

# Part I

## Descriptive Statistics

Chapter 1	Introduction
Chapter 2	Frequency Distributions and Graphs
Chapter 3	Transformed Scores I: Percentiles
Chapter 4	Measures of Central Tendency
Chapter 5	Measures of Variability
Chapter 6	Additional Techniques for Describing Batches of Data
Chapter 7	Transformed Scores II: $z$ and $T$ Scores
Chapter 8	The Normal Distribution



# Chapter 1

## Introduction

### PREVIEW

#### Why study statistics?

What are three important reasons why a knowledge of statistics is essential for anyone majoring in psychology, sociology, or education?

#### Descriptive and Inferential Statistics

What is the difference between descriptive and inferential statistics? Why must behavioral science researchers use inferential statistics?

#### Populations, Samples, Parameters, and Statistics

What is the difference between a population and a sample?

Why is it important to specify clearly the population from which a sample is drawn?

What is the difference between a parameter and a statistic?

#### Measurement Scales

What types of scales are used to measure variables in the behavioral sciences?

What is the difference between qualitative and quantitative data?

#### Independent and Dependent Variables

What is the difference between observational and experimental studies?

#### Sara's Study

An example that provides a common thread tying together all of the subsequent chapters.

#### Summation Notation

Why is summation notation used by statisticians?

What are the eight rules involving summation notation?

#### Summary

#### Exercises

#### Thought Questions

#### Computer Exercises

#### Bridge to SPSS

## Why Study Statistics?

---

This book is written primarily for undergraduates majoring in psychology, sociology, and education. There are three reasons why a knowledge of statistics is essential for those who wish to pursue the study of these behavioral sciences:

***To understand the professional literature.*** Most professional literature in the behavioral sciences includes results that are based on statistical analyses. Therefore, you will be unable to understand important articles in scientific journals and books unless you understand statistics. It is possible to seek out secondhand reports that are designed for the statistically ignorant, but those who prefer this alternative to obtaining firsthand information should not be majoring in the fields of behavioral science.

***To understand the rationale underlying research in the behavioral sciences.*** Statistics is not just a catalog of procedures and formulas. It offers the rationale upon which much of behavioral science research is based—namely, drawing inferences about a population based on data obtained from a sample. Those familiar with statistics understand that research consists of a series of educated guesses and fallible decisions, not right or wrong answers. Those without a knowledge of statistics, on the other hand, cannot understand the strengths and weaknesses of the techniques used by behavioral scientists to collect information and draw conclusions.

***To carry out behavioral science research.*** In order to do competent research in the behavioral sciences, it is necessary to design the statistical analysis *before* the data are collected. Otherwise, the research procedures may be so poorly planned that not even an expert statistician can make any sense out of the results. To be sure, it is possible (and often advisable) to consult someone more experienced in statistics for assistance. Without some statistical knowledge of your own, however, you will find it difficult or impossible to convey your needs to someone else and to understand the replies.

Save for these introductory remarks, we do not regard it as our task to persuade you that statistics is important in psychology, sociology, and education. If you are seriously interested in any of these fields, you will find this out for yourself. Accordingly, this book contains no documented examples selected from the professional literature to prove to you that statistics really is used in these fields. Instead, we have devised one detailed, realistic example with numerical values that reveal the issues involved as clearly as possible.

The example we will use throughout this text is based on a hypothetical experiment performed by Sara, a first-year doctoral student. The participants are students who are enrolled in her statistics classes. We will describe this experiment in greater detail later in this chapter and return to it in all of the subsequent chapters.

We have tried to avoid a “cookbook” approach that places excessive emphasis on computational recipes. The various statistical procedures and the essential underlying concepts have been explained at length, and insofar as possible in standard English, so that you will know not only what to do but why you are doing it. Do not, however, expect to learn the material in this book from a single reading. The concepts involved in statistics, especially inferential statistics, are so challenging that it is often said that the only way to completely understand statistics is to teach it (or write a book about it). On the other hand, there is no reason to approach statistics with fear and trembling. You do not have to be a mathematical expert to obtain a good working knowledge of statistics. What *is* needed is mathematical comprehension sufficient to cope with high school-level algebra and a willingness to work at new concepts until they are understood.

## **Descriptive and Inferential Statistics**

One purpose of statistics is to summarize or describe the characteristics of a set of data in a clear and convenient fashion. This is accomplished by what are called *descriptive statistics*. For example, your grade point average serves as a convenient summary of all of the grades that you have received in college. Part I of this book is devoted to descriptive statistics.

A second function of statistics is to make possible the solution of an extremely important problem. Behavioral scientists can never measure *all* of the cases in which they are interested. For example, a clinical psychologist studying the effects of various kinds of therapies cannot obtain data on all of the mental patients in the world; a social psychologist studying gender differences in attitudes cannot measure all of the millions of men and women in the United States; an experimental psychologist cannot observe the maze behavior of all rats. Behavioral scientists want to know what is happening in a given *population*—a large group (theoretically an infinitely large group) of people, animals, objects, or responses that are alike in at least one respect (for example, all men in the United States). They cannot measure the entire population, however, because it is so large that it would be much too time consuming and expensive to do so. What to do?

One reasonable procedure is to measure a relatively small number of cases drawn from the population (that is, a *sample*). A sample of 100 people can readily be interviewed or given a written questionnaire. However, conclusions that apply only to the 100 people who happened to be included in the sample are unlikely to be of much interest. The behavioral scientist hopes to advance scientific knowledge by drawing general conclusions—for example, about the populations of men and women from which the samples of 50 men and 50 women were drawn. *Inferential statistics* makes it possible to draw inferences about what is happening in the population based on what is observed in a sample from that population. (This point will be discussed at greater length in

Chapter 9.) The remaining parts of this book are devoted to inferential statistics, which makes frequent use of some of the descriptive statistics discussed in Part I.

## Populations, Samples, Parameters, and Statistics

---

As the above discussion indicates, the term *population* as used in statistics does not necessarily refer to people. For example, the population of interest may be that of all white rats of a given genetic strain, or all responses of a single subject's eyelid in a conditioning experiment.

Whereas the population consists of all of the cases of interest, a *sample* consists of any subgroup drawn from the specified population. It is important that the population be clearly specified. For example, a group of 100 New York University freshmen might be a well-drawn sample from the population of all NYU freshmen or a poorly drawn sample from the population of all undergraduates in the United States. It is strictly proper to apply (that is, *generalize*) the research results only to the specified population. (A researcher *may* justifiably argue that her results are more widely generalizable, but she is on her own if she does so because the rules of statistical inference do not justify this.)

A *statistic* is a numerical quantity (such as an average) that summarizes some characteristic of a sample. A *parameter* is the corresponding value of that characteristic in the population. For example, if the average studying time of a sample of 100 NYU freshmen is 7.4 hours per week, then 7.4 is a statistic. If the average studying time of the population of all NYU freshmen is 9.6 hours per week, then 9.6 is the corresponding population parameter. Usually the values of population parameters are unknown because the population is too large to measure in its entirety, and appropriate techniques of inferential statistics are used to estimate the values of population parameters from sample statistics. If the sample is properly selected, the sample statistics will often give good estimates of the parameters of the population from which the sample was drawn; if the sample is poorly chosen, erroneous conclusions are likely to occur. Whether you are doing your own research or reading about that of someone else, you should always check to be sure that the population to which the results are generalized is proper in light of the sample from which the results were obtained.

## Measurement Scales

---

You may have noticed that we have used the term *data* several times without talking about where the data come from. It should come as no surprise, however, that in the behavioral sciences the data generally come from measuring

some aspect of the behavior of a human or animal. Unlike physics, in which there are quite a few important *constants* (values that are always the same, such as the speed of light or the mass of an electron), the behavioral sciences deal mainly with the measurement of *variables*, which can take on a range of different values. An additional complication faced by the behavioral scientist is that some of the variables of most interest are difficult to measure (e.g., anxiety).

### Interval Scales

How precisely you can measure a behavioral or psychological variable depends in part on the type of scale you use. The most precise scales are the kinds that are used for physical measurement. For instance, the temperature of the skin at your fingertips can be related to the amount of stress that you experience (high stress can cause the constriction of peripheral blood vessels, resulting in a decrease in skin temperature). Using either the Fahrenheit or Celsius temperature scale allows a precise measurement of skin temperature. Because degrees on either scale are fixed units that are always the same size, you can be sure that the difference between, say, 32 and 33 degrees Celsius is exactly the same as the difference between 18 and 19 degrees Celsius. These two temperature scales are therefore called *interval* scales.

### Ratio Scales

Another desirable property that scales may have is the ratio property, which requires that a measurement of zero on the scale indicates that there is really nothing left of what is being measured. If the scale has a true zero point in addition to the interval property, a measurement of 6 units, for instance, will actually indicate twice as much of what is being measured as 3 units. Therefore, such scales are called *ratio* scales. A measurement of zero degrees on the Kelvin temperature scale means that there is no temperature at all (this is absolute zero), which is not the case for the Fahrenheit or Celsius scales. Therefore, Kelvin is a ratio scale, whereas the latter two are just interval scales.

It is only the interval property that is needed for precise measurement, so it is common to make no distinction between interval and ratio scales, referring instead to interval/ratio data. We will use the less formal term *quantitative data*, as opposed to qualitative data, which we discuss next.

### Nominal Scales

The crudest form of measurement is to classify items by assigning names to them (categorization), which does not involve any numerical precision at all. Such a scale is called a *nominal* scale. For example, a person's occupation can only be "measured" on a nominal scale (e.g., accountant, lawyer, carpenter, sales clerk). We can count the number of people that fall into each category,

but (unlike in interval or ratio scales) there is no obvious order to the categories, and certainly no regular intervals between them. We will refer to such categorical data as being *qualitative*, as distinguished from quantitative.

### Ordinal Scales

Sometimes it is possible to order your categories, even though the intervals are not precise. The most common example of this is called a Likert scale (after its creator, Rensis Likert), on which respondents rate their agreement with some statement by choosing, for instance, among “strongly agree,” “agree,” “uncertain,” “disagree,” and “strongly disagree.” Because the order of the categories is clear but there is no way to be sure that they are equally spaced, this type of scale lacks the interval property and is therefore called an *ordinal scale*. Although it is a somewhat controversial practice, many behavioral researchers simply assign numbers to the categories (e.g., strongly agree is 1, agree is 2) and then treat the data as though they came from an interval/ratio scale.

Another less common way that an ordinal scale can be created is by rank ordering. It may not be possible to measure, in a precise way, the creativity of paintings produced by students in an art class, but a panel of judges could rank them from most to least creative, with perhaps a few paintings tied at the same rank.

Sometimes researchers have quantitative measurements that vary in such an odd way that it becomes more useful just to rank them and use the ranks in place of the original measurements, even though ranks are less precise. We will explain this somewhat unusual practice when we focus on statistical tests with ordinal data in Chapter 21. Until that chapter, we will deal only with purely qualitative (nominal) and purely quantitative (interval/ratio) data.

## Independent and Dependent Variables

Most behavioral research can be classified into one of the following two categories: *observational* or *experimental*.

In the simplest experiment, a researcher creates two conditions. The participants assigned to one of the conditions get some form of treatment, such as a pill intended to cure depression. Those assigned to the other condition get something that superficially resembles the treatment, such as a fake pill (placebo); they are part of a control group. These two conditions are the two different levels of an *independent* variable, or one that is created by the experimenter. Commonly, an independent variable is one whose levels are qualitative (e.g., a real pill versus a placebo). The *dependent* variable, or the variable that is measured by the experimenter and is expected to change from one level of the independent variable to another, is usually quantitative (such as a self-rating of depression). We will begin to describe such experiments in Chapter 10.

Behavioral researchers often study the relationships among variables when

it is not convenient, or even possible, to manipulate any of the variables of interest. If one simply measures the relationship between two dependent variables (e.g., self-esteem in teenagers and their family's annual income to see if those from more affluent families tend to have higher—or lower!—self-esteem), this is an observational study. We will begin to describe this kind of research, where both the independent and dependent variables may be quantitative, in Chapter 12. Research in which both the independent and dependent variables (or two dependent variables) are qualitative involves what are called *nonparametric statistical procedures* and will be discussed in Part V of this text. Part V will also cover the special case in which one or two of a researcher's variables have been measured on an ordinal scale.

## Sara's Study

---

This text covers a number of statistical procedures, some of which look very different from the others. In order to emphasize the links between these seemingly disparate statistical formulas and methods, we will refer throughout the chapters ahead to the same basic set of data collected by a hypothetical first-year doctoral student named Sara (she appears later in her career as the hypothetical researcher Dr. Tonin in *Explaining Psychological Statistics* [Cohen, 2000]). Her study, whose data are available on the Web at <http://www.psych.nyu.edu/cohen/statstext.html>, consists of both observational and experimental aspects.

The participants in Sara's study were students who attended one of two recitation classes that she conducted each week as the teaching assistant (TA) for an undergraduate class in statistics. (Of course, all of her students voluntarily signed proper informed consent forms, and her study 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 85 students who came to either Sara's morning or afternoon recitation class filled in a brief background questionnaire on which they provided contact information, some qualitative data (gender and undergrad major), some quantitative data (number of math courses already completed, latest math SAT score, and the score they received on a diagnostic math background quiz they were all required to take before registering for statistics), and some ordinal data (a rating of their math phobia on a scale from 0 to 10).

The rest of Sara's data were collected as part of an experiment that she conducted on one day in the middle of the semester. The combined results of the two class sessions on that day add up to a total of 100 students who participated in the experiment. (Due to late registration and other factors, not all of Sara's students had shown up on the first day of classes.) First, Sara explained how each student could take his or her own pulse. She then provided a one-minute interval during which they counted the number of beats and wrote

down that number as their (baseline) heart rate in beats per minute (bpm). Then each student filled out an anxiety questionnaire consisting of ten items, each rated (0 to 4) on a 5-point Likert scale. The questionnaire items inquired about anxiety and how the student was feeling at the present time (e.g., “Would you say that you are now feeling tense and restless? Circle one: Not at all; Somewhat; Moderately; Quite a bit; Extremely”). Total scores could range from 0 to 40, and provided a measure of baseline anxiety.

Next, Sara announced a pop quiz. She handed out a page containing eleven multiple-choice statistics questions on material covered during the preceding 2 weeks, and asked the students to keep this page face down while taking and recording their (prequiz) pulse and filling out an anxiety questionnaire for a second time. Then Sara 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 extra credit. Sara’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 at the end of the 15 minutes, Sara asked the students to provide heart rate and anxiety data (postquiz) one more time. Finally, Sara 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.

We will use Sara’s data set in most chapters to illustrate the main statistical procedure being taught, and in all chapters to create exercises that can be solved by a computer running statistical software. We hope that all statistics students will eventually become proficient at performing statistical procedures by computer. However, we still believe that there is important educational value in asking students to apply basic statistical formulas directly to small sets of numbers to see how the formulas work and to thus gain a greater understanding of the statistical results being generated by computer programs. The first step toward understanding the statistical formulas that will be presented in this text is to become familiar with the “workhorse” of statistics, the procedure known as *summation*.

## Summation Notation

---

Mathematical formulas and symbols often appear forbidding. In fact, when you get to know them, you will see that they are just very convenient ways to clearly and concisely convey information that would be much more awkward to express in words. In statistics, a particularly important symbol is the one used to represent the *sum* of a set of numbers—that is, the value obtained by adding up all of the numbers.

To illustrate the use of summation notation, let us suppose that eight stu-

dents take a 10-point quiz. Letting  $X$  stand for the variable in question (quiz scores), let us further suppose that the results are as follows:

$$\begin{array}{cccc} X_1 = 7 & X_2 = 9 & X_3 = 6 & X_4 = 10 \\ X_5 = 6 & X_6 = 5 & X_7 = 3 & X_8 = X_N = 4 \end{array}$$

Notice that  $X_1$  represents the first score on  $X$ ;  $X_2$  stands for the second score on  $X$ ; and so on. Also, the *number of scores* is denoted by  $N$ ; in this example,  $N = 8$ . The last score may be represented by either  $X_8$  or  $X_N$ . The *sum of all the  $X$  scores* is represented by

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

where  $\Sigma$ , the Greek capital letter sigma, stands for “the sum of” and is called the *summation sign*. The subscript below the summation sign indicates that the sum begins with the first score ( $X_i$  where  $i = 1$ ), and the superscript above the summation sign indicates that the sum continues up to and including the last score ( $X_i$  where  $i = N$  or 8). Thus,

$$\begin{aligned} \sum_{i=1}^N X_i &= X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 \\ &= 7 + 9 + 6 + 10 + 6 + 5 + 3 + 4 \\ &= 50 \end{aligned}$$

In some instances, the sum of only a subgroup of the numbers may be needed. For example, the symbol

$$\sum_{i=3}^6 X_i$$

represents the sum beginning with the third score ( $X_i$  where  $i = 3$ ) and ending with the sixth score ( $X_i$  where  $i = 6$ ). Thus,

$$\begin{aligned} \sum_{i=3}^6 X_i &= X_3 + X_4 + X_5 + X_6 \\ &= 6 + 10 + 6 + 5 \\ &= 27 \end{aligned}$$

Most of the time, however, the sum of *all* the scores is needed in the statistical analysis. In such situations it is customary to omit the indices  $i$  and  $N$  from the notation, as follows:

$$\sum X = \text{sum of all the } X \text{ scores}$$

The fact that there is no written indication as to where to begin and end the summation is taken to mean that all the  $X$  scores are to be summed.

### Summation Rules

Certain rules involving summation notation will prove useful in subsequent chapters. Let us suppose that the eight students previously mentioned take a second quiz, denoted by  $Y$ . The results of both quizzes can be summarized conveniently as follows:

Subject ( $S$ )	Quiz 1 ( $X$ )	Quiz 2 ( $Y$ )
1	7	8
2	9	6
3	6	4
4	10	10
5	6	5
6	5	10
7	3	9
8	4	8

We have already seen that  $\sum X = 50$ . The sum of the scores on the second quiz is equal to

$$\begin{aligned}\sum Y &= Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + Y_6 + Y_7 + Y_8 \\ &= 8 + 6 + 4 + 10 + 5 + 10 + 9 + 8 \\ &= 60\end{aligned}$$

The following rules are illustrated using the small set of data shown, and you should verify each one carefully.

**Rule 1.**  $\sum(X + Y) = \sum X + \sum Y$

Illustration	$S$	$X$	$Y$	$X + Y$
	1	7	8	15
	2	9	6	15
	3	6	4	10
	4	10	10	20
	5	6	5	11
	6	5	10	15
	7	3	9	12
	8	4	8	12
		$\sum X = 50$	$\sum Y = 60$	$\sum (X + Y) = 110$
		$\sum X + \sum Y = 110$		

This rule should be intuitively obvious; the same total should be reached regardless of the order in which the scores are added.

**Rule 2.**  $\sum(X - Y) = \sum X - \sum Y$

Illustration	S	X	Y	X - Y
	1	7	8	-1
	2	9	6	3
	3	6	4	2
	4	10	10	0
	5	6	5	1
	6	5	10	-5
	7	3	9	-6
	8	4	8	-4
		$\sum X = 50$	$\sum Y = 60$	$\sum(X - Y) = -10$
		$\sum X - \sum Y = -10$		

As with the first rule, it makes no difference whether you subtract first and then sum  $[\sum(X - Y)]$  or obtain the sums of  $X$  and  $Y$  first and then subtract  $(\sum X - \sum Y)$ .

Unfortunately, matters are not so simple when multiplication and squaring are involved.

**Rule 3.**  $\sum XY \neq \sum X \sum Y$

That is, first multiplying each  $X$  score by the corresponding  $Y$  score and then summing ( $\sum XY$ ) is *not* equal to summing the  $X$  scores ( $\sum X$ ) and summing the  $Y$  scores ( $\sum Y$ ) first and then multiplying once ( $\sum X \sum Y$ ).

Illustration	S	X	Y	XY
	1	7	8	56
	2	9	6	54
	3	6	4	24
	4	10	10	100
	5	6	5	30
	6	5	10	50
	7	3	9	27
	8	4	8	32
		$\sum X = 50$	$\sum Y = 60$	$\sum XY = 373$
		$\sum X \sum Y = (50)(60) = 3,000$		

Observe that  $\sum XY = 373$ , while  $\sum X \sum Y = 3,000$ .

**Rule 4.**  $\sum X^2 \neq (\sum X)^2$

That is, first squaring all of the  $X$  values and then summing ( $\sum X^2$ ) is *not* equal to summing first and then squaring a single quantity  $[(\sum X)^2]$ .

Illustration	$S$	$X$	$X^2$
	1	7	49
	2	9	81
	3	6	36
	4	10	100
	5	6	36
	6	5	25
	7	3	9
	8	4	16
		$\sum X = 50$	$\sum X^2 = 352$
		$(\sum X)^2 = (50)^2 = 2,500$	

Here,  $\sum X^2 = 352$ , while  $(\sum X)^2 = 2,500$ .

**Rule 5.** If  $k$  is a *constant* (a fixed numerical value), then

$$\sum k = Nk$$

Illustration	Suppose that $k = 3$ . Then,
$S$	$k$
1	3
2	3
3	3
4	3
5	3
6	3
7	3
8	3
	$\sum k = 24$
	$Nk = (8)(3) = 24$

**Rule 6. If  $k$  is a constant,**

$$\sum(X + k) = \sum X + \sum k = \sum X + Nk$$

$S$	Illustration $X$	Suppose that $k = 5$ . Then, $k$	$X + k$
1	7	5	12
2	9	5	14
3	6	5	11
4	10	5	15
5	6	5	11
6	5	5	10
7	3	5	8
8	4	5	9
	$\sum X = 50$	$\sum k = Nk = 40$	$\sum(X + k) = 90$
	$\sum X + Nk = 50 + (8)(5) = 90$		

This rule follows directly from Rules 1 and 5.

**Rule 7. If  $k$  is a constant,**

$$\sum(X - k) = \sum X - Nk$$

The illustration of this rule is similar to that of Rule 6 and is left to the reader as an exercise.

**Rule 8. If  $k$  is a constant,**

$$\sum kX = k \sum X$$

$S$	Illustration $X$	Suppose that $k = 2$ . Then, $k$	$kX$
1	7	2	14
2	9	2	18
3	6	2	12
4	10	2	20
5	6	2	12
6	5	2	10
7	3	2	6
8	4	2	8
	$\sum X = 50$		$\sum kX = 100$
	$k \sum X = (2)(50) = 100$		

## Summary

*Descriptive statistics* are used to summarize and make understandable large quantities of data. *Inferential statistics* are used to draw inferences about numerical quantities (called *parameters*) concerning *populations* based on numerical quantities (called *statistics*) obtained from *samples*. Some behavioral variables cannot be measured precisely (e.g., religion) but can only be measured qualitatively using categories (e.g., Protestant, Catholic, Jewish), which comprise a *nominal* scale. If the categories can be placed in order (e.g., the belts awarded for different levels of skill in the martial arts—black belt, brown belt, etc.), an *ordinal* scale has been created. If the scale involves precise measurement resulting in units of equal size, the data are considered to be *quantitative*, whether the scale has a true zero point (*ratio* scale) or not (*interval* scale). Experiments involve measuring dependent variables that are expected to vary somewhat as a function of the different levels of one or more independent variables created by the researcher. *Observational* research involves comparing dependent variables to each other since we are not manipulating variables.

The summation sign,  $\Sigma$ , is used to indicate “the sum of” and occurs frequently in statistical work. Remember that  $\Sigma X$  is a shorthand version of

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

(where  $N$  = number of subjects or cases).

Summation Rules:

1.  $\Sigma(X + Y) = \Sigma X + \Sigma Y$
2.  $\Sigma(X - Y) = \Sigma X - \Sigma Y$
3.  $\Sigma XY$  (multiply first, then add)  $\neq \Sigma X \Sigma Y$  (add first, then multiply)
4.  $\Sigma X^2$  (square first, then add)  $\neq (\Sigma X)^2$  (add first, then square)

If  $k$  is a constant,

5.  $\Sigma k = Nk$
6.  $\Sigma(X + k) = \Sigma X + Nk$
7.  $\Sigma(X - k) = \Sigma X - Nk$
8.  $\Sigma kX = k \Sigma X$

## Exercises

The exercises in this section refer to the following set of data.<sup>1</sup>

### Hypothetical Scores on a 20-Point Psychology Test for Students Drawn at Random from Four Universities

University A ( $N = 50$ )	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	18	12	13	12	17	8	16	12	15
University B ( $N = 50$ )	17	8	12	12	3	12	7	14	1	11
	12	11	9	14	10	13	7	13	8	12
	9	12	17	11	6	10	10	3	9	8
	6	13	5	16	10	9	19	5	12	10
	16	11	14	11	13	12	2	17	10	14
University C ( $N = 10$ )	9	11	6	5	4	9	0	4	5	7
University D ( $N = 5$ )	14	8	17	6	10					

- Express the following words in symbols.
  - Add up all the scores on test  $X$ , then add up all the scores on test  $Y$ , and then add the two sums together.
  - Add up all the scores on test  $G$ . To this, add the following: the sum obtained by squaring all the scores on test  $P$  and then adding them up.
  - Square all the scores on test  $X$ . Add them up. From this, subtract 6 times the sum you get when you multiply each score on  $X$  by the corresponding score on  $Y$  and add them up. To this, add 4 times the quantity obtained by adding up all the scores on test  $X$  and squaring the result. To this, add twice the sum obtained by squaring each  $Y$  score and then adding them up. (Compare the amount of space needed to express this equation in words with the amount of space needed to express it in symbols. Do you see why summation notation is necessary?)
- Five students are enrolled in an advanced course in psychology. Two quizzes are given early in the semester, each worth a total of 10 points. The results are as follows:

Student	Quiz 1 ( $X$ )	Quiz 2 ( $Y$ )
1	0	2
2	2	6
3	1	7
4	3	6
5	4	9

<sup>1</sup> These data, the problems in this section, and most of the problems in subsequent Exercises sections come from Ewen (2000).

- (a) Compute each of the following:

$$\begin{array}{lll} \Sigma X = \underline{\hspace{2cm}} & (\Sigma X)^2 = \underline{\hspace{2cm}} & \Sigma(X - Y) = \underline{\hspace{2cm}} \\ \Sigma Y = \underline{\hspace{2cm}} & (\Sigma Y)^2 = \underline{\hspace{2cm}} & \Sigma X - \Sigma Y = \underline{\hspace{2cm}} \\ \Sigma X^2 = \underline{\hspace{2cm}} & \Sigma(X - Y) = \underline{\hspace{2cm}} & \Sigma XY = \underline{\hspace{2cm}} \\ \Sigma Y^2 = \underline{\hspace{2cm}} & \Sigma X + \Sigma Y = \underline{\hspace{2cm}} & \Sigma X \Sigma Y = \underline{\hspace{2cm}} \\ \Sigma_{i=1}^3 X_i = \underline{\hspace{2cm}} & \Sigma_{i=2}^5 Y_i = \underline{\hspace{2cm}} & \Sigma_{i=2}^4 X_i Y_i = \underline{\hspace{2cm}} \end{array}$$

- (b) Using the results of part (a), show that each of the following rules listed in this chapter is true:

Rule 1:  $\underline{\hspace{2cm}} = \underline{\hspace{2cm}}$

Rule 2:  $\underline{\hspace{2cm}} = \underline{\hspace{2cm}}$

Rule 3:  $\underline{\hspace{2cm}} \neq \underline{\hspace{2cm}}$

Rule 4:  $\underline{\hspace{2cm}} \neq \underline{\hspace{2cm}}$  ( $X$  data)

$\underline{\hspace{2cm}} \neq \underline{\hspace{2cm}}$  ( $Y$  data)

- (c) After some consideration, the instructor decides that Quiz 1 was excessively difficult and decides to add 4 points to each student's score. This can be represented in symbols by using
- $k$
- to stand for the constant amount in question, 4 points.

Using Rule 6, compute  $\Sigma(X + k) = \underline{\hspace{2cm}} + \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$ .

Compute  $\Sigma X + k = \underline{\hspace{2cm}} + \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$ . (Note that this result is different from the preceding one.)

Now add 4 points to each student's score on Quiz 1 and obtain the sum of these new scores.

- (d) Had the instructor been particularly uncharitable, he might have decided that Quiz 2 was too easy and subtracted 3 points from each student's score on that quiz. Although this is a new problem, the letter
- $k$
- can again be used to represent the constant; here,
- $k = 3$
- .

Using Rule 7, compute  $\Sigma(Y - k) = \underline{\hspace{2cm}} - \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$ .

Compute  $\Sigma Y - k = \underline{\hspace{2cm}} - \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$ . (Note that this result is different from the preceding one.)

Now subtract 3 points from each student's score on Quiz 2 and obtain the sum of these new scores.

- (e) Suppose that the instructor decides to double all of the original scores on Quiz 1.

Using Rule 8, compute  $\Sigma kX = \underline{\hspace{2cm}} \cdot \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$ .

Now double each student's score on Quiz 1 and obtain the sum of these new scores.

3. Compute the following:

- (a) For University C:

$\Sigma X = \underline{\hspace{2cm}} \quad \Sigma X^2 = \underline{\hspace{2cm}} \quad (\Sigma X)^2 = \underline{\hspace{2cm}}$

(b) For University D:

$$\sum X = \underline{\hspace{2cm}} \quad \sum X^2 = \underline{\hspace{2cm}} \quad (\sum X)^2 = \underline{\hspace{2cm}}$$

If you would like some additional practice, you may verify that

$$\text{For University A, } \sum X = 617; \sum X^2 = 8385; (\sum X)^2 = 380,689$$

$$\text{For University B, } \sum X = 526; \sum X^2 = 6316; (\sum X)^2 = 276,676$$

4. For each of the following (separate) sets of data, compute the values needed in order to fill in the answer spaces. Then answer the additional questions that follow.

*Data set 1:*

S	X	Y	N =	_____
1	1	2	$\sum X =$	_____ $\sum Y =$ _____
2	3	5	$\sum X^2 =$	_____ $\sum Y^2 =$ _____
3	1	0	$(\sum X)^2 =$	_____ $(\sum Y)^2 =$ _____
4	0	1	$\sum XY =$	_____ $\sum X \sum Y =$ _____
5	2	3	$\sum(X + Y) =$	_____ $\sum(X - Y) =$ _____

*Data set 2:*

S	X	Y	N =	_____
1	7.14	0	$\sum X =$	_____ $\sum Y =$ _____
2	8.00	2.60	$\sum X^2 =$	_____ $\sum Y^2 =$ _____
3	0	4.32	$(\sum X)^2 =$	_____ $(\sum Y)^2 =$ _____
4	4.00	2.00	$\sum XY =$	_____ $\sum X \sum Y =$ _____
5	4.00	6.00	$\sum(X + Y) =$	_____ $\sum(X - Y) =$ _____
6	1.00	1.15		
7	2.25	1.00		
8	10.00	3.00		

If every X score is multiplied by 3.2,  
what is the new  $\sum X$  in each set?

*set 1*      *set 2*

\_\_\_\_\_      \_\_\_\_\_

If 7 is subtracted from every Y score,  
what is the new  $\sum Y$  in each set?

\_\_\_\_\_      \_\_\_\_\_

If 1.8 is added to every X score,  
what is the new  $\sum X$  in each set?

\_\_\_\_\_      \_\_\_\_\_

If every Y score is divided by 4,  
what is the new  $\sum Y$  in each set?

\_\_\_\_\_      \_\_\_\_\_

(Hint: Use the appropriate summation rule in each case so as to make the calculations easier.)

5. Compute the values needed to fill in the blanks.

*Data set 3:*

S	X	Y	$N =$ _____	
1	97	89	$\sum X =$ _____	$\sum Y =$ _____
2	68	57	$\sum X^2 =$ _____	$\sum Y^2 =$ _____
3	85	87	$(\sum X)^2 =$ _____	$(\sum Y)^2 =$ _____
4	74	76	$\sum XY =$ _____	$\sum X \sum Y =$ _____
5	92	97	$\sum(X + Y) =$ _____	$\sum(X - Y) =$ _____
6	92	79		
7	100	91		
8	63	50		
9	85	85		
10	87	84		
11	81	91		
12	93	91		
13	77	75		
14	82	77		

## Thought Questions

1. What is the difference between (a) a population and a sample? (b) a parameter and a statistic? (c) descriptive statistics and inferential statistics?
2. What is the difference between a constant and a variable?
3. What is the difference between a ratio scale and an interval scale?
4. What important property do interval scales have that ordinal and nominal scales do *not* have?
5. If we want to know how many people have each of five different psychological disorders—Major Depressive Disorder, Bipolar Disorder, Generalized Anxiety Disorder, Obsessive-Compulsive Disorder, and Phobic Disorder—what kind of measurement scale are we using?
6. A poll of sportswriters ranks the 25 best college football teams in the country, where #1 is the best team, #2 is the second best team, and so on. What kind of measurement scale is this?
7. What kind of measurement scales are used for each of the following variables in Sara's study? (a) The gender of each student. (b) The undergraduate major of each student. (c) The number of math courses each student has

already completed. (d) The latest SAT score of each student. (e) The ratings of the math phobia of each student.

8. What is the difference between an independent variable and a dependent variable?

## Computer Exercises

1. Read Sara's data into your statistical software package. For those not using the Statistical Package for the Social Sciences (SPSS), we have provided the data in two convenient formats: a tab-delimited text file, and an Excel spreadsheet (Microsoft Office 95). For the convenience of SPSS users, we have also included the data as an SPSS.sav file, though your instructor may want you to know how to read text or Excel files into SPSS.
2. Label the values of the categorical (i.e., qualitative) variables according to the following codes: For gender, 1 = Female and 2 = Male; for undergrad major, 1 = Psychology, 2 = Pre-med, 3 = Biology, 4 = Sociology, and 5 = Economics. Your instructor may ask you to fill in missing-value codes for any data that are missing (e.g., blank cell in the Excel spreadsheet).
3. A good many statistical functions can be performed in Excel. As a first step, use the Sum function to add up the scores for each of the quantitative variables in the Excel file of Sara's data.
4. Create a new variable that adds 10 points to everyone's math background quiz score. How does the sum of this variable compare to the sum of the original variable? What general rule is being illustrated by this comparison?
5. Create a new variable that is 10 times the statistics quiz score. How does the sum of this variable compare to the sum of the original variable? What general rule is being illustrated by this comparison?

## Bridge to SPSS

SPSS is certainly not the only "statpack" available, and your instructor may prefer to teach you how to use SAS, Minitab, or another of SPSS's many competitors. However, because SPSS for Windows is the most popular statpack in the social and behavioral sciences, and is particularly easy to learn for introductory statistics students, we provide this section after the computer exercises in each chapter. (This should help students translate the terminology used by SPSS into the language that we are using to describe the same topics in this text.) We also show you briefly how to obtain from SPSS the statistics discussed in the chapter, along with a few tricks and shortcuts. However, we understand

that even though you may be using SPSS in your statistics course, you may not be using the latest version, which will probably be 14.0 by the time this text is printed. Therefore, we only describe aspects of SPSS that have not changed between version 10, released in the year 2000, and version 14.

The rightmost column in SPSS's Variable View spreadsheet is labeled **Measure**, and it allows you to classify each of your variables as being measured by one of the following three types of scales: nominal, ordinal, or scale. The first two terms are used the same way by SPSS that we have defined them in this chapter. *Scale* is SPSS's term for interval/ratio data. In general, numerical data are set to **scale** by default, whereas string data—which contain letters instead of, or in addition to, numbers—are set to **nominal**. In practice, these scale designations are not very important, because SPSS uses them only to determine the way some charts are displayed.

For simplicity, Sara's data set is presented entirely in terms of numbers, even for the categorical variables of gender and undergrad major. To assign meaningful labels to the arbitrary numbers we have used to represent the different levels of the categorical variables, go down the Values column of Variable View until you reach the row for a categorical variable. Then click in the right side of that cell to open the Value Labels box. For gender, you would type **1** for Value, and then tab to Value Label, where you can type **female**. Click the **Add** button, and provide a label (male) for value **2**, then **Add** again, and **OK**. The process is similar for undergrad major. Note that if you use the Missing column to define a particular value of a variable, say 99, as meaning that the value is missing, rather than just leaving the cell blank (e.g., you could use 99 to mean "missing" for the math background quiz, because none of the real values can be that high), you can then attach a value label to that value, such as "never took the math quiz."

To create new variables that are based on ones already in your spreadsheet, click on the **Transform** menu, then **Compute**. In the Compute Variable box that opens up, Target Variable is a name that you make up (and type into that box) for the new variable (but no more than eight characters and no embedded spaces); when you have filled in a Numeric Expression and then click **OK**, the new variable will suddenly appear in the rightmost column of your Data View spreadsheet. We will leave it to your instructor, or an SPSS guide book, to teach you various ways to create Numeric Expressions that transform your existing variables into new ones.