PART A

Introducing Statistics

CHAPTER 1

Statistics and Probability Are Not Intuitive

> If something has a 50% chance of happening, then 9 times out of 10 it will.
>
> Yogi Berra

The word *intuitive* has two meanings. One is "easy to use and understand," which is my goal for this book—hence its title. The other meaning is "instinctive, or acting on what one feels to be true even without reason." This fun (really!) chapter shows how our instincts often lead us astray when dealing with probabilities.

WE TEND TO JUMP TO CONCLUSIONS

A three-year-old girl told her male buddy, "You can't become a doctor; only girls can become doctors." To her, this statement made sense, because the three doctors she knew were all women.

When my oldest daughter was four, she "understood" that she was adopted from China, whereas her brother "came from Mommy's tummy." When we read her a book about a woman becoming pregnant and giving birth to a baby girl, her reaction was, "That's silly. Girls don't come from Mommy's tummy. Girls come from China." With only one person in each group, she made a general conclusion. Like many scientists, when new data contradicted her conclusion, she questioned the accuracy of the new data rather than the validity of the conclusion.

The ability to generalize from a sample to a population is hardwired into our brains and has even been observed in eight-month-old babies (Xu & Garcia, 2008). To avoid our natural inclination to make overly strong conclusions from limited data, scientists need to use statistics.

WE TEND TO BE OVERCONFIDENT

How good are people at judging how confident they are? You can test your own ability to quantify uncertainty using a test devised by Russo and Schoemaker (1989). Answer each of these questions with a range. Pick a range that you think

has a 90% chance of containing the correct answer. Don't use Google to find the answer. Don't give up and say you don't know. Of course, you won't know the answers precisely! The goal is not to provide precise answers, but rather is to correctly quantify your uncertainty and come up with ranges of values that you think are 90% likely to include the true answer. If you have no idea, answer with a super wide interval. For example, if you truly have no idea at all about the answer to the first question, answer with the range zero to 120 years old, which you can be 100% sure includes the true answer. But try to narrow your responses to each of these questions to a range that you are 90% sure contains the right answer:

- Martin Luther King Jr.'s age at death
- Length of the Nile River, in miles or kilometers
- Number of countries in OPEC
- Number of books in the Old Testament
- Diameter of the moon, in miles or kilometers
- Weight of an empty Boeing 747, in pounds or kilograms
- Year Mozart was born
- Gestation period of an Asian elephant, in days
- Distance from London to Tokyo, in miles or kilometers
- Deepest known point in the ocean, in miles or kilometers

Compare your answers with the correct answers listed at the end of this chapter. If you were 90% sure of each answer, you would expect about nine intervals to include the correct answer and one to exclude it.

Russo and Schoemaker (1989) tested more than 1,000 people and reported that 99% of them were overconfident. The goal was to create ranges that included the correct answer 90% of the time, but most people created narrow ranges that included only 30% to 60% of the correct answers. Similar studies have been done with experts estimating facts in their areas of expertise, and the results have been similar.

Since we tend to be too sure of ourselves, scientists must use statistical methods to quantify confidence properly.

WE SEE PATTERNS IN RANDOM DATA

Table 1.1 presents simulated data from 10 basketball players (1 per row) shooting 30 baskets each. An "X" represents a successful shot and a "–" represents a miss. Is this pattern random? Or does it show signs of nonrandom streaks? Look at Table 1.1 before continuing.

Most people see streaks of successful shots and conclude this is not random. Yet Table 1.1 was generated randomly. Each spot had a 50% chance of being "X" (a successful shot) and a 50% chance of being "–" (an unsuccessful shot), without taking into consideration previous shots. We see clusters perhaps because our brains have evolved to find patterns and do so very well. This ability may have served our ancestors well to avoid predators and poisonous plants, but it is

–	–	X	–	X	–	X	X	X	–	–	–	–	X	X	X	–	X	X	–	X	X	–	–	–	X	X	–	–	–	X	X
X	–	–	X	–	X	X	–	–	X	X	–	–	X	–	X	–	X	–	–	–	X	X	X	X	–	–	X	X	–	–	–
X	X	X	X	–	X	X	–	X	–	X	–	X	X	X	–	–	–	–	–	X	–	X	–	X	X	X	–	–	–	–	X
–	X	–	X	–	–	X	X	–	X	X	–	X	X	–	–	X	X	X	X	–	–	–	–	X	X	–	X	–	X	–	–
–	X	–	X	–	X	X	–	–	–	X	X	–	–	–	–	–	–	X	–	X	–	X	–	–	X	–	–	X	–	X	X
–	–	X	X	X	–	X	–	X	–	–	–	X	X	X	X	–	X	X	X	X	–	–	–	–	–	X	X	–	X	X	X
X	–	–	X	X	–	–	X	X	X	X	–	X	X	X	–	–	X	–	–	X	X	X	X	X	–	X	X	X	–	–	–
X	–	X	–	–	–	X	X	X	X	X	–	–	X	X	–	X	X	–	X	X	X	–	X	X	–	X	–	–	X	–	X
X	X	X	–	–	X	X	X	X	X	–	X	–	X	–	X	X	–	X	–	X	X	X	X	–	X	X	–	X	X	X	X
–	–	–	X	X	X	–	–	X	X	X	–	X	X	X	–	–	X	–	–	X	–	X	X	X	X	X	–	–	–	X	–

Table 1.1. Random patterns don't seem random.

Table 1.1 represents 10 basketball players (1 per row) shooting 32 baskets each. An "X" represents a successful shot, and a "–" represents a miss.

important that we recognize this built-in mental bias. Statistical rigor is essential to avoid being fooled by seeing apparent patterns among random data.

WE DON'T REALIZE THAT COINCIDENCES ARE COMMON

In November 2008, I attended a dinner for the group Conservation International. The actor Harrison Ford is on their board, and I happened to notice that he wore an ear stud. The next day, I watched an episode of the TV show *Private Practice*, and one character pointed out that another character had an ear stud that looked just like Harrison Ford's. The day after that, I happened to read (in a book on serendipity!) that the Nobel Prize–winning scientist Baruch Blumberg looks like Indiana Jones, a movie character played by Harrison Ford (Meyers, 2007).

What is the chance that I would encounter Harrison Ford, or a mention to him, three times in three days? Tiny, but that doesn't mean much. While it is highly unlikely that any particular coincidence will occur, it is almost certain that some seemingly astonishing set of unspecified events will happen often, since we notice so many things each day. Remarkable coincidences are always noted in hindsight and never predicted with foresight.

WE DON'T EXPECT VARIABILITY TO DEPEND ON SAMPLE SIZE

Gelman (1998) looked at the relationship between the populations of counties and the age-adjusted, per-capita incidence of kidney cancer (a fairly rare cancer, with an incidence of about 15 cases per 100,000 adults in the United States). First, he focused on counties with the lowest per-capita incidence of kidney cancer. Most of these counties had small populations. Why? One might imagine that something about the environment in these rural counties leads to lower rates of kidney cancer. Then, he focused on counties with the highest incidence of kidney cancer. These also tended to be the smallest counties. Why? One might imagine that lack of

medical care in these tiny counties leads to higher rates of kidney cancer. But it seems pretty strange that both the highest and lowest incidences of kidney cancer be in counties with small populations?

The reason is simple, once you think about it. In large counties, there is little variation around the average rate. Among small counties, however, there is much more variability. Consider an extreme example of a tiny county with only 1,000 residents. If no one in that county had kidney cancer, that county would be among those with the lowest (zero) incidence of kidney cancer. But if only one of those people had kidney cancer, that county would then be among those with the highest rate of kidney cancer. In a really tiny county, it only takes one case of kidney cancer to flip from having one of the lowest rates to having one of the highest rates. In general, just by chance, the incidence rates will vary much more among counties with tiny populations than among counties with large populations. Therefore, counties with both the highest and the lowest incidences of kidney cancer tend to have smaller populations than counties with average incidences of kidney cancer.

Random variation can have a bigger effect on averages within small groups than within large groups. This simple principle is logical, yet is not intuitive to many people.

WE HAVE INCORRECT INTUITIVE FEELINGS ABOUT PROBABILITY

Imagine that you can choose between two bowls of jelly beans. The small bowl has nine white and one red jelly bean. The large bowl has 93 white beans and 7 red beans. Both bowls are well mixed, and you can't see the beans. Your job is to pick one bean. You win a prize if your bean is red. Should you pick from the small bowl or the large one?

When you choose from the small bowl, you have a 10% chance of picking a red jelly bean. When you pick from the large bowl, the chance of picking a red one is only 7%. So your chances of winning are higher if you choose from the small bowl. Yet about two-thirds of people prefer to pick from the larger bowl (Denes-Raj & Epstein, 1994). Many of these people do the math and know that the chance of winning is higher with the small bowl, but they feel better about choosing from the large bowl because it has more red beans and offers more chances to win. Of course, the large bowl also has more white beans and more chances to lose. Our brains have simply not evolved to deal sensibly with probability, and most people make the illogical choice.

WE FIND IT HARD TO COMBINE PROBABILITIES

Here is a classic brainteaser called the Monty Hall problem, named after the host of the game show *Let's Make a Deal*: You are a contestant on a game show and are presented with three doors. Behind one is a fancy new car. You must choose one door, and you get to keep whatever is behind it. You pick a door. At this point, the

host chooses one of the other two doors to open and shows you that there is no car behind it. He now offers you the chance to change your mind and choose the other door (the one he has not opened).

Should you switch?

Before reading on, you should think about the problem and decide whether you should switch. There are no tricks or traps. Exactly one door has the prize, all doors appear identical, and the host—who knows which door leads to the new car—has a perfect poker face and gives you no clues. There is never a car behind the door the moderator chooses to open. Don't cheat. Think it through before continuing.

When you first choose, there are three doors and each is equally likely to have the car behind it. So your chance of picking the winning door is one-third. Let's separately think through the two cases: originally picking a winning door or originally picking a losing door.

If you originally picked the winning door, then neither of the other doors has a car behind it. The host opens one of these two doors. If you now switch doors, you will have switched to the other losing door.

What happens if you originally picked a losing door? In this case, one of the remaining doors has a car behind it and the other one doesn't. The host knows which is which. He opens the door without the car. If you now switch, you will win the car.

Let's recap. If you originally chose the correct door (an event that has a one-third chance of occurring), then switching will make you lose. If you originally picked either of the two losing doors (an event that has a two-thirds chance of occurring), then switching will definitely make you win. Switching from one losing door to the other losing door is impossible, because the host will have opened the other losing door.

Your best choice is to switch! Of course, you can't be absolutely sure that switching doors will help. One-third of the time you will be switching away from the prize. But the other two-thirds of the time you will be switching to the prize. If you repeat the game many times, you will win twice as often by switching doors every time. If you only get to play once, you have twice the chance of winning by switching doors.

Almost everyone (including mathematicians and statisticians) intuitively reaches the wrong conclusion and thinks that switching won't be helpful (Vos Savant, 1997).

WE DON'T DO BAYESIAN CALCULATIONS INTUITIVELY

Imagine this scenario: You are screening blood donors for the presence of human immunodeficiency virus (HIV). Only a tiny fraction (0.1%) of the blood donors has HIV. The antibody test is highly accurate but not quite perfect. It correctly identifies 99% of infected blood samples but also incorrectly concludes that 1% of noninfected samples have HIV. When this test identifies a blood sample as having HIV present, what is the chance that the donor does, in fact, have HIV, and what is the chance the test result is an error (a false positive)?

Try to come up with the answer before reading on.

Let's imagine that 100,000 people are tested. Of these, 100 (0.1%) will have HIV, and the test will be positive for 99 (99%) of them. The other 99,900 people will not have HIV, but the test will incorrectly return a positive result in 1% of cases. So there will be 999 false positive tests. Altogether, there will be 99 + 999 = 1,098 positive tests, of which only 99/1,098 = 9% will be true positives. The other 91% of the positive tests will be false positives. So if a test is positive, there is only a 9% chance that there is actually HIV in that sample.

Most people, including most physicians, intuitively think that a positive test almost certainly means that HIV is present. Our brains are not adept at combining what we already know (the prevalence of HIV) with new knowledge (the test is positive).

Now imagine that the same test is used in a population of intravenous drug users, of which 10% have HIV. Again, let's imagine that 100,000 people are tested. Of these, 10,000 (10%) will have HIV, and the test will be positive for 9,900 (99%) of them. The other 90,000 people will not have HIV, but the test will incorrectly return a positive result in 1% of cases. So there will be 900 false positive tests. Altogether, there will be 9,900 + 900 = 10,800 positive tests, of which 9,900/10,800 = 92% will be true positives. The other 8% of the positive tests will be false positives. So if a test is positive, there is a 92% chance that there is HIV in that sample.

The interpretation of the test result depends greatly on what fraction of the population has the disease. This example gives you a taste of what is called Bayesian logic (a subject that will be discussed again in Chapters 2 and 18).

WE ARE FOOLED BY MULTIPLE COMPARISONS

Austin and colleagues (2006) sifted through a database of health statistics of 10 million residents of Ontario, Canada. They examined 223 different reasons for hospital admission and recorded the astrological sign of each patient (computed from his or her birth date). They then asked if people with certain astrological signs are more likely to be admitted to the hospital for certain conditions.

The results seem impressive. Seventy-two diseases occurred more frequently in people with one astrological sign than in people with all the other astrological signs put together, with the difference being statistically significant. Essentially, a result that is *statistically significant* would occur by chance less than 5% of the time (you'll learn more about what *statistically significant* means in Chapter 16).

Sounds impressive, doesn't it? Indeed, those data might make you think that there is a convincing relationship between astrology and health. But there is a problem. It is misleading to focus on the strong associations between one disease and one astrological sign without considering all the other combinations. Austin et al. (2006) tested the association of 223 different reasons for hospital admissions with 12 astrological signs and so tested 2,676 distinct hypotheses (223 × 12 = 2,676). Therefore, they would expect to find 134 statistically significant associations just by chance (5% of 2,676 = 134) but in fact only found 72.

Note that this study wasn't really done to ask about the association between astrological sign and disease. It was done to demonstrate the difficulty of interpreting statistical results when many comparisons are performed.

Chapters 22 and 23 explore multiple comparisons in more depth.

WE TEND TO IGNORE ALTERNATIVE EXPLANATIONS

Imagine you are doing a study on the use of acupuncture in treating osteoarthritis. Patients who come in with severe arthritis pain are treated with acupuncture. They are asked to rate their arthritis pain before and after the treatment. The pain decreases in most patients after treatment, and statistical calculations show that such consistent findings are exceedingly unlikely to happen by chance. Therefore, the acupuncture must have worked. Right?

Not necessarily. The decrease in recorded pain may not be caused by the acupuncture. Here are five alternative explanations (adapted from Bausell, 2007):

- If the patients believe in the therapist and treatment, that belief may reduce the pain considerably. The pain relief may be a placebo effect and have nothing to do with the acupuncture itself.
- The patients want to be polite and may tell the experimenter what he or she wants to hear (that the pain decreased). Thus, the decrease in reported pain may be because the patients are not accurately reporting pain after therapy.
- Before, during, and after the acupuncture treatment, the therapist talks with the patients. Perhaps he or she recommends a change in aspirin dose, a change in exercise, or the use of nutritional supplements. The decrease in reported pain might be due to these aspects of the treatment, rather than the acupuncture.
- The experimenter may have altered the data. For instance, what if three patients experience worse pain with acupuncture, whereas the others get better? The experimenter carefully reviews the records of those three patients and decides to remove them from the study because one of those people actually has a different kind of arthritis than the others, and two had to climb stairs to get to the appointment because the elevator didn't work that day. The data, then, are biased or skewed because of the omission of these three participants.
- The pain from osteoarthritis varies significantly from day to day. People tend to seek therapy when pain is at its worst. If you start keeping track of pain on the day when it is the worst, it is quite likely to get better, even with no treatment. The next section explores this idea of *regression to the mean*.

WE ARE FOOLED BY REGRESSION TO THE MEAN

Figure 1.1 illustrates simulated blood pressures before and after a treatment. Figure 1.1A includes 24 pairs of values. The "before" and "after" groups are about the same. In some cases, the value goes up after treatment, and in others it goes down. If these were real data, you'd conclude that there is no evidence at all that the treatment had any effect on the outcome (blood pressure).

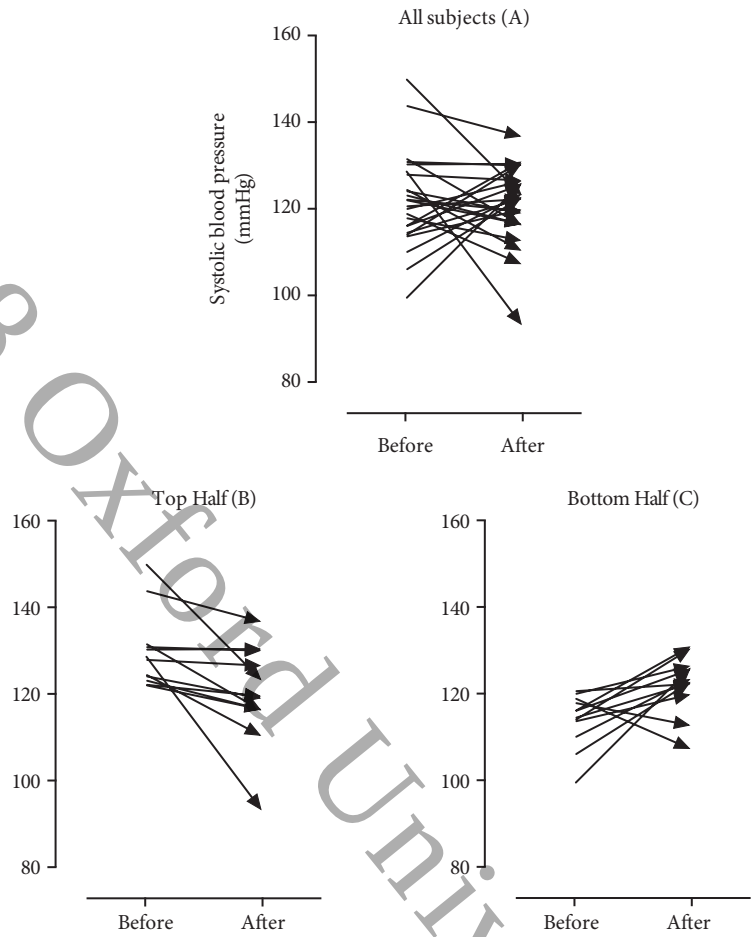

Figure 1.1. Regression to the mean.

All data in (A) were drawn from random distributions (Gaussian; mean = 120, SD = 15) without regard to the designations "before" and "after" and without regard to any pairing. (A) shows 48 random values, divided arbitrarily into 24 before–after pairs (which overlap enough that you can't count them all). (B) shows only the 12 pairs with the highest before values. In all but one case, the after values are lower than the before values. (C) shows the pairs with the lowest before measurements. In 10 of the 12 pairs, the after value is higher than the before value. If you only saw the graph in (B) or (C), you'd probably conclude that whatever treatment came between the before and after measurements had a large impact on blood pressure. In fact, these graphs simply show random values, with no systematic change between before and after. The apparent change is called *regression to the mean*.

Now imagine the study was designed differently. You've made the before measurements and want to test a treatment for high blood pressure. There is no point in treating individuals whose blood pressure is not high, so you select the people with the highest pressures to study. Figure 1.1B illustrates data for only

those 12 individuals with the highest before values. In every case but one, the after values are lower. If you performed a statistical test (e.g., a paired t test; see Chapter 31), the results would convince you that the treatment decreased blood pressure. Figure 1.1C illustrates a similar phenomenon with the other 12 pairs, those with low before values. In all but two cases, these values go up after treatment. This evidence would convince you that the treatment increases blood pressure.

But these are random data! The before and after values came from the same distribution. What happened?

This is an example of *regression to the mean*: the more extreme a variable is upon its first measurement, the more likely it is to be closer to the average the second time it is measured. People who are especially lucky at picking stocks one year are likely to be less lucky the next year. People who get extremely high scores on one exam are likely to get lower scores on a repeat exam. An athlete who does extremely well in one season is likely to perform more poorly the next season. This probably explains much of the *Sports Illustrated* cover jinx—many believe that appearing on the cover of *Sports Illustrated* will bring an athlete bad luck (Wolff, 2002).

WE LET OUR BIASES DETERMINE HOW WE INTERPRET DATA

Kahan and colleagues (2013) asked a bunch of people to analyze some simple data, similar to what you'll see in Chapter 27. While all the tables had the same values, sometimes the table was labeled to test the effectiveness of a skin cream for treating a rash (Table 1.2), and sometimes to test the effectiveness of a law prohibiting carrying concealed hand guns in public (Table 1.3).

	RASH GOT BETTER	RASH GOT WORSE	TOTAL
Patients who **did use** the cream	223	75	298
Patients who **did not use** the cream	107	21	128

Table 1.2. One of the tables that Kahan and colleagues used. After viewing this table, people were asked to determine whether the hypothetical experiment showed that the skin condition of people treated with the cream was more likely to "get better" or "get worse" compared to those who were not treated.

	DECREASE IN CRIME	INCREASE IN CRIME	TOTAL
Cities that **did ban** carrying concealed handguns in public	223	75	298
Cities that **did not ban** carrying concealed handguns in public	107	21	128

Table 1.3. The other table that Kahan and colleagues used. After viewing this table, people were asked to determine whether the made-up data show that cities that enacted a ban on carrying concealed handguns were more likely to have an increase or decrease in crime compared to cities that did not ban concealed handguns.

The experimental subjects were not asked for subtle interpretation of the data but rather were simply asked whether or not the data support a particular hypothesis. The math in Tables 1.2 and 1.3 is pretty straightforward:

- In the top row, the rash got better in 223/298 = 74.8% of the treated people, and the crime rate went down in 74.8% of the cities that banned concealed handguns.
- In the bottom row, the rash got better in 107/128 = 83.6% of the untreated people, and the crime rate when down in 83.6% of the cities that did not ban carrying concealed handguns.
- Most people had a decrease in rash, and most cities had a decrease in crime. But did the intervention matter? Since 74.8% is less than 83.6%, the data clearly show people who used the cream were less likely to have an improved rash than people who did not use the cream. Cities that passed handgun law had a smaller decrease in crime than cities who did not pass such a law.

When the data were labeled to be about the effectiveness of a skin cream, liberal Democrats and conservative Republicans (the two main political parties in the United States) did about the same. But when the data were labeled to be about the effectiveness of a gun safety policy, the results depended on political orientation. Liberal democrats tended to find that the data showed that gun safety laws reduced crime. Conservatives tended to conclude the opposite (this study was done in the United States, where conservatives tend to be against gun safety legislation.)

This study shows that when people have a preconceived notion about the conclusion, they tend to interpret the data to support that conclusion.

WE CRAVE CERTAINTY, BUT STATISTICS OFFERS PROBABILITIES

Many people expect statistical calculations to yield definite conclusions. But in fact, every statistical conclusion is stated in terms of probability. Statistics can be very difficult to learn if you keep looking for definitive conclusions. As statistician Myles Hollander reportedly said, "Statistics means never having to say you're certain!" (quoted in Samaniego, 2008).

CHAPTER SUMMARY

- Our brains do a really bad job of interpreting data. We see patterns in random data, tend to be overconfident in our conclusions, and mangle interpretations that involve combining probabilities.
- Our intuitions tend to lead us astray when interpreting probabilities and when interpreting multiple comparisons.
- Statistical (and scientific) rigor is needed to avoid reaching invalid conclusions.

TERM INTRODUCED IN THIS CHAPTER

- Regression to the mean (p. 9)

Answers to the 10 questions in the "We Tend to Be Overconfident" section.
Martin Luther King Jr.'s age at death: 39
Length of the Nile River: 4,187 miles or 6,738 kilometers
Number of countries in OPEC: 13
Number of books in the Old Testament: 39
Diameter of the moon: 2,160 miles or 3,476 kilometers
Weight of an empty Boeing 747: 390,000 pounds or 176,901 kilograms
Year Mozart was born: 1756
Gestation period of an Asian elephant: 645 days
Distance from London to Tokyo: 5,989 miles or 9,638 kilometers
Deepest known point in the ocean: 6.9 miles or 11.0 kilometers