IN THIS CHAPTER

- Getting the big picture of the field of statistics
- Overviewing the steps of the scientific method
- Seeing the role of statistics at each step

Chapter 1
Statistics in a Nutshell

The most common description of statistics is that it's the process of analyzing data — number crunching, in a sense. But statistics is not just about analyzing the data. It's about the whole process of using the scientific method to answer questions and make decisions. That process involves designing studies, collecting good data, describing the data with numbers and graphs, analyzing the data, and then making conclusions. In this chapter, I review each of these steps and show where statistics plays the all-important role.

Designing Studies

Once a research question is defined, the next step is designing a study in order to answer that question. This amounts to figuring out what process you'll use to get the data you need. In this section, I overview the two major types of studies: observational studies and experiments.

Surveys

An *observational study* is one in which data are collected on individuals in a way that doesn't affect them. The most common observational study is the survey. *Surveys* are questionnaires that

are presented to individuals who have been selected from a population of interest. Surveys take on many different forms: paper surveys sent through the mail; websites; call-in polls conducted by TV networks; and phone surveys. If conducted properly, surveys can be very useful tools for getting information. However, if not conducted properly, surveys can result in bogus information. Some problems include improper wording of questions, which can be misleading, people who were selected to participate but do not respond, or an entire group in the population who had no chance of even being selected. These potential problems mean a survey has to be well thought-out before it's given.

A downside of surveys is that they can only report relationships between variables that are found; they cannot claim cause and effect. For example, if in a survey researchers notice that the people who drink more than one Diet Coke per day tend to sleep fewer hours each night than those who drink at most one per day, they cannot conclude that Diet Coke is causing the lack of sleep. Other variables might explain the relationship, such as number of hours worked per week. See all the information about surveys, their design, and potential problems in Chapter 12.

Experiments

An *experiment* imposes one or more treatments on the participants in such a way that clear comparisons can be made. Once the treatments are applied, the response is recorded. For example, to study the effect of drug dosage on blood pressure, one group might take 10 mg of the drug, and another group might take 20 mg. Typically, a control group is also involved, where subjects each receive a fake treatment (a sugar pill, for example).

Experiments take place in a controlled setting, and are designed to minimize biases that might occur. Some potential problems include: researchers knowing who got what treatment; a certain condition or characteristic wasn't accounted for that can affect the results (such as weight of the subject when studying drug dosage); or lack of a control group. But when designed correctly, if a difference in the responses is found when the groups are compared, the researchers can conclude a cause and effect relationship. See coverage of experiments in Chapter 13.

It is perhaps most important to note that no matter what the study, it has to be designed so that the original questions can be answered in a credible way.

Collecting Data

Once a study has been designed, be it a survey or an experiment, the subjects are chosen and the data are ready to be collected. This phase of the process is also critical to producing good data.

Selecting a good sample

First, a few words about selecting individuals to participate in a study (much, much more is said about this topic in Chapter 12). In statistics, we have a saying: "Garbage in equals garbage out." If you select your subjects in a way that is *biased* — that is, favoring certain individuals or groups of individuals — then your results will also be biased.

Suppose Bob wants to know the opinions of people in your city regarding a proposed casino. Bob goes to the mall with his clipboard and asks people who walk by to give their opinions. What's wrong with that? Well, Bob is only going to get the opinions of a) people who shop at that mall; b) on that particular day; c) at that particular time; d) and who take the time to respond. That's too restrictive — those folks don't represent a cross-section of the city. Similarly, Bob could put up a website survey and ask people to use it to vote. However, only those who know about the site, have Internet access, and want to respond will give him data. Typically, only those with strong opinions will go to such trouble. So, again, these individuals don't represent all the folks in the city.

In order to minimize bias, you need to select your sample of individuals *randomly* — that is, using some type of "draw names out of a hat" process. Scientists use a variety of methods to select individuals at random (more in Chapter 12), but getting a random sample is well worth the extra time and effort to get results that are legitimate.

Avoiding bias in your data

Say you're conducting a phone survey on job satisfaction of Americans. If you call them at home during the day between 9 a.m. and 5 p.m., you'll miss out on all those who work during the day; it could be that day workers are more satisfied than night workers, for example. Some surveys are too long — what if someone stops answering questions halfway through? Or what if they give you misinformation and tell you they make $100,000 a year instead of $45,000? What if they give you an answer that isn't on your list of possible answers? A host of problems can occur when collecting survey data; Chapter 12 gives you tips on avoiding and spotting them.

Experiments are sometimes even more challenging when it comes to collecting data. Suppose you want to test blood pressure; what if the instrument you are using breaks during the experiment? What if someone quits the experiment halfway through? What if something happens during the experiment to distract the subjects or the researchers? Or they can't find a vein when they have to do a blood test exactly one hour after a dose of a drug is given? These are just some of the problems in data collection that can arise with experiments; Chapter 13 helps you find and minimize them.

Describing Data

Once data are collected, the next step is to summarize it all to get a handle on the big picture. Statisticians describe data in two major ways: with pictures (that is, charts and graphs) and with numbers, called *descriptive statistics*.

Descriptive statistics

Data are also summarized (most often in conjunction with charts and/or graphs) by using what statisticians call descriptive statistics. *Descriptive statistics* are numbers that describe a data set in terms of its important features.

If the data are categorical (where individuals are placed into groups, such as gender or political affiliation), they are typically

summarized using the number of individuals in each group (called the *frequency*) or the percentage of individuals in each group (the *relative frequency*).

Numerical data represent measurements or counts, where the actual numbers have meaning (such as height and weight). With numerical data, more features can be summarized besides the number or percentage in each group. Some of these features include measures of center (in other words, where is the "middle" of the data?); measures of spread (how diverse or how concentrated are the data around the center?); and, if appropriate, numbers that measure the relationship between two variables (such as height and weight).

Some descriptive statistics are better than others, and some are more appropriate than others in certain situations. For example, if you use codes of 1 and 2 for males and females, respectively, when you go to analyze that data, you wouldn't want to find the average of those numbers — an "average gender" makes no sense. Similarly, using percentages to describe the amount of time until a battery wears out is not appropriate. A host of basic descriptive statistics are presented, compared, and calculated in Chapter 2.

Charts and graphs

Data are summarized in a visual way using charts and/or graphs. Some of the basic graphs used include pie charts and bar charts, which break down variables such as gender and which applications are used on teens' cellphones. A bar graph, for example, may display opinions on an issue using 5 bars labeled in order from "Strongly Disagree" up through "Strongly Agree."

But not all data fit under this umbrella. Some data are numerical, such as height, weight, time, or amount. Data representing counts or measurements need a different type of graph that either keeps track of the numbers themselves or groups them into numerical groupings. One major type of graph that is used to graph numerical data is a histogram. In Chapter 3, you delve into pie charts, bar graphs, histograms and other visual summaries of data.

Analyzing Data

After the data have been collected and described using pictures and numbers, then comes the fun part: navigating through that black box called the *statistical analysis*. If the study has been designed properly, the original questions can be answered using the appropriate analysis, the operative word here being *appropriate*. Many types of analyses exist; choosing the wrong one will lead to wrong results.

In this book, I cover the major types of statistical analyses encountered in introductory statistics. Scenarios involving a fixed number of independent trials where each trial results in either success or failure use the binomial distribution, described in Chapter 4. In the case where the data follow a bell-shaped curve, the normal distribution is used to model the data, covered in Chapter 5.

Chapter 7 deals with confidence intervals, used when you want to make estimates involving one or two population means or proportions using a sample of data. Chapter 8 focuses on testing someone's claim about one or two population means or proportions — these analyses are called hypothesis tests. If your data set is small and follows a bell-shape, the *t*-distribution might be in order; see Chapter 9.

Chapter 10 examines relationships between two numerical variables (such as height and weight) using correlation and simple linear regression. Chapter 11 studies relationships between two categorical variables (where the data place individuals into groups, such as gender and political affiliation). You can find a fuller treatment of these topics in *Statistics For Dummies* (Wiley), and analyses that are more complex than that are discussed in the book *Statistics II For Dummies*, also published by Wiley.

Making Conclusions

Researchers perform analysis with computers, using formulas. But neither a computer nor a formula knows whether it's being used properly, and they don't warn you when your results are incorrect. At the end of the day, computers and formulas can't tell you what the results mean. It's up to you.

One of the most common mistakes made in conclusions is to overstate the results, or to generalize the results to a larger group than was actually represented by the study. For example, a professor wants to know which Super Bowl commercials viewers liked best. She gathers 100 students from her class on Super Bowl Sunday and asks them to rate each commercial as it is shown. A Top 5 list is formed, and she concludes that Super Bowl viewers liked those five commercials the best. But she really only knows which ones *her students* liked best — she didn't study any other groups, so she can't draw conclusions about all viewers.

Statistics is about much more than numbers. It's important to understand how to make appropriate conclusions from studying data, and that's something I discuss throughout the book.