IN THIS CHAPTER

- Introducing statistical concepts
- Generalizing from samples to populations
- Testing hypotheses
- Looking at two types of errors

Chapter 1
Data, Statistics, and Decisions

Statistics, first and foremost, is about *decision-making*. Statisticians look at data and wonder what the numbers are saying.

R helps you crunch the data and compute the numbers. As a bonus, R can also help you comprehend statistical concepts.

Developed specifically for statistical analysis, R is a computer language that implements many of the analytical tools statisticians have developed for decision-making. I wrote this book to show how to use these tools in your work.

The Statistical (and Related) Notions You Just Have to Know

The analytical tools that R provides are based on statistical concepts in the remainder of this chapter. These concepts are based on common sense.

Samples and populations

If you watch TV on election night, you know that one of the main events is the prediction of the outcome immediately after the polls close (and before all the votes are counted).

The idea is to talk to a *sample* of voters right after they vote. If they're truthful about how they marked their ballots, and if the sample is representative of the *population* of voters, analysts can use the sample data to draw conclusions about the population.

That, in a nutshell, is what statistics is all about — using the data from samples to draw conclusions about populations.

Here's another example. Imagine that your job is to find the average height of 10-year-old children in the United States. Because you probably wouldn't have the time or the resources to measure every child, you'd measure the heights of a representative sample. Then you'd average those heights and use that average as the estimate of the population average.

Estimating the population average is one kind of *inference* that statisticians make from sample data. I discuss inference in more detail in the upcoming section "Inferential Statistics: Testing Hypotheses."

REMEMBER

Here's some important terminology: Properties of a population (like the population average) are called *parameters*, and properties of a sample (like the sample average) are called *statistics*. If your only concern is the sample properties (like the heights of the children in your sample), the statistics you calculate are *descriptive*. If you're concerned about estimating the population properties, your statistics are *inferential*.

REMEMBER

Now for an important convention about notation: Statisticians use Greek letters (μ, σ, ρ) to stand for parameters, and English letters ($\bar{X}$, s, r) to stand for statistics. Figure 1-1 summarizes the relationship between populations and samples, and between parameters and statistics.

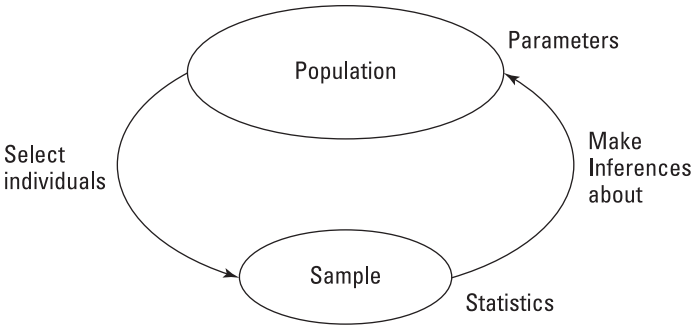

FIGURE 1-1: The relationship between populations, samples, parameters, and statistics.

Variables: Dependent and independent

A *variable* is something that can take on more than one value — like your age, the value of the dollar against another currency, or the number of games your favorite sports team wins. Something that can have only one value is a *constant*. Scientists tell us that the speed of light is a constant, and we use the constant π to calculate the area of a circle.

Statisticians work with *independent* variables and *dependent* variables. In any study or experiment, you'll find both kinds. Statisticians assess the relationship between them.

REMEMBER

A dependent variable is what a researcher *measures*. In an experiment, an independent variable is what a researcher *manipulates*. In some contexts, a researcher can't manipulate an independent variable. Instead, he notes naturally occurring values of the independent variable and how they affect a dependent variable.

REMEMBER

In general, the objective is to find out whether changes in a dependent variable are associated with changes in an independent variable.

REMEMBER

In examples that appear throughout this book, I show you how to use R to calculate characteristics of groups of scores, or to compare groups of scores. Whenever I show you a group of scores, I'm talking about the values of a dependent variable.

Types of data

When you do statistical work, you can run into four kinds of data. And when you work with a variable, the way you work with it depends on what kind of data it is:

The first kind is *nominal* data. If a set of numbers happens to be nominal data, the numbers are labels — their values don't signify anything.

The next kind is *ordinal* data. In this data-type, the numbers are more than just labels. The order of the numbers is important. If I ask you to rank ten foods from the one you like best (one), to the one you like least (ten), we'd have a set of ordinal data.

But the difference between your third-favorite food and your fourth-favorite food might not be the same as the difference between your ninth-favorite and your tenth-favorite. This type of data lacks equal intervals and equal differences.

The third kind of data, *interval*, gives us equal differences. The Fahrenheit scale of temperature is a good example. The difference between 30° and 40° is the same as the difference between 90° and 100°. Each degree is an interval.

On the Fahrenheit scale, a temperature of 80° is not twice as hot as 40°. For ratio statements ("twice as much as," "half as much as") to make sense, "zero" has to mean the complete absence of the thing you're measuring. A temperature of 0°F doesn't mean the complete absence of heat — it's just an arbitrary point on the Fahrenheit scale. (The same holds true for Celsius.)

The fourth kind of data, *ratio*, provides a meaningful zero point. On the Kelvin Scale of temperature, zero means "absolute zero," where all molecular motion (the basis of heat) stops. So 200° Kelvin is twice as hot as 100° Kelvin. Another example is length. Eight inches is twice as long as four inches. "Zero inches" means "a complete absence of length."

REMEMBER

An independent variable or a dependent variable can be either nominal, ordinal, interval, or ratio. The analytical tools you use depend on the type of data you work with.

A little probability

When statisticians make decisions, they use probability to express their confidence about those decisions. They can never be absolutely certain about what they decide. They can only tell you how *probable* their conclusions are.

What do we mean by probability? In my experience, the best way to understand probability is with examples.

If you toss a coin, what's the probability that it turns up heads? If the coin is fair, you might figure that you have a 50-50 chance of heads and a 50-50 chance of tails. And you'd be right. In terms of the kinds of numbers associated with probability, that's ½.

Think about rolling a fair die (one member of a pair of dice). What's the probability that you roll a 4? Well, a die has six faces and one of them is 4, so that's ⅙.

Still another example: Select one card at random from a standard deck of 52 cards. What's the probability that it's a diamond? A deck of cards has four suits, so that's ¼.

In general, the formula for the probability that a particular event occurs is

$$\Pr(\text{event}) = \frac{\text{Number of ways the event can occur}}{\text{Total number of possible events}}$$

At the beginning of this section, I say that statisticians express their confidence about their conclusions in terms of probability, which is why I brought all this up in the first place. This line of thinking leads to *conditional* probability — the probability that an event occurs given that some other event occurs. Suppose that I roll a die, look at it (so that you don't see it), and tell you that I rolled an odd number. What's the probability that I've rolled a 5? Ordinarily, the probability of a 5 is ⅙, but "I rolled an odd number" narrows it down. That piece of information eliminates the three even numbers (2, 4, 6) as possibilities. Only the three odd numbers (1,3, 5) are possible, so the probability is ⅓.

What's the big deal about conditional probability? What role does it play in statistical analysis? Read on.

Inferential Statistics: Testing Hypotheses

Before a statistician does a study, he draws up a tentative explanation — a *hypothesis* that tells why the data might come out a certain way. After gathering all the data, the statistician has to decide whether or not to reject the hypothesis.

That decision is the answer to a conditional probability question — what's the probability of obtaining the data, given that this hypothesis is correct? Statisticians have tools that calculate the probability. If the probability turns out to be low, the statistician rejects the hypothesis.

Back to coin-tossing for an example: Imagine that you're interested in whether a particular coin is fair — whether it has an equal chance of heads or tails on any toss. Let's start with "The coin is fair" as the hypothesis.

To test the hypothesis, you'd toss the coin a number of times — let's say, a hundred. These 100 tosses are the sample data. If the coin is fair (as per the hypothesis), you'd expect 50 heads and 50 tails.

If it's 99 heads and 1 tail, you'd surely reject the fair-coin hypothesis: The conditional probability of 99 heads and 1 tail given a fair coin is very low. Of course, the coin could still be fair and you could, quite by chance, get a 99-1 split, right? Sure. You never really know. You have to gather the sample data (the 100 toss-results) and then decide. Your decision might be right, or it might not.

Null and alternative hypotheses

Think again about that coin-tossing study I just mentioned. The sample data are the results from the 100 tosses. I said that we can start with the hypothesis that the coin is fair. This starting point is called the *null hypothesis*. The statistical notation for the null hypothesis is H_o. According to this hypothesis, any heads-tails split in the data is consistent with a fair coin. Think of it as the idea that nothing in the sample data is out of the ordinary.

An alternative hypothesis is possible — that the coin isn't a fair one and it's biased to produce an unequal number of heads and tails. This hypothesis says that any heads-tails split is consistent

with an unfair coin. This alternative hypothesis is called, believe it or not, the *alternative hypothesis.* The statistical notation for the alternative hypothesis is H_1.

Now toss the coin 100 times and note the number of heads and tails. If the results are something like 90 heads and 10 tails, it's a good idea to reject H_0. If the results are around 50 heads and 50 tails, don't reject H_0.

REMEMBER

Notice that I did *not* say "accept H_0." The way the logic works, you *never* accept a hypothesis. You either reject H_0 or don't reject H_0.

Two types of error

Whenever you evaluate data and decide to reject H_0 or to not reject H_0, you can never be absolutely sure. You never really know the "true" state of the world.

Because you're never certain about your decisions, you can make an error either way you decide. As I mention earlier, the coin could be fair and you just happen to get 99 heads in 100 tosses. That's not likely, and that's why you reject H_0 if that happens. It's also possible that the coin is biased, yet you just happen to toss 50 heads in 100 tosses. Again, that's not likely and you don't reject H_0 in that case.

Although those errors are not likely, they are possible. They lurk in every study that involves inferential statistics. Statisticians have named them *Type I* errors and *Type II* errors.

If you reject H_0 and you shouldn't, that's a Type I error. In the coin example, that's rejecting the hypothesis that the coin is fair, when in reality it is a fair coin.

If you don't reject H_0 and you should have, that's a Type II error. It happens if you don't reject the hypothesis that the coin is fair, and in reality it's biased.

How do you know if you've made either type of error? You don't — at least not right after you make the decision to reject or not reject H_0. (If it's possible to know, you wouldn't make the error in the first place!) All you can do is gather more data and see if the additional data is consistent with your decision.