARY

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

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

EXERCISES

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

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

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

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

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

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

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

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

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

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

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

C

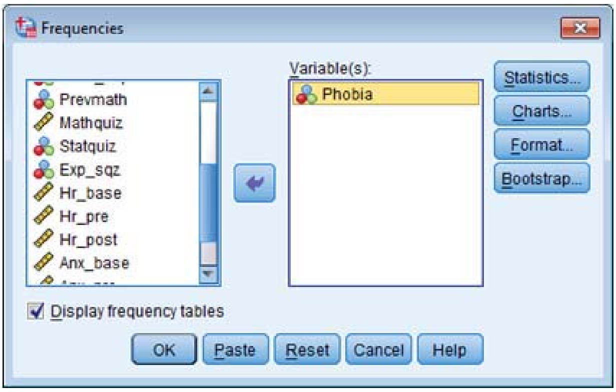
ANALYSIS BY SPSS

Creating Frequency Distributions

To create a frequency distribution, follow these six steps:

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

Figure 2.11



Variable(s): Phobia

☑ Display frequency tables

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

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

Table 2.13

Phobia

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

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

Percentile Ranks and Missing Values

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

Table 2.14

mathquiz

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

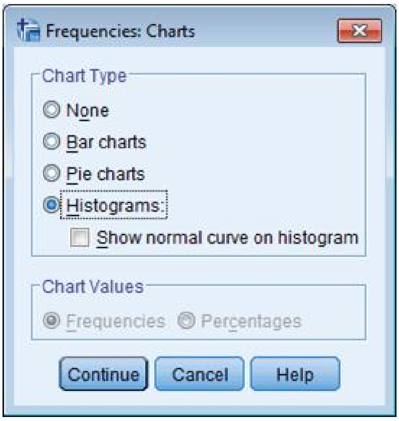
The first row labeled Total has a Frequency entry of 85, because the sum of the entries in the Frequency column will be 85—that's how many students had *mathquiz* scores. The next row indicates how many cases had missing data for that variable, and the last row tells you the total number of cases in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

Graphing Your Distribution

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

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

Figure 2.12



Frequencies: Charts

Chart Type

○ None

○ Bar charts

○ Pie charts

● Histograms:

□ Show normal curve on histogram

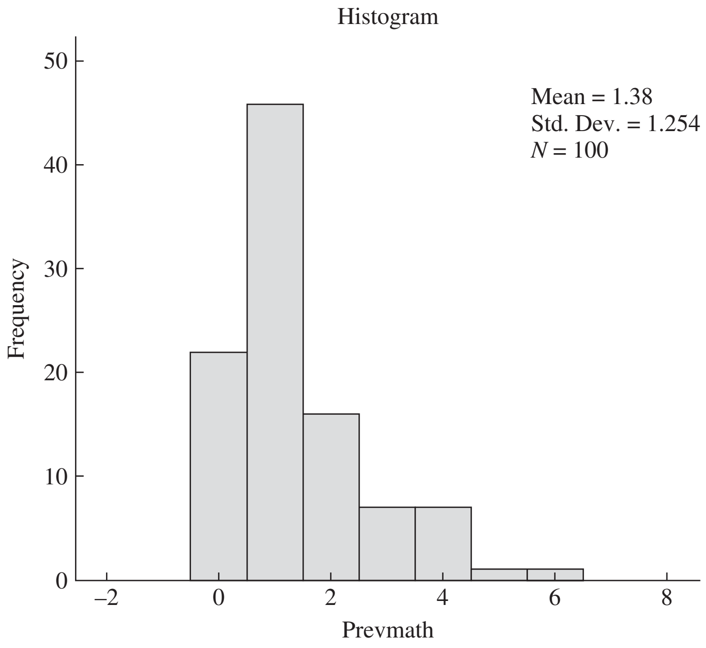
Chart Values

● Frequencies ○ Percentages

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

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

Figure 2.13



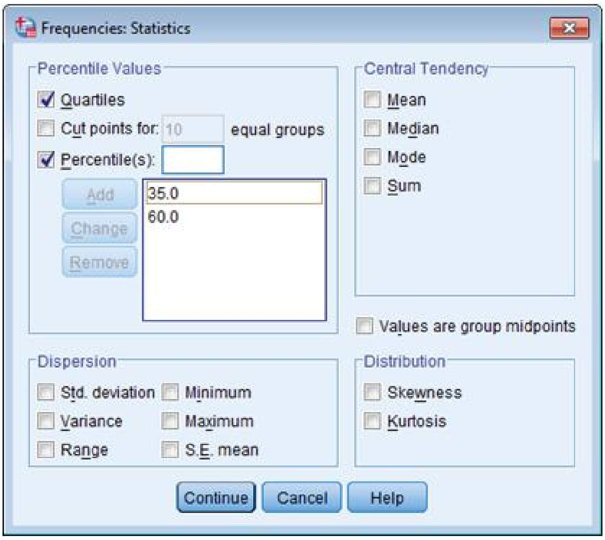
Obtaining Percentiles

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

The Split File Function

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

Figure 2.14



Frequencies: Statistics

Percentile Values

☑ Quartiles

□ Cut points for: 10 equal groups

☑ Percentile(s):

35.0
60.0

Central Tendency

□ Mean

□ Median

□ Mode

□ Sum

□ Values are group midpoints

Dispersion

□ Std. deviation □ Minimum

□ Variance □ Maximum

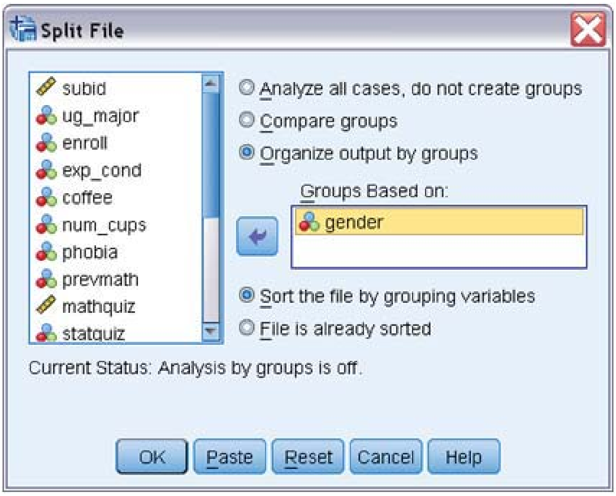
□ Range □ S.E. mean

Distribution

□ Skewness

□ Kurtosis

Figure 2.15



subid
ug_major
enroll
exp_cond
coffee
num_cups
phobia
prevmath
mathquiz
statquiz

○ Analyze all cases, do not create groups
○ Compare groups
● Organize output by groups

Groups Based on:
gender

● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

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

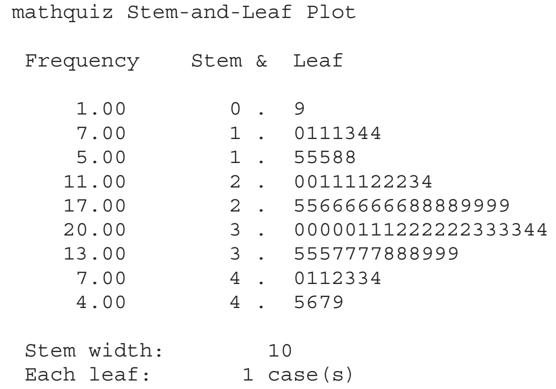
Stem-and-Leaf Plots

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

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

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

Figure 2.16



```
mathquiz Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

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

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

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

To create stem-and-leaf displays:

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

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

EXERCISES

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

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.

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.

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

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

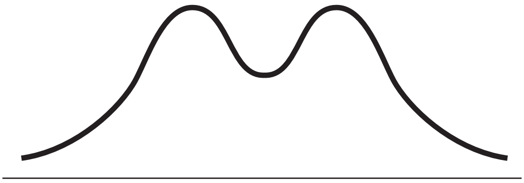


Figure 3.1

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.

Table 3.2

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

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 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

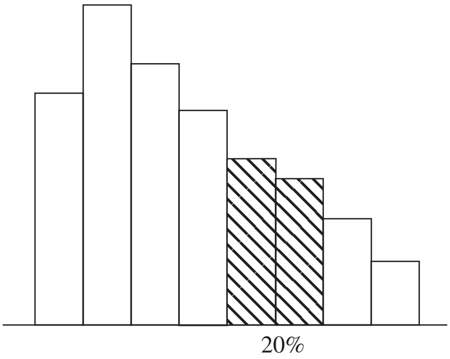


Figure 3.2

Area of a Frequency Histogram

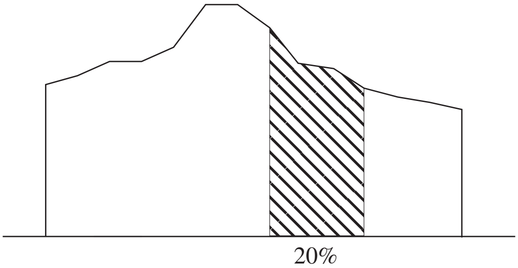


Figure 3.3

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.

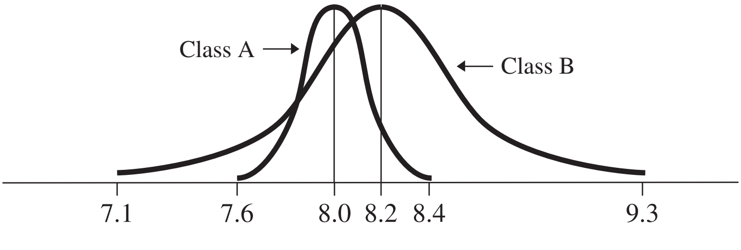
The Range

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



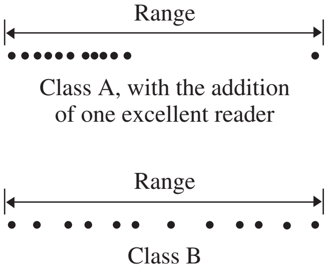


Figure 3.5

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

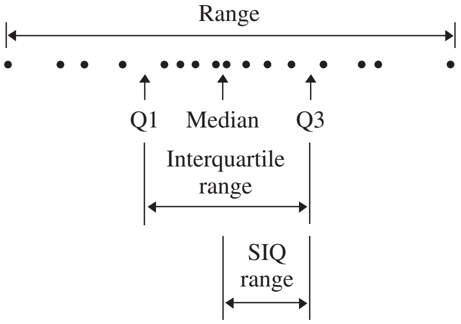
There is one measure of variability that can be used with open-ended distributions and is virtually unaffected by extreme scores because, like the median, it is based on percentiles. It is called the *interquartile* (*IQ*) *range*, and it is found by subtracting the 25th percentile from the 75th percentile. 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:

$$\text{SIQ range} = \frac{\text{Q3} - \text{Q1}}{2}$$ **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:

$$\text{Mean deviation} = \frac{\sum|X_i - \mu|}{N}$$ **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:

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

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

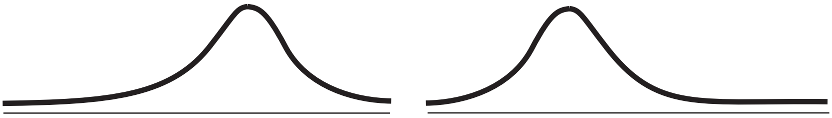
In applying these formulas to the sample of three numbers (1, 3, 8), you do not have to recalculate *SS*, which was the numerator when we found σ^2 by Formula 3.5A. Given that $SS = 26$, Formula 3.7A tells you that $s^2 = SS/(n-1) = 26/2 = 13$, which is considerably larger than σ^2 (8.67). The increase 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,

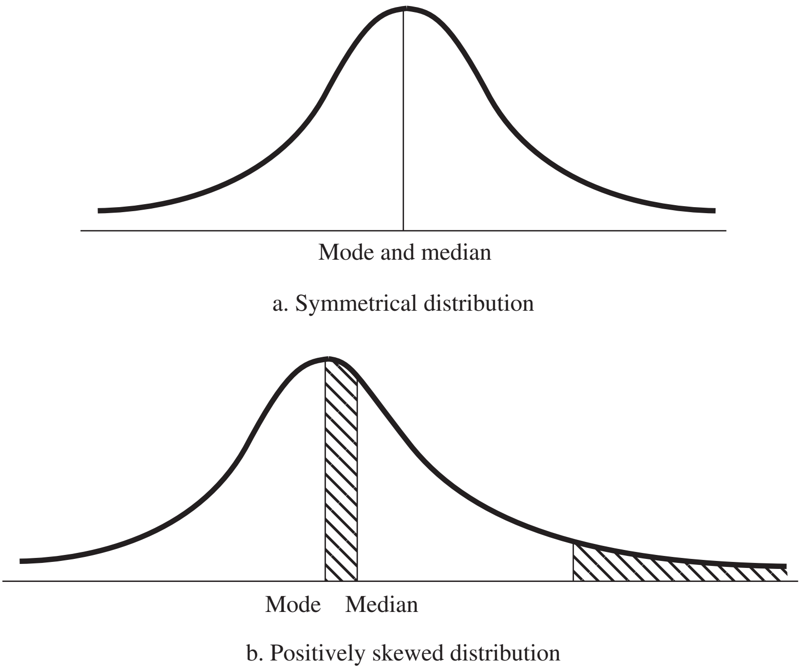

Mode and median
a. Symmetrical distribution
Mode
Median
b. Positively skewed distribution

Figure 3.8

Median of a Skewed Distribution

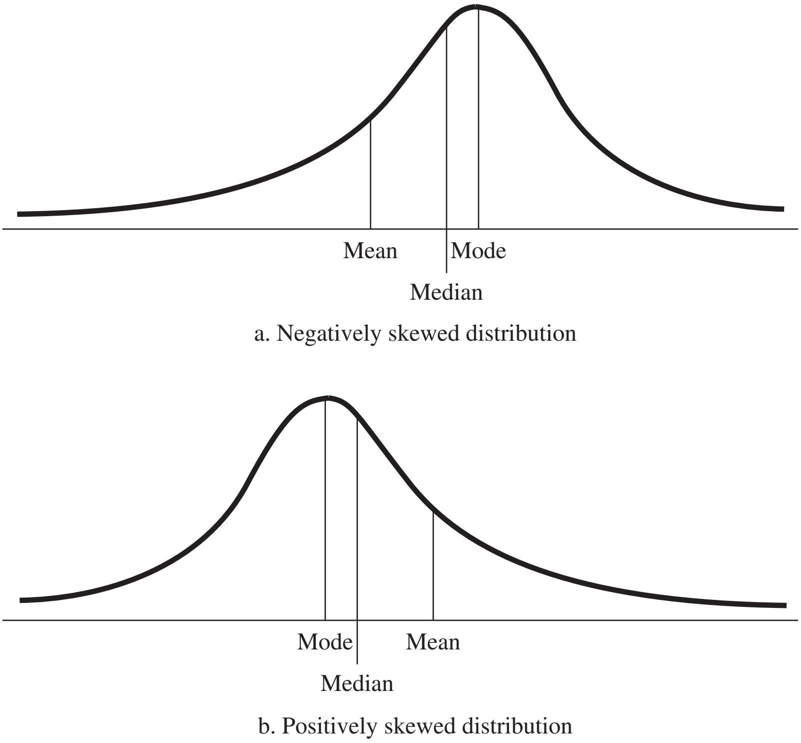

Mean
Mode
Median
a. Negatively skewed distribution
Mode
Mean
Median
b. Positively skewed distribution

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

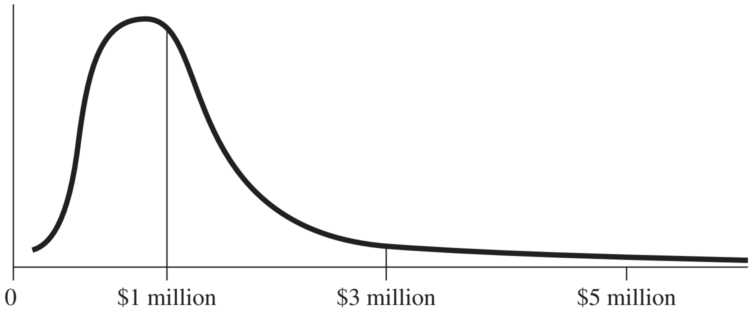
Let us consider an example of a skewed distribution for which choosing a measure of central tendency has practical consequences. There has been much publicity in recent years about the astronomical salaries paid to a few superstar athletes. However, the bulk of the professional athletes in any particular 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.

A 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:

$$\mu = \frac{\sum X}{N}$$ **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):

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

$$SS = \sum X^2 - N\mu^2$$

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:

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

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:

$$\sigma^2 = \frac{\sum X^2}{N} - \mu^2$$

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:

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

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., $\mu_{\text{new}} = \mu_{\text{old}} + C$). 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 *i*th waiter and μ is the mean of the tips for all the waiters. The property above can be stated in symbols as $\sum(X_i - \mu) = 0$.

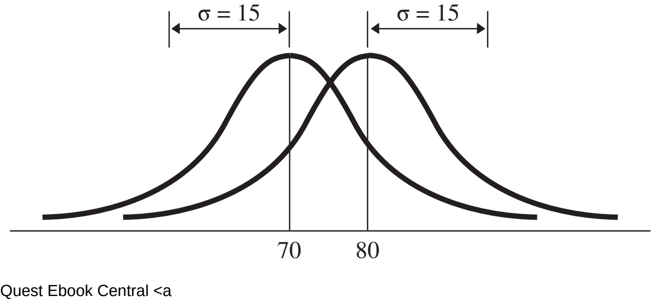
In terms of the waiters, this property says that the sum of the gains must equal the sum of the losses. This makes sense—the gains of the waiters who come out ahead in this system come entirely from the losses of the waiters who come out behind. Note, however, that the *number* of gains does not have 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:

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

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

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

Rearranging the order of terms gives the following expression:

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

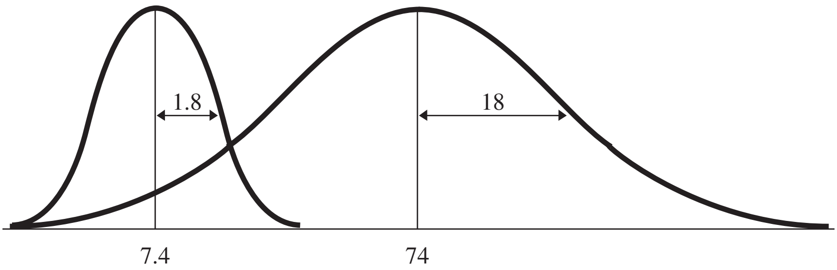
The above proof works the same way if you are subtracting, rather than adding, a constant.

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:

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

Figure 3.12

Multiplying a Distribution by a Constant



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

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

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:

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

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.)

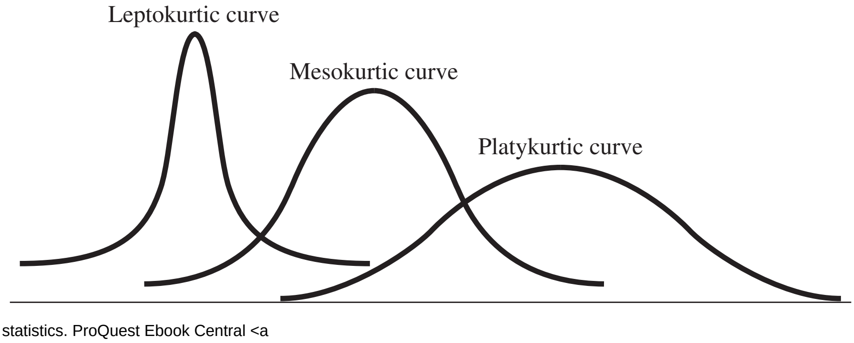
Just as the measure of skewness is based on cubed deviations from the mean, the measure of kurtosis is based on deviations from the mean raised to the fourth power. This measure of kurtosis must then be divided by the standard deviation raised to the fourth power, to create a dimensionless measure of 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., $\mu_{\text{new}} = \mu_{\text{old}} \pm C$).
   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., $\mu_{\text{new}} = C\mu_{\text{old}}$).
   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., $\sigma_{\text{new}} = \sigma_{\text{old}}$).
   b. If every score is multiplied (or divided) by a constant, the standard deviation will be multiplied (or divided) by that constant (i.e., $\sigma_{\text{new}} = C\sigma_{\text{old}}$).
   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

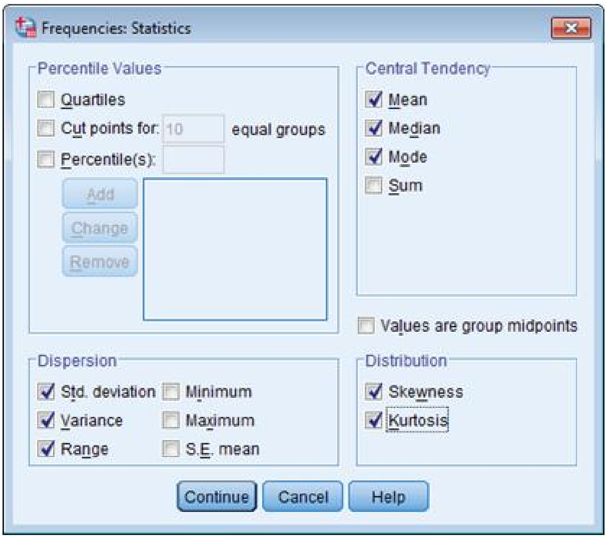
Summary Statistics

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



Frequencies: Statistics

Percentile Values

□ Quartiles

□ Cut points for: 10 equal groups

□ Percentile(s):

Central Tendency

☑ Mean

☑ Median

☑ Mode

□ Sum

□ Values are group midpoints

Dispersion

☑ Std. deviation □ Minimum

☑ Variance □ Maximum

☑ Range □ S.E. mean

Distribution

☑ Skewness

☑ Kurtosis

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.

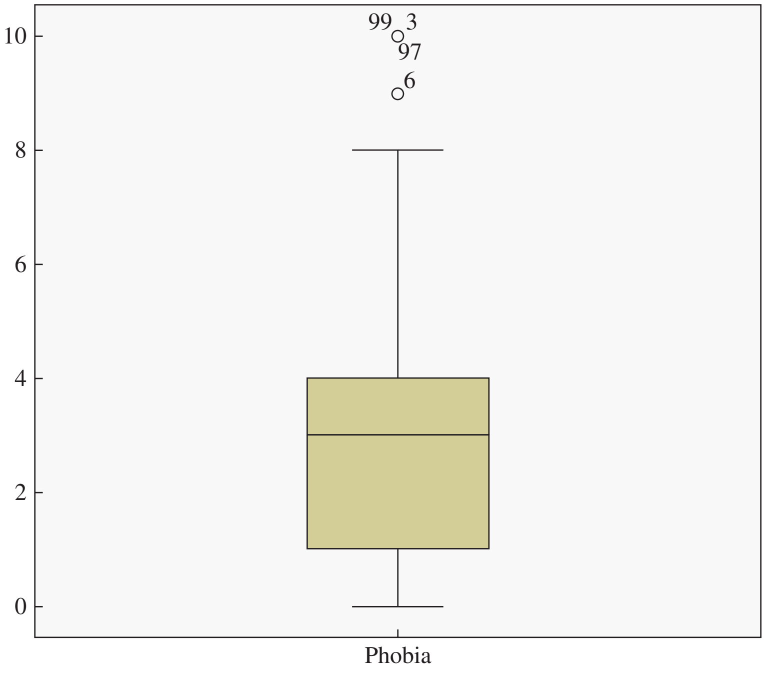
Boxplots

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 × 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.

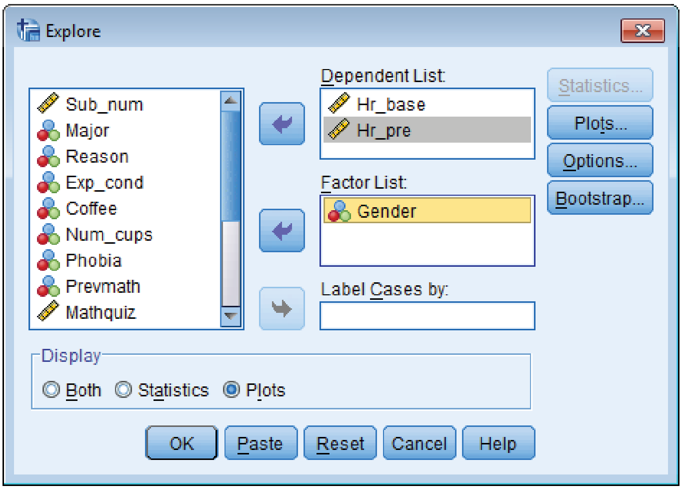
To create Boxplots:

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



Explore

Sub_num
Major
Reason
Exp_cond
Coffee
Num_cups
Phobia
Prevmath
Mathquiz

Dependent List:
Hr_base
Hr_pre

Factor List:
Gender

Label Cases by:

Display
○ Both ○ Statistics ● Plots

Figure 3.16

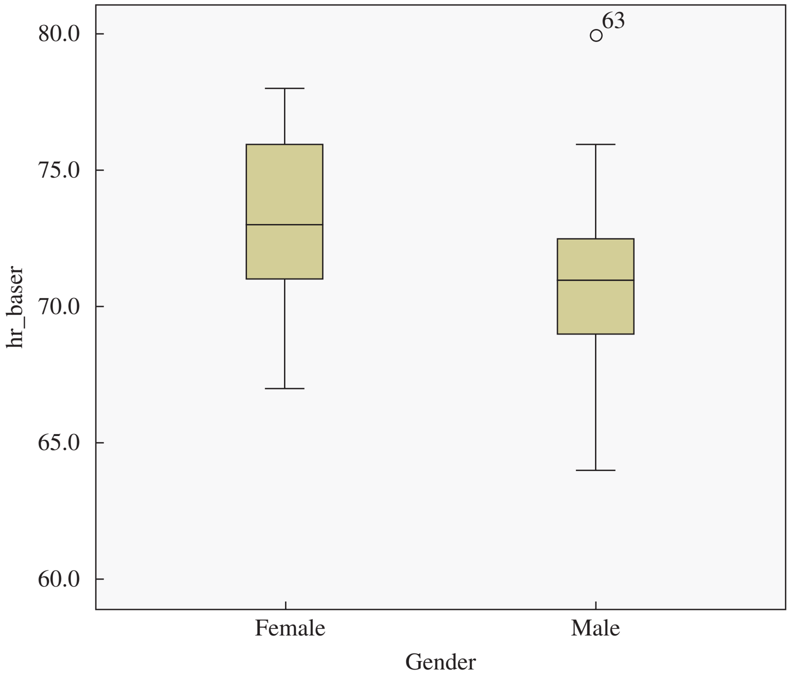
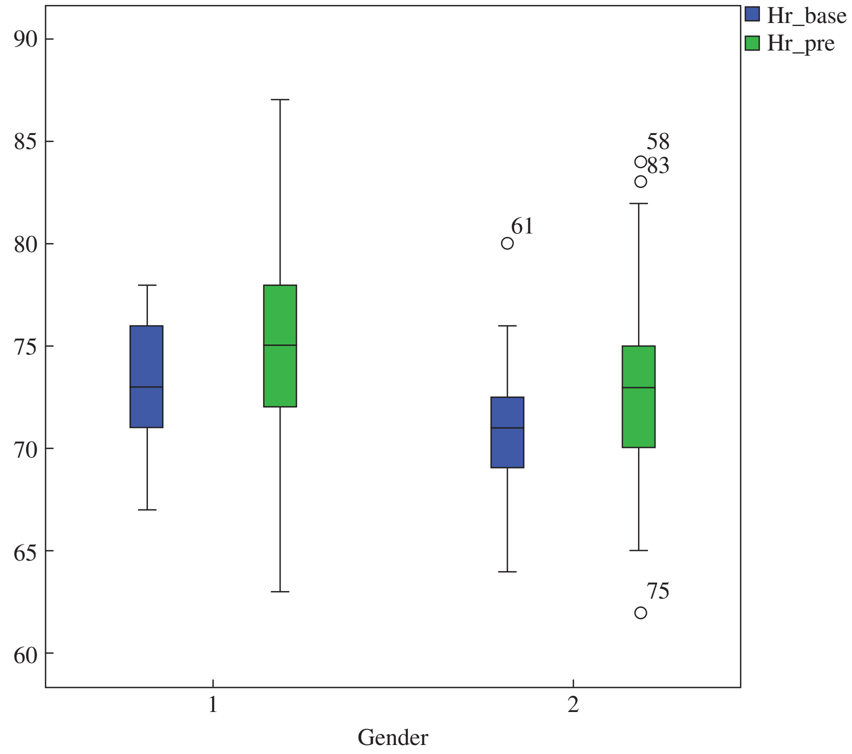


Figure 3.17

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

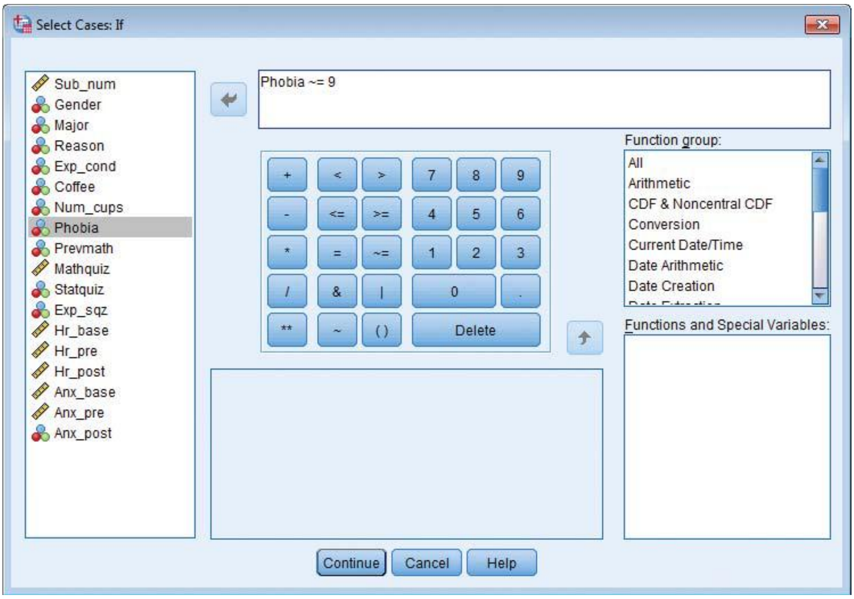
Having identified a few outliers in your data that are not based on mistakes, you may want to run your analyses with and without the outliers to see just how much difference that makes. For example, you may want SPSS to compute the mean and *SD* of the phobia variable without the four students who rated 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.



Select Cases: If

Phobia ~= 9

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 prequiz 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:

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

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

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

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

$$SS = \sum (X_i - \mu)^2$$ **Formula 3.3**

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

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

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

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

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

$$\sigma^2 = MS = \frac{SS}{N}$$ **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:

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

Standardized Scores and the Normal Distribution

4

Chapter

You will need to use the following from previous chapters:

Symbols
Σ: Summation sign
μ: Population mean
σ: Population standard deviation
σ^2: Population variance

Concepts
Percentile ranks
Mathematical distributions
Properties of the mean and standard deviation

A

CONCEPTUAL FOUNDATION

A friend meets you on campus and says, "Congratulate me! I just got a 70 on my physics test." At first, it may be hard to generate much enthusiasm about this grade. You ask, "That's out of 100, right?" and your friend proudly says, "Yes." You may recall that a 70 was not a very good grade in high school, even in physics. But if you know how low exam grades often are in college physics, you might be a bit more impressed. The next question you would probably want to ask your friend is, "What was the average for the class?" Let's suppose your friend says 60. If your friend has long been afraid to take this physics class and expected to do poorly, you should offer congratulations. Scoring 10 points above the mean isn't bad.

On the other hand, if your friend expected to do well in physics and is doing a bit of bragging, you would need more information to know if your friend has something to brag about. Was 70 the highest grade in the class? If not, you need to locate your friend more precisely within the class distribution to know just how impressed you should be. Of course, it is not important in this case to be precise about your level of enthusiasm, but if you were the teacher trying to decide whether your friend should get a B+ or an A−, more precision would be helpful.

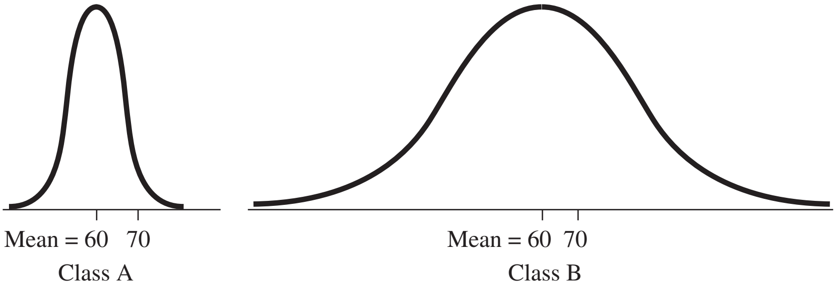
z Scores

To see just how different a score of 70 can be in two different classes, even if both of the classes have a mean of 60, take a look at the two class distributions in Figure 4.1. (To simplify the comparison, I am assuming that the classes are large enough to produce smooth distributions.) As you can see, a score of 70 in class A is excellent, being near the top of the class, whereas the same score is not so impressive in class B, being near the middle of the distribution. The difference between the two class distributions is visually obvious—class B is much more spread out than class A. Having read the previous chapter, you should have an idea of how to quantify this difference in variability. The most useful way to quantify the variability is to calculate the standard deviation (σ). The way the distributions are drawn in Figure 4.1, σ would be about 5 points for class A and about 20 for class B.

An added bonus from calculating σ is that, in conjunction with the mean (μ), σ provides us with an easy and precise way of locating scores in a

Figure 4.1

Distributions of Scores on a Physics Test



distribution. In both classes a score of 70 is 10 points above the mean, but in class A those 10 points represent two standard deviations, whereas 10 points in class B is only half of a standard deviation. Telling someone how many standard deviations your score is above or below the mean is more informative than telling your actual (raw) score. This is the concept behind the *z score*. In any distribution for which μ and σ can be found, any raw score can be expressed as a *z* score by using Formula 4.1:

$$z = \frac{X - \mu}{\sigma}$$ **Formula 4.1**

Let us apply this formula to the score of 70 in class A:

$$z = \frac{70 - 60}{5} = \frac{10}{5} = +2$$

and in class B:

$$z = \frac{70 - 60}{20} = \frac{10}{20} = +.5$$

In a compact way, the *z* scores tell us that your friend's exam score is more impressive if your friend is in class A ($z = +2$) rather than class B ($z = +.5$). Note that the plus sign in these *z* scores is very important because it tells you that the scores are above rather than below the mean. If your friend had scored a 45 in class B, her *z* score would have been:

$$z = \frac{45 - 60}{20} = \frac{-15}{20} = -.75$$

The minus sign in this *z* score informs us that in this case your friend was three quarters of a standard deviation *below* the mean. The *sign* of the *z* score tells you whether the raw score is above or below the mean; the *magnitude* of the *z* score tells you the raw score's distance from the mean in terms of standard deviations.

z scores are called standardized scores because they are not associated with any particular unit of measurement. The numerator of the *z* score is associated with some unit of measurement (e.g., the difference of someone's height from the mean height could be in inches), and the denominator is associated with the same unit of measurement (e.g., the standard deviation for height might be 3 inches), but when you divide the two, the result is dimensionless. A major advantage of standardized scores is that they provide a neutral way to compare raw scores from different distributions. To continue the previous example, suppose that your friend scores a 70 on the physics exam in class B and a 60 on a math exam in a class where μ = 50 and σ = 10. In which class was your friend further above the mean? We have already found that your friend's *z* score for the physics exam in class B was +.5. The *z* score for the math exam would be:

$$z = \frac{60 - 50}{10} = \frac{10}{10} = +1$$

Thus, your friend's *z* score on the math exam is higher than her *z* score on the physics exam, so she seems to be performing better (in terms of class standing on the last exam) in math than in physics.

Finding a Raw Score From a *z* Score

As you will see in the next section, sometimes you want to find the raw score that corresponds to a particular *z* score. As long as you know μ and σ for the distribution, this is easy. You can use Formula 4.1 by filling in the given *z* score and solving for the value of *X*. For example, if you are dealing with class A (as shown in Figure 4.1) and you want to know the raw score for which the *z* score would be −3, you can use Formula 4.1 as follows: −3 = (*X* − 60)/5, so −15 = *X* − 60, so *X* = −15 + 60 = 45. To make the calculation of such problems easier, Formula 4.1 can be rearranged in a new form that I will designate Formula 4.2:

$$X = z\sigma + \mu$$ **Formula 4.2**

Now if you want to know, for instance, the raw score of someone in class A who obtained a *z* score of −2, you can use Formula 4.2, as follows:

$$X = z\sigma + \mu = -2(5) + 60 = -10 + 60 = 50$$

Note that you must be careful to retain the minus sign on a negative *z* score when working with a formula, or you will come up with the wrong raw score. (In the previous example, *z* = +2 would correspond to a raw score of 70, as compared to a raw score of 50 for *z* = −2.) Some people find negative *z* scores a bit confusing, probably because most measurements in real life (e.g., height, IQ) cannot be negative. It may also be hard to remember that a *z* score of zero is not bad; it is just average (i.e., if *z* = 0, the raw score = μ). Formula 4.2 will come in handy for some of the procedures outlined in Section B. The structure of this formula also bears a strong resemblance to the formula for a confidence interval, for reasons that will be made clear when confidence intervals are defined in Chapter 6.

Sets of *z* Scores

It is interesting to see what happens when you take a group of raw scores (e.g., exam scores for a class) and convert all of them to *z* scores. To keep matters simple, we will work with a set of only four raw scores: 30, 40, 60, and 70. First, we need to find the mean and standard deviation for these numbers. The mean equals (30 + 40 + 60 + 70)/4 = 200/4 = 50. The standard deviation can be found by Formula 3.13B, after first calculating ΣX^2: $30^2 + 40^2 + 60^2 + 70^2 = 900 + 1600 + 3600 + 4900 = 11{,}000$. The standard deviation is found as follows:

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \mu^2} = \sqrt{\frac{11,000}{4} - 50^2} = \sqrt{2750 - 2500} = \sqrt{250} = 15.81$$

Each raw score can now be transformed into a z score using Formula 4.1:

$$z = \frac{30 - 50}{15.81} = \frac{-20}{15.81} = -1.265$$
$$z = \frac{40 - 50}{15.81} = \frac{-10}{15.81} = -.6325$$
$$z = \frac{60 - 50}{15.81} = \frac{+10}{15.81} = +.6325$$
$$z = \frac{70 - 50}{15.81} = \frac{+20}{15.81} = +1.265$$

By looking at these four z scores, it is easy to see that they add up to zero, which tells us that the mean of the z scores will also be zero. This is not a coincidence. The mean for any complete set of z scores will be zero. This follows from Property 1 of the mean, as discussed in the previous chapter: If you subtract a constant from every score, the mean is decreased by the same constant. To form z scores, you subtract a constant (namely, μ) from all the scores before dividing. Therefore, this constant must also be subtracted from the mean. But because the constant being subtracted *is* the mean, the new mean is μ (the old mean) minus μ (the constant), or zero.

It is not obvious what the standard deviation of the four z scores will be, but it will be instructive to find out. We will use Formula 3.13B again, substituting z for X:

$$\sigma_z = \sqrt{\frac{\sum z^2}{N} - \mu_z^2}$$

The term that is subtracted is the mean of the z scores squared. But as you have just seen, the mean of the z scores is always zero, so this term drops out. Therefore, $\sigma_z = \sqrt{\Sigma z^2/N}$. The term $\Sigma\, z^2$ equals $(-1.265)^2 + (-.6325)^2 + (.6325)^2 + (1.265)^2 = 1.60 + .40 + .40 + 1.60 = 4.0$. Therefore, σ equals $\sqrt{\Sigma z^2/N} = \sqrt{(4/4)} = \sqrt{1} = 1$. As you have probably guessed, this is also no coincidence. The standard deviation for a complete set of z scores will always be 1. This follows from two of the properties of the standard deviation described in the last chapter. Property 1 implies that subtracting the mean (or any constant) from all the raw scores will not change the standard deviation. Then, according to Property 2, dividing all the scores by a constant will result in the standard deviation being divided by the same constant. The constant used for division when creating z scores *is* the standard deviation, so the new standard deviation is σ (the old standard deviation) divided by σ (the constant divisor), which always equals 1.

Properties of *z* Scores

I have just derived two important properties that apply to any set of z scores: (1) the mean will be zero, and (2) the standard deviation will be 1. Now we must consider an important limitation of z scores. I mentioned that z scores can be useful in comparing scores from two different distributions (in the example discussed earlier in the chapter, your friend performed relatively better in math than in physics). However, the comparison is reasonable only if the two distributions are similar in shape. Consider the distributions for classes D and E, shown in Figure 4.2. In the negatively skewed distribution of class D, a z score of +2 would put you very near the top of the distribution. In class E, however, the positive skewing implies that although there may

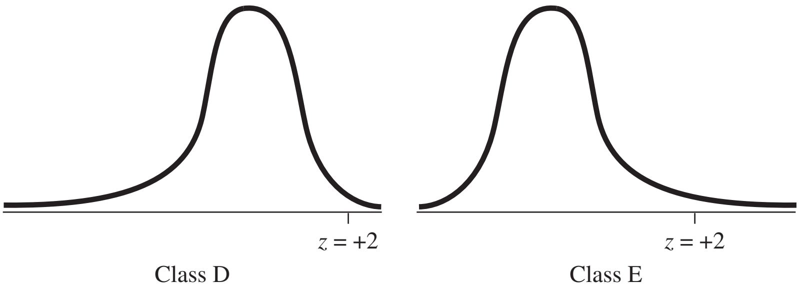


Figure 4.2

Comparing z Scores in Differently Skewed Distributions

not be a large percentage of scores above $z = +2$, there are some scores that are much higher.

Another property of z scores is relevant to the previous discussion. Converting a set of raw scores into z scores will not change the shape of the original distribution. For instance, if all the scores in class E were transformed into z scores, the distribution of z scores would have a mean of zero, a standard deviation of 1, and exactly the same positive skew as the original distribution. We can illustrate this property with a simple example, again involving only four scores: 3, 4, 5, 100. The mean of these scores is 28, and the standard deviation is 41.58 (you should calculate these yourself for practice). Therefore, the corresponding z scores (using Formula 4.1) are −.6, −.58, −.55, and +1.73. First, note the resemblance between the distribution of these four z scores and the distribution of the four raw scores: In both cases there are three numbers close together with a fourth number much higher. You can also see that the z scores add up to zero, which implies that their mean is also zero. Finally, you can calculate $\sigma = \sqrt{\sum z^2/N}$ to see that $\sigma = 1$.

SAT, *T*, and IQ Scores

For descriptive purposes, standardized scores that have a mean of zero and a standard deviation of 1.0 may not be optimal. For one thing, about half the z scores will be negative (even more than half if the distribution has a positive skew), and minus signs can be cumbersome to deal with; leaving off a minus sign by accident can lead to a gross error. For another thing, most of the scores will be between 0 and 2, requiring two places to the right of the decimal point to have a reasonable amount of accuracy. Like minus signs, decimals can be cumbersome to deal with. For these reasons, it can be more desirable to standardize scores so that the mean is 500 and the standard deviation is 100. Because this scale is used by the Educational Testing Service (Princeton, New Jersey) to report the results of the Scholastic Assessment Test, standardized scores with $\mu = 500$ and $\sigma = 100$ are often called *SAT scores*. (Recently, ETS changed the scale it uses to report the results of the Graduate Record Examination from the same one as the SAT to one that has a mean of 150 and an *SD* of 10.)

Probably the easiest way to convert a set of raw scores into SAT scores is to first find the z scores with Formula 4.1 and then use Formula 4.3 to transform each z score into an SAT score:

$$\text{SAT} = 100z + 500 \qquad \textbf{Formula 4.3}$$

Thus a z score of −3 will correspond to an SAT score of $100(-3) + 500 = -300 + 500 = 200$. If $z = +3$, the $\text{SAT} = 100(+3) + 500 = 300 + 500 = 800$.

(Notice how important it is to keep track of the sign of the z score.) For any distribution of raw scores that is not extremely skewed, nearly all of the z scores will fall between −3 and +3; this means (as shown previously) that nearly all the SAT scores will be between 200 and 800. There are so few scores that would lead to an SAT score below 200 or above 800 that generally these are the most extreme scores given; thus, we don't have to deal with any negative SAT scores. (Moreover, from a psychological point of view, it must feel better to score 500 or 400 on the SAT than to be presented with a zero or negative z score.) Because z scores are rarely expressed to more than two places beyond the decimal point, multiplying by 100 also ensures that the SAT scores will not require decimal points at all. Less familiar to students, but commonly employed for reporting the results of psychological tests, is the *T score*. The T score is very similar to the SAT score, as you can see from Formula 4.4:

$$T = 10z + 50 \qquad \textbf{Formula 4.4}$$

A full set of T scores will have a mean of 50 and a standard deviation of 10. If z scores are expressed to only one place past the decimal point, the corresponding T scores will not require decimal points.

The choice of which standardized score to use is usually a matter of convenience and tradition. The current convention regarding intelligence quotient (IQ) scores is to use a formula that creates a mean of 100. The Stanford-Binet test uses the formula $16z + 100$, resulting in a standard deviation of 16, whereas the Wechsler test uses $15z + 100$, resulting in a standard deviation of 15.

The Normal Distribution

It would be nice if all variables measured by psychologists had identically shaped distributions because then the z scores would always fall in the same relative locations, regardless of the variable under study. Although this is unfortunately not the case, it is useful that the distributions for many variables somewhat resemble one or another of the well-known mathematical distributions. Perhaps the best understood distribution with the most convenient mathematical properties is the normal distribution (mentioned in Chapter 2). Actually, you can think of the normal distribution as a family of distributions. There are two ways that members of this family can differ. Two normal distributions can differ either by having different means (e.g., heights of men and heights of women) and/or by having different standard deviations (e.g., heights of adults and IQs of adults). What all normal distributions have in common is the same shape—and not just any bell-like shape, but rather a very precise shape that follows an exact mathematical equation (see Advanced Material at the end of Section B).

Because all normal distributions have the same shape, a particular z score will fall in the same relative location on any normal distribution. Probably the most useful way to define relative location is to state what proportion of the distribution is above (i.e., to the right of) the z score and what proportion is below (to the left of) the z score. For instance, if $z = 0$, .5 of the distribution (i.e., 50%) will be above that z score and .5 will be below it. (Because of the symmetry of the normal distribution, the mean and the median fall at the same location, which is also the mode.) A statistician can find the proportions above and below any z score. In fact, these proportions have been found for all z scores expressed to two decimal places (e.g., 0.63, 2.17, etc.) up to some limit, beyond which the proportion on one side is too small to deal with easily. These proportions have been put into tables of the standard normal distribution, such as Table A.1 in Appendix A of this text.

The Standard Normal Distribution

Tables that give the proportion of the normal distribution below and/or above different z scores are called tables of the *standard normal distribution*; the standard normal distribution is just a normal distribution for which $\mu = 0$ and $\sigma = 1$. It is the distribution you get when you transform all of the scores from any normal distribution into z scores. Of course, you could work out a table for any particular normal distribution. For example, a table for a normal distribution with $\mu = 60$ and $\sigma = 20$ would show that the proportion of scores above 60 is .5, and there would be entries for 61, 62, and so forth. However, it should be obvious that it would be impractical to have a table for every possible normal distribution. Fortunately, it is easy enough to convert scores to z scores (or vice versa) when necessary and use a table of the standard normal distribution (see Table A.1 in the Appendix). It is also unnecessary to include negative z scores in the table. Because of the symmetry of the normal distribution, the proportion of scores above a particular positive z score is the same as the proportion below the corresponding negative z score (e.g., the proportion above $z = +1.5$ equals the proportion below $z = -1.5$; see Figure 4.3).

Suppose you want to know the proportion of the normal distribution that falls between the mean and one standard deviation above the mean (i.e., between $z = 0$ and $z = +1$). This portion of the distribution corresponds to the shaded area in Figure 4.4. Assume that all of the scores in the distribution fall between $z = -3$ and $z = +3$. (The fraction of scores not included in this region of the normal distribution is so tiny that you can ignore it without fear of making a noticeable error.) Thus for the moment, assume that the shaded area plus the cross-hatched areas of Figure 4.4 represent 100% of the distribution, or 1.0 in terms of proportions. The question about proportions can now be translated into areas of the normal distribution. If you knew what proportion of the area of Figure 4.4 is shaded, you would know what proportion of the scores in the entire distribution were between $z = 0$ and $z = +1$. (Recall from Chapter 2 that the size of an "area under the curve" represents a proportion of the scores.) The shaded area looks like it is about one third of the entire distribution, so you can guess that in any normal distribution about one third of the scores will fall between the mean and $z = +1$.

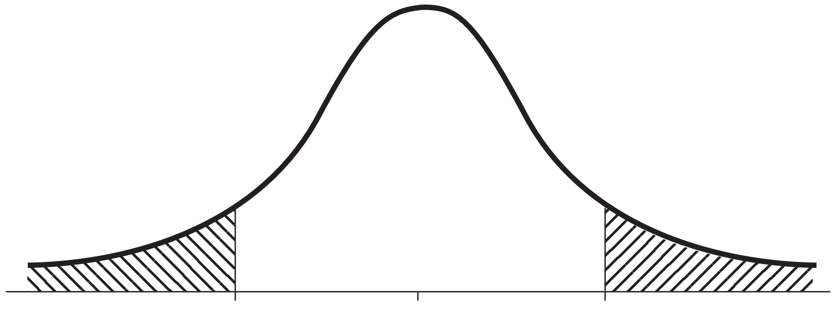


Figure 4.3

Areas Beyond $z = \pm 1.5$

Figure 4.4

Proportion of the Normal Distribution Between the Mean and $z = +1.0$

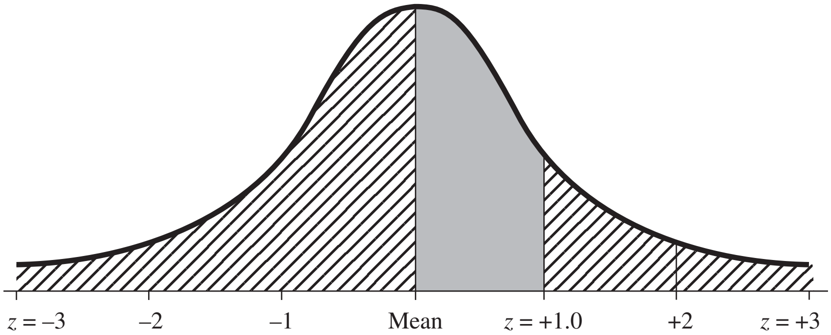


Table 4.1

z	Mean to z	Beyond z
.98	.3365	.1635
.99	.3389	.1611
1.00	.3413	.1587
1.01	.3438	.1562
1.02	.3461	.1539

Table of the Standard Normal Distribution

Fortunately, you do not have to guess about the relative size of the shaded area in Figure 4.4; Table A.1 can tell you the exact proportion. A small section of that table has been reproduced in Table 4.1. The column labeled "Mean to z" tells you what you need to know. First, go down the column labeled z until you get to 1.00. The column next to it contains the entry .3413, which is the proportion of the normal distribution enclosed between the mean and z when $z = 1.00$. This tells you that the shaded area in Figure 4.4 contains a bit more than one third of the scores in the distribution—it contains .3413. The proportion between the mean and $z = -1.00$ is the same: .3413. Thus about 68% (a little over two thirds) of any normal distribution is within one standard deviation on either side of the mean. Section B will show you how to use all three columns of Table A.1 to solve various practical problems.

Introducing Probability: Smooth Distributions Versus Discrete Events

The main reason that finding areas for different portions of the normal distribution is so important to the psychological researcher is that these areas can be translated into statements about probability. Researchers are often interested in knowing the probability that a totally ineffective treatment can accidentally produce results as promising as the results they have just obtained in their own experiment. The next two chapters will show how the normal distribution can be used to answer such abstract questions about probability rather easily.

Before we can get to that point, it is important to lay out some of the basic rules of probability. These rules can be applied either to discrete events or to a smooth, mathematical distribution. An example of a discrete event is picking one card from a deck of 52 playing cards. Predicting which five cards might be in an ordinary poker hand is a more complex event, but it is composed of simple, discrete events (i.e., the probability of each card). Applying the rules of probability to discrete events is useful in figuring out the likely outcomes of games of chance (e.g., playing cards, rolling dice, etc.) and in dealing with certain types of nonparametric statistics. I will postpone any discussion of discrete probability until Part VII, in which nonparametric statistics are introduced. Until Part VII, I will be dealing with parametric statistics, which are based on measurements that lead to smooth distributions. Therefore, at this point, I will describe the rules of probability only as they apply to smooth, continuous distributions, such as the normal curve.

A good example of a smooth distribution that resembles the normal curve is the distribution of height for adult females in a large population. Finding probabilities involving a continuous variable like height is very different from dealing with discrete events like selecting playing cards. With a deck of cards there are only 52 distinct possibilities. On the other hand, how many different measurements for height are there? It depends on how precisely height is measured. With enough precision, everyone in the population can be determined to have a slightly different height from everyone else. With an infinite population (which is assumed when dealing with the true normal distribution), there are infinitely many different height measurements. Therefore, instead of trying to determine the probability of any particular height being selected from the population, it is only feasible to consider the probability associated with a range of heights (e.g., 60 to 68 inches or 61.5 to 62.5 inches).

Probability as Area Under the Curve

In the context of a continuous variable measured on an infinite population, an “event” can be defined as selecting a value within a particular range of a distribution (e.g., picking an adult female whose height is between 60 and 68 inches). Having defined an event in this way, we can next define the probability that a particular event will occur if we select one adult female at random. The probability of some event can be defined as the proportion of times this event occurs out of an infinite number of random selections from the distribution. This proportion is equal to the area of the distribution under the curve that is enclosed by the range in which the event occurs. This brings us back to finding areas of the distribution. If you want to know the probability that the height of the next adult female selected at random will be between one standard deviation below the mean and one standard deviation above the mean, you must find the proportion of the normal distribution enclosed by $z = -1$ and $z = +1$. I have already pointed out that this proportion is equal to about .68, so the probability is .68 (roughly two chances out of three) that the height of the next woman selected at random will fall in this range. If you wanted to know whether the next randomly selected adult female would be between 60 and 68 inches tall, you would need to convert both of these heights to z scores so that you could use Table A.1 to find the enclosed area according to procedures described in Section B. Probability rules for dealing with combinations of two or more events (e.g., selections from a normal distribution) will be described at the end of Section B.

Real Distributions Versus the Normal Distribution

It is important to remember that this text is dealing with methods of *applied* statistics. We are taking theorems and laws worked out by mathematicians for ideal cases and applying them to situations involving humans or animals or even abstract entities (e.g., hospitals, cities, etc.). To a mathematical statistician, a population has nothing to do with people; it is simply an infinite set of numbers that follow some distribution. Usually some numbers are more popular in the set than others, so the curve of the distribution is higher over those numbers than others (although you can have a uniform distribution, in which all of the numbers are equally popular). These distributions are determined by mathematical equations. On the other hand, the distribution that a psychologist is dealing with (or speculating about) is a set of numbers that is not infinite. The numbers would come from measuring each individual in some very large, but finite, population (e.g., adults in the United States) on some variable of interest (e.g., need for achievement, ability to recall words from a list). Thus, we can be sure that such a population of numbers will not follow some simple mathematical distribution exactly. This leads to some warnings about using Table A.1, which is based on a perfect, theoretical distribution: the normal distribution.

First, we can be sure that even if the human population were infinite, none of the variables studied by psychologists would produce a perfect normal distribution. I can state this with confidence because the normal distribution never ends. No matter what the mean and standard deviation, the normal distribution extends infinitely in both directions. On the other hand, the measurements psychologists deal with have limits. For instance, if a psychophysiologist is studying the resting heart rates of humans, the distribution will have a lowest and a highest value and therefore will differ from a true normal distribution. This means that the proportions found in Table A.1 will not apply exactly to the variables and populations in the problems of Section B. However, for many real-life distributions the deviations from Table A.1 tend to be small, and the approximation involved can be a very useful tool. More importantly, when you group many scores together before finding the distribution, the distribution tends to look like the normal distribution, even if the distribution of individual scores does not. Because experiments are usually done with groups rather than individuals, the normal distribution plays a pervasive role in evaluating the results of experiments. This is the topic I will turn to next. But first, one more warning.

Because *z* scores are usually used only when a normal distribution can be assumed, some students get the false impression that converting to *z* scores somehow makes any distribution more like the normal distribution. In fact, as I pointed out earlier, converting to *z* scores does not change the shape of a distribution at all. Certain transformations *will* change the shape of a distribution (as described in Section C of this chapter), and in some cases will normalize the distribution, but converting to *z* scores is not one of them. (The *z* score is a linear transformation, and linear transformations don't change the shape of the distribution. These kinds of transformations will be discussed further in Chapter 9.)

z Scores as a Research Tool

You can use *z* scores to locate an individual within a normal distribution and to see how likely it is to encounter scores randomly in a particular range. However, of interest to psychological research is the fact that determining how unusual a score is can have more general implications. Suppose you know that heart rate at rest is approximately normally distributed, with a mean of 72 beats per minute (bpm) and a standard deviation of 10 bpm. You also know that a friend of yours, who drinks an unusual amount of coffee every day—five cups—has a resting heart rate of 95 bpm. Naturally, you suspect that the coffee is related to the high heart rate, but then you realize that some people in the ordinary population must have resting heart rates just as high as your friend's. Coffee isn't necessary as an explanation of the high heart rate because there is plenty of variability within the population based on genetic and other factors. Still, it may seem like quite a coincidence that your friend drinks so much coffee *and* has such a high heart rate. How much of a coincidence this really is depends in part on just how unusual your friend's heart rate is. If a fairly large proportion of the population has heart rates as high as your friend's, it would be reasonable to suppose that your friend was just one of the many with high heart rates that have nothing to do with coffee consumption. On the other hand, if a very small segment of the population has heart rates as high as your friend's, you must believe either that your friend happens to be one of those rare individuals who naturally have a high heart rate or that the coffee is elevating his heart rate. The more unusual your friend's heart rate, the harder it is to believe that the coffee is not to blame.

You can use your knowledge of the normal distribution to determine just how unusual your friend's heart rate is. Calculating your friend's z score (Formula 4.1), we find:

$$z = \frac{X - \mu}{\sigma} = \frac{95 - 72}{10} = \frac{23}{10} = 2.3$$

From Table A.1, the area beyond a z score of 2.3 is only about .011, so this is quite an unusual heart rate; only a little more than 1% of the population has heart rates that are as high or higher. The fact that your friend drinks a lot of coffee could be just a coincidence, but it also suggests that there may be a connection between drinking coffee and having a high heart rate (such a finding may not seem terribly shocking or interesting, but what if you found an unusual association between coffee drinking and some serious medical condition?).

The above example suggests an application for z scores in psychological research. However, a researcher would not be interested in finding out whether coffee has raised the heart rate of one particular individual. The more important question is whether coffee raises the heart rates of humans in general. One way to answer this question is to look at a random series of individuals who are heavy coffee drinkers and, in each case, find out how unusually high the heart rate is. Somehow all of these individual probabilities would have to be combined to decide whether these heart rates are just too unusual to believe that the coffee is uninvolved. There is a simpler way to attack the problem. Instead of focusing on one individual at a time, psychological researchers usually look at a group of subjects as a whole. This is certainly not the only way to conduct research, but because of its simplicity and widespread use, the group approach is the basis of statistics in introductory texts, including this one.

Sampling Distribution of the Mean

It is at this point in the text that I will begin to shift the focus from individuals to groups of individuals. Instead of the heart rate of an individual, we can talk about the heart rate of a group. To do so we have to find a single heart rate to characterize an entire group. Chapter 3 showed that the mean, median, and mode are all possible ways to describe the central tendency of a group, but the mean has the most convenient mathematical properties and leads to the simplest statistical procedures. Therefore, for most of this text, the mean will be used to characterize a group; that is, when I want to refer to a group by a single number, I will use the mean.

A researcher who wanted to explore the effects of coffee on resting heart rate might begin by assembling a group of heavy coffee drinkers and find the mean of their heart rates. Then the researcher could see if the group mean was unusual or not. However, to evaluate how unusual a group mean is, you cannot compare the group mean to a distribution of individuals. You need, instead, a distribution of groups (all the same size). This is a more abstract concept than a population distribution that consists of individuals, but it is a critical concept for understanding the statistical procedures in the remainder of this text.

If we know that heart rate has a nearly normal distribution with a mean of 72 and a standard deviation of 10, what can we expect for the mean heart rate of a small group? There is a very concrete way to approach this question. First, you have to decide on the size of the groups you want to deal with—this makes quite a difference, as you will soon see. For our first example, let us say that we are interested in studying groups that have 25 participants each. So we take 25 people at random from the general population and find the mean heart rate for that group. Then we do this again and again, each time recording the mean heart rate. If we do this many times, the mean heart rates will start to pile up into a distribution. As we approach an infinite number of group means, the distribution becomes smooth and continuous. One convenient property of this distribution of means is that it will be a normal distribution, provided that the variable has a normal distribution for the individuals in the population.

Because the groups that we have been hypothetically gathering are supposed to be random samples of the population, the group means are called *sample means* and are symbolized by $\overline{X}$. The distribution of sample means is called a *sampling distribution*. More specifically, it is called the *sampling distribution of the mean*. (Had we been taking the median of each group of 25 and piling up these medians into a distribution, it would be called the sampling distribution of the *median*.) Just as the population distribution gives us a picture of how the individuals are spread out on a particular variable, the sampling distribution shows us how the sample means (or medians or whatever is being used to summarize each sample) would be spread out if we grouped the population into very many samples. To make things simple, I will assume for the moment that we are always dealing with variables that have a normal distribution in the population. Therefore, the sampling distribution of the mean will always be a normal distribution, which implies that we need only know its mean and standard deviation to know everything about it.

First, consider the mean of the sampling distribution of the mean. This term may sound confusing, but it really is very simple. The mean of all the group means will always be the same as the mean of the individuals (i.e., the population mean, μ). It should make sense that if you have very many random samples from a population, there is no reason for the sample means to be more often above or below the population mean. For instance, if you are looking at the average heights for groups of men, why should the average heights of the groups be any different from the average height of individual men? However, finding the standard deviation of the sampling distribution is a more complicated matter. Whereas the standard deviation of the individuals within each sample should be roughly the same as the standard deviation of the individuals within the population as a whole, the standard deviation of the sample means is a very different kind of thing.

Standard Error of the Mean

The means of samples do not vary as much as the individuals in the population. To make this concrete, consider again a very familiar variable: the height of adult men. It is obvious that if you were to pick a man off the street at random, it is somewhat unlikely that the man would be over 6 feet tall, but not very unlikely (in some countries, the chance would be better than .2). On the other hand, imagine selecting a group of 25 men *at random* and finding their average height. The probability that the 25 men would average over 6 feet in height is extremely small. Remember that the group was selected at random. It is not difficult to find 25 men whose average height is over 6 feet tall (you might start at the nearest basketball court), but if the selection is truly random, men below 5 feet 6 inches will be just as likely to be picked as men over 6 feet tall. The larger the group, the smaller the chance that the group mean will be far from the population mean (in this case, about 5 feet 9 inches). Imagine finding the average height of men in each of the 50 states of the United States. Could the average height of men in Wisconsin be much different from the average height of men in Pennsylvania or Alabama? Such extremely large groups will not vary much from each other or from the population mean. That sample means vary less from each other than do individuals is a critical concept for understanding the statistical procedures in most of this book. The concept is critical because we will be judging whether groups are unusual, and the fact that groups vary less than individuals do implies that it takes a smaller deviation for a group to be unusual than for an individual to be unusual. Fortunately, there is a simple formula that can be used to find out just how much groups tend to vary.

Because sample means do not vary as much as individuals, the standard deviation for the sampling distribution will be less than the standard deviation for a population. As the samples get larger, the sample means are clustered more closely, and the standard deviation of the sample means therefore gets smaller. This characteristic can be expressed by a simple formula, but first I will introduce a new term. The standard deviation of the sampling distribution of the mean is called the *standard error of the mean* and is symbolized as $\sigma_{\overline{X}}$. For any particular sampling distribution, all of the samples must be the same size, symbolized by n (for the number of observations in each sample). How many different random samples do you have to select to make a sampling distribution? The question is irrelevant because nobody really creates a sampling distribution this way. The kinds of sampling distributions that I will be discussing are mathematical ideals based on drawing an infinite number of samples all of the same size. (This approach creates a sampling distribution that is analogous to the population distribution, which is based on an infinite number of individuals.)

Now I can show how the standard error of the mean decreases as the size of the samples increases. This relationship is expressed as Formula 4.5:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$ **Formula 4.5**

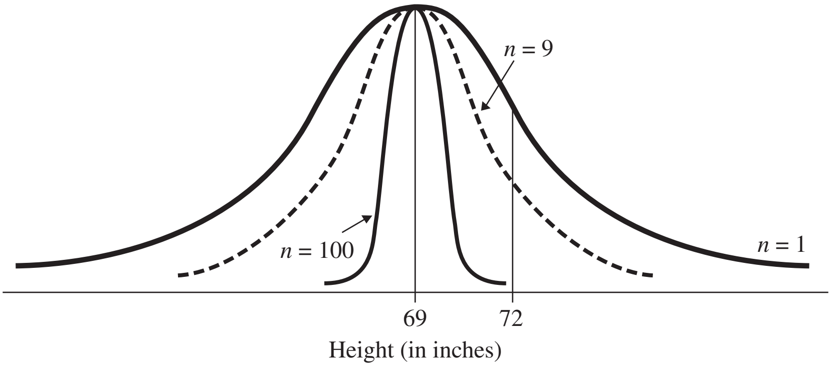
To find the standard error, you start with the standard deviation of the population and then divide by the square root of the sample size. This means that for any given sample size, the more the individuals vary, the more the groups will vary (i.e., if σ gets larger, $\sigma_{\overline{X}}$ gets larger). On the other hand, the larger the sample size, the *less* the sample means will vary (i.e., as n increases, $\sigma_{\overline{X}}$ decreases). For example, if you make the sample size 4 times larger, the standard error is cut in half (e.g., σ is divided by 5 for a sample size of 25, but it is divided by 10 if the sample size is increased to 100).

Sampling Distribution Versus Population Distribution

In Figure 4.5, you can see how the sampling distribution of the mean compares with the population distribution for a specific case. We begin with the population distribution for the heights of adult men. It is a nearly normal distribution with $\mu = 69$ inches and $\sigma = 3$ inches. For $n = 9$, the sampling distribution of the mean is also approximately normal. We know this because there is a statistical law that states that if the population

Figure 4.5

Sampling Distributions of the Mean for Different Sample Sizes



distribution is normal, the sampling distribution of the mean will also be normal. Moreover, there is a theorem that states that when the population distribution is not normal, the sampling distribution of the mean will be closer to the normal distribution than the population distribution (I'm referring to the Central Limit Theorem, which will be discussed further in the next chapter). So, if the population distribution is close to normal to begin with (as in the case of height for adults of the same gender), we can be sure that the sampling distribution of the mean for this variable will be very similar to the normal distribution.

Compared to the population distribution, the sampling distribution of the mean will have the same mean but a smaller standard deviation (i.e., standard error). The standard error for height when n is 9 is 1 inch ($\sigma_{\overline{X}} = \sigma/\sqrt{n} = 3/\sqrt{9} = 3/3 = 1$). For $n = 100$, the sampling distribution becomes even narrower; the standard error equals 0.3 inch. Notice that the means of groups tend to vary less from the population mean than do the individuals and that large groups vary less than small groups.

Referring to Figure 4.5, you can see that it is not very unusual to pick a man at random who is about 72 inches, or 6 feet, tall. This is just one standard deviation above the mean, so nearly one in six men is 6 feet tall or taller. On the other hand, to find a group of nine randomly selected men whose average height is over 6 feet is quite unusual; such a group would be three standard errors above the mean. This corresponds to a z score of 3, and the area beyond this z score is only about .0013. And to find a group of 100 randomly selected men who averaged 6 feet or more in height would be extremely rare indeed; the area beyond $z = 10$ is too small to appear in standard tables of the normal distribution. Section B will illustrate the various uses of z scores when dealing with both individuals and groups.

A SUMMARY

1. To localize a score within a distribution or compare scores from different distributions, *standardized scores* can be used. The most common standardized score is the *z score*. The z score expresses a raw score in terms of the mean and standard deviation of the distribution of raw scores. The magnitude of the z score tells you how many standard deviations away from the mean the raw score is, and the sign of the z score tells you whether the raw score is above (+) or below (−) the mean.
2. If you take a set of raw scores and convert each one to a z score, the mean of the z scores will be zero and the standard deviation will be 1. The shape of the distribution of z scores, however, will be exactly the same as the shape of the distribution of raw scores.
3. z scores can be converted to *SAT scores* by multiplying by 100 and then adding 500. SAT scores have the advantages of not requiring minus

signs or decimals to be sufficiently accurate. *T scores* are similar but involve multiplication by 10 and the addition of 50.

4. The *normal distribution* is a symmetrical, bell-shaped mathematical distribution whose shape is precisely determined by an equation (see Advanced Material at the end of Section B). The normal distribution is actually a family of distributions, the members of which differ according to their means and/or standard deviations.
5. If all the scores in a normal distribution are converted to *z* scores, the resulting distribution of *z* scores is called the *standard normal distribution*, which is a normal distribution that has a mean of zero and a standard deviation of 1.
6. The proportion of the scores in a normal distribution that falls between a particular *z* score and the mean is equal to the amount of area under the curve between the mean and *z*, divided by the total area of the distribution (defined as 1.0). This proportion is the probability that one random selection from the normal distribution will have a value between the mean and that *z* score. The areas between the mean and *z* and the areas beyond *z* (into the tail of the distribution) are given in Table A.1.
7. Distributions based on real variables measured in populations of real subjects (whether people or not) can be similar to, but not exactly the same as, the normal distribution. This is because the true normal distribution extends infinitely in both the negative and positive directions.
8. Just as it is sometimes useful to determine if an individual is unusual with respect to a population, it can also be useful to determine how unusual a group is compared to other groups that could be randomly selected. The group mean (more often called the *sample mean*) is usually used to summarize the group with a single number. To find out how unusual a sample is, the sample mean ($\overline{X}$) must be compared to a distribution of sample means, called, appropriately, the *sampling distribution of the mean*.
9. This sampling distribution could be found by taking very many samples from a population and gathering the sample means into a distribution, but there are statistical laws that tell you just what the sampling distribution of the mean will look like if certain conditions are met. If the population distribution is normal, and the samples are *independent random samples*, all of the same size, the sampling distribution of the mean will be a normal distribution with a mean of μ (the same mean as the population) and a standard deviation called the *standard error of the mean*.
10. The larger the sample size, *n*, the smaller the standard error of the mean, which is equal to the population standard deviation divided by the square root of *n*.

EXERCISES

*1. If you convert each score in a set of scores to a *z* score, which of the following will be true about the resulting set of *z* scores?
   a. The mean will equal 1.
   b. The variance will equal 1.
   c. The distribution will be normal in shape.
   d. All of the above.
   e. None of the above.

2. The distribution of body weights for adults is somewhat positively skewed—there is much more room for people to be above average than below. If you take the mean

weights for random groups of 10 adults each and form a new distribution, how will this new distribution compare to the distribution of individuals?

a. The new distribution will be more symmetrical than the distribution of individuals.
b. The new distribution will more closely resemble the normal distribution.
c. The new distribution will be narrower (i.e., have a smaller standard deviation) than the distribution of individuals.
d. All of the above.
e. None of the above.

*3. Assume that the mean height for adult women (μ) is 65 inches, and that the standard deviation (σ) is 3 inches.
a. What is the z score for a woman who is exactly 5 feet tall? Who is 5 feet 5 inches tall?
b. What is the z score for a woman who is 70 inches tall? Who is 75 inches tall? Who is 64 inches tall?
c. How tall is a woman whose z score for height is −3? −1.33? −0.3? −2.1?
d. How tall is a woman whose z score for height is +3? +2.33? +1.7? +.9?

4. a. Calculate μ and σ for the following set of scores and then convert each score to a z score: 64, 45, 58, 51, 53, 60, 52, 49.
b. Calculate the mean and standard deviation of these z scores. Did you obtain the values you expected? Explain.

*5. What is the SAT score corresponding to
a. $z = -0.2$?
b. $z = +1.3$?
c. $z = -3.1$?
d. $z = +1.9$?

6. What is the z score that corresponds to an SAT score of
a. 520?
b. 680?
c. 250?
d. 410?

*7. Suppose that the verbal part of the SAT contains 30 questions and that $\mu = 18$ correct responses, with $\sigma = 3$. What SAT score corresponds to
a. 15 correct?
b. 10 correct?
c. 20 correct?
d. 27 correct?

8. Suppose the mean for a psychological test is 24 with $\sigma = 6$. What is the T score that corresponds to a raw score of
a. 0?
b. 14?
c. 24?
d. 35?

*9. Use Table A.1 to find the area of the normal distribution between the mean and z, when z equals
a. .18
b. .50
c. .88
d. 1.25
e. 2.11

10. Use Table A.1 to find the area of the normal distribution beyond z, when z equals
a. .09
b. .75
c. 1.05
d. 1.96
e. 2.57

11. Assuming that IQ is normally distributed with a mean of 100 and a standard deviation of 15, describe completely the sampling distribution of the mean for a sample size (n) equal to 20.

*12. If the population standard deviation (σ) for some variable equals 17.5, what is the value of the standard error of the mean when
a. $n = 5$?
b. $n = 25$?
c. $n = 125$?
d. $n = 625$?
If the sample size is cut in half, what happens to the standard error of the mean for a particular variable?

13. a. In one college, freshman English classes always contain exactly 20 students. An English teacher wonders how much these classes are likely to vary in terms of their verbal scores on the SAT. What would you expect for the standard deviation (i.e., standard error) of class means on the verbal SAT?
b. Suppose that a crew for the space shuttle consists of seven people, and we are interested in the average weights of all possible shuttle crews. If the standard deviation for weight is 30 pounds, what is the standard deviation for the mean weights of shuttle crews (i.e., the standard error of the mean)?

*14. If for a particular sampling distribution of the mean we know that the standard error is 4.6, and we also know that $\sigma = 32.2$, what is the sample size (n)?

B
BASIC STATISTICAL PROCEDURES

As you have seen, *z* scores can be used for descriptive purposes to locate a score in a distribution. Later in this section, I will show that *z* scores can also be used to describe groups, although when we are dealing with groups, we usually have some purpose in mind beyond pure description. For now, I want to expand on the descriptive power of *z* scores when dealing with a population of individuals and some variable that follows the normal distribution in that population. As mentioned in Chapter 2, one of the most informative ways of locating a score in a distribution is by finding the percentile rank (PR) of that score (i.e., the percentage of the distribution that is below that score). To find the PR of a score within a small set of scores, the techniques described in Chapter 2 are appropriate. However, if you want to find the PR of a score with respect to a very large group of scores whose distribution resembles the normal distribution (and you know both the mean and standard deviation of this reference group), you can use the following procedure.

Finding Percentile Ranks

I'll begin with the procedure for finding the PR of a score that is above the mean of a normal distribution. The variable we will use for the examples in this section is the IQ of adults, which has a fairly normal distribution and is usually expressed as a standardized score with $\mu = 100$ and (for the Stanford-Binet test) $\sigma = 16$. To use Table A.1, however, IQ scores will have to be converted back to *z* scores. I will illustrate this procedure by finding the PR for an IQ score of 116. First find *z* using Formula 4.1:

$$z = \frac{116 - 100}{16} = \frac{16}{16} = +1.0$$

Next, draw a picture of the normal distribution, always placing a vertical line at the mean ($z = 0$), and at the *z* score in question ($z = +1$, for this example). The area of interest, as shown by the crosshatching in Figure 4.6, is the portion of the normal distribution to the left of $z = +1.0$. The entire crosshatched area does not appear as an entry in Table A.1 (although some standard normal tables include a column that would correspond to the shaded area). Notice that the crosshatched area is divided in two portions by the mean of the distribution. The area to the left of the mean is always half of the normal distribution and therefore corresponds to a proportion of .5. The area between the mean and $z = +1.0$ can be found in Table A.1 (under "Mean to *z*"), as demonstrated in Section A. This proportion is .3413. Adding .5 to .3413, we get .8413, which is the proportion represented by the

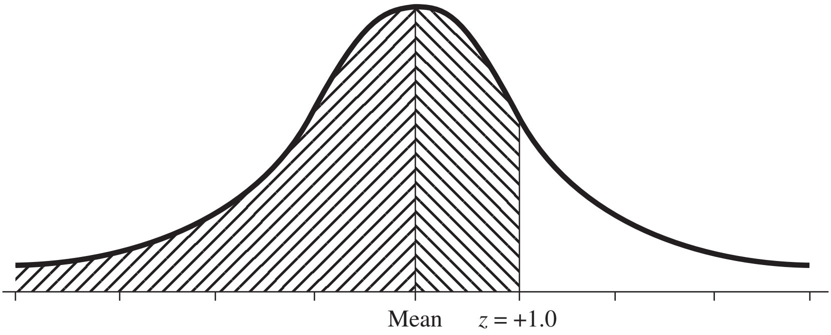
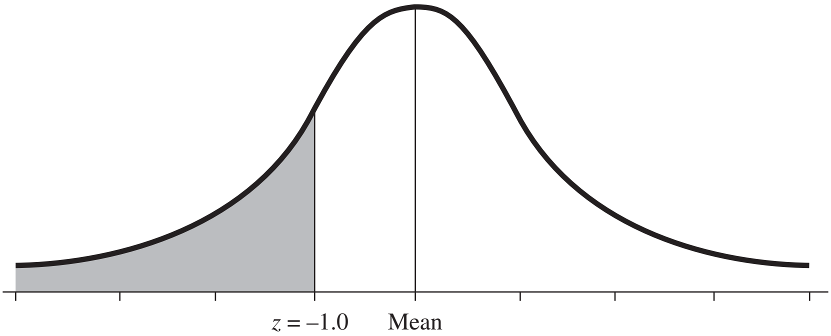


Figure 4.6

Percentile Rank: Area Below $z = +1.0$

Figure 4.7

Area Beyond $z = -1.0$



crosshatched area in Figure 4.6. To convert a proportion to a percentage, we need only multiply by 100. Thus the proportion .8413 corresponds to a PR of 84.13. Now we know that 84.13% of the population have IQ scores lower than 116. I emphasize the importance of drawing a picture of the normal distribution to solve these problems. In the problem above, it would have been easy to forget the .5 area to the left of the mean without a picture to refer to.

It is even easier to find the PR of a score below the mean if you use the correct column of Table A.1. Suppose you want to find the PR for an IQ of 84. Begin by finding z:

$$z = \frac{84 - 100}{16} = \frac{-16}{16} = -1.0$$

Next, draw a picture and shade the area to the left of $z = -1.0$ (see Figure 4.7). Unlike the previous problem, the shaded area this time consists of only one section, which *does* correspond to an entry in Table A.1. First, you must temporarily ignore the minus sign of the z score and find 1.00 in the first column of Table A.1. Then look at the corresponding entry in the column labeled "Beyond z," which is .1587. This is the proportion represented by the shaded area in Figure 4.7 (i.e., the area to the left of $z = -1.0$). The PR of 84 = .1587 × 100 = 15.87; only about 16% of the population have IQ scores less than 84.

The area referred to as Beyond z (in the third column of Table A.1) is the area that begins at z and extends *away* from the mean in the direction of the closest tail. In Figure 4.7, the area between the mean and $z = -1.0$ is .3413 (the same as between the mean and $z = +1.0$), and the area beyond $z = -1$ is .1587. Notice that these two areas add up to .5000. In fact, for any particular z score, the entries for Mean to z and Beyond z will add up to .5000. You can see why by looking at Figure 4.7. The z score divides one half of the distribution into two sections; together those two sections add up to half the distribution, which equals .5.

Finding the Area Between Two z Scores

Now we are ready to tackle more complex problems involving two different z scores. I'll start with two z scores on opposite sides of the mean (i.e., one z is positive and the other is negative). Suppose you have devised a teaching technique that is not accessible to someone with an IQ below 76 and would be too boring for someone with an IQ over 132. To find the proportion of the population for whom your technique would be appropriate, you must first find the two z scores and locate them in a drawing.

$$z = \frac{76 - 100}{16} = \frac{-24}{16} = -1.5$$
$$z = \frac{132 - 100}{16} = \frac{32}{16} = +2.0$$

From Figure 4.8 you can see that you must find two areas of the normal distribution, both of which can be found under the column "Mean to z." For $z = -1.5$ you ignore the minus sign and find that the area from the mean to z is .4332. The corresponding area for $z = +2.0$ is .4772. Adding these two areas together gives a total proportion of .9104. Your teaching technique would be appropriate for 91.04% of the population.

Finding the area enclosed between two z scores becomes a bit trickier when both of the z scores are on the same side of the mean (i.e., both are positive or both are negative). Suppose that you have designed a remedial teaching program that is only appropriate for those whose IQs are below 80 but would be useless for someone with an IQ below 68. As in the problem above, you can find the proportion of people for whom your remedial program is appropriate by first finding the two z scores and locating them in your drawing.

$$z = \frac{80 - 100}{16} = \frac{-20}{16} = -1.25$$
$$z = \frac{68 - 100}{16} = \frac{-32}{16} = -2.0$$

The shaded area in Figure 4.9 is the proportion you are looking for, but it does not correspond to any entry in Table A.1. The trick is to notice that if you take the area from $z = -2$ to the mean and remove the section from $z = -1.25$ to the mean, you are left with the shaded area. (You could also find the area beyond $z = -1.25$ and then remove the area beyond $z = -2.0$.) The area between $z = 2$ and the mean was found in the previous problem to be .4772. From this we subtract the area between $z = 1.25$ and the mean, which is .3944. The proportion we want is .4772 − .3944 = .0828. Thus the remedial teaching program is suitable for use with 8.28% of the population. Note that you cannot subtract the two z scores and then find an area corresponding to the difference of the two z scores; z scores just don't work that way (e.g., the area between z scores of 1 and 2 is much larger than the area between z scores of 2 and 3).

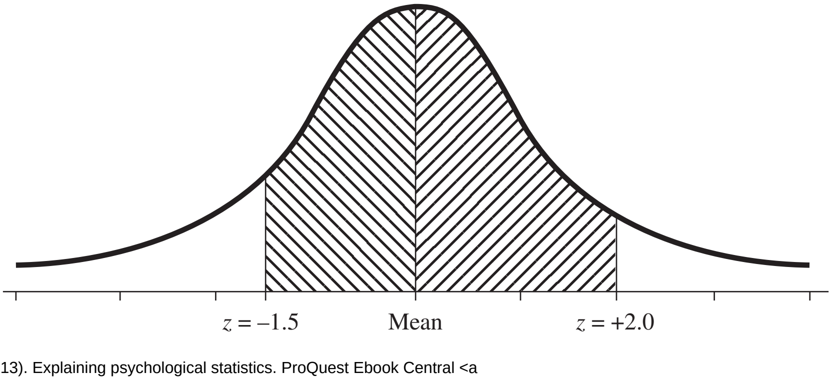
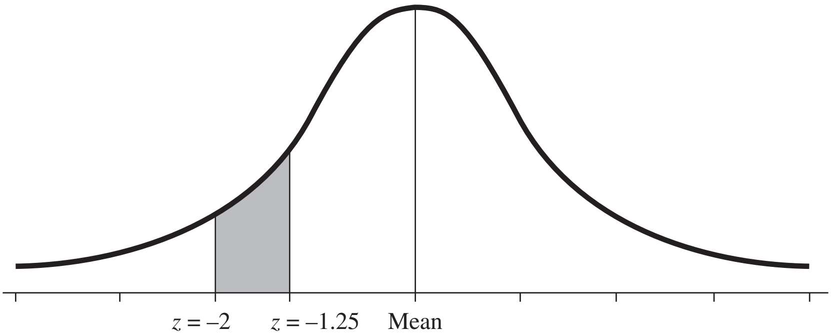


Figure 4.8

The Area Between Two z Scores on Opposite Sides of the Mean

Figure 4.9

The Area Between Two z Scores on the Same Side of the Mean



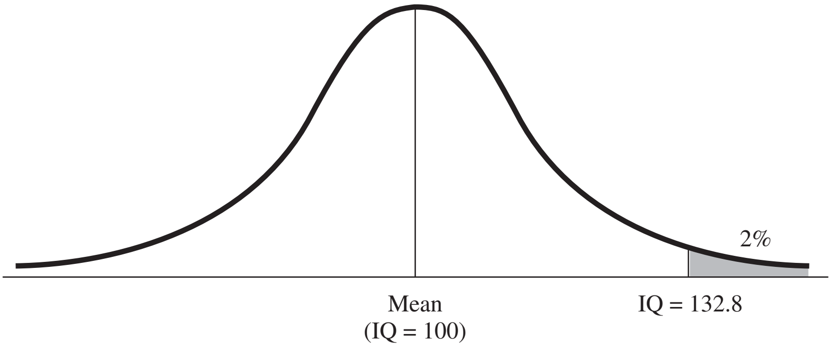
Finding the Raw Scores Corresponding to a Given Area

Often a problem involving the normal distribution will be presented in terms of a given proportion, and it is necessary to find the range of raw scores that represents that proportion. For instance, a national organization called MENSA is a club for people with high IQs. Only people in the top 2% of the IQ distribution are allowed to join. If you were interested in joining and you knew your own IQ, you would want to know the minimum IQ score required for membership. Using the IQ distribution from the problems above, this is an easy question to answer (even if you are not qualified for MENSA). However, because you are starting with an area and trying to find a raw score, the procedure is reversed. You begin by drawing a picture of the distribution and shading in the area of interest, as in Figure 4.10. (Notice that the score that cuts off the upper 2% is also the score that lands at the 98th percentile; that is, you are looking for the score whose PR is 98.) Given a particular area (2% corresponds to a proportion of .0200), you cannot find the corresponding IQ score directly, but you can find the z score using Table A.1. Instead of looking down the z column, you look for the area of interest (in this case, .0200) first and then see which z score corresponds to it. From Figure 4.10, it should be clear that the shaded area is the *area beyond* some as yet unknown z score, so you look in the "Beyond z" column for .0200. You will not be able to find this exact entry, as is often the case, so look at the closest entry, which is .0202. The z score corresponding to this entry is 2.05, so $z = +2.05$ is the z score that cuts off (about) the top 2% of the distribution. To find the raw score that corresponds to $z = +2.05$, you can use Formula 4.2:

$$X = z\sigma + \mu = +2.05(16) + 100 = 32.8 + 100 = 132.8$$

Figure 4.10

Score Cutting Off the Top 2% of the Normal Distribution



Rounding off, you get an IQ of 133—so if your IQ is 133 or above, you are eligible to join MENSA.

Areas in the Middle of a Distribution

One of the most important types of problems involving normal distributions is to locate a given proportion in the middle of a distribution. Imagine an organization called MEZZA, which is designed for people in the middle range of intelligence. In particular, this organization will only accept those in the middle 80% of the distribution—those in the upper or lower 10% are not eligible. What is the range of IQ scores within which your IQ must fall if you are to be eligible to join MEZZA? The appropriate drawing is shown in Figure 4.11. From the drawing you can see that you must look for .1000 in the column labeled "Beyond *z*." The closest entry is .1003, which corresponds to $z = 1.28$. Therefore, $z = +1.28$ cuts off (about) the upper 10%, and $z = -1.28$ the lower 10% of the distribution. Finally, both of these *z* scores must be transformed into raw scores, using Formula 4.2:

$$X = -1.28(16) + 100 = -20.48 + 100 = 79.52$$
$$X = +1.28(16) + 100 = +20.48 + 100 = 120.48$$

Thus (rounding off) the range of IQ scores that contain the middle 80% of the distribution extends from 80 to 120.

From Score to Proportion and Proportion to Score

The above procedures relate raw scores to areas under the curve, and vice versa, by using *z* scores as the intermediate step, as follows:

Raw score ↔ (Formula) ↔ *z* score ↔ (Table A.1) ↔ Area

When you are given a raw score to start with, and you are looking for a proportion or percentage, you move from left to right in the preceding diagram. A raw score can be converted to a *z* score using Formula 4.1. Then an area (or proportion) can be associated with that *z* score by looking down the appropriate column of Table A.1. Drawing a picture will make it clear which column is needed. When given a proportion or percentage to start with, you move from right to left. First, use Table A.1 backwards (look up the area in the appropriate column to find the *z* score), and then use Formula 4.2 to transform the *z* score into a corresponding raw score.

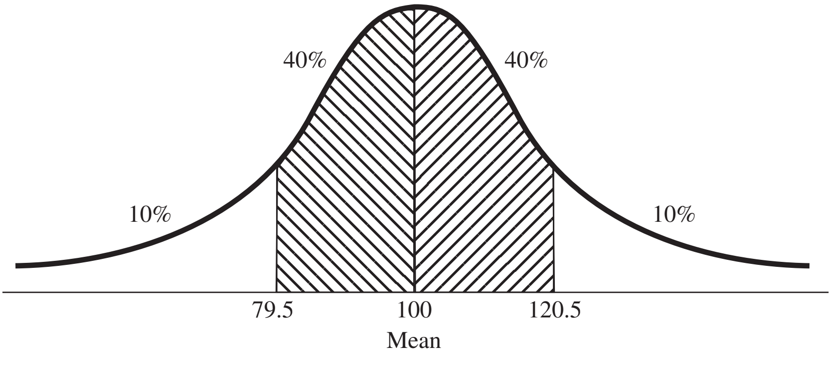


Figure 4.11

Scores Enclosing the Middle 80% of the Normal Distribution

Describing Groups

You can use z scores to find the location of one group with respect to all other groups of the same size, but to do that you must work with the sampling distribution of the mean. The z score has to be modified only slightly for this purpose, as you will see. As an example of a situation in which you may want to compare groups, imagine the following. Suppose there is a university that encourages women's basketball by assigning all of the female students to one basketball team or another at random. Assume that each team has exactly the required five players. Imagine that a particular woman wants to know the probability that the team she is assigned to will have an average height of 67 inches or more. I will show how the sampling distribution of the mean can be used to answer that question.

We begin by assuming that the heights of women at this large university form a normal distribution with a mean of 65 inches and a standard deviation of 3 inches. Next, we need to know what the distribution would look like if it were composed of the means from an infinite number of basketball teams, each with five players. In other words, we need to find the sampling distribution of the mean for $n = 5$. First, we can say that the sampling distribution will be a normal one because we are assuming that the population distribution is (nearly) normal. Given that the sampling distribution is normal, we need only specify its mean and standard deviation.

The z Score for Groups

The mean of the sampling distribution is the same as the mean of the population, that is, μ. For this example, $\mu = 65$ inches. The standard deviation of the sampling distribution of the mean, called the standard error of the mean, is given by Formula 4.5. For this example the standard error, $\sigma_{\overline{X}}$, equals:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{5}} = \frac{3}{2.24} = 1.34$$

So, the sampling distribution of the mean in this case is an approximately normal distribution with a mean of 65 and a standard deviation (i.e., standard error) of 1.34. Now that we know the parameters of the distribution of the means of groups of five (e.g., the basketball teams), we are prepared to answer questions about any particular group, such as the team that includes the inquisitive woman in our example. Because the sampling distribution is normal, we can use the standard normal table to determine, for example, the probability of a particular team having an average height greater than 67 inches. However, we first have to convert the particular group mean of interest to a z score—in particular, a z score with respect to the sampling distribution of the mean, or more informally, a z score for groups. The z score for groups closely resembles the z score for individuals and is given by Formula 4.6:

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \qquad \textbf{Formula 4.6}$$

in which $\sigma_{\overline{X}}$ is a value found using Formula 4.5.

To show the parallel structures of the z score for individuals and the z score for groups, Formula 4.1 for the individual z score follows:

$$z = \frac{X - \mu}{\sigma}$$

Comparing Formula 4.1 with Formula 4.6, you can see that in both cases we start with the score of a particular individual (or mean of a particular sample), subtract the mean of those scores (or sample means), and then divide by the standard deviation of those scores (or sample means). Note that we could put a subscript X on the σ in Formula 4.1 to make it clear that that formula is dealing with individual scores, but unless we are dealing with more than one variable at a time (as in Chapter 9), it is common to leave off the subscript for the sake of simplicity.

In the present example, if we want to find the probability that a randomly selected basketball team will have an average height over 67 inches, it is necessary to convert 67 inches to a z score for groups, as follows:

$$z = \frac{67 - 65}{1.34} = 1.49$$

The final step is to find the area beyond $z = 1.49$ in Table A.1; this area is approximately .068. As Figure 4.12 shows, most of the basketball teams have mean heights that are less than 67 inches; an area of .068 corresponds to fewer than 7 chances out of 100 (or about 1 out of 15) that the woman in our example will be on a team whose average height is at least 67 inches.

Using the z score for groups, you can answer a variety of questions about how common or unusual a particular group is. For the present example, because the standard error is 1.34 and the mean is 65, we know immediately that a little more than two thirds of the basketball teams will average between 63.66 inches (i.e., 65 − 1.34) and 66.34 inches (i.e., 65 + 1.34) in height. Teams with average heights in this range would be considered fairly common, whereas teams averaging more than 67 inches or less than 63 inches in height would be relatively uncommon.

The most common application for determining the probability of selecting a random sample whose mean is unusually small or large is to test a research hypothesis, as you will see in the next chapter. For instance, you could gather a group of heavy coffee drinkers, find the average heart rate

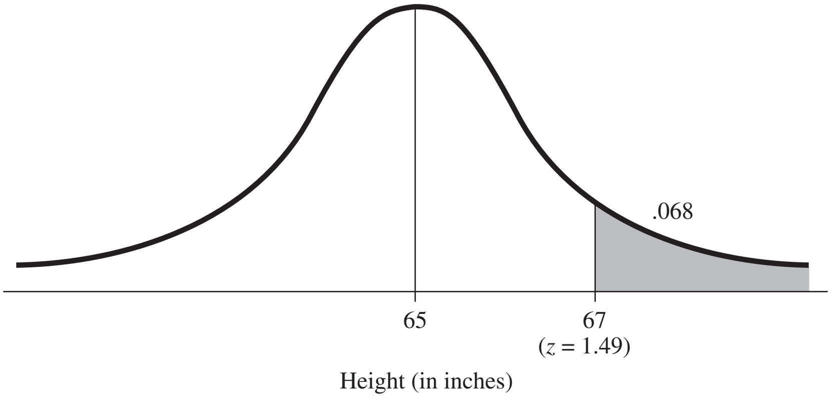


Figure 4.12

Area of the Sampling Distribution Above a z Score for Groups for that group, and use the preceding procedures to determine how unusual it would be to find a random group (the same size) with an average heart rate just as high. The more unusual the heart rate of the coffee-drinking group turns out to be, the more inclined you would be to suspect a link between coffee consumption and heart rate. (Remember, however, that the observation that heavy coffee drinkers do indeed have higher heart rates does not imply that drinking coffee *causes* an increase in heart rate; there are alternative explanations that can only be ruled out by a *true* experiment, in which the experimenter decides at random who will be drinking coffee and who will not.) However, the area that we look up beyond a particular z score does not translate very well to a statement about probability, unless we can make a few important assumptions about how the sample was selected, and about the population it was selected from. I will discuss these assumptions in the next chapter, in the context of drawing inferential conclusions from z scores. In the meantime, I will conclude this section by explaining some important rules of probability.

Probability Rules

As I pointed out in Section A, statements about areas under the normal curve can be translated directly to statements about probability. For instance, if you select one person at random, the probability that that person will have an IQ between 76 and 132 is about .91, because that is the amount of area enclosed between those two IQ scores, as we found earlier (see Figure 4.8). To give you a more complete understanding of probability and its relation to problems involving distributions, I will lay out some specific rules. To represent the probability of an event symbolically, I will write p(A), where A stands for some event. For example, $p(\text{IQ} > 110)$ stands for the probability of selecting someone with an IQ greater than 110.

Rule 1

Probabilities range from 0 (the event is certain *not* to occur) to 1 (the event is *certain* to occur) or from 0 to 100 if probability is expressed as a percentage instead of a proportion. As an example of $p = 0$, consider the case of adult height. The distribution ends somewhere around $z = -15$ on the low end and $z = +15$ on the high end. So for height, the probability of selecting someone for whom z is greater than +20 (or less than −20) is truly zero. An example of $p = 1$ is the probability that a person's height will be between $z = -20$ and $z = +20$ [i.e, $p(-20 < z < +20) = 1$].

Rule 2: The Addition Rule

If two events are *mutually exclusive*, the probability that either one event *or* the other will occur is equal to the sum of the two individual probabilities. Stated as Formula 4.7, the addition rule for mutually exclusive events is:

$$p(\text{A or B}) = p(\text{A}) + p(\text{B})$$ **Formula 4.7**

Two events are mutually exclusive if the occurrence of one rules out the occurrence of the other. For instance, if we select one individual from the IQ distribution, this person cannot have an IQ that is both above 120.5 and also below 79.5—these are mutually exclusive events. As I demonstrated in the discussion of the hypothetical MEZZA organization, the probability of each of these events is .10. We can now ask: What is the probability that a randomly selected individual will have an IQ above 120.5 *or* below 79.5? Using Formula 4.7, we simply add the two individual probabilities: .1 + .1 = .2. In terms of a single distribution, two mutually exclusive events are represented by two areas under the curve that do *not* overlap. (In contrast, the area from $z = -1$ to $z = +1$ and the area above $z = 0$ are *not* mutually exclusive because they *do* overlap.) If the areas do not overlap, we can simply add the two areas to find the probability that an event will be in one area *or* the other. The addition rule can be extended easily to any number of events, if all of the events are mutually exclusive (i.e., no event overlaps with any other). For a set of mutually exclusive events the probability that one of them will occur, p(A or B or C, etc.), is the sum of the probabilities for each event, that is, $p(\mathrm{A}) + p(\mathrm{B}) + p(\mathrm{C})$, and so on.

The Addition Rule for Overlapping Events

The addition rule must be modified if events are not mutually exclusive. If there is some overlap between two events, the overlap must be subtracted after the two probabilities have been added. Stated as Formula 4.8, the addition rule for two events that are *not* mutually exclusive is:

$$p(\text{A or B}) = p(\text{A}) + p(\text{B}) - p(\text{A and B}) \qquad \textbf{Formula 4.8}$$

where p(A and B) represents the overlap (the region where A and B are both true simultaneously). For example, what is the probability that a single selection from the normal distribution will be either within one standard deviation of the mean *or* above the mean? The probability of the first event is the area between $z = -1$ and $z = +1$, which is about .68. The probability of the second event is the area above $z = 0$, which is .5. Adding these we get .68 + .5 = 1.18, which is more than 1.0 and therefore impossible. However, as you can see from Figure 4.13, these events are not mutually exclusive; the area of overlap corresponds to the interval from $z = 0$ to $z = +1$. The area of overlap, that is, p(A and B), equals about .34, and because it is actually being added in twice (once for each event), it must be subtracted once from the total. Using Formula 4.8, we find that p(A or B) = .68 + .5 − .34 = 1.18 − .34 = .84 (rounded off).

A Note About Exhaustive Events

Besides being mutually exclusive, two events can also be *exhaustive* if one or the other *must* occur (together they exhaust all the possible events). For

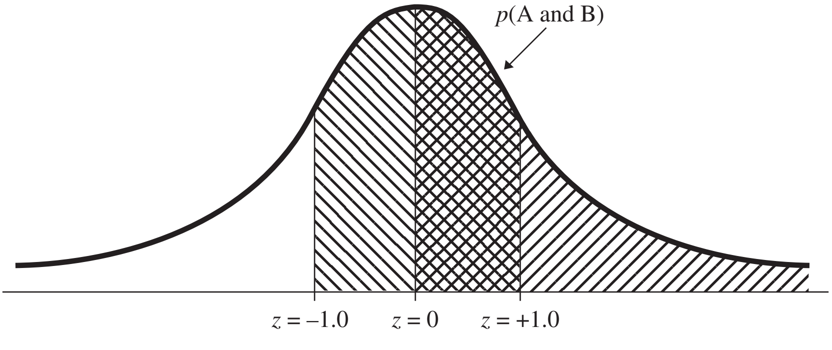


Figure 4.13

The Area Corresponding to Two Overlapping Events instance, consider the event of being above the mean and the event of being below the mean; these two events are not only mutually exclusive, they are exhaustive as well. The same is true of being within one standard deviation from the mean and being at least one standard deviation away from the mean. When two events are both mutually exclusive and exhaustive, one event is considered the *complement* of the other, and the probabilities of the two events must add up to 1.0. If the events A and B are mutually exclusive and exhaustive, we can state that $p(\text{B}) = 1.0 - p(\text{A})$.

Just as two events can be mutually exclusive but not exhaustive (e.g., $z > +1.0$ and $z < -1.0$), two events can be exhaustive without being mutually exclusive. For example, the two events $z > -1.0$ and $z < +1.0$ are exhaustive (there is no location in the normal distribution that is not covered by one event or the other), but they are not mutually exclusive; the area of overlap is shown in Figure 4.14. Therefore, the two areas represented by these events will not add up to 1.0, but rather somewhat more than 1.0.

Rule 3: The Multiplication Rule

If two events are *independent*, the probability that both will occur (i.e., A *and* B) is equal to the two individual probabilities multiplied together. Stated as Formula 4.9, the multiplication rule for independent events is:

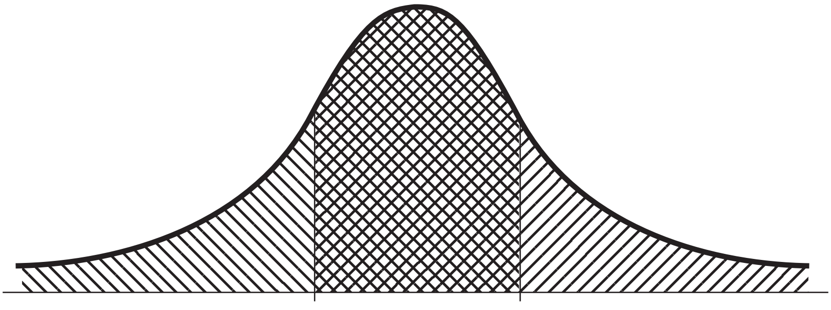
$$p(\text{A and B}) = p(\text{A})p(\text{B}) \qquad \textbf{Formula 4.9}$$

Two events are said to be independent if the occurrence of one in no way affects the probability of the other. The most common example of independent events is two flips of a coin. As long as the first flip does not damage or change the coin in some way—and it's hard to imagine how flipping a coin could change it—the second flip will have the same probability of coming up heads as the first flip. (If the coin is unbiased, $p(\text{H}) = p(\text{T}) =.5$.) Even if you have flipped a fair coin and have gotten 10 heads in a row, the coin will not be altered by the flipping; the chance of getting a head on the eleventh flip is still .5. It may seem that after 10 heads, a tail would become more likely than usual, so as to even out the total number of heads and tails. This belief is a version of the *gambler's fallacy*; in reality, the coin has no memory—it doesn't keep track of the previous 10 heads in a row. The multiplication rule can be extended easily to any number of events that are all independent of each other ($p(\text{A and B and C, etc.}) = p(\text{A})p(\text{B})p(\text{C})$, etc.).

Consider two independent selections from the IQ distribution. What is the probability that if we choose two people at random, their IQs will both be within one standard deviation of the mean? In this case, the probability

Figure 4.14

The Area Corresponding to Two Exhaustive, but Not Mutually Exclusive, Events



of both individual events is the same, about .68 (assuming we replace the first person in the pool of possible choices before selecting the second; see the next paragraph). Formula 4.9 tells us that the probability of both events occurring jointly, that is, p(A and B), equals (.68)(.68) = .46. When the two events are *not* independent, the multiplication rule must be modified. If the probability of an event changes because of the occurrence of another event, we are dealing with a *conditional probability*.

Conditional Probability

A common example of events that are *not* independent are those that involve successive samplings from a finite population *without replacement*. Let us take the simplest possible case: A bag contains three marbles; two are white and one is black. If you grab marbles from the bag without looking, what is the probability of picking two white marbles in a row? The answer depends on whether you select marbles *with replacement* or *without replacement*. In selecting with replacement, you take out a marble, look at it, and then replace it in the bag before picking a second marble. In this case, the two selections are independent; the probability of picking a white marble is the same for both selections: 2/3. The multiplication rule tells us that when the two events are independent (e.g., sampling *with* replacement), we can multiply their probabilities. Therefore, the probability of picking two white marbles in a row with replacement is (2/3)(2/3) = 4/9, or about .44.

On the other hand, if you are sampling *without* replacement, the two events will *not* be independent because the first selection will alter the probabilities for the second. The probability of selecting a white marble on the first pick is still 2/3, but if the white marble is *not* replaced in the bag, the probability of selecting a white marble on the second pick is only 1/2 (there is one white marble and one black marble left in the bag). Thus the conditional probability of selecting a white marble, *given that* a white marble has already been selected and not replaced, that is, $p(\mathrm{W} \mid \mathrm{W})$, is 1/2 (the vertical bar between the Ws represents the word "given"). To find the probability of selecting two white marbles in a row when not sampling with replacement, we need to use Formula 4.10 (the multiplication rule for dependent events):

$$p(\text{A and B}) = p(\text{A})p(\text{B}|\text{A}) \qquad \textbf{Formula 4.10}$$

In this case, both A and B can be symbolized by W (picking a white marble): $p(\mathrm{W})p(\mathrm{W} \mid \mathrm{W}) = (2/3)\ (1/2) = 1/3$, or .33 (less than the probability of picking two white marbles when sampling with replacement). The larger the population, the less difference it will make whether you sample with replacement or not. (With an infinite population, the difference is infinitesimally small.) For the remainder of this text I will assume that the population from which a sample is taken is so large that sampling without replacement will not change the probabilities enough to have any practical consequences. Conditional probability will have a large role to play, however, in the logical structure of null hypothesis testing, as described in the next chapter.

B SUMMARY

1. If a variable is normally distributed and you know both the mean and standard deviation of the population, it is easy to find the proportion of the distribution that falls above or below any raw score or between any two raw scores. Conversely, for a given proportion at the top, bottom, or middle of the distribution, you can find the raw score or scores that form the boundary of that proportion.

2. It is easier to work with Table A.1 if you begin by drawing a picture of the normal distribution and then draw a vertical line in the middle and one that corresponds roughly to the *z* score or area with which you are working.
3. To find the proportion below a given raw score first convert the raw score to a *z* score with Formula 4.1. If the *z* score is negative, the area below (i.e., to the left of) the *z* score is the area in the "Beyond *z*" column. If the *z* score is positive, the area below the *z* score is the area in the "Mean to *z*" column *plus* .5. Multiply the proportion by 100 to get the percentile rank for that score. To find the proportion above a particular raw score, you can alter the procedure just described appropriately or find the proportion below the score and subtract from 1.0.
4. To find the proportion between two raw scores,
   a. If the corresponding *z* scores are opposite in sign, you must find two areas—the area between the mean and *z* for each of the two *z* scores—and add them.
   b. If both *z* scores are on the same side of the mean, you can find the area between the mean and *z* for each and then *subtract* the smaller from the larger area (alternatively, you can find the area beyond *z* for both and subtract). Reminder: You cannot subtract the two *z* scores first and then look for the area.
5. To find the raw score corresponding to a given proportion, first draw the picture and shade in the appropriate area. If the area is the top *X%*, convert to a proportion by dividing by 100. Then find that proportion (or the closest one to it) in the "Beyond *z*" column. The corresponding *z* score can then be converted to a raw score using Formula 4.2. If the area is the bottom *X%* (i.e., left side of distribution), the procedure is the same, but when you are using Formula 4.2 don't forget that the *z* score has a negative sign.
6. To find the raw scores enclosing the middle *X%* of the distribution, first subtract *X%* from 100% and then divide by 2 to find the percentage in each tail of the distribution. Convert to a proportion by dividing by 100 and then look for this proportion in the "Beyond *z*" column of Table A.1 (e.g., if you are trying to locate the middle 90%, you would be looking for a proportion of .05). Transform the corresponding *z* score to a raw score to find the upper (i.e., right) boundary of the enclosed area. Then, put a minus sign in front of the same *z* score and transform it to a raw score again to find the lower boundary.
7. If you are working with the mean of a sample instead of an individual score, and you want to know the proportion of random samples of the same size that would have a larger (or smaller) mean, you have to convert your sample mean to a *z* score for groups. Subtract the population mean from your sample mean, and then divide by the standard error of the mean, which is the population standard deviation divided by the square root of the sample size. The area above or below this *z* score can be found from Table A.1 just as though the *z* score were from an individual score in a normal distribution.
8. To use Table A.1 for a *z* score from a sample mean, you must assume that the sampling distribution of the mean is a normal distribution.
9. The Rules of Probability

   Rule 1: The probability of an event, for example, $p(\text{A})$, is usually expressed as a proportion, in which case $p(\text{A})$ can range from zero (the event A is certain *not* to occur) to 1.0 (the event A *is* certain to occur).

   Rule 2: The addition rule for mutually exclusive events states that if the occurrence of event A precludes the occurrence of event B, the probability of either A *or* B occurring, $p(\text{A or B})$, equals $p(\text{A}) + p(\text{B})$. The addition rule must be modified as follows if the events are *not* mutually exclusive: $p(\text{A or B}) = p(\text{A}) + p(\text{B}) - p(\text{A and B})$. Also, if two events are both mutually exclusive and *exhaustive* (one of the two events must occur), $p(\text{A}) + p(\text{B}) = 1.0$, and therefore, $p(\text{B}) = 1.0 - p(\text{A})$.

Rule 3: The multiplication rule for independent events states that the probability that two independent events will both occur, $p(\text{A and B})$, equals $p(\text{A})p(\text{B})$. Two events are *not* independent if the occurrence of one event changes the probability of the other. The probability of one event, given that another has occurred, is called a *conditional probability*.

10. The probability of two *dependent* events both occurring is given by a modified multiplication rule: The probability of one event is multiplied by the conditional probability of the other event, *given that the first* event has occurred. When you are sampling from a finite population without replacement, successive selections will not be independent. However, if the population is very large, *sampling without replacement* is barely distinguishable from *sampling with replacement* (any individual in the population has an exceedingly tiny probability of being selected twice), so successive selections can be considered independent even without replacement.

Advanced Material: The Mathematics of the Normal Distribution

The true normal curve is determined by a mathematical equation, just as a straight line or a perfect circle is determined by an equation. The equation for the normal curve is a mathematical function into which you can insert any *X* value (usually represented on the horizontal axis) to find one corresponding *Y* value (usually plotted along the vertical axis). Because *Y*, the height of the curve, is a function of *X*, *Y* can be symbolized as *f*(*X*). The equation for the normal curve can be stated as follows:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

where π is a familiar mathematical constant, and so is *e* ($e = 2.7183$, approximately). The symbols μ and σ^2 are called the *parameters* of the normal distribution, and they stand for the ordinary mean and variance. These two parameters determine which normal distribution is being graphed.

The preceding equation is a fairly complex one, but it can be simplified by expressing it in terms of *z* scores. This gives us the equation for the standard normal distribution and shows the intimate connection between the normal distribution and *z* scores:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Because the variance of the standard normal distribution equals 1.0, $2\pi\sigma^2 = 2\pi$, and the power that *e* is raised to is just $-\frac{1}{2}$ times the *z* score squared. The fact that *z* is being squared tells us that the curve is symmetric around zero. For instance, the height for $z = -2$ is the same as the height for $z = +2$ because in both cases, $z^2 = +4$. Because the exponent of *e* has a minus sign, the function is largest (i.e., the curve is highest) when *z* is smallest, namely, zero. Thus, the mode of the normal curve (along with the mean and median) occurs in the center, at $z = 0$. Note that the height of the curve is never zero; that is, the curve never touches the X axis. Instead, the curve extends infinitely in both directions, always getting closer to the X axis. In mathematical terms, the X axis is the *asymptote* of the function, and the function touches the asymptote only at infinity.

The height of the curve ($f(X)$, or Y) is called the density of the function, so the preceding equation is often referred to as a *probability density function*. In more concrete terms, the height of the curve can be thought of as representing the relative likelihood of each X value; the higher the curve, the more likely is the X value at that point. However, as I pointed out in Section A, the probability of any *exact* value occurring is infinitely small, so when one talks about the probability of some X value being selected, it is common to talk in terms of a range of X values (i.e., an interval along the horizontal axis). The probability that the next random selection will come from that interval is equal to the proportion of the total distribution that is contained in that interval. This is the area under the curve corresponding to the given interval, and this area can be found mathematically by *integrating* the function over the interval using the calculus. Conveniently, the areas between the mean and various z scores have already been calculated and entered into tables, such as Table A.1. Thanks to modern software, these areas are also easily obtained with great accuracy by statistical calculators on the web, or from statistical packages, like SPSS.

EXERCISES

*1. Suppose that a large Introduction to Psychology class has taken a midterm exam, and the scores are normally distributed (approximately) with $\mu = 75$ and $\sigma = 9$. What is the percentile rank (PR) for a student
   a. Who scores 90?
   b. Who scores 70?
   c. Who scores 60?
   d. Who scores 94?

2. Find the area between
   a. $z = -0.5$ and $z = +1.0$
   b. $z = -1.5$ and $z = +0.75$
   c. $z = +0.75$ and $z = +1.5$
   d. $z = -0.5$ and $z = -1.5$

*3. Assume that the resting heart rate in humans is normally distributed with $\mu = 72$ bpm (i.e., beats per minute) and $\sigma = 8$ bpm.
   a. What proportion of the population has resting heart rates above 82 bpm? Above 70 bpm?
   b. What proportion of the population has resting heart rates below 75 bpm? Below 50 bpm?
   c. What proportion of the population has resting heart rates between 80 and 85 bpm? Between 60 and 70 bpm? Between 55 and 75 bpm?

4. Refer again to the population of heart rates described in Exercise 3:
   a. Above what heart rate do you find the upper 25% of the people? (That is, what heart rate is at the 75th percentile, or third quartile?)
   b. Below what heart rate do you find the lowest 15% of the people? (That is, what heart rate is at the 15th percentile?)
   c. Between which two heart rates do you find the middle 75% of the people?

*5. A new preparation course for the math SAT is open to those who have already taken the test once and scored in the middle 90% of the population. In what range must a test-taker's previous score have fallen for the test-taker to be eligible for the new course?

6. A teacher thinks her class has an unusually high IQ, because her 36 students have an average IQ ($\overline{X}$) of 108. If the population mean is 100 and $\sigma = 15$,
   a. What is the z score for this class?
   b. What percentage of classes ($n = 36$, randomly selected) would be even higher on IQ?

*7. An aerobics instructor thinks that his class has an unusually low resting heart rate.

If $\mu = 72$ bpm and $\sigma = 8$ bpm, and his class of 14 pupils has a mean heart rate ($\overline{X}$) of 66,

a. What is the *z* score for the aerobics class?

b. What is the probability of randomly selecting a group of 14 people with a mean resting heart rate lower than the mean for the aerobics class?

8. Imagine that a test for spatial ability produces scores that are normally distributed in the population with $\mu = 60$ and $\sigma = 20$.

a. Between which two scores will you find the middle 80% of the people?

b. Considering the means of groups, all of which have 25 participants, between what two scores will the middle 80% of these means be?

*9. Suppose that the average person sleeps 8 hours each night and that $\sigma = .7$ hour.

a. If a group of 50 joggers is found to sleep an average of 7.6 hours per night, what is the *z* score for this group?

b. If a group of 200 joggers also has a mean of 7.6, what is the *z* score for this larger group?

c. Comparing your answers to parts a and b, can you determine the mathematical relation between sample size and *z* (when $\overline{X}$ remains constant)?

10. Referring to the information in Exercise 7, if an aerobics class had a mean heart rate ($\overline{X}$) of 62, and this resulted in a group *z* score of −7.1, how large must the class have been?

*11. Suppose that the mean height for a group of 40 women who had been breastfed for at least the first 6 months of life was 66.8 inches.

a. If $\mu = 65.5$ inches and $\sigma = 2.6$ inches, what is the *z* score for this group?

b. If height had been measured in centimeters, what would the *z* score be? (*Hint*: Multiply $\overline{X}$, μ, and σ by 2.54 to convert inches to centimeters.)

c. Comparing your answers to parts a and b, what can you say about the effect on *z* scores of changing units? Can you explain the significance of this principle?

12. Suppose that the mean weight of adults (μ) is 150 pounds with $\sigma = 30$ pounds. Consider the mean weights of all possible space shuttle crews ($n = 7$). If the space shuttle cannot carry a crew that weighs more than a total of 1190 pounds, what is the probability that a randomly selected crew will be too heavy? (Assume that the sampling distribution of the mean would be approximately normal.)

*13. Consider a normally distributed population of resting heart rates with $\mu = 72$ bpm and $\sigma = 8$ bpm:

a. What is the probability of randomly selecting someone whose heart rate is either below 58 or above 82 bpm?

b. What is the probability of randomly selecting someone whose heart rate is either between 67 and 75 bpm, above 80 bpm, or below 60 bpm?

c. What is the probability of randomly selecting someone whose heart rate is either between 66 and 77 bpm or above 74 bpm?

14. Refer again to the population of heart rates described in the previous exercise:

a. What is the probability of randomly selecting two people in a row whose resting heart rates are both above 78 bpm?

b. What is the probability of randomly selecting *three* people in a row whose resting heart rates are all below 68 bpm?

c. What is the probability of randomly selecting two people, one of whom has a resting heart rate below 70 bpm, while the other has a resting heart rate above 76 bpm?

*15. What is the probability of selecting each of the following at random from the population (assume $\sigma = 16$):

a. One person whose IQ is either above 110 or below 95?

b. One person whose IQ is either between 95 and 110 or above 105?

c. Two people with IQs above 90?

d. One person with an IQ below 90 and one person with an IQ above 115?

16. An ordinary deck of playing cards consists of 52 different cards, 13 in each of four suits (hearts, diamonds, clubs, and spades).

a. What is the probability of randomly drawing two hearts in a row if you replace the first card before picking the second?

b. What is the probability of randomly drawing two hearts in a row if you draw *without* replacement?

c. What is the probability of randomly drawing one heart and then one spade in two picks *without* replacement?

C ANALYSIS BY SPSS

After SPSS has calculated the mean and standard deviation (*SD*) for a variable in your spreadsheet, you could find the *z* scores for that variable by using **Transform/Compute Variable**. For instance, if you are creating *z* scores for *mathquiz*, you would type the following in the *Numeric Expression* space of the Compute Variable box: **(mathquiz—xx.xx) / yy.yy**, where xx.xx and yy.yy are the values SPSS gave you for the mean and *SD* of *mathquiz*, respectively. It would make sense to name the target variable something like *z_mathquiz*. Note that the standard deviation SPSS gives you is the unbiased one, so your *z* scores will be a bit different from what you would get by using the biased *SD*. Fortunately, this difference is only noticeable when dealing with small samples. A bigger concern is retaining enough digits beyond the decimal point for both the mean and *SD* you use in the Compute box. If your numbers are all less than 1.0, for instance, the two decimals shown in the preceding example would not give you much accuracy. You can avoid this issue by asking SPSS to compute *z* scores for any of your variables automatically using the following four steps.

Creating *z* Scores

1. Select **Descriptive Statistics** from the **Analyze** menu, and then click on **Descriptives** . . .
2. Under the list of variables that appears on the left side of the **Descriptives** dialog box, check the little box that precedes the phrase "Save standardized values as variables."
3. Move over the variables for which you would like to see *z* scores.
4. Click on the **Options** button if you want to see any statistics other than those that are selected by default. When back to the original dialog box, click **OK**.

Two very different things will happen as the result of following the above procedure. First, you will get the usual Descriptives output for your chosen variables (plus any additional statistics you checked in the Options box). Second, for each of those variables, SPSS will have added a new variable at the end of your spreadsheet, containing the *z* scores for that variable. SPSS names the new *z*-score variables by just putting the letter "z" at the beginning of the original name, so, for example, *mathquiz* becomes *zmathquiz*. Note again that these *z* scores will be based on the unbiased standard deviation of your variable. Conveniently, the *un*biased standard deviation of these *z* scores will be 1.0.

Obtaining Standard Errors

If you want to calculate the *standard error of the mean* (SEM) for a particular variable, you just have to divide the standard deviation for that variable by the square root of the size of the sample (*n*). However, if you would like SPSS to do it for you, you can use the preceding list of steps for creating *z* scores with a little modification. Start with step #1, skip step #2, and in step #3 move over the variables for which you would like to see SEMs. At step #4, check the little box for *S.E. mean*. Applied to the variable *mathquiz*, the SPSS results of the steps just described are shown in Table 4.2.

Table 4.2

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
mathquiz	85	9	49	29.07	1.028	9.480
Valid N (listwise)	85					

You will find the SEM right next to the Mean, labeled *Std. Error*. From the organization of the output, SPSS is telling you that the *Statistic* for estimating the population mean and its standard error are 29.07 and 1.028, respectively. If you knew the actual population mean, you could subtract it from the mean in the table, and then divide that difference by the Std. Error to obtain the z score for your sample. I will expand on this point in the next chapter. The entry in the lower-left corner of the table, "Valid N (listwise)," tells you how many cases in your dataset have valid (i.e., not missing) values for every variable that you moved over in the Descriptives box.

Obtaining Areas of the Normal Distribution

There are some convenient normal distribution calculators available for free on the web, but in the unlikely case that you are offline and do have access to SPSS, you can use SPSS to obtain areas under the normal distribution with far more accuracy (i.e., more digits past the decimal point) than you could get from a printed table, like Table A.1 in your text. And you can obtain areas beyond *z* scores that are larger than those included in any printed tables.

1. Start by opening a new (i.e., empty) data sheet, and entering the *z* score of interest in the first cell. For convenience, you can assign the simple variable name "z" to this first column.
2. Then open the Compute Variable box by selecting Compute (the first choice) from the Transform menu.
3. In the Target Variable space, type an appropriate variable name like "area."
4. In the Numeric Expression space, type "CDFNORM" and then, in parentheses, the name of the first variable—for example, CDFNORM (z). (Note: I am using uppercase letters for emphasis, but SPSS does not distinguish between upper- and lowercase letters in variable or function names).

CDFNORM is a function name that stands for the *C*umulative *D*ensity *F*unction for the *NORM*al distribution; therefore it returns a value equal to all of the area to the left of (i.e., below) the *z* score you entered. If you multiply this value by 100, you get the percentile rank associated with the *z* score in question, *if you are dealing with a normal distribution* (this works for both positive and negative *z* scores, as long as you include the minus sign for any negative z score). Note that the larger the *z* score you enter in the first cell of your SPSS datasheet, the more "decimals" you will need to display the answer accurately. The fourth column in Variable View lets you set the number of digits that will be displayed to the right of the decimal point—as long as this number is less than the number in the third column (Width) for that variable. For instance, if you are looking for the area associated with a *z* score between 3 and 4, you will want to set the "decimals" number to at least 6.

Data Transformations

If you create a set of *z* scores corresponding to one of your variables, the distribution of the *z* scores will have exactly the same shape as the distribution of the original variable. If you want to change the shape of your distribution, usually to make it resemble the normal distribution, you need to use a transformation that is *not* linear. For example, a transformation that is often used to greatly reduce the positive skew of a distribution is to take the logarithm of each value. This is another task that is best handled by first selecting Compute Variable from the Transform menu. One of the most positively skewed variables in Ihno's data set is *prevmath*, so I'll use that variable for my example. After you have opened the Compute Variable box, type a new variable name, like *log_prevmath* in the Target Variable space, and then type "Lg10 (*prevmath* + 1)" in the Numeric Expression space and click OK. If you look at the distribution of the logs of the *prevmath* scores (and, even better, request a skewness measure), you'll see that it is much less skewed than the distribution for *prevmath*. Note that I had to add 1 to *prevmath*, because there are quite a few scores of zero, and you can't take the log of zero. Note also that "Lg10" yields logs to the base 10, but the natural logs, obtained by typing "Ln" instead of "Lg10," will produce a distribution that has exactly the same shape as do logs to the base 10. Finally, if you wanted to *replace* the original *prevmath* values with their log-transformed values, rather than adding a new variable, you would just type *prevmath* in the Target Variable space. Because this action eliminates the original values of your variable from the spreadsheet, SPSS warns you of this by asking "Change existing variable?" and you can then click OK or Cancel.

EXERCISES

1. Create new variables consisting of the z scores for the anxiety and heart rate measures at baseline in Ihno's data set. Request means and *SD*s of the z-score variables to demonstrate that the means and *SD*s are 0 and 1, respectively, in each case.
2. Create a z-score variable corresponding to the math background quiz score, and then transform the z-score variable to a T score, an SAT score, and an IQ score. Repeat for the stats quiz.
3. Use SPSS to find the following areas under the normal curve (your answer should include six digits past the decimal point):
   a. The area below a z score of +3.1.
   b. The area above a z score of +3.3.
   c. The area below a z score of −3.7.
   d. The area between the mean and a z score of +.542
   e. The area between the mean and a z score of −1.125
4. Use SPSS to find the percentile ranks for the following z scores (your answer should include two digits past the decimal point):
   a. 3.1
   b. 3.3
   c. 3.7
   d. .542
   e. −1.125
5. Find the mean, *SD*, standard error, and skewness for the *phobia* variable. Then, create a new variable that is the square root of the *phobia* variable, and find those statistics again. What happened to the skewness of *phobia* after taking the square root?
6. Find the mean, *SD*, standard error, and skewness for the *statsquiz* variable. Then, create a new variable that is the natural log of the *statsquiz* variable, and find those statistics again. What happened to the skewness of *statsquiz*? Explain the lesson that you learned from this exercise.

KEY FORMULAS

The z score corresponding to a raw score:

$$z = \frac{X - \mu}{\sigma}$$ **Formula 4.1**

The raw score that corresponds to a given z score:

$$X = z\sigma + \mu$$ **Formula 4.2**

The SAT score corresponding to a raw score, if the z score has already been calculated:

$$\text{SAT} = 100z + 500$$ **Formula 4.3**

The T score corresponding to a raw score, if the z score has already been calculated:

$$T = 10z + 50 \qquad \textbf{Formula 4.4}$$

The standard error of the mean:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{Formula 4.5}$$

The z score for groups (the standard error of the mean must be found first):

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \qquad \textbf{Formula 4.6}$$

The addition rule for mutually exclusive events:

$$p(\text{A or B}) = p(\text{A}) + p(\text{B}) \qquad \textbf{Formula 4.7}$$

The addition rule for events that are *not* mutually exclusive:

$$p(\text{A or B}) = p(\text{A}) + p(\text{B}) - p(\text{A and B}) \qquad \textbf{Formula 4.8}$$

The multiplication rule for independent events:

$$p(\text{A and B}) = p(\text{A})p(\text{B}) \qquad \textbf{Formula 4.9}$$

The multiplication rule for events that are *not* independent (based on *conditional probability*):

$$p(\text{A and B}) = p(\text{A})p(\text{B}|\text{A}) \qquad \textbf{Formula 4.10}$$