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

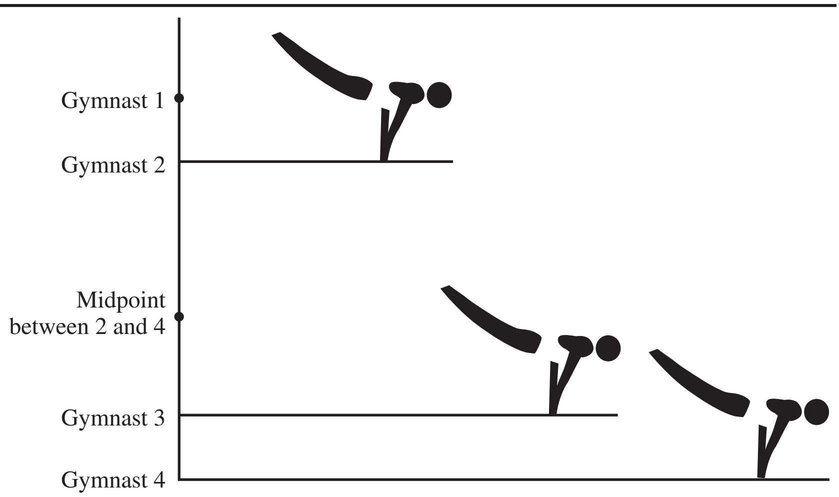
A quantitative level of measurement is being used when the different values of a scale can be placed in order. For instance, an elementary school teacher may rate the handwriting of each student in a class as excellent, good, fair, or poor. Unlike the categories of a nominal scale, these designations 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).

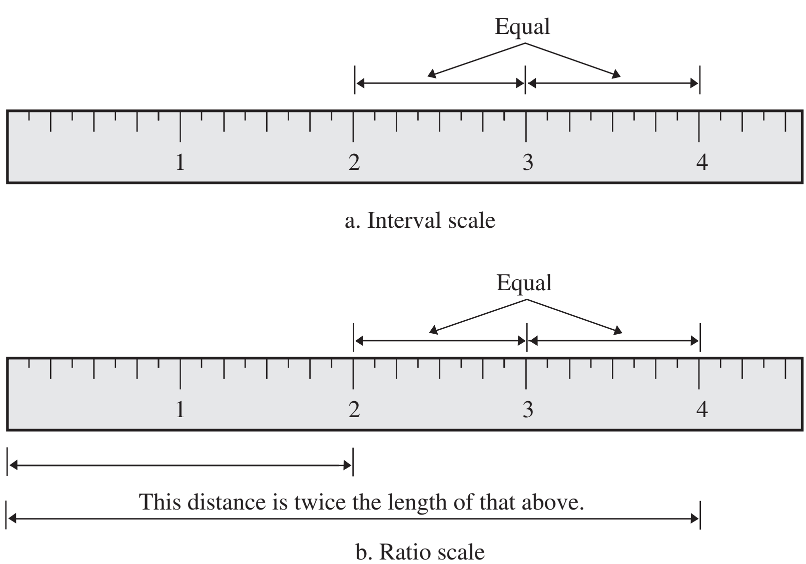
Interval and Ratio Scales

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

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.)

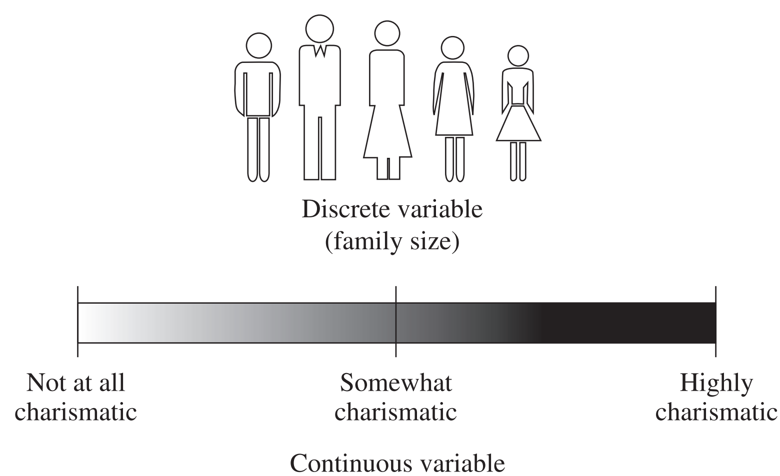
An example of a *discrete variable* is the size of a family. This variable can be measured on a ratio scale by simply counting the family members, but it does not vary continuously—a family can have two or three children, but there is no meaningful value in between. The size of a family will always be 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

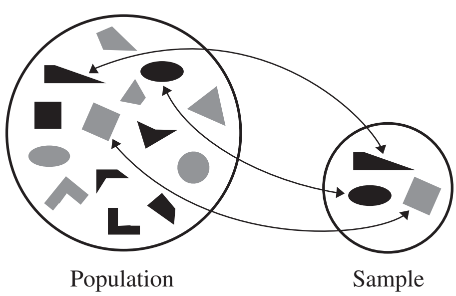


Figure 1.4

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 $\sum(X_i X_i) \neq \sum(X_i)\sum(X_i)$. Because $X_i X_i$ 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.)

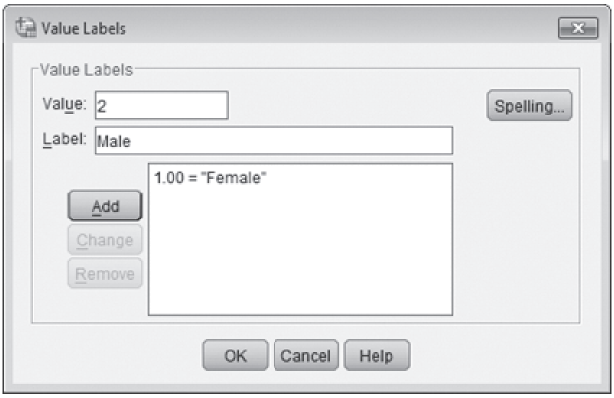


Figure 1.5

Data Coding

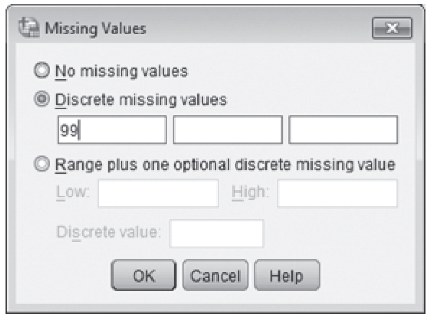
Ihno could have entered the words male and female as the values for gender, but for reasons that will be explained as we go, it was more useful for her to enter all of her data in the form of numbers, even for the nominal variables. She used the number 1 to represent females and 2 for males, which is 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



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*.