IN THIS CHAPTER

- Introducing statistical concepts
- Generalizing from samples to populations
- Getting into probability
- Testing hypotheses
- Two types of error

Chapter 1

Data, Statistics, and Decisions

Statistics? That's all about crunching numbers into arcane-looking formulas, right? Not really. Statistics, first and foremost, is about decision-making. Some number-crunching is involved, of course, but the primary goal is to use numbers to make decisions. Statisticians look at data and wonder what the numbers are saying. What kinds of trends are in the data? What kinds of predictions are possible? What conclusions can you make?

To make sense of data and answer these questions, statisticians have developed a wide variety of analytical tools.

About the number-crunching part: If you had to do it via pencil-and-paper (or with the aid of a pocket calculator), you'd soon grow discouraged with the amount of computation involved and the errors that might creep in. Software like Python helps you crunch the data and compute the numbers. As a bonus, working with Python can also help you comprehend statistical concepts.

Although Python is an all-purpose computing language, many of its libraries make it ideal for statistical work. I wrote this book to show you how to use these libraries and the statistical tools they make available.

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

The analytical tools you find in Python are based on statistical concepts I help you explore in the remainder of this chapter. As you'll see, these concepts are based on common sense.

Samples and populations

If you watch TV on election night, you know that one exciting occurrence that takes place before the main event is the prediction of the outcome immediately after the polls close (and before all the votes are counted). How is it that pundits almost always get it right?

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 later 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.

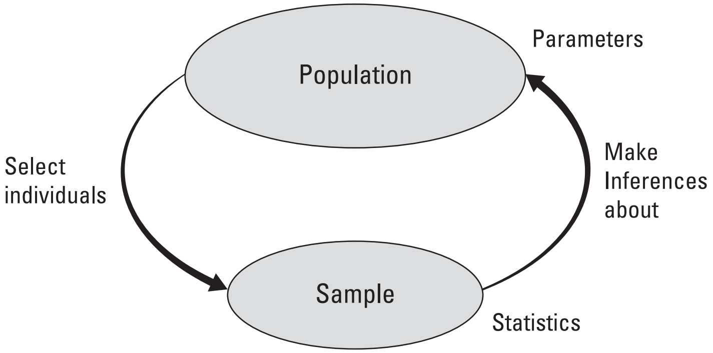

FIGURE 1-1: The relationship between populations, samples, parameters, and statistics.

Variables: Dependent and independent

A *variable* is something that can take on different values at different times — like your age, the value of the dollar against other currencies, 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.

Imagine a computerized training method designed to increase a person's IQ. How would a researcher find out whether this method does what it's supposed to do? First, that person would randomly assign a sample of people to one of two groups. One group would receive the training method, and the other would complete another kind of computer-based activity — like reading text on a website. Before and after each group completes its activities, the researcher measures each person's IQ. What happens next? I discuss that topic in the later section "Inferential Statistics: Testing Hypotheses."

For now, understand that the independent variable here is Type of Activity. The two possible values of this variable are IQ Training and Reading Text. The dependent variable is the change in IQ from Before to After.

REMEMBER

A dependent variable is what a researcher *measures*. In an experiment, an independent variable is what a researcher *manipulates*. In other contexts, a researcher can't manipulate an independent variable. Instead, they note 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 an independent variable are associated with changes in a dependent variable.

REMEMBER

In the examples that appear throughout this book, I show you how to use Python 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. On a sports team, the jersey numbers are nominal. They just identify the players.

The next kind is *ordinal* data. In this data type, the numbers are more than just labels. As the name *ordinal* might tell you, the order of the numbers is important. If I were to ask you to rank ten foods from the one you like best (1) to the one you like least (10), 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. So this type of data lacks equal intervals and equal differences.

Interval data 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°. So each degree is an interval.

People are sometimes surprised to find out that 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 4 inches. *Zero inches* means "a complete absence of length."

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 I mean by *probability*? Mathematicians and philosophers might give you complex definitions. In my experience, however, the best way to understand probability is in terms of examples.

Here's a simple example: 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 1 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.

These examples tell you that if you want to know the probability that an event occurs, count how many ways that event can happen and divide by the total number of events that can happen. In the first two examples (heads, 4), the event you're interested in happens in only one way. For the coin, you divide 1 by 2. For the die, you divide 1 by 6. In the third example (diamond), the event can happen in 1 of 13 ways (ace through king), so you divide 13 by 52 (to get ¼).

Now for a slightly more complicated example. Toss a coin and roll a die at the same time. What's the probability of tails and a 4? Think about all the possible events that can happen when you toss a coin and roll a die at the same time. You could have tails and 1 through 6, or heads and 1 through 6. That adds up to 12 possibilities. The tails-and-4 combination can happen only one way. So the probability is $^1\!/_{12}$.

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 any statistician begins a study, they draw 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 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 100. 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.

Juries make these types of decisions. In the United States, the starting hypothesis is that the defendant is not guilty ("innocent until proven guilty"). Think of the evidence as data. Jury members consider the evidence and answer a conditional probability question: What's the probability of the evidence, given that the defendant is not guilty? Their answer determines the verdict.

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 loaded 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_o. If the results are around 50 heads and 50 tails, don't reject H_o.

Similar ideas apply to the IQ example I gave earlier. One sample receives the computer-based IQ training method, and the other participates in a different computer-based activity — like reading text on a website. Before and after each group completes its activities, the researcher measures each person's IQ. The null hypothesis, H_o, is that one group's improvement isn't different from the other. If the improvements are greater with the IQ training than with the other activity — so much greater that it's unlikely that the two aren't different from one another — reject H_o. If they're not, don't reject H_o.

REMEMBER

Notice that I did *not* say "accept H_o." The way the logic works, you *never* accept a hypothesis. You either reject H_o or don't reject H_o. In a jury trial, the verdict is either "guilty" (reject the null hypothesis of "not guilty") or "not guilty" (don't reject H_o). "Innocent" (acceptance of the null hypothesis) is not a possible verdict.

Notice also that in the coin-tossing example, I said "around 50 heads and 50 tails." What does *around* mean? Also, I said that if it's 90-10, reject H_o. What about 85-15? 80-20? 70-30? Exactly how much different from 50-50 does the split have to be for you to reject H_o? In the IQ training example, how much greater does the IQ improvement have to be to reject H_o?

I won't answer these questions now. Statisticians have formulated decision rules for situations like this, and I'll help you explore those rules throughout this book.

Two types of error

Whenever you evaluate data and decide to reject H_o or not reject H_o, you can never be absolutely sure. You never really know the "true" state of the world. In the coin-tossing example, that means you can't be certain whether the coin is fair. All you can do is make a decision based on the sample data. If you want to know for sure about the coin, you have to have the data for the entire population of tosses — which means you have to keep tossing the coin until the end of time.

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_o 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_o in that case.

Although those errors aren't likely, they're 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_o 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's a fair coin.

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

How do you know whether 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_o. (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 whether the additional data is consistent with your decision.

If you think of H_o as a tendency to maintain the status quo and not interpret anything as being out of the ordinary (no matter how it looks), a Type II error means you've missed out on something big. In fact, some iconic mistakes are Type II errors.

Here's what I mean. On New Year's Day in 1962, a rock group consisting of three guitarists and a drummer auditioned in the London studio of a major recording company. Legend has it that the recording executives didn't like what they heard, didn't like what they saw, and believed that guitar groups were on their way out. Although the musicians played their hearts out, the group failed the audition.

Who was that group? The Beatles!

And *that's* a Type II error.