CHAPTER 1 A REVIEW OF SOME ELEMENTARY STATISTICAL CONCEPTS

Chapter Objectives

This chapter reviews elementary statistical concepts that are important for understanding intermediate methods, such as linear regression.

In chapter 1, we will

- Briefly discuss descriptive statistics, such as measures of central tendency and dispersion.
- Link the normal distribution and statistical methods.
- Discuss important concepts for linear regression, such as statistical inference, the use of samples, and variable standardization.
- Reconsider inferential methods, such as the *t* test.
- Run some basic statistical analyses in SPSS and Stata.

Elementary statistics can be confusing, especially to people who are uncomfortable with numbers. Many of us were first introduced to statistics in a pre-algebra or algebra course. However, your initial introduction probably occurred in elementary school. Do you remember the first time you heard the word *mean* used to indicate the average of a set of numbers? This likely took place in some math class very early on in your education. How about graphing exercises? Do you recall being given two sets of points and being asked to plot them on graph paper? You were introduced to the *x*-axis and the *y*-axis, or the coordinate axes.

Around the same time, or perhaps a little later, you became familiar with elementary *probability*. This likely took the form of a question such as, "What is the probability of a die being thrown and landing on a five?" You first learned that you needed to count the number of possible outcomes (there are six faces on a typical die, so there are six possible outcomes). This was the denominator. Then you counted the particular

outcome. This was the numerator. Putting these two counts together, you learned that the probability of the roll coming up as five is 1/6, or approximately 0.167. This latter value is known as a *proportion*. Proportions—and probabilities—must fall between zero and one. They can easily be transformed into percentages by moving the decimal place over two spaces to the right (or multiplying by 100) and placing a percentage sign next to the number. What does this mean, though? Well, one way to consider it is to say that we expect a five to come up about 16.7 percent of the time when we roll a typical die numerous times. Of course, you can probably confirm this by rolling the die many, many times. Some statisticians refer to such a view as a *frequentist* interpretation of or approach to statistics.

Probabilities are normally presented using, not surprisingly, the letter *P*. One way to represent the probability of a five from a roll of a die is with $P(5)$. So we may write $P(5) = 0.167$ or $P(5) = 1/6$. You might recall that some statistical tests, such as *t* tests (see the description later in the chapter) or analyses of variance (ANOVAs), are often accompanied by *p* values. As we shall learn, *p* values are a type of probability value used in many statistical tests.

By combining the principles of probability and elementary statistical concepts, we may develop the basic foundations for statistical analysis. In general, there are two types of statistical analyses: *descriptive* and *inferential*. The former set of methods normally is used to describe or summarize one or more variables (recall that the term *variables* is used to indicate phenomena that can take on more than one value; this contrasts with *constants*, or phenomena that take on only one value). Some common terms that you are probably familiar with are *measures of central tendency* and *measures of dispersion*. We will see several of these measures a little later. Then there are the many graphical techniques that may be used to "see" the variable. You might recognize techniques such as histograms, stem-and-leaf plots, dot plots, and box-and-whisker plots.

Inferential statistics are designed to infer or deduce something about a population from a sample. Suppose, for instance, that we are interested in determining who is likely to win the next presidential election in the United States. We will assume there are only two candidates from which to choose: Clinton and Rice. Of course, it would be enormously expensive to ask each person who is likely to vote in the next election his or her choice of president. Therefore, we may take a sample of likely voters and ask them for whom they plan to vote. Can we infer anything about the population of voters on the basis of our sample? The answer is that it depends on a number of factors. Did we collect a good sample? Were the people who responded honest? Do people change their minds as the election approaches? We do not have time to get into the many issues involved in sampling and survey responses, so we will have to assume that our sample is a good representation of the population from which it is drawn and that people are generally honest in their responses to our inquiries. Most important for our purposes

is this: Inferential statistics include a set of techniques designed to help us answer questions about a population from a sample.

Another way of dividing up statistics is to compare techniques that deal with one variable from those that deal with two or more variables. Most readers of this presentation will likely be familiar with techniques designed for one variable. These include, as we shall see later, most of the descriptive statistical methods. The bulk of this presentation, at least in later chapters, concerns a technique designed for analyzing two or more variables simultaneously. A key question that motivates us is whether two or more variables are associated in some way. As the values of one variable increase, do the values of the other variable also tend to increase? Or do they decrease? In elementary statistics, students are introduced to covariances and correlations, two techniques designed to answer these questions generally. However, recall that you are not necessarily saying that one variable somehow changes another variable. Remember the maxim, *correlation does not equal causation*? We will try to avoid the term *causation* in this presentation because it involves many thorny philosophical issues (see Pearl, 2000). Nonetheless, one of our main concerns is whether one or more variables are associated with another variable in a systematic way. Determining the characteristics of this association is one of the main goals of the linear regression model that we shall learn about later.

Measures of Central Tendency and Dispersion

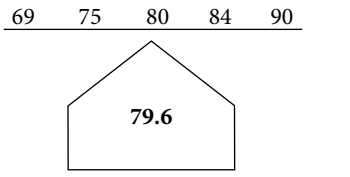
Means

Now that we have some background information on elementary statistics, let's learn more about the most important measures, including how they are used and how they are computed. We will begin with measures of central tendency. Suppose we have collected data on a variable such as weight in kilograms. Our intrepid researchers have carefully placed each person in the sample on a scale and recorded their body weights. To simplify things, we will assume the researchers rounded the weights to the nearest kilogram. What would be your best guess of the average weight among the sample? It is not always the best, but the most frequent measure is the *arithmetic mean*, which is computed using the following formula:

$$E[X] = \overline{x} = \sum x_i / n$$

The term on the left-hand side of the equation is $E[X]$. This is a shorthand way of saying that this is the *expected value* of the variable X. It is often used to represent the mean. To be more precise, we might also list this term as E [weight in kg], but usually, as long as it is clear that X = weight in kilograms, using $E[X]$ is sufficient. The middle term—read as *x-bar*—may also be familiar as a common symbol for the mean.

Figure 1.1: What Is a Mean?

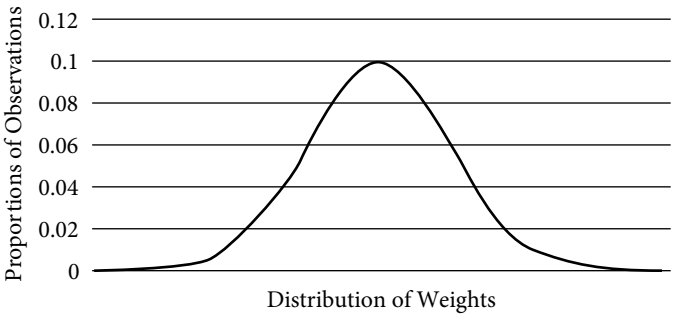


The formula for computing the mean is rather simple. We add all the values of the variable and divide this sum by the number of observations. Note that the rather cumbersome symbol that looks like an overgrown *E* in the right-hand part of the equation is the summation sign; it tells us to add whatever is next to it. The symbol x_i denotes the specific values of the *x* variable, or the individual weights we have measured. The subscript *i* indicates each observation. The symbol *n* represents the sample size. Sometimes the individual observations are represented as *i* . . . *n*. If you know that $n = 5$, then you know there are five individual observations in your sample. In statistics, we often use uppercase roman letters to represent population values and lowercase roman letters to represent sample values. Therefore, when we say $E[X] = \overline{x}$, we are implying that our sample mean estimates the population expected value, or the population mean.

Here is a simple example: We have a sample of people's weights (in kilograms) that consist of the following set: [84, 75, 80, 69, 90]. The sum of this set is [84 + 75 + 80 + 69 + 90] = 398; therefore, the mean is 398/5 = 79.6. Another way of thinking about this mean value is that it represents the center of gravity. Suppose we have a plank of wood that is magically weightless (or of uniform weight across its span). We order the people from lightest to heaviest—trying to space them out proportionally to their weights—and ask them to sit on our plank of wood. The mean is the point of balance, or the place at which we would place a fulcrum underneath to balance the people on the plank (as illustrated in Figure 1.1).

There are some additional things you should know about the mean. First, it is measured in the same units as the observations. If your observations are not all measured in the same unit (for example, some people's weights are in kilograms, others in pounds), then the mean cannot be interpreted. Second, the mean provides a good measure of central tendency if the variable is measured continuously and is normally distributed. What do these terms mean? A variable is measured continuously—or we say the variable is *continuous*—if it can conceivably take on any real number. Of course, we usually cannot be so precise when we measure things, so it is not uncommon to round our measures to whole numbers or integers. We also often measure things using positive numbers only; it makes little sense, for instance, to try to measure a person's weight using negative numbers. The other type of variable is known as *discrete* or *categorical*; these variables

Figure 1.2: The Normal Distribution



have a finite number of possible values. For example, we normally use only two categories to measure gender: female and male. Hence, this is a discrete variable.

Normally Distributed Data

We say a variable is normally distributed if it follows a bell-shaped curve. Order the values of the variable from lowest to highest and then plot them by their frequencies or the percentage of observations that have a particular value (we must assume that there are many values of our variable). We may then view the "shape" of the distribution of the variable. Figure 1.2 shows an example of a bell-shaped distribution, usually termed a *normal* or *Gaussian distribution*, using a simulated sample of weights. (It is known as Gaussian after the famous German mathematician Carl Friedrich Gauss, who purportedly discovered it.)

We will return to means and the normal distribution frequently here and throughout the book. To give you a hint of what is to come, the linear regression model is designed, in part, to predict means for particular sets of observations in the sample. For instance, if we have information on the heights of our sample members, we may wish to use this information to predict their weights. Our predictions could include predicting the mean weight of people who are 72 cm tall. We may use a linear regression model to do this.

However, suppose that our variable does not follow a normal distribution. May we still use the mean to represent the average value? The simple answer is yes, as long as the distribution does not deviate too far from the normal distribution. In many situations in the social and behavioral sciences, though, variables do not have normal distributions. A good example of this involves annual income. When we ask a sample of people about their annual incomes, we usually find that a few people earn a lot more than others. Measures of income typically provide *skewed* distributions, with long right tails. If asked to find a good measure of central tendency for income, there are several solutions

available. First, we may take the *natural* (*Naperian*) *logarithm* of the variable. You might recall from an earlier mathematics course that using the natural logarithm (or, as an alternative, the base 10 logarithm) pulls in extreme values. If this is not clear, try taking your calculator and using the LN function with some large and small values (for example, 10 and 1,500). You will easily see the effect this has on the values of a variable. If you are lucky, you may find that taking the natural logarithm of a variable with a long right tail transforms it into a normal distribution. The square root or cube root of a variable may also work to "normalize" a skewed distribution. We will see examples of this in chapter 10.

Medians

Second, there are several direct measures of central tendency appropriate for skewed distributions (or other distributions plagued by extreme values such as outliers; see chapter 12), such as the *trimmed mean* and the *median*. The trimmed mean cuts off some percentage of values from the upper and lower ends of the distribution, usually 5 percent, and uses the remaining values to compute the mean value. The median should be familiar to you. It is the middle value of the distribution. To find it, we first order the values of the variable from lowest to highest. Then we choose either the middle value if there is an odd number of observations or the average of the middle two values if there is an even number of observations. If you are familiar with percentiles (or quartiles or deciles), then you might recall that the median is the 50th percentile of a distribution. The median is known as a *robust* statistic because it is relatively unaffected by extreme values. As an example, suppose we have two variables, one that follows a normal distribution (or something close to it) and another that has an extreme value:

Variable 1: [45, 50, 55, 60, 65, 70, 75]
Variable 2: [46, 51, 54, 59, 66, 71, 375]

Variable 1 has a mean of 60 and a median of 60, so we make the same estimate of its central value regardless of which measure is used (the mean and median being equal). In contrast, Variable 2 has a mean of 103 but a median of 59. Although we might debate the point, we think most people would agree that the median is a better representative of the average value than the mean for Variable 2.

Standard Deviations

The next issue to address from elementary statistics involves measures of dispersion. As the name implies, these measures consider the spread of the distribution of a variable. Most readers are familiar with the term *standard deviation*, as it is the most common measure for continuous variables. However, before seeing the formula for the standard

deviation, it is useful to consider some other measures of dispersion. The most basic measure is the *sum of squares*, or SS[*X*]:

$$\mathrm{SS}[X] = \sum(x_1 - \overline{x})^2$$

This formula first computes *deviations from the mean* $(x_i - \overline{x})$, squares each one, and adds them up. If you have learned about ANOVA models, the sum of squares should be familiar. Perhaps you even recall that there are various forms of the sum of squares. We will learn more about these in chapter 4.

A second measure of dispersion that may be more familiar to you is the *variance*, or Var[X]. It is often labeled as s^2. The formula is:

$$\mathrm{Var}[\mathrm{X}] = s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

Notice that another way of computing the variance is to take the sum of squares and divide it by the sample size minus one. One of the drawbacks of the variance is that it is measured not in units of the variable but rather in squared units of the variable. To transform it into the same units as the variable, it is, therefore, a simple matter to take the square root of the variance. This measure is the standard deviation (which is why the standard deviation is noted by the letter *s*, while the variance is labeled as s^2):

$$SD[\mathrm{X}] = s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n^{-1}}}$$

A variable's distribution, assuming it is normal, is often represented by its mean and standard deviation. In shorthand form, this is listed as $x \sim N(\overline{x}, s)$ (the wavy line means "distributed as"). Obviously, a variable that is measured in the same units as another and that shares the same mean is less dispersed if its standard deviation is smaller. Although not often used, another promising measure of dispersion is the *coefficient of variation* (CV), which is computed by dividing the standard deviation by the mean of the variable $(s/\overline{x})$. It is often multiplied by 100. The CV is valuable because it shows how much a variable varies about its mean.

An important point to remember is that symbols such as *s* and s^2 are used to represent sample statistics. Greek symbols—or, as we have seen up until this point, uppercase roman letters—are often used to represent population statistics. For example, the symbol for the population mean is the Greek letter *mu* (μ), whereas the symbol for the population standard deviation is the Greek letter *sigma* (σ). However, we will see when we get into the symbology of linear regression that Greek letters are used to represent other quantities as well.

Standard Error

Another useful measure of dispersion or variability refers not to the variable directly but rather to its mean. When we compute the variance or the standard deviation, we are concerned with the spread of the distribution of the variable. Typically, we have only one sample from a population—but many, many samples are possible. Let's imagine that we did take many, many samples from a population and compute a mean for each sample. We would end up with a sample of means from the population rather than simply a sample of observations. We could then compute a mean of these means, or an overall mean, which should pretty accurately reflect—assuming we do a good job of drawing the samples—the actual mean of the population of observations. Nonetheless, these numerous means will also have a distribution. It is possible to plot these means to see whether they follow a normal distribution. In fact, an important theorem from mathematical statistics states that the sample means follow a normal distribution even if they come from a nonnormally distributed variable in the population (see chapter 3). This is a very valuable finding because it allows us to make important claims about the linear regression model. We will learn about these claims in later chapters.

Our concern here is not whether the distribution of sample means is normally distributed, at least not directly. Rather, we need to consider a measure of the dispersion of these means. Statistical theory suggests that a good measure (estimate) of dispersion is the *standard error* of the mean. Why should we use the standard error? Consider the following example. In 2012, it was estimated that 42.2 million Americans, or about 15 percent of the U.S. population, used food stamps. Imagine that we wanted to know how children and adults using food stamps were similar or different from one another in terms of physical health. It is simply not practical to get information from every family using food stamps. Instead, we need to sample from this group. If we took a sample of 10,000 people from the population of 42.2 million, there would be literally trillions of possible combinations of respondents in our data. Each possible sample would be a little different from the next one. The standard error tells us how much difference we can expect between our sample and the true population mean. Think of it as a way to account for not having the full population of food stamp recipients.

The standard error is computed using the sample standard deviation as:

$$SE(\text{mean}) = \frac{s}{\sqrt{n}}$$

Standard errors are very important in linear regression analysis. Later, we will discuss another type of standard error—known as the standard error of the slope coefficient—which we use to make inferences about the regression model.

Standardizing a Variable

One of the difficult issues when we are presented with different continuous variables is that they are rarely measured in the same units. Education is measured in years, income is measured in dollars, body weight is measured in pounds or kilograms, and food intake is measured in kilocalories. It is convenient to have a way to adjust variables so their measurement units are similar. You might recall that this is one of the purposes of z scores. Assuming that we have a normally distributed set of variables, we may transform them into z scores so they are comparable. A z-score transformation uses the following formula:

$$z\,\text{score} = \frac{(x_i - \overline{x})}{s}$$

Each observation of a variable is put through this formula to yield z scores, or what are commonly known as *standardized values*. The unit of measurement for z scores is standard deviations. The mean of a set of z scores is zero, whereas its standard deviation is one.

You may remember that z scores are used to determine what percentage of a distribution falls a certain distance from the mean. For example, 95 percent of the observations from a normal distribution fall within 1.96 standard deviations of the mean. This translates into 95 percent of the observations using standardized values falling within 1.96 z scores of the mean. With a slight modification, this phenomenon is helpful when we wish to make inferences from the results of the linear regression model to the population. The plot of z scores from a normally distributed variable is known as the *standard normal distribution*. As mentioned earlier, one of the principal advantages of computing z scores is that they provide a tool for comparing variables that are measured in different units; this will come in handy as we learn about linear regression models. Of course, we must be intuitively familiar with standard deviations to be able to make these comparisons.

Covariance and Correlation

Our next task involves moving from a single variable to two variables. An important use of statistics is to consider the association or relationship between two variables. As mentioned earlier, an interesting question we might ask is whether two variables shift or change together. To give an obvious and not-very-interesting example, is it fair to say that height and weight shift together? Are taller people, on average, heavier than shorter

people? The answer, again on average, is most certainly yes. In statistical language, we say that height and weight *covary* or are *correlated*. The two measures most commonly used to assess the association between two continuous variables are, not surprisingly, the *covariance* and the *correlation*. To be precise, the correlation used most often is the Pearson's product-moment correlation (there are actually many types of correlations; the type attributed to the statistician Karl Pearson is the most common).

A covariance is a measure of the joint variation of two continuous variables. In less technical terms, we claim that two variables covary when there is a high probability that large values of one are accompanied by large or small values of the other. For instance, height and weight covary because large values of one tend to accompany large values of the other in a population or in most samples. This is not a perfect association because there is clearly substantial variation in heights and weights among people. The equation for the covariance is:

$$\text{Cov}\ (x, y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

The equation computes deviations from the means of both variables, multiplies them, adds up these products for each observation, and then divides this sum by the sample size minus one. Don't forget that this implies that the *x*s and *y*s come from the same unit, whether it is a person, place, or thing.

One of the problems with the covariance is that it depends on the measurement scheme of both variables. It would be helpful to have a measure of association that did not depend on these measurement units but rather offered a way to compare various associations of different sets of variables. The correlation coefficient accomplishes this task. Among the several formulas we might use to compute the correlation, the following equations are perhaps the most intuitive:

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}[x] \times \text{Var}[y]}}$$

$$\text{Corr}(x, y) = \frac{\sum (z_x)(z_y)}{n-1}$$

The first equation shows that the correlation is simply the covariance divided by a joint measure of variability: the variances of each variable multiplied, with the square root of this quantity representing what we might call the joint or pooled standard deviation. The second equation shows the relationship between *z* scores and correlations. We might even say, without violating too many tenets of the statistical community, that the correlation is a standardized measure of association.

A couple of interesting properties of correlations are, first, that they always fall between −1 and +1, with positive numbers indicating a positive association and negative numbers indicating a negative association (as one variable increases, the other tends to decrease). A correlation of zero means there is no statistical association, at least not one that can be measured assuming a straight line association, between the two variables. Second, the correlation is unchanged if we add a constant to the values of the variables or if we multiple the values by some constant number. However, these constants must have the same sign, negative or positive.

As mentioned earlier, there are several other types of correlations (or what we refer to generally as *measures of association*) in addition to Pearson's. For instance, a Spearman's correlation is based on the ranks of the values of variables rather than the actual values. Similar to the median when compared to the mean, it is less sensitive to extreme values. There are also various measures of association designed for categorical variables, such as gamma, Cramer's *V*, lambda, eta, and odds ratios. Odds ratios, in particular, are popular for estimating the association between two binary (two-category) variables. Odds ratios are discussed in much greater detail in chapter 13.

Comparing Means from Two Samples

Another important topic that we will discuss before moving into linear regression analysis involves comparing means from two distributions. Of course, we may compare many statistics from distributions, including standard deviations, correlations, and standard errors, but an important issue in applied statistics is determining whether the mean from one sample (or subsample) is different from the mean of another sample (or subsample). For example, I may wish to know whether the mean income of adults from Salt Lake City, UT, is higher than the mean income of adults from Seattle, WA. If I have good samples of adults from these two cities. I can consider a couple of things. First, I can take the difference between the means. Let's say that the average annual income among a Salt Lake City sample is $35,000 and the average annual income among a Seattle sample is $32,500. It appears as though the average income in Salt Lake City is higher. However, we must consider something else: We have two samples, so we must consider the possibility of sampling error. Namely, our samples likely have different means than the true population means, so we should take this into account. A *t* test is designed to consider these issues by, first, taking the difference between the two means and, second, by considering the sampling error with what is known as the *pooled standard deviation*. This provides an estimate of the overall variability in the means.

The name *t* test is used because the *t* value that results from the test follows what is termed a *Student's t distribution*. This distribution looks a lot like the normal distribution; in fact, it is almost indistinguishable when the sample size is greater than 50.

Table 1.1: Descriptive Statistics from Salt Lake City and Seattle

Statistic	Salt Lake City	Seattle
M	$35,000	$32,500
SD	$500	$750
n	50	50

At smaller sample sizes, the *t* distribution has fatter tails and is a bit flatter in the middle than the normal distribution.

As mentioned earlier, the *t* test has two components: the difference between the means and an estimate of the pooled standard deviation. The following equation shows the form the *t* test takes:

$$t = \frac{\overline{x} - \overline{y}}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2) - 2}}$$

The s_p in the above equations is the pooled standard deviation. The *n*s are the sample sizes and the s^2 represents the variances for the two groups. A key assumption that this type of *t* test makes is that the variances are equal for the two groups represented by the means. Using our example of incomes in Salt Lake City and Seattle above, let's calculate the *t* test. Table 1.1 gives the basic descriptive statistics for each town so that we may calculate this statistic.

Initially, we can plug in the mean values into the *t* test and sample sizes for each city, as seen here:

$$t = \frac{35{,}000 - 32{,}500}{s_p \times \sqrt{\frac{1}{50} + \frac{1}{50}}}$$

Of course, we need to calculate the pooled standard deviation (s_p). We need to use the sample size (n) and standard deviation (s^2) for each town to calculate it.

$$s_p = \sqrt{\frac{(50-1)500^2 + (50-1)750^2}{(50+50) - 2}}$$

No doubt, this looks like a daunting calculation, but if we take it a piece at a time, it can be easily solved.

$$s_p = \sqrt{\frac{(49)500^2 + (49)750^2}{98}}$$

$$s_p = \sqrt{\frac{12250000 + 27562500}{98}}$$

$$s_p = 6186.87$$

We can now solve the full equation:

$$t = \frac{35{,}000 - 32{,}500}{6186.87 \times \sqrt{.002}} = \frac{2{,}500}{247.47} = 10.1$$

A t value of 10.1 indicates that there is a strong statistically significant difference between the two (we will discuss this later in the chapter) because it is larger than the critical t of 1.96. As a result, we can say that the average income in Salt Lake City is indeed higher than in Seattle. Without doing this test, we would not have known that.

Sometimes, we find that the assumption of equal variances is not true; the variances differ to a large degree between two samples. When this occurs, the following test, which is known as Welch's t test, is used:

$$t' = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\text{Var}[x]}{n_1} + \frac{\text{Var}[y]}{n_2}}}$$

Unfortunately, we must use special tables if we wish to compute this value by hand and determine the probability that there is a difference between the two means. Fortunately, though, many statistical software packages provide both types of mean comparison tests, along with another test that is designed to show whether we should use the standard t test or Welch's t test.

An important assumption that we are forced to make if we wish to use these mean-comparison procedures is that the variables follow a normal distribution. The t test, for example, does not provide accurate results if the variable from either sample does not follow a normal distribution. There are other tests, such as those designed to compare ranks or medians (for example, the Wilcoxon-Mann-Whitney test), which are appropriate for nonnormal variables.

There are many other types of comparisons that we might wish to make. Clearly, comparing two means does not exhaust our interest. Suppose we wish to compare three means, four means, or even ten means. We might, for instance, have samples of incomes

from adults in Salt Lake City, Utah; Seattle, Washington; Reno, Nevada; Portland, Oregon; and Boise, Idaho. We may use ANOVA procedures to compare means that are drawn from multiple samples. Using multiple comparison procedures, we may also determine whether one of the means is significantly different from one of the others. Books that describe ANOVA techniques provide an overview of these procedures. As we will learn in subsequent chapters, we may also use linear regression analysis to compute and compare means for different groups that are part of a larger sample.

Samples, Inferences, Significance Tests, and Confidence Intervals

The last substantive topic to cover in this chapter involves the core of inferential statistics: How do we know that what we found actually reflects what is occurring in the population? The cynical—but perhaps most honest—answer is that we never know whether what we found says anything accurate about the population. After all, statistics has been called the science of uncertainty for good reason. We can offer only degrees of confidence that our results reflect characteristics of the population.

However, what do we mean by a *population*? Populations may be divided into *target* populations and *study* populations. Target populations are the group about which you wish to learn something. As this might include a group in the future ("I wish to know the risk of heart attacks at 60 years old among people who are now 40 years old"), we typically find a population that closely resembles the target population; we call this the study population. There are clearly many types of populations. For instance, we might be interested in the population of sea lions off the coast of San Diego in July 2005; the population of poodles in New York City; or the population of voters in Massachusetts during the 2014 election. Yet quite a few people, when they hear the term *population* used in statistics, assume it means the U.S. population, the world's population, or some other extremely large group.

A *sample* is a set of items selected from the population. There are many books and articles on the various types of samples that one might draw from a population. The most commonly described in elementary statistics is the *simple random sample*. This means that each item in or member of the population has an equal chance of being in the sample. There are also clustered samples, stratified samples, and many other types. Most of the classic theoretical work that has gone into inferential statistics is based on the simple random sample, however.

It should be obvious by now that one of the valuable results of statistics is the ability to say something about a population from a sample, but recall the lesson we learned when discussing the standard error of the mean: We usually take only one sample from

a population, but we could conceivably draw many. Therefore, any sample statistic we compute or test we run must consider the uncertainty involved in sampling. The solution to this dilemma of uncertainty has been the frequent use of standard errors for test statistics, including the mean, the standard deviation, correlations, and medians. As we will see in chapter 2, there is also a standard error for slope coefficients in linear regression models.

These standard errors may be thought of as quantitative estimates of the uncertainty involved in test statistics. They typically are used in one of two ways. First, recall from elementary statistics that when we use a *t* test, we compare the *t* value to a table of *p* values. All else being equal, a larger *t* value equates to a smaller *p* value. This approach is known generally as *significance testing* because we wish to determine whether our results are significantly different from some other possible result. It is important to note, though, that the term *significant* does not mean *important*. Rather, it was originally used to mean that the results signified or showed something (see Salsburg, 2001). We often confuse or mislead when we claim that a significance test demonstrates that we have found something special.

Showing where *p* values come from is beyond the scope of this presentation. It is perhaps simpler to provide a general interpretation. Earlier we calculated a *t* test to compare the mean income levels in Salt Lake City ($n = 50$) and Seattle ($n = 50$). If we look up a table of *t* values (available online or in most introductory statistics textbooks), we find, using a sample size of 100, that a *t* value of 10.1 corresponds to a *p* value of less than .001. This leaves approximately .001 of the area under the curve that represents the *t* distribution. One way to interpret this *p* value is with the following long-winded statement: Assuming we took many, many samples from the population of adults in Salt Lake City and Seattle, and there was actually *no difference* in mean income from these populations, we would expect to find a difference of $2,500 or something larger (in Salt Lake City) only one time, on average, out of every 1,000 samples we drew.

If you remember using null and alternative hypotheses, such a statement may sound familiar. In fact, we can translate the above inquiry into the following hypotheses:

H_0: Mean Income, Salt Lake City = Mean Income, Seattle
H_a: Mean Income, Salt Lake City > Mean Income, Seattle

Astute readers may notice that we have set up a *one-tailed* significance test. This is important because one- and two-tailed significance tests imply different comparisons. A *two-tailed* significance test, for example, is a test of whether the mean is higher in Salt Lake City *or in* Seattle. We will return to the interpretation of *p* values in chapter 2.

The second way that standard errors are used is to compute *confidence intervals* (CIs). There are many applied statisticians who prefer CIs because they provide a range

of values within which some measure is likely to fall. Here, we contrast a point estimate and an interval estimate. Means and correlations are examples of point estimates: They are single numbers computed from the sample that estimate population values. An interval estimate provides a range of values that (presumably) contains the actual population value. A CI offers a range of possible or plausible values. Those who prefer CIs argue that they provide a better representation of the uncertainty inherent in statistical analyses.

The general formula for a CI is:

$$\text{Point estimate} \pm [(\text{confidence level}) \times (\text{standard error})]$$

The confidence level represents the percentage of the time, based on a *z* statistic or a *t* statistic, you wish to be able to say that the interval includes the point estimate. For example, assume we have collected data on violent crime rates from a representative sample of 100 cities in the United States. We wish to estimate a suitable range of values for the mean violent crime rate in the population. Our sample yields a mean of 104.9 with a standard deviation of 23.1. The 95 percent CI is computed as:

$$95\%\ \text{CI} = 104.9 \pm \left[1.96 \times \left(\frac{23.1}{\sqrt{100}} \right) \right] = \{100.4, 109.4\}$$

The value of 1.96 for the confidence level comes from a table of standard normal values, or *z* values. It corresponds to a *p* value of .05 (two-tailed test). The standard error formula was presented earlier in this chapter.

How do we interpret the interval of 100.4 – 109.4? There are two ways that are generally used:

1. Given a sample mean of 104.9, we are 95 percent confident that the population mean of violent crime rates falls in the interval of 100.4 and 109.4.
2. If we were to draw many samples from the population of cities in the United States, and we claimed that the population mean fell within the interval of 100.4 – 109.4, we would be accurate about 95 percent of the time.

In subsequent chapters, we will discuss how it is also possible to construct CIs for point estimates from a linear regression model.

An Example Using SPSS

The file *GSS 2010.sav* is an SPSS data set that contains many variables from the 2010 edition of the General Social Survey. We are going to use SPSS to show some of the test statistics we have discussed in this chapter. Move the variable *mntlhlth*, which measures

Table 1.2: SPSS Output

Descriptive Statistics

	N Statistic	Mean		Std. Deviation Statistic	Variance Statistic
		Statistic	Std. Error		
Days of poor mental health past	1,151	3.83	.216	7.315	53.515
Valid *N* (listwise)	1,151				

Figure 1.3: Stata Output

Days of Poor Mental Health Past

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	obs	1151
25%	0	0	Sum of Wgt.	1151
50%	0		Mean	3.825369
		Largest	Std. Dev.	7.315374
75%	4	30		
90%	15	30	Variance	53.51469
95%	25	30	Skewness	2.426593
99%	30	30	Kurtosis	8.238027

number of poor mental health days in the past 30 days, into the variable box. Click *Options* (bottom right corner) and place a check mark in the boxes labeled Mean, Std. deviation, Variance, and S.E. mean. Then click *Continue*. After it returns you to the *Descriptive Statistics* window, click *OK*. (If you wish to see how SPSS uses syntax files, try clicking *Paste* and it will bring up a new screen with the coding.) SPSS's output screen is shown in Table 1.2.

An Example Using Stata

Using the same data in Stata format, *GSS 2010.dta,* we can run the same analyses as we did with SPSS. After opening the data set, use the summarize command with the detail option to see descriptive statistics of poor mental health days. The detail option allows you to see additional statistics, such as the median and variance. After entering the following text in Stata's command line—summarize mntlhlth, detail—the output screen in Stata can be found in Figure 1.3.

Key Terms

The following is a list of key terms and concepts in chapter 1. You can find definitions to these words throughout the chapter.

- Categorical
- Causation
- Confidence intervals
- Constants
- Correlation
- Covary/correlate
- Descriptive
- Deviations from the mean
- Discrete
- E[X] = expected value of the variable X
- Frequentist
- Inferential
- Measures of central tendency
- Measures of dispersion
- Mean (arithmetic)
- Median
- Natural logarithm
- Normal/Gaussian distribution
- One-tailed significance test
- Population
- Probability
- Proportion
- Robust statistics
- Sample
- Sigma—summation
- Significance testing
- Simple random sample
- Standard deviation
- Sum of squares
- Trimmed mean
- *t* test
- Two-tailed significance test
- Variables
- Variance
- x_i
- *x*-bar

Analysis Exercise

1. As an exercise, see whether you can compute the standard error of the mean from the mean and the standard deviation (abbreviated Std.) from the variance listed in the table.

   SPSS: Use *Analyze-Correlate-Bivariate* to estimate the correlation between **educ** and **mntlhlth** in the GSS 2010 data (make sure the Pearson box is checked). Can you figure out how to estimate the covariance between these two variables? (Hint: use *Options* on the *Bivariate Correlations* screen.) If you are interested in confidence intervals for the mean, try using *Analyze-Descriptive Statistics-Explore* and place Public Expenditures in the *Dependent List* box.

Stata: Use the correlate or pwcorr command to estimate the correlation between **educ** and **mntlhlth** (for example, `correlate educ mntl-hlth`). Can you figure out how to estimate the covariance between these two variables? If you are interested in confidence intervals for the mean, use the ci command.

Solution

You should find that the correlation is .748 and the covariance is 37.065. For public expenditures, you should obtain 95 percent CIs of 37.18 and 50.05. Using the GSS 2010 data, look for statistically significant differences by gender in personal income with a *t* test.

SPSS: Go to Analyze → Compare Means → Independent Samples T-test to run the analysis. Use *p income* as your test variable and gender (1 = male and 0 = female) as your grouping variable. What does this table show? What is the *t* value? What is the *p* value?

Stata: In the GSS 2010 data file, you will find a variable called *gender*, which is coded as 0 = male and 1 = female (use the command codebook gender to see some information about this variable). We will use it to compare personal income (labeled *pincome*) for males and females. The command `t test pincome, by(gender)` should produce a table with which to accomplish this task. What does this table show? What is the *t* value? What is the *p* value? Unlike SPSS, you can try using the subcommand welch to request Welch's version of the *t* test for unequal variances (*t* test pincome, by(gender) welch). What does this test show?