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 meaning is "easy to use and understand." That 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." Using this definition, statistical reasoning is far from intuitive. This fun (really!) chapter demonstrates how our instincts often lead us astray when dealing with probabilities.*

WE TEND TO JUMP TO CONCLUSIONS

A 3-year-old girl told her male buddy, "You can't become a doctor; only girls can become doctors." To her this made sense, because the three doctors she knew were all women.

When my oldest daughter was 4, 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 n = 1 in each group, she made a general conclusion. When new data contradicted that conclusion, she questioned the accuracy of the new data rather than the validity of her conclusion.

The ability to generalize from a sample to a population is hard wired into our brains and has even been observed in 8-month-old babies (Xu & Garcia, 2008).

Scientists need statistical rigor to avoid giving in to the impulse to make overly strong conclusions from limited data.

WE TEND TO BE OVERCONFIDENT

Sometimes the phrase "90% confident" is not a result of a statistical calculation, but rather a way to quantify a subjective feeling of uncertainty. 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 that you are 90% confident contains the correct answer. Don't use Google to find the answer. Don't give up and say you don't know. Of course you don't know the answers precisely! The goal is not to provide correct answers, but rather to correctly quantify your uncertainty and come up with ranges of answers that you think are 90% likely to include the true answer. If you have no idea, answer with a superwide interval. For example, if you truly have no idea at all about the answer to the first question, answer with the range 0 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 meet the goal of being 90% confident, you will have created nine intervals that include the correct answer and one that excludes it.

Russo and Schoemaker (1989) tested more than 1,000 people and reported that 99% of them were overconfident. Almost everyone was too confident and answered with narrow ranges that miss the correct answer far more than 10% of the time. The goal was to create ranges that were correct 90% of the time, but most people created ranges that were too narrow and 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 are similar.

These results emphasize that you must distinguish computed confidence from informal confidence intervals that are informal guesstimates (even from an expert).

WE SEE PATTERNS IN RANDOM DATA

Most basketball fans believe in "hot hands"—that players occasionally have streaks of successful shots. People think that once a player has successfully made a shot, he is more likely to make the next shot, and that clusters of successful shots will happen more often than predicted by chance.

Gilovich (1985) analyzed data from the Philadelphia 76ers during the 1980–1981 basketball season. Players and fans both strongly agreed that a player

–	–	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 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? Most people tend to see patterns, but in fact the arrangement is entirely random. Each spot in the table had a 50% chance of having an "X."

was more likely to make a shot after making the last one than after missing it, and more likely to miss a shot after missing the prior shot. The data clearly show this is not the case. Additionally, the number of streaks of four, five, or six successful shots in a row was no larger than predicted by chance. The sequence of successes and failures was entirely random, yet almost everyone saw patterns.

Table 1.1 demonstrates the problem. 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 patterns. It just doesn't seem random.

In fact, Table 1.1 was generated randomly. Each spot had a 50% chance of being "X" (successful shot) and 50% chance of being "–" (not successful), without taking into consideration previous shots. The pattern is entirely random, as if it were the result of flipping a coin.

Although I know that the arrangement is entirely random, I can't help but see patterns. The X's *seem* to cluster together more than expected by chance alone, although they really don't. Our brains evolved to find patterns and do so very well. Too well! Statistical rigor is needed to avoid being fooled by apparent patterns among random data.

It is important that we recognize this built-in handicap. Our brains tend to find patterns among random data, so statistical methods are needed to make correct conclusions. Conversely, this makes it impossible to informally generate random numbers or assign subjects randomly to treatments. Attempts at informal randomization never have long enough runs of the same value. If you want random numbers, don't make them up. Flip a coin, throw dice, or use a computer program.

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 this set of coincidences would happen? Tiny. But that doesn't mean much. It is very unlikely that any particular coincidence will occur. But it is very likely that some astonishing set of unspecified events will occur. That is why remarkable coincidences are always noted in hindsight and never predicted with foresight.

WE HAVE INCORRECT INTUITIVE FEELINGS ABOUT PROBABILITY

Imagine that you can choose between two bowls of jelly beans. The small bowl has 9 white and 1 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 1 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 clearly, 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, it also has more white beans and more chances to lose. Our brains simply are not evolved to deal with probability sensibly, and most people make the illogical choice.

Another example: Many people rated cancer as riskier when it was described as killing 1,286 of 10,000 people than when it was described as killing 24.14 of 100 people, although the latter is double the risk (Yamagishi, 1997).

WE AVOID THINKING ABOUT AMBIGUOUS SITUATIONS

Imagine that you have to choose between two urns. The first urn contains exactly 50 red jelly beans and 50 black jelly beans, randomly mixed together. The second urn also contains 100 jelly beans. Some are red and some are black, but you don't know how many of each. You can reach in and pick a jelly bean at random, but can't see a bean until you choose it. You will win a prize if you happen to pick a red jelly bean, and can choose which urn to select from. Which urn should you choose from?

Now the rules change, and you will win a prize if you happen to pick a black bean. Which urn should you choose from?

Almost everyone chooses the first urn in both cases (Ellsberg, 1961). There is nothing in the problem that tells you whether there are more red jelly beans or more black ones in the second urn, so you are equally likely to pick either. Yet almost everyone prefers to choose from the first urn.

Choosing from the first urn requires that you think about probability—the chance of randomly picking from a 50:50 mixture of red and black jelly beans. Choosing from the second urn is more complicated because it combines ambiguity (you simply don't know whether it contains more red jelly beans, more black ones, or an equal number of each) and probability. Thinking about the second urn makes us feel uncomfortable. Use of functional magnetic-resonance imaging to map blood flow in the brain demonstrates that different parts of the brain deal with risk (probability) and ambiguity. When one thinks about an ambiguous situation (analogous to the second urn above), activity in the fear center in the amygdala increases and activity in the reward center in the caudate decreases (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005). Our brains don't like thinking about ambiguous situations, and this prevents us from logically comparing the two situations.

WE FIND IT HARD TO COMBINE PROBABILITIES

Here is a classic brain teaser called the Monty Hall problem, named after the host of a game show. You are a contestant on a game show and are presented with three doors. Behind one is a fancy new car. Behind the others are worthless prizes. 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; 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, and the host opens one of these. If you switch, you'll switch 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 one doesn't. The host knows which door the car is behind and opens the other one. This means that the remaining closed door must be the winning door. If you originally picked one of the two wrong doors, then switching will certainly lead you to win.

Let's recap. If you originally chose the correct door (which has a one-third chance), then switching will make you lose. If you originally picked either of the two losing doors (which has a two-thirds chance), 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). It is very hard to simultaneously think through two (or more) parallel tracks.

WE DON'T DO BAYESIAN CALCULATIONS INTUITIVELY

Imagine this scenario. You are screening blood samples for the presence of human immunodeficiency virus (HIV). The prevalence of HIV is quite low (0.1%) among the selected donors. The antibody test is quite 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 (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 in 99 (99%) of them. The other 99,900 people do 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 and 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 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 wired to combine what we already know (the prevalence of HIV) with the new knowledge (the test is positive).

If the same test is used in a different situation, the results would be different. Imagine the same test is used in a population of IV-drug users in which you expect the prevalence of HIV to be 10%. Again, let's imagine that 100,000 people are tested. Of these, 10,000 (10%) will have HIV, and the test will be positive in 9,900 (99%) of them. The other 90,000 people do 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 and 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 the prevalence of the disease. To reach the correct conclusion, one must combine a baseline frequency with new data. This example gives you a taste of what is called Bayesian logic (which will be discussed more thoroughly in Chapter 18).

WE ARE FOOLED BY MULTIPLE COMPARISON

Austin, Mamdani, Juurlink, and Hux (2006) "mined" a database of health statistics of 10 million residents of Ontario, Canada. They examined 223 different reasons for hospital admission and tested each to see whether it occurred more often in people born under each astrological sign. Seventy-two diseases (reasons for hospital admission) occurred statistically significantly more frequently in one astrological sign than in all the others put together. This means that in each of those 72 cases, the results would occur by chance alone less than 5% of the time (you'll learn more about what "statistical significance" means later in this book).

Sounds impressive, doesn't it? Makes you think that there really is some relationship between astrology and health. But the study wasn't really done to investigate any association between astrological sign and disease; rather, it was done as a warning about the difficulties of interpreting statistical results when many comparisons are performed.

It is misleading to focus on the strong associations between one disease and one astrological sign without considering the others. Austin et al. (2006) examined 223 different reasons for hospital admissions and asked whether each occurred more often in each of 12 astrological signs. Therefore, they made 223 × 12 = 2,676 distinct comparisons. If there truly is no association between astrological sign and disease (and there is no reason to think there is), you'd expect that just by chance, 5% of these comparisons would have P values less than 0.05. Because 5% of 2,676 = 134, one would expect to find about 134 significant associations just by chance. So it is hardly impressive that they found 72 significant associations. That's fewer significant results than you'd expect to find purely by chance.

One of the comparisons was truly striking. People born under the sign of Taurus had 27% more admissions for diverticulitis of the colon. The chance of observing this large a difference in incidence rates by chance alone is 0.0006. This means that if there truly were no association between diverticulitis and being born under the sign of Taurus, the chance of seeing such a striking difference in hospital admissions rates, by chance alone, is 0.06%.

This sounds impressive. Could it be real?

By chance alone, you'd expect to see a P value less than 0.0006 in 1 of 1,667 comparisons (1/0.0006). Since these investigators made nearly 3,000 comparisons of different diseases with different astrological signs, a P value less than 0.0006 is not surprising. You expect such a small P value purely based on chance.

Our brains evolved to spot patterns and are good at it. So we notice when a particular disease is more frequent among those born under a particular astrological sign. It doesn't seem natural to correct for multiple comparisons, but this is essential if you don't want to be fooled by chance associations.

Chapters 22 and 23 explore multiple comparisons in more depth.

WE TEND TO IGNORE ALTERNATIVE EXPLANATIONS

Imagine this scenario (adapted from Bausell, 2007). You are doing a study of acupuncture for 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, but statistical calculations show that such consistent findings are exceedingly unlikely to happen by chance. Therefore, the acupuncture must have worked. Right?

Not really. The decrease in recorded pain may not be caused by the acupuncture. Here are five alternative explanations:

- Placebos reduce pain considerably. 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.
- 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 recommends a change in aspirin dose, a change in exercise, or nutritional supplements. The decrease in reported pain might be due to these aspects of the treatment, rather than the acupuncture.
- What if three patients experience worse pain with acupuncture, whereas the others get better? The experimenter reviews the records of those three patients carefully 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. These kinds of manipulations of the data, although well intentioned, are fraudulent and may explain all the pain relief observed in the study.

- 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 *regression to the mean*.

WE ARE FOOLED BY REGRESSION TO THE MEAN

Figure 1.1 illustrates simulated pressures. All values were randomly chosen in the same manner. The graph is divided into two columns (before treatment and after treatment) but the values were randomly chosen without regard to the labels. Figure 1.1A illustrates 24 pairs of values. The "before" and "after" groups are about the same. In some cases the value goes up 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).

Now imagine the study were 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 the 12 individuals with the highest before values. In every case but 1, the after values are lower. If you performed a statistical test (paired t test; see Chapter 31), the results would seem to be extremely convincing. The graph on the bottom-right illustrates the other 12 pairs, those with low values when measured before. In all but 2 pairs, the values go up. Again, these values alone would seem to be convincing evidence that the treatment brings down the value measured (blood pressure).

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

Variation in blood pressure (and almost any other variable) has two components. Some of the variability is biological. However, this example was constructed to have no systematic (biological) difference between the before and after values. The rest of the variation is random. All the variation in this example is random. For Figure 1.1C, we selected subjects who happened to have the highest blood pressures. When blood pressure is assessed again, there is no reason to expect that the random factor will again lead to a high pressure. So, on average, the after measurements are lower. This is not because of any effect of the treatment, but is purely a matter of chance. When we selected only the people who happened to have low blood pressure, the treatment appeared to cause a substantial increase.

When you select individuals because some measurement is particularly high, a later measurement is likely to be lower. This effect is called *regression to the mean*. People who are especially lucky at picking stocks one year are likely to be less lucky then next year. People who get extremely high scores on one exam are likely to get lower scores on a repeat exam. An athletes that does extremely well in one season is likely to perform more poorly the next season. Athletic

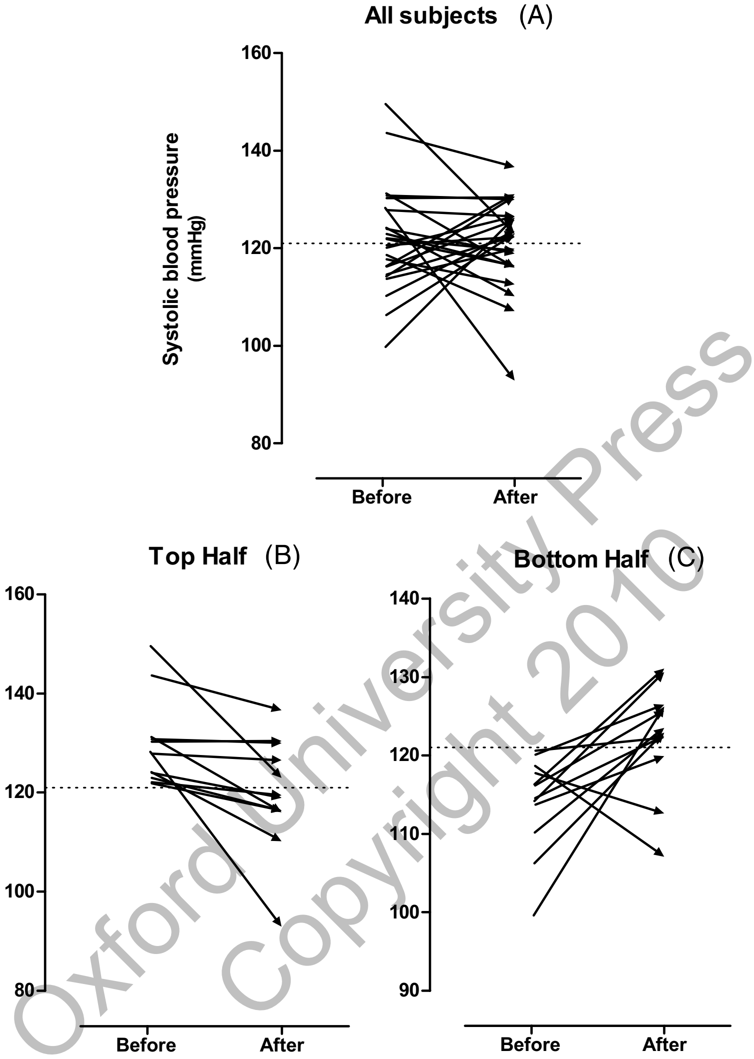

Figure 1.1. Regression to the mean.

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) includes 48 random values, divided arbitrarily into 24 before–after pairs (which overlap enough that you can't count them all). (B) includes only the 12 pairs with the highest before values. In all but 1 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 before and after had a large impact on blood pressure. In fact, these graphs simply illustrate random values, with no change between before and after. The apparent change is called *regression to the mean*.

performance certainly requires great skill, but random factors also play a major role and will cause regression to the mean. 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).

Answers to the ten questions in the 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