1 A Review of Basic Statistical Concepts

The record of a month's roulette playing at Monte Carlo can afford us material for discussing the foundations of knowledge.

—Karl Pearson

I know too well that these arguments from probabilities are imposters, and unless great caution is observed in the use of them, they are apt to be deceptive.

—Plato (in *Phaedo*)

Introduction

It is hard to find two quotations from famous thinkers that reflect more divergent views of probability and statistics. The eminent statistician Karl Pearson (the guy who invented the correlation coefficient) was so enthralled with probability and statistics that he seems to have believed that understanding probability and statistics is a cornerstone of human understanding. Pearson argued that statistical methods can offer us deep insights into the nature of reality. The famous Greek philosopher Plato also had quite a bit to say about the nature of reality. In contrast to Pearson, though, Plato was skeptical of the "fuzzy logic" of probabilities and central tendencies. From Plato's viewpoint, we should only trust what we can know with absolute certainty. Plato probably preferred deduction (e.g., If *B* then *C*) to induction (In my experience, bees seem to like flowers).

Even Plato seemed to agree, though, that if we observe "great caution," arguments from probabilities may be pretty useful. In contrast, some modern nonstatisticians might agree with what the first author's father, Bill Pelham, used to say about statistics and probability theory: "Figures

can't lie, but liars sure can figure." His hunch, and his fear, was that "you can prove anything with statistics." To put this a little differently, a surprising number of thoughtful, intelligent students are thumbs-down on statistics. In fact, some students only take statistics because they *have* to (e.g., to graduate with a major in psychology, to earn a second or third PhD). If you fall into this category, our dream for you is that you enjoy this book so much that you will someday talk about the next time that you *get* to take—or teach—a statistics class.

One purpose of this first chapter, then, is to convince you that Karl Pearson's rosy view of statistics is closer to the truth than is Bill Pelham's jaded view. It is possible, though, that you fully agree with Pearson, but you just don't like memorizing all those formulas Pearson and company came up with. In that case, the purpose of this chapter is to serve as a quick refresher course that will make the rest of this book more useful. In either event, no part of this book requires you to memorize a lot of complex statistical formulas. Instead, the approach emphasized here is heavily conceptual rather than heavily computational. The approach emphasized here is also hands-on. If you can count on your fingers, you can count your blessings because you are fully capable of doing at least some of the important calculations that lie at the very heart of statistics. The hands-on approach of this book emphasizes logic over rote calculation, capitalizes on your knowledge of everyday events, and attempts to pique your innate curiosity with realistic research problems that can best be solved by understanding statistics. If you know whether there is any connection between rain and umbrellas, if you love or hate weather forecasters, and if you find games of chance interesting, we hope that you enjoy at least some of the demonstrations and data analysis activities that are contained in this book.

Before we jump into a detailed discussion of statistics, however, we would like to briefly remind you that (a) statistics is a branch of mathematics and (b) statistics is its own very precise language. This is very fitting because we can trace numbers and, ultimately, statistics back to the beginning of human language and thus to the beginning of human written history. To appreciate fully the power and elegance of statistics, we need to go back to the ancient Middle East.

How Numbers and Language Revolutionized Human History

About 5,000 years ago, once human beings had began to master agriculture, live in large city states, and make deals with one another, an unknown Sumerian trader or traders invented the **cuneiform** writing system to keep track of economic transactions. Because we live in a world surrounded by numbers and written language, it is difficult for us

to appreciate how ingenious it was for someone to realize that *writing things down* solves a myriad of social and economic problems. When Basam and Gabor got into their semimonthly fistfight about whether Gabor owed Basam *five* more or *six* more geese to pay for a newly weaned goat, our pet theory is that it was an exasperated neighbor who finally got sick of all the fighting and thus proposed the cuneiform writing system. The cuneiform system involved making marks with a stylus in wet clay that was then dried and fired as a permanent record of economic transactions. This system initially focused almost exclusively on who had traded what with whom—and, most important, in what quantity. Thus, some Sumerian traders made the impressive leap of impressing important things in clay. This early cuneiform writing system was about as sophisticated as the scribbles of your 4-year-old niece, but it quickly caught on because it was *way* better than spoken language alone.

For example, it apparently wasn't too long before the great-great-great-grandchild of that original irate neighbor got a fantastically brilliant idea. Instead of drawing a stylized duck, duck, duck, duck to represent four ducks, this person realized that *four-ness* itself (like two-ness and thirty-seven-ness) was a concept. He or she thus created *abstract characters* for numbers that saved ancient Sumerians a *lot* of clay. We won't insult you by belaboring how much easier it is to write and verify the cuneiform version of "17 goats" than to write "goat, goat, goat, goat, goat, goat, goat, goat, goat, goat, goat, goat, goat, goat, goat, goat . . ." oh yeah ". . . goat," but we can summarize a few thousand years of human technological and scientific development by reminding you that incredibly useful concepts such as zero, fractions, π (pi), and logarithms, which make possible great things such as penicillin, the Sistine Chapel, and iPhones, would have never come about were it not for the development of abstract numbers and language.

It is probably a bit more fascinating to textbook authors than to textbook readers to recount in great detail what happened over the course of the next 5,000 years, but suffice it to say that written language, numbers, and mathematics revolutionized—and sometimes limited—human scientific and technological development. For example, one of the biggest ruts that brilliant human beings ever got stuck into has to do with numbers. If you have ever given much thought to Roman numerals, it may have dawned on you that they are an inefficient pain in the butt. Who thought it was a great idea to represent 1,000 as M while representing 18 as XVIII? And why the big emphasis on five (V, that is) in a base-10 number system? The short answer to these questions is that whoever formalized Roman numbers got a little too obsessed with counting on his or her fingers and never fully got over it. For example, we hope it's obvious that the Roman numerals I and II are stand-ins for human fingers. It is probably less obvious

that the Roman V ("5") is a stand-in for the "V" that is made by your thumb and first finger when you hold up a single hand and tilt it outward a bit (sort of the way you would to give someone a "high five"). If you do this with both of your hands and move your thumbs together until they cross in front of you, you'll see that the X in Roman numerals is, essentially, V + V. Once you're done making shadow puppets, we'd like to tell you that, as it turns out, there are some major drawbacks to Roman numbers because the Roman system does not perfectly preserve place (the way we write numbers in the ones column, the tens column, the hundreds column, etc.).

If you try to do subtraction, long division, or any other procedure that requires "carrying" in Roman numerals, you quickly run into serious problems, problems that, according to at least some scholars, sharply limited the development of mathematics and perhaps technology in ancient Rome. We can certainly say with great confidence that, labels for popes and Super Bowls notwithstanding, there is a good reason why Roman numerals have fallen by the wayside in favor of the nearly universal use of the familiar Arabic base-10 numbers. In our familiar system of representing numbers, a 5-digit number can never be smaller than a 1-digit number because a numeral's *position* is even more important than its shape. A bank in New Zealand (NZ) got a painful reminder of this fact in May 2009 when it accidentally deposited $10,000,000.00 (yes, ten *million*) NZ dollars rather than $10,000.00 (ten *thousand*) NZ dollars in the account of a couple who had applied for an overdraft. The couple quickly fled the country with the money (all three extra zeros of it).[1] To everyone but the unscrupulous couple, this mistake may seem tragic, but we can assure you that bank errors of this kind would be more common, rather than less common, if we still had to rely on Roman numerals.

If you are wondering how we got from ancient Sumer to modern New Zealand—or why—the main point of this foray into numbers is that life as we know and love it depends heavily on numbers, mathematics, and even statistics. In fact, we would argue that to an ever increasing degree in the modern world, sophisticated thinking requires us to be able understand statistics. If you have ever read the influential book *Freakonomics*, you know that the authors of this book created quite a stir by using statistical analysis (often multiple regression) to make some very interesting points about human behavior (Do real estate agents work as hard for you as they claim? Do Sumo wrestlers always try to win? Does cracking down on crime in conventional ways reduce it? The respective answers appear to be no, no, and no, by the way.) So statistics are important. It is impossible to be a sophisticated, knowledgeable modern person without having at least a passing knowledge of modern statistical methods. Barack Obama appears to have appreciated this fact prior to his election in 2008 when he

assembled a dream team of behavioral economists to help him get elected—and then to tackle the economic meltdown. This dream team relied not on classical economic models of what people *ought* to do but on empirical studies of what people actually do under different conditions. For example, based heavily on the work of psychologist Robert Cialdini, the team knew that one of the best ways to get people to vote on election day is to remind them that many, many other people plan to vote (Can you say "baaa"?).[2]

So if you want a cushy job advising some future president, or a more secure retirement, you would be wise to increase your knowledge of statistics. As it turns out, however, there are two distinct branches of statistics, and people usually learn about the first branch before they learn about the second. The first branch is descriptive statistics, and the second branch is inferential statistics.

Descriptive Statistics

Statistics are a set of mathematical procedures for summarizing and interpreting observations. These observations are typically numerical or categorical facts about specific people or things, and they are usually referred to as **data.** The most fundamental branch of statistics is **descriptive statistics,** that is, statistics used to summarize or describe a set of observations.

The branch of statistics used to interpret or draw inferences about a set of observations is fittingly referred to as **inferential statistics.** Inferential statistics are discussed in the second part of this chapter. Another way of distinguishing descriptive and inferential statistics is that descriptive statistics are the *easy* ones. Almost all the members of modern, industrialized societies are familiar with at least some descriptive statistics. Descriptive statistics include things such as means, medians, modes, and percentages, and they are everywhere. You can scarcely pick up a newspaper or listen to a newscast without being exposed to heavy doses of descriptive statistics. You might hear that LeBron James made 78% of his free throws in 2008–2009 or that the Atlanta Braves have won 95% of their games this season when they were leading after the eighth inning (and 100% of their games when they outscored their opponents). Alternately, you might hear the results of a shocking new medical study showing that, as people age, women's brains shrink 67% less than men's brains do. You might hear a meteorologist report that the average high temperature for the past 7 days has been over 100 °F. The reason that descriptive statistics are so widely used is that they are so useful. They take what could be an extremely large and cumbersome set of observations and boil them down to one or two highly representative numbers.

In fact, we're convinced that if we had to live in a world without descriptive statistics, much of our existence would be reduced to a hellish nightmare. Imagine a sportscaster trying to tell us exactly how well LeBron James has been scoring this season without using any descriptive statistics. Instead of simply telling us that James is averaging nearly 30 points per game, the sportscaster might begin by saying, "Well, he made his first shot of the season but missed his next two. He then made the next shot, the next, and the next, while missing the one after that." That's about as efficient as "goat, goat, goat, goat. . . ." By the time the announcer had documented all of the shots James took this season (without even mentioning *last* season), the game we hoped to watch would be over, and we would never have even heard the score. Worse yet, we probably wouldn't have a very good idea of how well James is scoring this season. A sea of specific numbers just doesn't tell people very much. A simple mean puts a sea of numbers in a nutshell.

CENTRAL TENDENCY AND DISPERSION

Although descriptive statistics are everywhere, the descriptive statistics used by laypeople are typically incomplete in a very important respect. Laypeople make frequent use of descriptive statistics that summarize the **central tendency** (loosely speaking, the average) of a set of observations ("But my old pal Michael Jordan once averaged 32 points in a season"; "A follow-up study revealed that women also happen to be exactly 67% less likely than men to spend their weekends watching football and drinking beer"). However, most laypeople are relatively unaware of an equally useful and important category of descriptive statistics. This second category of descriptive statistics consists of statistics that summarize the **dispersion,** or **variability,** of a set of scores. Measures of dispersion are not only important in their own (descriptive) right, but as you will see later, they are also important because they play a very important role in inferential statistics.

One common and relatively familiar measure of dispersion is the **range** of a set of scores. The range of a set of scores is simply the difference between the highest and the lowest value in the entire set of scores. ("The follow-up study also revealed that virtually *all* men showed the same amount of shrinkage. The smallest amount of shrinkage observed in all the male brains studied was 10.0 cc, and the largest amount observed was 11.3 cc. That's a range of only 1.3 cc. In contrast, many of the women in the study showed no shrinkage whatsoever, and the largest amount of shrinkage observed was 7.2 cc. That's a range of 7.2 cc.") Another very common, but less intuitive, descriptive measure of dispersion is the **standard deviation.** It's a special kind of average itself—namely, an average

measure of how much each of the scores in the sample *differs* from the sample mean. More specifically, it's the square root of the average squared deviation of each score from the sample mean, or

$$S = \sqrt{\frac{\Sigma(x-m)^2}{n}}.$$

Σ (sigma) is a summation sign, a symbol that tells us to perform the functions that follow it for all the scores in a sample and then to add them all together. That is, this symbol tells us to take each individual score in our sample (represented by x), to subtract the mean (m) from it, and to square this difference. Once we have done this for all our scores, sigma tells us to add all these squared difference scores together. We then divide these summed scores by the number of observations in our sample and take the square root of this final value.

For example, suppose we had a small sample of only four scores: 2, 2, 4, and 4. Using the formula above, the standard deviation turns out to be

$$\frac{(2-3)^2+(2-3)^2+(4-3)^2+(4-3)^2}{4},$$

which is simply

$$\frac{1+1+1+1}{4},$$

which is exactly 1.

That's it. The standard deviation in this sample of scores is exactly 1. If you look back at the scores, you'll see that this is pretty intuitive. The mean of the set of scores is 3.0, and every single score deviates from this mean by exactly 1 point. There is a computational form of this formula that is much easier to deal with than the definitional formula shown here (especially if you have a lot of numbers in your sample). However, we included the definitional formula so that you could get a sense of what the standard deviation means. Loosely speaking, it's the average ("standard") amount by which all the scores in a distribution differ (deviate) from the mean of that same set of scores. Finally, we should add that the specific formula we presented here requires an adjustment if you hope to use a sample of scores to estimate the standard deviation in the population of scores from which these sample scores were drawn. It is this adjusted standard deviation that researchers are most likely to use in actual research (e.g., to make inferences about the population standard deviation). Conceptually, however, the adjusted formula (which requires you to divide by $n - 1$ rather

than *n*) does *exactly* what the unadjusted formula does: It gives you an idea of how much a set of scores varies around a mean.

Why are measures of dispersion so useful? Like measures of central tendency, measures of dispersion summarize a very important property of a set of scores. For example, consider the two groups of four men whose heights are listed as follows:

	Group 1	**Group 2**
Tallest guy	6′2″	6′9″
Tall guy	6′1″	6′5″
Short guy	5′11″	5′10″
Shortest guy	5′10″	5′0″

A couple of quick calculations will reveal that the mean height of the men in both groups is exactly 6 feet. Now suppose you were a heterosexual woman of average height and needed to choose a blind date by drawing names from one of two hats. One hat contains the names of the four men in Group 1, and the other hat contains the names of the four men in Group 2. From which hat would you prefer to choose your date? If you followed social conventions regarding dating and height, you would probably prefer to choose your date from Group 1. Now suppose you were choosing four teammates for an intramural basketball team and had to choose one of the two *groups* (in its entirety). In this case, we assume that you would choose Group 2 (and try to get the ball to the big guy when he posts up under the basket). Your preferences reveal that *dispersion* is a very important statistical property because the only way in which the two groups of men differ is in the dispersion (i.e., the variability) of their heights. In Group 1, the standard deviation is 1.58 inches. In Group 2, it's 7.97 inches.[3]

Another example of the utility of measures of dispersion comes from a 1997 study of parking meters in Berkeley, California. The study's author, Ellie Lamm, strongly suspected that some of the meters in her hometown had been shortchanging people. To put her suspicions to the test, she conducted an elegantly simple study in which she randomly sampled 50 parking meters, inserted two nickels in each (enough to pay for 8 minutes), and timed with a stopwatch the actual amount of time each meter delivered. Lamm's study showed that, on average, the amount of time delivered was indeed very close to 8 minutes. The *central tendency* of the 50 meters was to give people what they were paying for.

However, a shocking 94% of the meters (47 of 50) were off one way or the other by at least 20 seconds. In fact, the *range* of delivered time was about 12 minutes! The low value was just under *2 minutes*, and the high

was about *14 minutes*. Needless to say, a substantial percentage of the meters were giving people way less time than they paid for. It didn't matter much that other meters were giving people *too much* time. There's an obvious asymmetry in the way tickets work. When multiplied across the city's then 3,600 parking meters, this undoubtedly created a lot of undeserved parking tickets.

Lamm's study got so much attention that she appeared to discuss it on the *David Letterman Show*. Furthermore, the city of Berkeley responded to the study by replacing their old, inaccurate mechanical parking meters with much more accurate electronic meters. Many thousands of people who had once gotten undeserved tickets were presumably spared tickets after the intervention, and vandalism against parking meters in Berkeley was sharply reduced. So this goes to prove that dispersion is sometimes more important than central tendency. Of course, it also goes to prove that research doesn't have to be expensive or complicated to yield important societal benefits. Lamm's study presumably cost her only $5 in nickels and perhaps a little bit for travel. That's good because Lamm conducted this study as part of her science fair project—when she was 11 years old.[4] We certainly hope she won a blue ribbon.

A more formal way of thinking about dispersion is that measures of dispersion complement measures of central tendency by telling you something about how *well* a measure of central tendency represents all the scores in a distribution. When the dispersion or variability in a set of scores is low, the mean of a set of scores does a great job of describing most of the scores in the sample. When the dispersion or the variability in a set of scores is high, however, the mean of a set of scores does *not* do such a great job of describing most of the scores in the sample (the mean is still the best available summary of the set of scores, but there will be a lot of people in the sample whose scores lie far away from the mean). When you are dealing with descriptions of people, measures of central tendency—such as the mean—tell you what the *typical* person is like. Measures of dispersion—such as the standard deviation—tell you how much you can expect specific people to differ from this typical person.

THE SHAPE OF DISTRIBUTIONS

A third statistical property of a set of observations is a little more difficult to quantify than measures of central tendency or dispersion. This third statistical property is the *shape* of a distribution of scores. One useful way to get a feel for a set of scores is to arrange them in order from the lowest to the highest and to graph them pictorially so that taller parts of the graph represent more frequently occurring scores (or, in the case of a theoretical or ideal distribution, more probable scores). Figure 1.1 depicts three different kinds of distributions: a rectangular distribution,

a bimodal distribution, and a normal distribution. The scores in a **rectangular distribution** are all about equally frequent or probable. An example of a rectangular distribution is the theoretical distribution representing the six possible scores that can be obtained by rolling a single six-sided die. In the case of a **bimodal distribution,** two distinct ranges of scores are more common than any other. A likely example of a bimodal distribution would be the heights of the athletes attending the annual sports banquet for a very large high school that has only two sports teams: women's gymnastics and men's basketball. If this example seems a little contrived, it should. Bimodal distributions are relatively rare, and they usually reflect the fact that a sample is composed of two meaningful subsamples. The third distribution depicted in Figure 1.1 is the most important. This is a **normal distribution:** a symmetrical, bell-shaped distribution in which most scores cluster near the mean and in which scores become increasingly rare as they become increasingly divergent from this mean. Many things that can be quantified are normally distributed. Distributions of height, weight, extroversion, self-esteem, and the age at which infants begin to walk are all examples of approximately normal distributions.

The nice thing about the normal distribution is that if you know that a set of observations is normally distributed, this further improves your ability to describe the entire set of scores in the sample. More specifically, you can make some very good guesses about the exact proportion of scores that fall within any given number of standard deviations (or fractions of a standard deviation) from the mean. As illustrated in Figure 1.2, about 68% of a set of normally distributed scores will fall within one standard deviation of the mean. About 95% of a set of normally distributed scores will fall within two standard deviations of the mean, and well over 99% of a set of normally distributed scores (99.8% to be exact) will fall within three standard deviations of the mean. For example, scores on modern intelligence tests (such as the Wechsler Adult Intelligence Scale) are normally distributed, have a mean of 100, and have a standard deviation of 15. This means that about 68% of all people have IQs that fall between 85 and 115. Similarly, more than 99% of all people (again, 99.8% of all people, to be more exact) should have IQs that fall between 55 and 145.

This kind of analysis can also be used to put a particular score or observation into perspective (which is a first step toward making *inferences* from particular observations). For instance, if you know that a set of 400 scores on an astronomy midterm (a) approximates a normal distribution, (b) has a mean of 70, and (c) has a standard deviation of exactly 6, you should have a very good picture of what this entire set of scores is like. And you should know exactly how impressed to be when you learn that your friend Amanda earned an 84 on the exam. She scored 2.33 standard deviations above the mean, which means that she probably scored in the top 1% of the class. How could you tell this? By

Figure 1.1 A Rectangular Distribution, a Bimodal Distribution, and a Normal Distribution

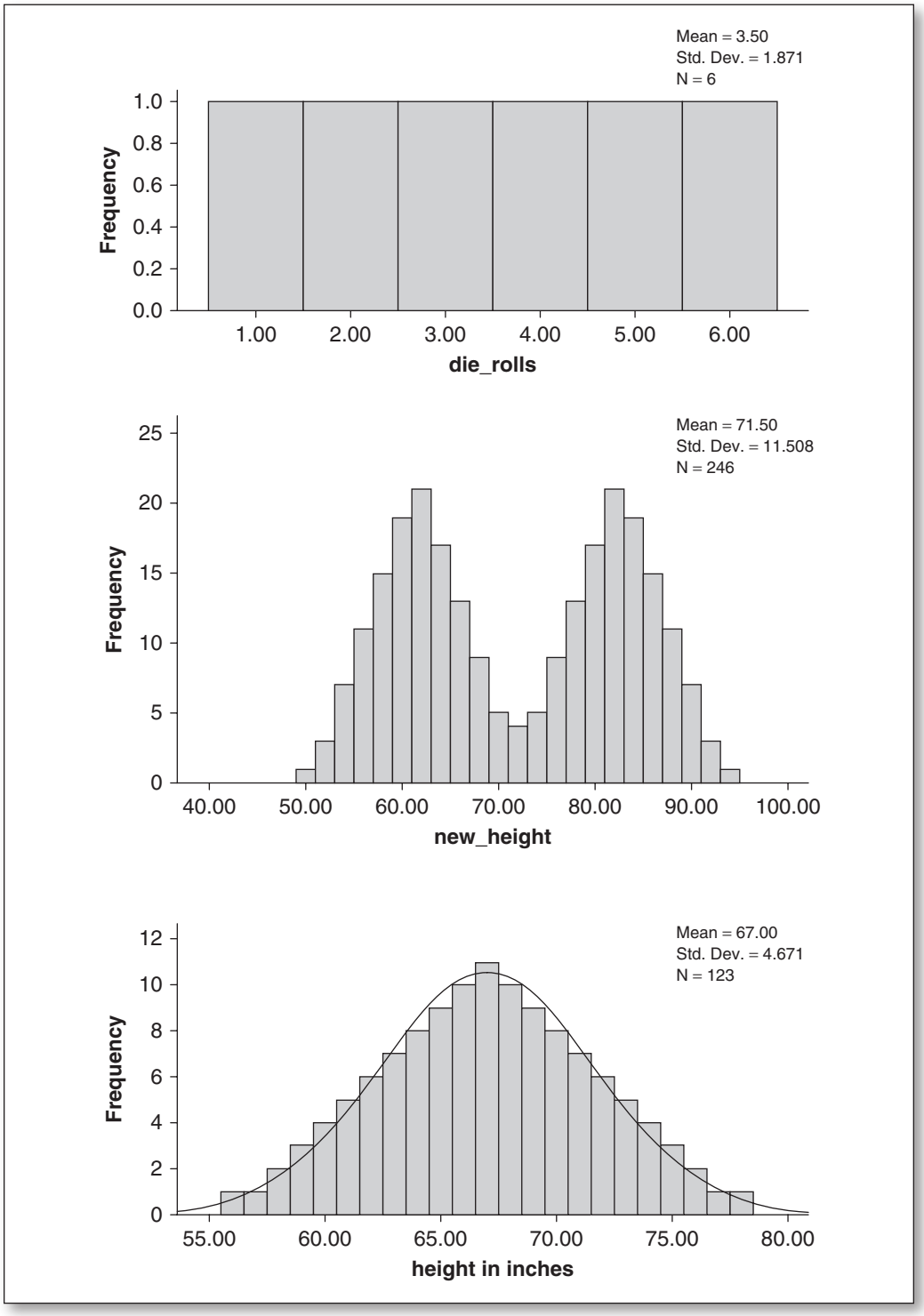
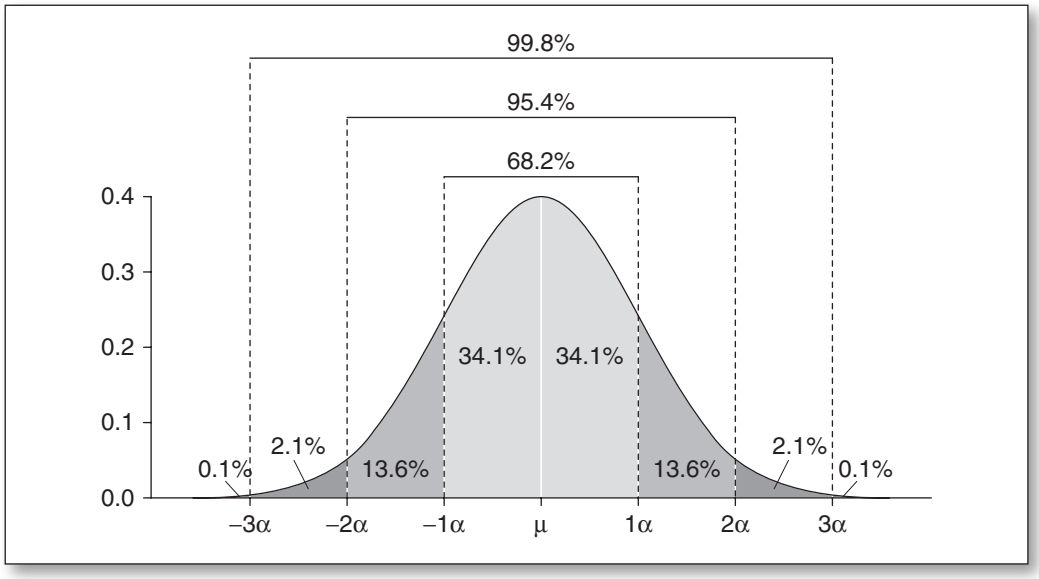


Figure 1.2 Percentage of Scores in a Perfectly Normal Distribution Falling Within 1, 2, and 3 Standard Deviations From the Mean



Source: Image courtesy of Wikipedia.

consulting a detailed table based on the normal distribution. Such a table would tell you that only about 2% of a set of scores are 2.33 standard deviations or more from the mean. And because the normal distribution is symmetrical, half of the scores that are 2.33 standard deviations or more from the mean will be 2.33 standard deviations or more *below* the mean. Amanda's score was in the half of that 2% that was well above the mean. Translation: Amanda kicked butt on the exam.

As you know if you have had any formal training in statistics, there is much more to descriptive statistics than what we have covered here. For instance, we skipped many of the specific measures of central tendency and dispersion, and we didn't describe all the possible kinds of distributions of scores. However, this overview should make it clear that descriptive statistics provide researchers with an enormously powerful tool for organizing and simplifying data. At the same time, descriptive statistics are only half of the picture. In addition to simplifying and organizing the data they collect, researchers also need to draw conclusions about populations from their sample data. That is, they need to move beyond the data themselves in the hopes of drawing general inferences about people. To do this, researchers rely on inferential statistics.

Inferential Statistics

The basic idea behind inferential statistical testing is that decisions about what to conclude from a set of research findings need to be made in a logical, unbiased fashion. One of the most highly developed forms of logic is mathematics, and statistical testing involves the use of objective, mathematical decision rules to determine whether an observed set of research findings is "real." The logic of statistical testing is largely a reflection of the skepticism and empiricism that are crucial to the scientific method. When conducting a statistical test to aid in the interpretation of a set of experimental findings, researchers begin by assuming that the **null hypothesis** is true. That is, they begin by assuming that their own predictions are *wrong*. In a simple, two-groups experiment, this would mean assuming that the experimental group and the control group are not really different after the manipulation—and that any apparent difference between the two groups is simply due to luck (i.e., to a failure of random assignment). After all, random assignment is good, but it is rarely perfect. It is always *possible* that any difference an experimenter observes between the behavior of participants in the experimental and control groups is simply due to chance. In the context of an experiment, the main thing statistical hypothesis testing tells us is exactly *how* possible it is (i.e., how likely it is) that someone would get results as impressive as, or more impressive than, those actually observed in an experiment if chance alone (and not an effective manipulation) were at work in the experiment.

The same logic applies, by the way, to the findings of *all* kinds of research (e.g., survey or interview research). If a researcher correlates a person's height with that person's level of education and observes a modest positive correlation (such that taller people tend to be better educated), it is always possible—out of dumb luck—that the tall people in this specific sample just happen to have been more educated than the short people. Statistical testing tells researchers exactly how likely it is that a given research finding would occur on the basis of luck alone (if nothing interesting is really going on). Researchers conclude that there is a true association between the variables they have manipulated or measured only if the observed association would rarely have occurred on the basis of chance.

Because people are not in the habit of conducting tests of statistical significance to decide whether they should believe what a salesperson is telling them about a new line of athletic shoes, whether there is intelligent life on other planets, or whether their friend's taste in movies is "significantly different" from their own, the concept of statistical testing is pretty foreign to most laypeople. However, anyone who has ever given much thought to how American courtrooms work should be extremely familiar with the logic of statistical testing. This is because the logic of statistical

testing is almost identical to the logic of what happens in an ideal courtroom. With this in mind, our discussion of statistical testing will focus on the simile of what happens in the courtroom. If you understand courtrooms, you should have little difficulty understanding statistical testing.

As mentioned previously, researchers performing statistical tests begin by assuming that the *null hypothesis* is correct—that is, that the researcher's findings reflect chance variation and are not real. The opposite of the null hypothesis is the **alternative hypothesis.** This is the hypothesis that any observed difference between the experimental and the control group *is* real. The null hypothesis is very much like the *presumption of innocence* in the courtroom. Jurors in a courtroom are instructed to assume that they are in court because an innocent person had the bad luck of being falsely accused of a crime. That is, they are instructed to be extremely skeptical of the prosecuting attorney's claim that the defendant is guilty. Just as defendants are considered "innocent until proven guilty," researchers' claims about the relation between the variables they have examined are considered incorrect unless the results of the study strongly suggest otherwise ("null until proven alternative," you might say). After beginning with the presumption of innocence, jurors are instructed to examine all the evidence presented in a completely rational, unbiased fashion. The statistical equivalent of this is to examine all the evidence collected in a study on a purely objective, *mathematical* basis. After examining the evidence against the defendant in a careful, unbiased fashion, jurors are further instructed to reject the presumption of innocence (to vote guilty) only if the evidence suggests *beyond a reasonable doubt* that the defendant committed the crime in question. The statistical equivalent of the principle of reasonable doubt is the **alpha level** agreed upon by most statisticians as the reasonable standard for rejecting the null hypothesis. In most cases, the accepted probability value at which alpha is set is .05. That is, researchers may reject the null hypothesis and conclude that their hypothesis is correct only when findings as extreme as those observed in the study (or more extreme) would have occurred by chance alone less than 5% of the time.

If prosecuting attorneys were statisticians, we could imagine them asking the statistical equivalent of the same kinds of questions they often ask in the courtroom: "Now, I'll ask you, the jury, to assume, as the defense claims, that temperature has no effect on aggression. If this is so, doesn't it seem like an *incredible coincidence* that in a random sample of 40 college students, the 20 students who just happened to be randomly assigned to the experimental group—that is, the 20 people who just happened to be placed in the uncomfortably hot room instead of the nice, comfortable, cool room—would give the stooge almost *three times* the amount of shock that was given by the people in the control group? Remember, Mr. Heat would have you believe that in comparison with the 20 participants in the control group, participants number 1, 4, 7, 9, 10, 11, 15, 17, 18, 21, 22, 24,

25, 26, 29, 33, 35, 36, 38, and 40, as a group, just *happened* to be the kind of people who are inherently predisposed to deliver extremely high levels of shock. Well, in case you're tempted to *believe* this load of bullsh—." "I object, your Honor! The question is highly inflammatory," the defense attorney interrupts. "Objection overruled," the judge retorts. "As I was saying, in case any one of you on the jury is tempted to take this claim seriously, I remind you that we asked Dr. R. A. Fisher, an eminent mathematician and manurist, to calculate the *exact probability* that something this unusual could happen due to a simple failure of random assignment. His careful calculations show that if we ran this experiment *thousands of times* without varying the way the experimental and control groups were treated, we would expect to observe results as unusual as these less than *one time in a thousand if the manipulation truly has no effect!* Don't you think the defense is asking you to accept a pretty incredible coincidence?"

A final parallel between the courtroom and the psychological laboratory is particularly appropriate in a theoretical field such as psychology. In most court cases, especially serious cases such as murder trials, successful prosecuting attorneys will usually need to do one more thing in addition to presenting a body of logical arguments and evidence pointing to the defendant. They will need to identify a plausible *motive,* a good reason why the defendant might have wanted to commit the crime. It is difficult to convict people solely on the basis of circumstantial evidence. A similar state of affairs exists in psychology. No matter how "statistically significant" a set of research findings is, most psychologists will place very little stock in it unless the researcher can come up with a plausible reason why one might expect to observe those findings. In psychology, these plausible reasons are called *theories.* It is quite difficult to publish a set of significant empirical findings unless you can generate a plausible theoretical explanation for them.

Having made this "friendly pass" through a highly technical subject, we will now try to enrich your understanding of inferential statistics by using inferential statistics to solve a couple of problems. In an effort to keep formulas and calculations as simple as possible, we have chosen some very simple problems. Analyzing and interpreting the data from most real empirical investigations require more extensive calculations than those you will see here, but of course these labor-intensive calculations are usually carried out by computers. In fact, a great deal of your training in this text will involve getting a computer to crunch numbers for you using the statistical software package SPSS. Regardless of how extensive the calculations are, however, the basic logic underlying inferential statistical tests is almost always the same—no matter which specific inferential test is being conducted and no matter who, or what, is doing the calculations.

PROBABILITY THEORY

As suggested in the thought experiment with American courtrooms, all inferential statistics are grounded firmly in the logic of probability theory. Probability theory deals with the mathematical rules and procedures used to predict and understand chance events. For example, the important statistical principle of regression toward the mean (the idea that extreme scores or performances are usually followed by less extreme scores or performances from the same person or group of people) can easily be derived from probability theory. Similarly, the odds in casinos and predictions about the weather can be derived from straightforward considerations of probabilities. What is a probability? From the classical perspective, the **probability** of an event is a very simple thing: It is (a) the number of all specific outcomes that qualify as the event in question divided by (b) the total number of all possible outcomes. The probability of rolling a 3 on a single roll with a standard six-sided die is 1/6, or .167, because there is (a) one and only one roll that qualifies as a 3 and (b) exactly six (equally likely) possible outcomes. For the same reason, the probability of rolling an odd number on the same die is 1/2 or .50—because three of the six possible outcomes qualify as odd numbers. It is important to remember that the probability of *any* event (or complex set of events, such as the observed results of an experiment) is the number of ways to observe that event divided by the total number of possible events.

With this in mind, suppose the Great Pumpkini told you that he had telekinetic powers that allow him to influence the outcome of otherwise fair coin tosses. How could you test his claim? One way would be to ask him to predict some coin tosses and to check up on the accuracy of his predictions. Imagine that you pulled out a coin, tossed it in the air, and asked Pumpkini to call it before it landed. He calls heads. Heads it is! Do you believe in Pumpkini's self-proclaimed telekinetic abilities? Of course not. You realize that this event could easily have occurred by chance. How easily? Fully half the time we performed the test. With this concern in mind, suppose Pumpkini agreed to predict exactly 10 coin tosses. Let's stop and consider a number of possible outcomes of this hypothetical coin-tossing test. To simplify things, let's assume that Pumpkini always predicts heads on every toss.

One pretty unremarkable outcome is that he'd make 5 of 10 correct predictions. Should you conclude that he does, indeed, have telekinetic abilities? Or that he is *half* telekinetic (perhaps on his mother's side)? Again, of course not. Making 5 of 10 correct predictions is no better than chance. To phrase this in terms of the results of the test, the number of heads we observed was no different than the *expected frequency* (the average, over the long run) of a random series of 10 coin tosses. In this case, the expected frequency is the probability of a head on a single toss (.50) multiplied by the total number of tosses (10). But what if Pumpkini made six or seven correct predictions

instead of only five? Our guess is that you still wouldn't be very impressed and would still conclude that Pumpkini does not have telekinetic abilities (in statistical terms, you would fail to reject the null hypothesis). OK, so what if he made a slightly more impressive eight correct predictions? What about nine? You should bear in mind that Pumpkini never said his telekinetic powers were absolutely flawless. Pumpkini can't *always* carry a glass of water across a room without spilling it, but his friends usually allow him to carry glasses of water unassisted. Despite your firmly entrenched (and justifiable) skepticism concerning psychic phenomena, we hope you can see that as our observations (i.e., the results of our coin-tossing test) depart further and further from chance expectations, you would start to become more and more convinced that something unusual is going on. At a certain point, you'd practically be forced to agree that Pumpkini is doing *something* to influence the outcome of the coin tosses.

The problem with casual analysis is that it's hard to know exactly *where* that certain point is. Some people might be easygoing enough to say they'd accept eight or more heads as compelling evidence of Pumpkini's telekinetic abilities. Other people might ask to see a perfect score of 10 (and still insist that they're not convinced). After all, extraordinary claims require extraordinary evidence. That's where inferential statistics come in. By making use of (a) some basic concepts in probability theory, along with (b) our knowledge of what a distribution of scores should look like when nothing funny is going on (e.g., when we are merely flipping a fair coin 10 times at random, when we are simply randomly assigning 20 people to either an experimental or a control condition), we can use inferential statistics to figure out exactly how likely it is that a given set of usual or not-so-usual observations would have been observed by chance. Unless it is pretty darn *un*likely that a set of findings would have been observed by chance, the logic of statistical hypothesis testing requires us to conclude that the set of findings represents a chance outcome.

To return to our coin-tossing demonstration, just how likely *is* it that a person would toss 9 or more heads by chance alone? One way to figure this out is to use our definition of probability and to figure out (a) all the specific ways there are to observe 9 or more heads in a string of 10 coin tosses and (b) all the specific outcomes (of any kind) that are possible for a string of 10 coin tosses. If we divide (a) by (b), we should have our answer. Let's begin with the number of ways there are to toss 9 or more heads. At the risk of sounding like the announcer who was describing LeBron James's scoring history without using statistics, notice that one way to do it would be to toss a tail on the first trial, followed by 9 straight heads. A second way to do it would be to toss a head on the first trial and a tail on the second trial, followed by 8 straight heads. If you follow this approach to its logical conclusion, you should see that there are exactly 10 specific ways to observe exactly 9 heads in a string of 10 coin tosses. And in case you actually want to see the 10 ways right in front of you, they

appear in Table 1.1—along with all of the unique ways there are to observe exactly *10* heads. As you already knew, there is only one of them. However, it's important to include this one in our list because we were interested in all of the specific ways to observe 9 *or more* heads in a series of 10 coin tosses.[5] So there are 11 ways.

But how many total unique outcomes are there for a series of 10 coin tosses? To count all of these would be quite a headache. So we'll resort to a less painful headache and figure it out logically. How many possible ways are there for 1 toss to come out? Two: heads or tails—which turns out to be 2^1 (2 to the first power). How about 2 tosses? Now we can observe 2^2 (2×2) or four possible ways—namely,

HH, HT, TH, or TT.

What about three tosses? Now we have 2^3 ($2 \times 2 \times 2$), or eight possible ways:

HHH, HHT, HTH, THH, HTT, THT, TTH, or TTT.

Notice that our answers always turn out to be 2 (the number of unique outcomes for an individual toss) raised to some *power*. The power to which 2 is raised is the number of trials or specific observations we are making. So the answer is 2^{10} ($2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2$) or 1,024 possible unique outcomes for a series of 10 coin tosses. This value of 1,024 includes every possible number of heads (from 0 to 10) and every possible order or position (1st through 10th) for all of these possible numbers of heads. So now we have our probability. The probability of observing 9 or more heads in a series of 10 truly random coin tosses is thus 11/1,024, or .011. So for every hundred times we conducted our coin-tossing study, you'd expect to see 9 or more heads only about once. That's only 1% of the time, and it's pretty impressive. (In fact, it's exactly as impressive as Amanda's score on the astronomy midterm, and we, for what it's worth, were very impressed with Amanda.) So if we had treated the study like a real experiment, if we had set alpha at .05, and if we had observed 9 heads, we would have had to conclude that Pumpkini does, in fact, possess the ability to influence the outcomes of otherwise fair coin tosses.

Now perhaps you're the literal type who is saying, "But wait a minute. I still wouldn't believe Pumpkini has telekinetic abilities, and I certainly don't think most scientists would, either." You are correct, of course, because the theory that you have been asked to accept flies in the face of everything you know about psychology and physics. A much more reasonable explanation for the observed findings is that Pumpkini has engaged in some form of trickery, such as using a biased coin. However, this simply means that, like any scientific practice, the practice of conducting statistical tests must be carried out using a little common sense. If someone is

Table 1.1 All the Possible Ways to Toss Nine or More Heads in 10 Tosses of a Fair Coin: A Single Tail Can Come on Any of the 10 Trials, or It Can Never Come at All

1.	**T**HHHHHHHHHH
2.	H**T**HHHHHHHHH
3.	HH**T**HHHHHHHH
4.	HHH**T**HHHHHHH
5.	HHHH**T**HHHHHH
6.	HHHHH**T**HHHHH
7.	HHHHHH**T**HHHH
8.	HHHHHHH**T**HHH
9.	HHHHHHHH**T**HH
10.	HHHHHHHHH**T**H
11.	HHHHHHHHHH**T**
	HHHHHHHHHHH

making a truly extraordinary claim, we might want to set alpha at .001, or even .0001, instead of .05. Of course, setting alpha at a very low value might require us to design a test with a much greater number of coin tosses (after all, 10 out of 10 tosses—the *best* you can possibly do—has a probability higher than .0001; it's 1/1,024, which is closer to .001), but the point is that we could easily design the test to have plenty of power to see what is going on. The exact design of our study is up to us (and, to some extent, to our critics). If people are sufficiently skeptical of a claim, they might also want to see a *replication* of a questionable or counterintuitive finding. If Pumpkini replicated his demonstration several times by correctly predicting 9 or more heads, and if we enacted some careful control procedures to prevent him from cheating (e.g., we let a group of skeptics choose and handle the coins), even the most ardent anti-telekinetician should eventually be persuaded. And if he or she weren't, we would argue that this person wasn't being very scientific.

The logic of the coin-tossing experiment is the same as the logic underlying virtually all inferential statistical tests. First, a researcher makes a set of observations. Second, these observations are compared with what we would expect to observe if nothing unusual were happening in the experiment (i.e., if the researcher's hypothesis were incorrect). This comparison

is ultimately converted into a *probability*—namely, the probability that the researcher would have observed a set of results at least this consistent with his or her hypothesis if the hypothesis were incorrect. Finally, if this probability is sufficiently low, we conclude that the researcher's hypothesis is probably correct. Because inferential statistics are a very important part of the research process, let's look at another highly contrived but informative question that could be answered only with the use of inferential statistics.

A STUDY OF CHEATING

Suppose we offered a group of exactly 50 students the chance to win a very attractive prize (say, a large amount of cash, or an autographed copy of this textbook) by randomly drawing a lucky orange ping-pong ball out of a large paper bag. Assume that each student gets to draw only one ball from the bag, that students return the drawn balls to the bag after each drawing, and that the bag contains exactly 10 balls, only 1 of which is orange. Because our university is trying to teach students the values of honesty and integrity, university regulations require us to administer the drawing on an honor system. Specifically, the bag of ping-pong balls is kept behind a black curtain, and students walk behind the curtain—one at a time, in complete privacy—to draw their balls at random from the bag. After drawing a ball, each student holds it up above the curtain for everyone else to see. Anyone who holds up an orange ball is a winner.

Suppose that we're the curious types who want to find out if there was a significant amount of cheating (peeking) during the drawing. At first blush, it would seem like there's nothing we could do. Unless we engage in a little cheating ourselves (e.g., by secretly videotaping the drawings), how can we figure out whether people were peeking as they selected their balls? We're at a complete loss to observe the unobservable—*unless* we rely on inferential statistics. By using inferential statistics, we could simply calculate the number of winners we'd expect to observe if *no one* was cheating. By making a comparison between this expected frequency and the number of winners we actually observed in our drawing, we could calculate the exact probability (based on chance alone) of obtaining a result as extreme as, or more extreme than, the result of our actual drawing. If the probability of having so many winners were sufficiently low, we might reluctantly reject the null hypothesis (our initial assumption that the students were all innocent until proven guilty) and conclude that a significant amount of cheating was happening during the drawing.

Let's find out. To begin with, we need to assume that our suspicions about cheating are completely unfounded and that no one peeked (as usual, we begin by assuming the null hypothesis). Assuming that no one was peeking, what's your best guess about how many of the 50 students should have selected a winning ball? If you are a little fuzzy on your probability

theory, remember that you can figure out the expected frequency of an event by multiplying (a) the probability of the event on a single trial by (b) the total number of trials in the series of events. This is how we knew that 5 was the expected number of observed heads in a series of 10 coin tosses. It was $.50 \times 10$. The answer here is also 5 (it's $.10 \times 50$). Now imagine that we had 6 winners. Or 9 winners, or 15—or 50. Hopefully, you can see, as you did in the coin-tossing study, that as our observed frequencies depart further and further from the frequency we'd expect by chance, we become more and more strongly convinced that our observed frequencies are *not* the product of chance.

For the purposes of actually seeing some inferential statistics in action, let's assume that we had exactly 10 winners in our drawing. Because our outcome was a categorical outcome ("success" or "failure" at the draw), and because we had a pretty large sample, we'd probably want to conduct a χ^2 (chi-square) test on these data. The formula for this test appears as follows:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}.$$

Recall that Σ (sigma) is a *summation sign* that tells you to add together all the appropriate examples of the basic calculation.

f_o refers to the *observed frequencies* of each of the events you care about (successes and failures when it comes to sampling a lucky orange ball).

f_e refers to the *expected frequencies* for each of these same events.

You could think of a χ^2 statistic as a "surprise index." Notice that the most important thing the formula does is to *compare* expected and observed frequencies. Specifically, expected frequencies are compared with (i.e., subtracted from) observed frequencies, and then a couple of simple transformations are made on these difference scores. The more our observed frequencies depart from what you'd expect if chance alone were operating (i.e., the more surprising our results are), the bigger our χ^2 statistic becomes. And as our χ^2 statistic grows, it tells us that it's less and less likely that we're observing a chance process (and, in this case, more and more likely that we're observing cheating).

The χ^2 value for 10 winners (out of 50) when only 5 were expected is computed as follows:

$$\chi^2 = \frac{(10 - 5)^2}{5} + \frac{(40 - 45)^2}{45}.$$

The 10 in the first half of the equation is the *observed* frequency of successes, and the two 5s both refer to the *expected* frequency of successes.

The 40 in the second half of the equation is the *observed* frequency of failures, and the two 45s both refer to the *expected* frequency of failures (this has to be the sample size, which is 50, minus the expected number of successes). When we do the math, we get 25/5 + 25/45, which works out to 5.55. Notice that this *isn't* a probability. The way most inferential statistics work is that you generate both the statistic itself (e.g., a correlation coefficient, a *t* value, an *F* ratio) and then use the exact value of the observed statistic to determine a probability value (one that corresponds to the value of your statistic). If you are doing your calculations on a computer, the software program you are using will always do this for you. That is, it will give you the exact *p* value (i.e., the exact probability) that corresponds to your results after they have been converted to the unambiguous language of your statistic. However, if you are doing your calculations by hand, as we have here, you will need to consult some kind of statistical table to see what the *critical values* are for your statistic. In the case of our study of cheating, the critical χ^2 value that corresponds to an alpha level of .05 is 3.841. Any χ^2 value that exceeds this score will have an associated *p* value that is lower than .05 and will thus be significant when alpha is set at .05. If we were a little bit more stringent, we might set alpha at .02 or .01. Our χ^2 table happens to include critical values for each of these levels of alpha (i.e., for each of these probability values). In a study such as ours, the critical χ^2 value for an alpha of .02 is 5.412, and the critical χ^2 value for an alpha of .01 is 6.635. By these criteria, our result is still significant even if alpha is set at .02. However, if we move to the still more stringent alpha level of .01, the number of winners we observed would no longer be significant (because we're effectively saying that it'd take more than 10 winners to convince us).

Suppose we followed standard practice and set alpha at .05. We'd have to conclude that some people cheated. Notice, however, that we couldn't draw any safe inferences about exactly *who* cheated. Presumably *about* 5 of our 10 winners just got lucky, and *about* 5 cheated. Realizing that *only* about 5 people cheated provides a different sort of perspective on our findings. Specifically, it highlights the fact that there is often more than one way to look at a set of observations. Notice also that an alternate, and equally correct, perspective on our observation is that people are significantly honest! It appears to be the case that about 45 of our 50 students were completely honest—even in a situation that allowed rampant cheating. Why did we say 45? Because we just decided that only about 5 people are likely to have cheated. In light of how hard it is to win the game by playing fairly, these 5 or so cheaters led to a significant amount of cheating. However, if we had started out with the hypothesis that 49 of 50, or 98%, of all people should be expected to cheat under these conditions, and if we had taken 49 (nearly absolute dishonesty) as our standard of comparison

rather than 5 (absolute honesty), we would have obtained an *extremely* large χ^2 value:

$$\chi^2 = \frac{(10 - 49)^2}{49} + \frac{(40 - 1)^2}{1},$$

which is 1552.04, and which corresponds to an infinitesimally small *p* value. Even if we set alpha at a very, very, very low level (say one in a billion, or .000000001), this would still be significant. In other words, it's important to keep in mind that we appear to have observed a lot more honesty than cheating.[6]

A final aspect of this exercise about drawing ping-pong balls from a bag is that it provides a useful metaphor for thinking about what researchers do when they draw inferences about people in their research. Notice that in the lottery involving ping-pong balls, we could not directly observe the phenomenon in which we were interested. The activities we cared about were shrouded behind a black curtain—just as the activities that psychologists often care about (e.g., dissonance reduction, feelings of passionate love, parallel distributed representations of language) are hidden inside the black box of people's minds. Inferential statistics work hand in hand with things like operational definitions to allow us to make scientific inferences. Operational definitions allow us to draw inferences about *processes* that we cannot observe (those that occur inside the person), and inferential statistics allow us to draw inferences about *people* we can't observe (those we didn't sample in our study). When we conclude that a research finding is significant, we are concluding that it is real and thus that it applies to people who did not take part in our study. This is one sense in which the ping-pong ball demonstration is a little different from most significance tests. Although it would probably be safe to generalize our findings about cheating to other college students, what we really cared about most in this particular test was finding out what was going on in *our particular* sample.

Virtually every inferential statistic that you will ever come across will be based on the logic that was explicated here. Of course, the particular distributions of responses that researchers examine vary enormously from one study to the next, and this, among other things, influences the particular statistics that researchers use to summarize and draw inferences about their data. Moreover, once a researcher has chosen a particular statistic, the specific calculations that she or he will have to carry out (or get a computer to carry out) will typically be a good bit more involved than those you have seen here. For example, in a two-way analysis of variance (ANOVA), there are separate calculations (and separate *degrees of freedom*) for each of the two possible main effects as well as for the two-way interaction. No matter what statistics they are computing, however,

researchers will always rely on the logic of probability theory to help them make their case that something significant is at the root of their empirical observations.

Things That Go Bump in the Light: Factors That Influence the Results of Significance Tests

ALPHA LEVELS AND TYPE I AND II ERRORS

Now that you have a better feel for what it means for a research finding to be statistically significant, we feel that it is our duty to warn you that when we look at significance testing in the cold, hard light of day, it has a couple of limitations. In other words, there are a few things that can go wrong when people are conducting statistical significance testing. First of all, it is important to remember that when a researcher conducts a statistical test and obtains a significant result, this does not *always* mean that his or her hypothesis is correct. Even if a study is perfectly executed with no systematic design flaws, it is always possible that the researcher's results *were* due to chance. In fact, the *p* value we observe in an experiment tells us exactly how likely it is that we would have obtained results like ours even if nothing but dumb luck were operating in our study. Statisticians refer to this worrisome possibility—incorrectly rejecting the null hypothesis when it is, in fact, correct—as a **Type I error.** The likelihood of making a Type I error is a direct function of where we set our alpha level. As suggested earlier, if we think it would be a practical or scientific disaster to reject the null hypothesis in error, we might want to set alpha at a very conservative level, such as .001. Then we would be taking only one chance in a thousand of falsely rejecting the null hypothesis.

So why not set alpha at .001 (or even lower) all the time? Because we have to strike a balance between being cautious and being so cautious that we become downright foolish. In statistical terms, if we always set alpha at an extraordinarily low level, we would decrease the likelihood of committing a Type I error at the expense of increasing the likelihood of committing a **Type II error.** A Type II error occurs when we fail to reject an incorrect null hypothesis—that is, when we fail to realize that our study has revealed something meaningful (usually that our hypothesis is correct). The reason it is useful to know about Type I and Type II errors is that there are things we can do to minimize our chances of making both of these troublesome mistakes. As suggested previously, one of the easiest ways to minimize Type I errors is to set alpha at a pretty low level. Over the years, most researchers have pretty well agreed that .05 is a reasonable level for alpha (i.e., a reasonable risk for making a Type I error). And of course, if we want to be a little more cautious, but we don't want to ask

anyone to adjust any alpha levels, we can always insist on seeing a replication. In the grand scheme of things, replications are what tell us whether an effect is real.

EFFECT SIZE AND SIGNIFICANCE TESTING

Although no one wants to make a Type I error, no one really wants to make a Type II error either. Several things influence the likelihood that a researcher will make a Type II error (and fail to detect a real effect). Some of these are things over which researchers have little or no control, and some of them are things over which researchers have almost complete control. One thing that researchers can't do too much about is their "effect size," the magnitude of the effect in which they happen to be interested. If you collected a sample of 20 people and measured their heights and their foot sizes, you could probably expect to observe a statistically significant correlation between height and foot size, even though your sample was pretty small. This is because there is a pretty robust tendency for big people to have big feet. Of course there are exceptions, but they are relatively rare. We doubt that you will ever meet a gymnast who squeezes into a size 14 (or an NBA center who slips comfortably into a size 9). On the other hand, if you gave a sample of 20 people a measure of extraversion and a measure of self-esteem, you might not necessarily observe a significant correlation. Although self-esteem and extraversion do tend to go hand in hand, this correlation is much more modest than the substantial correlation between height and foot size. To return to our example about peeking and ping-pong balls, it would have been much easier to detect an effect of cheating if cheating had been rampant. In fact, notice that in this study, it was quite easy to detect an effect of honesty—precisely because honesty was so rampant.

MEASUREMENT ERROR AND SIGNIFICANCE TESTING

Although it's obviously impossible to change the true size of an effect, one thing that researchers can sometimes do to maximize their chances of detecting a small effect is to conduct a within-subjects or repeated measures study. As we argue later in this text, within-subjects designs are usually more sensitive than between-subjects designs. One of the reasons this is the case is that within-subjects designs cut down on extraneous sources of variability that can mask an effect. A person in a cool room might deliver high levels of shock to a confederate just because this person happens to be an unusually aggressive person. However, if we could observe the behavior of the *same* person in both a hot and a cool room (and if we

could make sure the person didn't know that she or he was being studied), we would presumably see that the person would deliver even higher levels of shock when the temperature was cranked up a bit. Of course, another reason why within-subjects designs are more powerful than between-subjects designs is that they simply increase the number of observations in a study. If we measure the aggressive behavior of each of our 20 participants in both a hot and a cool room, it is almost as if we had 40 participants in our study rather than 20 (see Pelham, 1993, for a further discussion of the advantages of within-subjects designs).

SAMPLE SIZE AND SIGNIFICANCE TESTING

When researchers are unable to make use of within-subjects designs, they can still do a couple of things to maximize their chances of detecting a real effect. One simple, albeit potentially expensive, thing that researchers can do is to conduct studies with a lot of participants. Increasing your sample size in a study (whether it be an experiment, a quasi-experiment, a survey, or an archival study) can greatly increase the chances that you will detect a real effect. For example, suppose that the true correlation between extroversion and self-esteem among American adults is exactly .32. And suppose that you conducted a survey of 27 randomly sampled American adults and observed a correlation of exactly .32 in your study. Would this be statistically significant? Unfortunately not. In a sample of only 27 people, a correlation of .32 would have a *p* value slightly greater than .10—at best a marginally significant value. On the other hand, if you had sampled 102 people rather than 27, and if you happened to hit the nail on the head again by observing another correlation of exactly .32, this result would be significant even if you had set alpha at .001. That's because when you have a sample as large as the second, it's quite unusual to observe a correlation as large as .32 when the two variables in question are actually unrelated. If this doesn't quite seem right to you, consider your own intuitive conclusions when we asked you earlier what you'd think if Pumpkini were able to correctly predict 6 heads in 10 coin tosses. If he produced exactly the same proportion of heads (600) in 1,000 tosses, you should be much more impressed.

RESTRICTION OF RANGE AND SIGNIFICANCE TESTING

Limits in the *range or variability* of the variables you are measuring or manipulating (i.e., restriction of range) can also limit your ability to detect a true effect. Wording your dependent measures carefully, choosing the right population, and making sure that your independent variable is as potent and meaningful as possible (which means not shooting yourself in the foot by artificially *diminishing* your real effect size) are all potential

solutions to the problem of restriction of range, and thus, they are all potential solutions to the problem of avoiding Type II errors. The particular statistical analysis that you conduct can also play an important role in whether your research findings are significant. When you have a choice between conducting a powerful test (one that can detect even relatively small effects) and a less powerful test, you should always perform the more powerful of the two. For example, performing a correlation between two continuous variables (e.g., self-esteem and the number of minutes people spent reading positive feedback about themselves) is usually more powerful than performing a median split (e.g., on self-esteem) and then conducting an ANOVA or *t* test to see if the mean difference between the low group and the high group is significant. Similarly, making use of continuous ("How much did you like your partner?") rather than dichotomous ("Did you like your partner?") dependent measures in an experiment usually allows for more powerful statistical tests. As a second example, when you have a choice of conducting more than one separate between-subjects analysis (e.g., three different between-subjects ANOVAs, one on each of your three different dependent measures) versus a single within-subjects or mixed-model analysis on the same set of research findings (e.g., because you asked people to rate a target person on positive, neutral, and negative traits), you will usually be better served by the analysis that incorporates the within-subjects aspect of your design.

The issues discussed here can help you to conduct better research studies. Just as important, they can also help you to better interpret the findings of other people's studies. For example, if a team of experimenters claims that they failed to replicate an important effect, you would do well to ask a few questions about the nature of their manipulation, the nature of their sample, the wording of their dependent measures, and the number of participants they included in their between- or within-subjects study before you abandon your own research on the same topic. If Dr. Snittle noted that he failed to replicate Phillips's archival research on suicide by noting that none of the 23 people in his small Nebraska farming community committed suicide after reading about a front-page suicide, this wouldn't be much cause for concern. However, if Dr. Snittle learned to speak fluent Mandarin, traveled to China, gained access to media and suicide records in several very large Chinese provinces, duplicated Phillips's analytical strategies perfectly, and failed to replicate some aspect of Phillips's findings, we'd want to figure out why. Perhaps some aspect of Chinese culture (or Chinese media coverage) is responsible for the difference. This way of thinking about how to interpret statistics is consistent not only with common sense but also with the logic of the scientific method. It is important to remember that statistics are simply a tool. When effectively applied to an appropriate problem, statistics can be incredibly powerful and effective. However, when misapplied or misinterpreted, statistics—like real tools—can be useless or even dangerous.

The Changing State of the Art: Alternate Perspectives on Statistical Hypothesis Testing

During the past three quarters of a century, statistical hypothesis testing has become a methodological touchstone for evaluating specific research findings. When a provocative research finding proves to be statistically significant, it is considered scientifically meaningful. When an equally provocative research finding proves to be nonsignificant, it usually is not taken seriously in scientific circles. As we have just seen, however, an absolute reliance on significance testing—when divorced from basic considerations involving things such as effect size or sample size—can often lead researchers to inappropriate conclusions. Another way of putting this is that there is more to hypothesis testing than simple significance testing. Critics of significance testing have pointed out, for example, that even when a study is well designed, basing a decision about whether an effect is real solely on the basis of statistical "significance" is not always advisable. In actual practice, for example, when a researcher conducts a study whose results are promising but not significant, the researcher will often run additional participants—or modify the design and run the study again—rather than concluding that the original hypothesis is incorrect. In fact, some researchers have argued that the traditional use of significance testing is an inherently misleading process that should be abandoned in favor of other approaches (J. Cohen, 1994).

Although it seems unlikely that significance testing will be abandoned any time in the near future, most researchers would probably agree that it is often useful to complement significance testing with other indicators of the validity, meaningfulness, or repeatability of an effect. A complete review of the pros and cons of alternate approaches to significance testing is beyond the scope of this book. However, it is probably worth noting that researchers have recently begun to complement significance testing by making use of special statistics to assess the practical or theoretical meaningfulness of research findings. One way in which researchers have done this is to compute estimates of **effect sizes,** that is, indicators of the *strength* or magnitude of their effects. A second way is to compute estimates of (a) the overall amount of existing support for an effect or (b) the consistency or repeatability of the effect. The statistical approach most suited to this second category of questions is referred to as **meta-analysis.**

ESTIMATES OF EFFECT SIZE

When researchers want to assess the practical or theoretical rather than the statistical significance of a specific research finding—that is, when they want to know how big or meaningful an effect is—they typically calculate

an **effect size.** Although there are many useful indicators of effect size, the two most commonly reported indicators of effect size are probably *r* and *d.* The statistic *r* is the familiar correlation coefficient, and thus, you probably have had some practice interpreting this frequently used indicator of effect size. Psychological effects that are considered small, medium, and large correspond respectively to correlations of about .10, .30, and .50. The less familiar statistic *d* is more likely to be used to describe effect sizes from experiments or quasi-experiments because *d* is based on the difference between two treatment means. Specifically,

$$d = (\text{mean } 1 - \text{mean } 2)/D,$$

where *D* is simply the overall standard deviation of the dependent measure in the sample being studied (see Rosenthal & Rosnow, 1991, p. 302). Thus, *d* tells us *how different two means are in standard deviation units* (or fractions thereof). Because two means in a study can sometimes be more than one standard deviation apart, this means that *d,* unlike *r,* can sometimes be larger than 1. Otherwise, the interpretation of *d* is pretty similar to the interpretation of *r.* The respective values of *d* that correspond to small, medium, and large effects are about .20, .50, and .80 (see Rosenthal & Rosnow, 1991, p. 444).

Notice that we used the word *about* when we listed the specific values of *r* and *d* that correspond to different effect sizes. The reason we did so is that what makes an effect big or small is partly a judgment call. Moreover, how "big" an effect must be to qualify as meaningful varies quite a bit from one research area to another. If a cheap and easy-to-administer treatment (e.g., a daily vitamin C tablet) could reduce the risk of cancer and turned out to have a "small" effect size (e.g., $r = .10$ or less), this could easily translate into millions of saved dollars in medical expenses (and thousands of saved lives). Moreover, as we just noted, the size of an effect that researchers observe in a particular study is as much a function of how carefully the study is crafted as it is a function of the state of the world. Thus, considerations of effect size, like considerations of statistical significance, should reflect the theoretical or the practical significance of a given finding—regardless of its absolute magnitude. If our easy-to-administer experimental treatment gets blood from only 10% of the turnips that we treat, we will have to consider the relative value of blood and turnips before deciding how meaningful the treatment is.

For many years, when researchers wanted to know how strongly two variables were related, they would compute a **coefficient of determination** by squaring the correlation associated with a particular effect. So if researchers learned, for example, that people's attitudes about a politician correlated .40 with whether people voted for that politician, the researchers might note that attitudes about candidates account for only 16% of the variance in voting behavior ($.40 \times .40 = .16$, or 16%). Although this is a technically accurate way of summarizing the association between two variables, some researchers have noted that it provides a misleading picture of

the true strength of the relation between two variables. In particular, Rosenthal and Rubin (1982) developed the **binomial effect size display** as a more intuitive way to illustrate the magnitude and practical importance of a correlation. The binomial effect size display is referred to as binomial because it makes use of variables that can take on only two values (success or failure, survival or death, male or female) to illustrate effect sizes. As matters of convenience and simplicity, Rosenthal and Rubin demonstrate effect sizes using two dichotomous variables whose two values are equally likely. To simplify matters further, they express binomial effect sizes using samples in which exactly 100 people take on each of the two values of each of the two dichotomous variables.

Consider a hypothetical example involving attendance at a review session and performance on a difficult exam. Assume (a) that exactly 100 of 200 students attended the review session and (b) that exactly 100 of 200 students passed the exam. If we told you that the correlation between attending the review session and passing the exam was .20 (meaning that attendance at the review session accounts for only 4% of the variance in exam performance), you might not bother to attend the review session. However, if you examine the binomial effect size display that appears in Table 1.2, you can see that a correlation of .20 corresponds to 20 more people passing than failing the exam in the group of attendees (and 20 more people failing than passing the exam in the group of nonattendees). *More generally, when summarized using a binomial effect size display, a correlation coefficient corresponds to the difference in success rates that exists between two groups of interest on a dichotomous outcome.* If the correlation summarized in Table 1.2 had been .40, we would have seen that 70% of those attending the review (and only 30% of those failing to attend) passed the exam (70 – 30 = 40). Similarly, if we had observed a potential cookie thief for 200 days, if the person had been present in the kitchen for exactly 100 of the 200 days, and if cookies had disappeared on exactly 100 of the 200 days, then a correlation of .66 would mean that when the potential thief visited the kitchen, cookies disappeared on 83 out of 100 days (83 – 17 = 66). Even though the presence of this person accounts for only about 44% of the variance in cookie thefts ($.66^2 = .436$), notice that cookies are almost five times more likely to disappear when the person is present than when the person is absent ($.17 \times 5 = .85$).

Regardless of what format researchers use to illustrate effect sizes, reporting effect sizes provides a very useful complement to traditional significance testing. For example, suppose we know that the effect size for a specific research finding corresponds to a d of .43. If a researcher claims that he failed to replicate this finding, it would be useful to consider the effect size the researcher observed (rather than focusing solely on his observed p value) before concluding that his finding is different from the original (see Rosenthal & Rosnow, 1991, for a much more extensive discussion). In some cases, researchers have claimed that they failed to

Table 1.2 Performance on an Exam as a Function of Attendance at a Review Session

	Exam Performance		
Attendance at review	**Passed**	**Failed**	**Total**
Attended	60	40	100
Did not attend	40	60	100
Total	100	100	200

replicate findings when they observed effects that were just as large as, or larger than, those observed by previous researchers (e.g., when the second group of researchers had a much smaller sample than the first).

META-ANALYSIS

Estimates of effect size, such as *r* or *d,* provide a useful metric for describing and evaluating the magnitude of specific research findings. Regardless of how big a specific finding is, however, researchers are often interested in questions that have to do with the consistency or repeatability of the finding. Questions about the repeatability of a finding almost always have to do with a *group* of studies (and perhaps even an entire literature) rather than a single specific study. How many failed studies would have to exist to indicate that a set of findings is a statistical fluke rather than a bona fide phenomenon? If a phenomenon is bona fide, how consistently has it been observed from study to study? Even more important, what are the limiting conditions of the effect? That is, when is the effect most and least likely to be observed? Questions such as these can rarely be answered by any single study. Instead, researchers need systematic ways to summarize the findings of a *large number of studies.*

Fortunately, researchers have developed a special set of statistical techniques to summarize and evaluate entire sets of research findings. Not surprisingly, R. A. Fisher (1938), the person who popularized modern statistical testing, was one of the first researchers to address the question of how to combine the results of multiple studies. In the days since Fisher offered his preliminary suggestions, researchers have developed a wide array of techniques for summarizing and evaluating the results of multiple studies (see Rosenthal & Rosnow, 1991, for an excellent conceptual and computational review of such techniques). Statistical techniques that are designed for this purpose are typically referred to as *meta-analytic* techniques. The more commonly used term **meta-analysis** thus refers to the use of such techniques to analyze the results of *studies* rather than the responses

of individual *participants.* From this perspective, meta-analyses are to groups of *studies* what traditional statistical analyses are to groups of specific *participants.* Literally, meta-analysis refers to the analysis of analyses.

Prior to the development of meta-analysis, the only way researchers could summarize the results of a large group of studies was to logically analyze and verbally summarize all the studies. Meta-analyses complement such potentially imprecise analyses by providing precise mathematical summaries of different aspects of a set of research findings. For example, a meta-analysis of effect sizes can provide a good estimate of the average effect size that has been observed in all the published studies on a specific topic. Other meta-analytic techniques can be used to indicate how much *variability* in effect sizes has been observed from study to study on a specific topic (see Hedges, 1987). Finally, meta-analysis can be used to determine the kinds of studies that tend to yield especially large or small effect sizes (e.g., studies that did or did not make use of a particular control technique, studies conducted during a particular historical era, studies conducted in a particular part of the country). This final kind of meta-analysis can provide very useful theoretical and methodological information about the nature of a specific research finding.

As an example of this third approach, consider a couple of meta-analyses conducted by Alice Eagly. Eagly (1978) analyzed findings from a large number of studies of the effects of gender on conformity and social influence. Many researchers had argued that women are more easily influenced than men are. When Eagly looked at studies published prior to 1970 (i.e., prior to the beginning of the women's movement), this is exactly what she found. However, when she focused on studies published during the heyday of the women's movement (during the early to mid-1970s), Eagly observed very little evidence that women were more easily influenced than men. Furthermore, in a second meta-analysis, Eagly and Carli (1981) found that (a) the gender of the researcher conducting the study and (b) the specific topic of influence under investigation were good predictors of whether women were more conforming than were men. When studies were conducted by men or when the topic of influence was one with which women were likely to be unfamiliar (e.g., football), most studies showed that women were more conforming than were men. However, when studies were conducted by women or when the topic of influence was one with which men were likely to be unfamiliar (e.g., fashion), men often proved to be more conforming than were women.

Although meta-analysis may be used for many different purposes, the biggest contribution of meta-analysis to psychological research is probably an indirect one. The growing popularity of meta-analytic techniques has encouraged researchers to think about research findings in more sophisticated ways. Specifically, instead of treating alpha as an infallible litmus test for whether an effect is real, contemporary researchers are beginning to pay careful attention to the question of *when* a given effect is most (and least) likely to be observed. Ideally, when a meta-analysis suggests that an effect is magnified or diluted under certain conditions, researchers should

conduct a study in which they directly manipulate these conditions. Doing so boils down to designing *factorial* studies in which at least two independent variables are completely crossed. An example would be a single experiment on persuasion that randomly assigned half of all men and half of all women to read about a stereotypically masculine topic before seeing how much they conform to others' opinions on this topic. Of course, the other half of men and the other half of women would be randomly assigned to read about a stereotypically feminine topic before the researchers assessed conformity on this topic. If the results of the experiment confirmed the results of the meta-analysis, researchers could be even more confident of the conclusion suggested by the meta-analysis.

Summary

This chapter provided a brief review of statistics. We noted that statistical procedures can be broken down into *descriptive statistics* and *inferential statistics*. As the name suggests, descriptive statistics simply describe (i.e., illustrate, summarize) the basic properties of a set of data. Along these lines, measures of central tendency describe the typical or expected score in a given data set. In contrast, measures of dispersion reveal how much the entire set of scores varies around the typical score. The most common measures of central tendency are the mean, the median, and the mode, and the most common measures of dispersion are the range and standard deviation. Of course, psychologists interested in testing psychological theories are typically interested in inferential statistics as well as descriptive statistics. This second branch of statistics applies probability theory to determine whether and to what degree an observed data pattern truly differs from a chance pattern. Inferential statistics thus provide a basis for determining whether an observed research finding reveals a systematic association that is likely to be true in a population of interest or whether it merely reflects noise or error. For instance, if a treatment group differs from a control group to a degree that would not be expected by chance alone, then researchers will view this as evidence that the treatment is actually causing changes in the outcome. As another example, if people who tend to score high in self-esteem also tend to score high on a measure of aggression, inferential statistics can tell us whether this tendency for the scores to go hand in hand could have happened easily by chance or whether the tendency is strong and consistent enough that it probably reflects a true association between these two variables in the general population (or at least the population that most closely resembles the researcher's sample). Both of these examples reveal the logic of significance testing. More recently, statisticians have begun to complement traditional statistical tests with indicators of effect size. Whereas statistical significance tells you whether an effect is likely to exist, estimates of effect size tell you how large an effect is likely to be.

Appendix 1.1: Some Common Statistical Tests and Their Uses

In Cervantes's classic novel *Don Quijote*, there is a point at which Don Quijote expresses tremendous self-satisfaction when he learns that he has been speaking *prose* his entire life. Unlike Don Quijote, most statisticians are more impressed with proofs than with prose. Thus, many statistics texts offer readers flowcharts, formulae, or decision trees to help them decide what kind of statistical analysis to perform on different kinds of data. Because of our abiding love of prose as well as our pathological fear of decision trees, we offer an alternative in this appendix. That is, instead of a decision tree, we offer a series of definitions and concrete examples that are much richer than a decision tree. If you spend a little while reading over the list of analytic techniques covered in this appendix, we hope that you'll have a good sense of how to analyze most basic data sets while also gaining a good sense of all of the major topics we cover in this textbook. The one way in which this list does vaguely resemble a decision tree is that it is loosely organized in increasing order of the complexity or sophistication of the research question. It thus begins with a couple of simple descriptive statistics and progresses through a series of increasingly complex inferential statistical tests. Readers who are interested in a true decision tree can find an excellent one in Chapter 2 of Tabachnick and Fidell (2007, pp. 28–31).

Mean, median, and mode: Very often, researchers who have collected data on a continuous (i.e., interval or ratio) scale simply want to summarize what the typical score is like. When the scores are normally distributed with very little skew (and modest to high kurtosis), the mean is an excellent indicator of the typical score. Height, SAT scores, and the highway mileage for one's car are all normally distributed without too much skew or kurtosis (the Hummer and Prius notwithstanding). In contrast, variables such as personal income, number of criminal convictions, and number of depressive symptoms only approximate a normal distribution because they are typically highly skewed. In such cases, the median and/or mode are often better indicators of central tendency because the median and the mode are influenced very little by extreme outliers. When researches wish to make inferences about populations rather than merely summarize a set of scores, they will want to report the standard error of the mean and/or a 95% or 99% confidence interval for the mean—to give others some idea of how far from the observed sample mean the true population mean is likely to fall. The standard error of the mean is a function of the observed sample standard deviation (see below) and the sample size, and it grows smaller as the sample size gets larger. All else being equal, this means, for example, that a researcher who randomly samples 1,000 people will be able to make a more precise statement about the likely range of the population mean than will a

researcher who randomly samples only 100 people. We discuss measures of central tendency in great detail in Chapter 2. In that same chapter, we also discuss the important topic of variability (especially the **standard deviation**) while also covering important topics such as skewness and kurtosis. These last two topics set the stage for subsequent chapters on inferential statistics and data cleaning.

Standard deviation: The standard deviation is an indicator of the variability of a set of continuous scores around the mean. Whereas central tendency tells us what the typical score is like, the standard deviation tells us how well that typical score describes *all* of the scores in the distribution—because the standard deviation is an indicator of the average amount by which all of the scores in a distribution vary around the mean. We discuss both traditional and nontraditional (creative) uses of the standard deviation in Chapter 2.

Variance: The variance of a set of scores is simply the **standard deviation** of that set of scores squared.

Pearson's *r*: describes the strength and direction of the *linear* association between two continuous (interval or ratio) variables. An *r* of zero means that there is no linear association whatsoever between two variables. Absolute values of *r* closer to 1.0 indicate a stronger association. If *r* is negative, it means that as scores on one variable (*X*) increase, scores on the other variable (*Y*) decrease. If *r* is positive, it means that as scores on one variable (*X*) increase, scores on the other variable (*Y*) also increase. The concept of correlation is closely linked to prediction. Thus, for example, if height and weight are correlated $r = .70$, one can minimize errors of prediction by predicting that a person who is exactly one standard deviation above the mean in height is 0.70 standard deviation units above the mean in weight. We discuss the correlation coefficient in Chapter 3, including a brief discussion of how to assess curvilinear as well as linear associations between continuous variables.

Phi coefficient (φ): The phi coefficient is very similar to *r* except it has no sign because it is used to describe the strength of association between two nominal or categorical variables (variables that do not indicate quantity or amount). It thus ranges between zero and 1.0. We discuss both the **phi coefficient** and the **chi-square test of association** in Chapter 4.

Chi-square test of association: A chi-square (χ^2) test of association is conceptually identical to a phi coefficient (φ) because, like phi, this test of association indicates whether two categorical variables are related. In fact, it is very easy to convert a chi-square test of association to a phi coefficient using the simple formula $\varphi = \sqrt{(\chi^2_{obt} / N)}$.

Single-sample *t* test: This very simple test is designed to test see whether the mean score for a continuous, normally distributed variable (e.g., IQ score, height) in a specific sample differs from some known or hypothesized (e.g., theoretically meaningful) population value. For example, the SAT scores for the students at a particular high school might be compared with the known U.S. population mean for SAT scores to see if the students at that high school tend to be more academically prepared than the average American high school student. We discuss this test in the beginning of Chapter 6.

Independent samples *t* test: This common test is used to assess the reliability (statistical significance) of mean differences on a continuous variable between two independent groups or categories of people. Some examples of the use of this test are (a) drawing inferences about the results of a lab experiment that has one experimental and one control group, (b) assessing gender differences on a continuously measured emotional performance test, and (c) comparing people with versus without a disease on a suspected health consequence of the disease. The main assumption of the test is that the dependent measure is normally distributed (although the test is pretty robust to many, but not all, violations of this assumption). We discuss this test in some detail in the latter portion of Chapter 6.

One-way analysis of variance (ANOVA): This test is used to assess the reliability (statistical significance) of mean differences between three or more groups. This test is very similar to an independent samples *t* test (and shares the same assumption of a normally distributed dependent measure). However, the difference is that a one-way ANOVA controls for the experiment-wise error rate that occurs as researchers consider all of the many possible comparisons that can be made between specific groups when there are *multiple* experimental or naturally existing groups (three or more levels of the independent variable). Some examples of the use of this statistic are (a) outcomes in a lab experiment that has three different conditions (e.g., three dosage levels of a drug), (b) comparing kids in four different grades on an intellectual outcome, and (c) comparing the attitudes of soldiers from five different military ranks. When the researcher has a clear a priori reason to expect the various conditions to yield results that follow a specific pattern, the researcher can greatly increase statistical power by conducting a **planned comparison** based on this specific expectation. We discuss one-way ANOVA, including planned comparisons, in the first half of Chapter 7.

Factorial ANOVA: This technique is used to assess joint effects of two or more fully crossed categorical independent variables on a continuously scored outcome. It allows for the statistical separation of main effects of all independent variables and, if desired, interactions between two or more

independent variables. This technique is used frequently to draw conclusions about the results of laboratory experiments. It also assumes a normally distributed dependent measure. We discuss factorial ANOVA, including follow-up comparisons such as simple effects tests, in the second half of Chapter 7.

Analysis of covariance (ANCOVA): Sometimes researchers may wish to control for a confound or nuisance variable in an ANOVA and—rather than reporting the raw, between-groups means—hold the different, naturally occurring or experimental groups constant on that nuisance variable. This is both conceptually and mathematically identical to a simultaneous multiple regression analysis with at least one categorical variable (the independent variable or variables) and at least one continuous variable (the covariates in an ANCOVA). The main advantage of ANCOVA over a regression analysis is the fact that ANCOVA readily yields covariate-adjusted means, which look very much like regular means and thus are very easy to interpret.

One of the crucial assumptions of ANCOVA is **homogeneity of covariance,** meaning that the covariate for which the analysis makes a statistical adjustment should have roughly the same association with the dependent measure in all of the various experimental conditions. For example, a researcher studying gender differences in aggressive behavior in the lab might wish to control statistically for the fact that people who more strongly believe in the concept of defending one's honor (reported, say, on a 9-point scale) behave more aggressively than people who do not believe in the concept of defending one's honor (e.g., see D. Cohen & Nisbett, 1994). So long as the association between beliefs about honor and laboratory aggressive behavior was about equally strong for women and men, it would be appropriate to reduce the noise associated with this belief to see if a significant gender difference remained after controlling for the belief. The test for a gender difference could thus be more powerful than it would have been otherwise after controlling for any effects of this nuisance variable that is more or less independent of gender. In Chapter 11, we compare and contrast ANCOVA with ANOVA and with multiple regression analysis.We emphasize that whereas ANCOVA is mathematically identical to a simultaneous multiple regression analysis, the two techniques yield very different kinds of outputs. For example, it is often much easier for people to understand covariate-adjusted means (because they look just like traditional means) than to understand standardized regression coefficients.

Reliability analysis: A reliability analysis is used to determine the degree to which the multiple items in a scale all behave in the same fashion (i.e., are positively correlated with one another). Cronbach's alpha (α) is a very useful and easy-to-calculate statistic. However, a very high alpha statistic for a scale does not always guarantee that the items in the scale form a

single factor. On the other hand, low corrected item-total *r*s for specific items in a scale are a useful indicator that the specific items are not correlated with the other scale items (and thus should probably be excluded from the scale). Treating individual raters of a specific behavior or judgment as if they were specific items in a scale can allow an assessment of the reliability of individual raters. A rater whose item-total correlation is low is in disagreement with the average of all of the other raters and can either be retrained, if possible, or dropped. We discuss reliability analysis (along with principal components analysis and factor analysis) in Chapter 5.

Multiple regression analysis: Multiple regression analysis is used to assess the strength and direction of the unique linear association between multiple, continuous predictors of a continuous outcome (a criterion) and that outcome. Because each predictor is controlled statistically for the association between that predictor and all other predictors, an assessment of the relative strength of each predictor is possible. Each predictor is thus assessed controlling for the natural *confounds* between that predictor and all other predictors. The primary statistical indicator is a *B* or *b* (an unstandardized regression coefficient) or beta (β), a standardized regression coefficient that is conceptually very similar to *r*. Each coefficient has its own associated *p* value that indicates the reliability (statistical significance) of that unique association. Although both univariate (single-variable) normality and multivariate normality of the continuously measured predictors are assumed, this analysis is usually quite robust to the inclusion of one or more categorical variables, especially when these variables are not highly skewed. Gender, for example, is often dummy-coded without any problems in a multiple regression analysis that also includes several continuous, normally distributed predictors. We discuss the basics of multiple regression analysis, with a conceptual emphasis on how multiple regression identifies the unique association between different predictor variables and a criterion variable, in the first part of Chapter 9.

In most cases of multiple regression analysis, researchers expect the zero-order association between a predictor variable and the criterion of interest to grow smaller (and sometimes even disappear altogether) when all the other predictor variables are statistically held constant. However, a multiple regression analysis can also reveal that a predictor variable that appeared to be unrelated (at the simple or zero-order level) to a criterion variable is actually associated with the criterion variable once statistical adjustments are made for the effects of one or more additional predictors. This unusual situation is referred to as **suppression,** and it is discussed in great detail in Chapter 12.

Moderator analysis (multiple regression analysis of statistical interactions): If a researcher also wishes to know whether the association between a predictor and the criterion *differs* at different levels of some other predictor,

it is possible to conduct a **moderator** analysis by examining the cross-product(s) of the two or more predictors of interest and following up a significant interaction with simple slopes tests (analogous to simple main effect tests in ANOVA). A moderator analysis in multiple regression could thus be conducted to see if the association between the number of times people moved as children and their physical health as adults is stronger for introverts than for extraverts. (Introverts who moved a lot as children often have poorer than average adult health, but extraverts seem to show no such association.) A moderator variable can also be categorical while the other predictor and the criterion variable are both continuous. For example, the association between implicit self-esteem and explicit self-esteem (both continuous variables) seems to be stronger (more positive) for women than for men (Pelham, Koole, et al., 2005). We discuss moderator analyses in multiple regression, including simple slopes tests to elucidate the exact nature of a significant interaction, in Chapter 10.

Logistic regression: Logistic regression analysis is conceptually identical to a standard multiple regression analysis except that the criterion variable (and sometimes one or more of the predictors) is categorical rather than continuous. The primary output statistic is a predictor-adjusted odds ratio that is the rough conceptual equivalent of a *B* or a β. Unlike a simple odds ratio, however, the odds ratio from a logistic regression analysis refers to the association between one categorical variable and another while holding all other predictor variables in the model constant. We discuss the basics of logistic regression in the last section of Chapter 9.

Principal components analysis and factor analysis: These two closely related techniques are designed to uncover underlying dimensions along which a set of many separate responses vary. These numerous individual responses might be answers to individual personality questions, specific political or social attitudes, or self-reported liking for many different kinds of foods or many different specific types of music. For example, contemporary research in human personality suggests that hundreds of individual personality questions all boil down to five core personality dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (see Goldberg, 1990). A factor analysis of dozens of specific personality traits might thus reveal, for example, that trait terms such as *energetic*, *friendly*, *outgoing*, *outspoken*, and *loud* would all load heavily on the basic dimension of **extraversion.** In contrast, specific trait terms such as *reliable*, *punctual*, *obedient*, *organized*, and *honest* might all load heavily on the core dimension of **conscientiousness.** One key difference between principal components analysis and factor analysis is that principal components analysis is usually a bottom-up, purely empirical way to distill a large set of observations into a smaller number of dimensions. In contrast, factor analysis is more likely to be used when the researcher has

a clear a priori theory about how the different items ought to load together and how many factors there ought to be in a data set. We discuss principal components analysis and factor analysis in Chapter 5, including a discussion of how these methods are related to the concept of reliability (specifically, the internal consistency of a multiple-item scale).

Paired-samples *t* test: When two measures come from the same organism (or similar organisms), the two different measures are likely to be highly correlated with one another. For example, a specific child who excels in reading is also likely to excel in spelling. A similar lack of independence in specific observations often occurs when genetically related or experimentally yoked members of a pair are tested on the same outcome. When exactly two such repeated measures are obtained, a paired-samples *t* test can be used to assess the statistical significance of any observed behavioral or performance difference between the two measures. This test is also very useful when the same person fills out the same measure or task under different (manipulated) experimental conditions. For example, a child might be given the same intellectual measure by two different experimenters, one of whom expresses a positive expectancy about her performance and one of whom expresses no expectations. If appropriate experimental controls are used (e.g., counterbalancing the order of the two expectancies across participants), a paired-samples *t* test can reveal whether performance varies reliably with experimental condition, with a very high level of statistical power. Like the independent samples *t* test, this test assumes that the dependent measure (which in this case is the *difference* between two scores) is normally distributed. Importantly, the increased power that usually comes with this test comes precisely to the degree that the two measures of interest are strongly correlated with one another. It is this strong correlation between two related measures that effectively reduces the variance that serves as the error term for this analysis. We discuss the paired-samples *t* test in Chapter 8, with a special emphasis on how the correlation between two repeated measurements plays a crucial role in the power of a repeated measures *t* test to reveal differences between paired means.

Repeated measures ANOVA: If three or more within-subjects conditions (or measurement periods) rather than two are collected from the same (or related) participants, a repeated measures ANOVA is the appropriate statistical test for differences between the means. Just as a one-way ANOVA replaces an independent samples *t* test once you graduate from two to three or more groups, the one-way repeated measures ANOVA replaces the paired-samples *t* test once you graduate from two to three or more within-subjects conditions. For example, if children are exposed to three different expectancies rather than two, a repeated measures ANOVA could be conducted on the mean performance scores in the three within-subjects conditions. Repeated measures studies can also involve complete factorial designs. For example, a

completely within-subjects experiment might separately study reactions to sexist and nonsexist jokes that also vary independently in how funny the jokes are pretested to be. In its simplest form, this study would be a 2 (Level of Sexism: High vs. Low) × 2 (Level of Humor: High vs. Low) completely within-subjects study, analyzed using a within-subjects ANOVA. We discuss repeated measures ANOVAs in Chapter 8.

Mixed model ANOVA: If a study includes at least one within-subjects variable and at least one between-subjects variable (whether measured or manipulated), a mixed model ANOVA can test simultaneously for both between-subjects and within-subjects effects. Mixed model ANOVAs can also test for statistical interactions between one or more between-subjects variables and one or more within-subjects variables. For example, a cognitive psychologist might manipulate cognitive load on a between-subjects basis while assessing both implicit and explicit memory for studied material. She might predict, for example, that the cognitive load manipulation (e.g., rehearsing a 7-digit number) will have a large effect on explicit memory (recall memory) while having little or no effect on implicit memory (e.g., based on performance on a word fragment completion task). Thus, the researcher would expect to observe a Load × Memory–type interaction in this mixed model design. We discuss mixed model ANOVAs in Chapter 8.

Mediation analysis: In both laboratory experiments and passive observational studies, researchers often wish to know *why* one variable is related to another. For example, research on frustration and aggression suggests that one of the main reasons why frustration often leads to aggression is because frustration leads to anger, which then leads to aggression. In the language of mediation, this is to say that anger mediates the simple association between frustration and aggression. Mediation analyses are merely variations on a multiple regression analysis with an emphasis on assessing the degree to which the association between the original independent variable and the dependent variable disappears or gets weaker once you statistically control for the significant effects of the mediator on the dependent measure. Prototypically, if anger fully mediates the association between frustration and aggression, (a) frustration should predict aggression, (b) frustration should predict anger, (c) anger should predict aggression, and finally (d) the association between anger and aggression should completely disappear once you statistically control for the effects of anger on aggression (because frustration affects aggression indirectly through the route of increased anger). We discuss both **mediation analysis** and **path analysis** in Chapter 13.

Path analysis: Path analysis is a special version of multiple regression analysis that is designed to assess the plausibility of a proposed causal chain leading from one or more source variables to an ultimate ("downstream")

outcome variable. In fact, the simplest possible kind of path analysis is a three-variable mediation model with the mediator representing the causal step between just one source variable and just one outcome variable. In most cases, however, path analyses involve four or more variables, ideally measured at different carefully selected time points (e.g., in a longitudinal or prospective design). Moreover, researchers do not always expect every variable in the middle of a causal chain to mediate the associations between the source variables and the ultimate outcome variable. Instead, some of the source variables might be expected to have a direct (nonmediated) as well as an indirect (mediated) connection to the downstream variable. Path analysis is the historical and conceptual precursor of modern **structural equation modeling,** which can be thought of as a hybrid combination of path analysis and factor analysis. In fact, some researchers refer to structural equation modeling as confirmatory (aka "theory-driven") factor analysis. A detailed discussion of structural equation modeling is beyond the scope of this intermediate text.

Notes

1. For more details, see http://www.news.com.au/business/story/0,27753,25515799-462,00.html.

2. See the *Time* magazine story at http://www.time.com/time/magazine/article/0,9171,1889153,00.html.

3. We adapted this example of men of varying heights from an illuminating statistics lecture by Daniel Gilbert, who probably adapted it from a lecture by Plato.

4. http://imgs.sfgate.com/cgi-bin/article.cgi?f=/c/a/1998/12/28/MN9307.DTL&type=printable

5. Computing the probability of an event as extreme as *or more extreme than* an observed event (or set of events) is standard practice for most statistical tests. At first blush, paying attention to events even more extreme than an observed event may seem a little odd. However, if we care about events *as unusual as or more unusual than* our observed event—which we almost always do—it makes a lot of sense. If you think of the unusualness of a set of observations (e.g., a lot of heads tossed, a pair of means that are noticeably different) as a standard of experimental performance that a researcher hopes to meet or exceed, this may help make sense of this practice. If we set a high-jump bar at exactly 6 feet and Amanda clears it, the set of outcomes that Amanda, the judges, and the fans all care about is jumps of exactly 6 feet *or higher.* Furthermore, if we tried to calculate the probability of a specific observation or event, probabilities would almost always be pretty low—because the probability of any specific event is always quite low. For example, the probability of tossing a fair coin 20 times and observing *exactly* 10 heads is .176, even though this is the *most* likely of all the possible outcomes. Once we move to continuous rather than discrete events, this is even truer. The probability that a particular high jump would be *exactly* 6 feet—even for a very good jumper who was trying to jump exactly 6 feet—is extremely low.

6. Speaking of cheating, we cheated. Unless we increased our sample size to about 250 people, we couldn't actually conduct this second χ^2 analysis. That's because we're allowed to use the χ^2 statistic only in situations in which all our expected frequencies have a value of at least 5.0. With values lower than 5, the χ^2 values that are generated can be pretty unstable and pretty inaccurate. In an extreme case such as this one, however, it's safe to say that people were significantly honest. If nothing else, we could always choose to make a very conservative comparison and set 90% (instead of 98%) dishonesty as our standard of comparison. This would yield 5 rather than 1 as the expected number of nonwinners. In case you want to practice your calculations, the value you should get if you do the analysis this more conservative (but legal) way is χ^2 (1, $N = 50$) = 272.22. The 1 in the parentheses indicates the *degrees of freedom* you'd report in an actual research report in which you conducted this analysis. We come back to this in the section on reporting commonly used statistics.