CHAPTER 6

Samples, Uncertainty, and Statistical Inference

What You'll Learn

- All quantitative estimates are the sum of three terms: the true quantity of interest, bias, and noise.
- Statistical hypothesis testing allows analysts to assess whether an estimate was likely to have arisen from noise.
- Statistical significance and substantive significance are not the same and should not be conflated.

Introduction

Chapters 4 and 5 articulated tools that allow us to describe a relationship between variables within a data set. We need variation in both variables and then we can describe the correlation between those variables using regression. But often we want to go further. We want to use the relationships between variables that we find in the data we have (our *sample*) to make inferences about relationships that hold between those variables in the larger world (the *population* of interest). For instance, once we've found that crime is higher on warm days in 2018, we'd like to know whether we are justified in concluding that this relationship is likely to hold in other years and isn't simply an artifact of the 2018 data. That is, we want to know whether an observed relationship in a sample of days reflects a genuine phenomenon in the population of days or whether it happens to be true in the sample of data that we looked at by chance (or dumb luck). In this chapter, we discuss some tools that help us to adjudicate between these possibilities.

Estimation

To start thinking about this question, we need a common language to talk about the differences between the things we observe in our sample and the phenomena in the population that we'd like to learn about. To do so, we are going to use the following simple equation, which will come up so often throughout the rest of the book that, from now on, we are going to start calling it *our favorite equation*:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

We are going to explain each of these terms carefully as we go. But let's start with some basic definitions.

The *estimate* is the number we get as a result of our analysis. The *estimand* is the true quantity of interest in the population that we are trying to learn about. Our hope is that our estimate closely approximates our estimand. An estimate can differ from the estimand for two reasons: bias and noise. *Bias* refers to errors that occur for systematic reasons, and *noise* refers to idiosyncratic errors that occur because of chance.

Let's set the stage with a simple example that will allow us to define and understand these terms more clearly.

Suppose we conduct a poll to learn which of two candidates (a Republican and a Democrat) is going to win an upcoming election. We can think of this as a prediction problem: we are collecting data to forecast the future winner. But we can also think of this as pure description: we want to know the proportion of voters who support one candidate over the other.

In either case, a key challenge is that there are too many voters for us to ask all of them their opinions. Necessity forces us to poll a *sample*, constituting only a small proportion of the total *population* of voters. Thus, we need to figure out what we can conclude about political views in this larger population from evidence generated by a poll of only a relatively small sample.

In our example, we are interested in learning the proportion of voters in the population who support the Republican. Let's call that proportion, which is a number between 0 and 1, q. Since there are only two candidates, the proportion who support the Democrat is just $1 - q$. So q is our estimand. Until we actually hold the election, we don't get to observe q; we have to try to estimate it.

Suppose we poll a random sample of 100 voters and ask them whether they will support the Republican or the Democrat. We could estimate the number of voters in the population who support the Republican (which we can't observe) by calculating the proportion of people in our sample who support the Republican (which we can observe). Let's call our estimate from our sample $\hat{q}$, which we pronounce "q-hat." Following standard practice, we will notate estimands with a letter (it need not be q) and we will notate estimates of that estimand using that same letter with a hat over it. In this case, our hope is that our estimate, $\hat{q}$, is close to the estimand, q.

In this example, the *estimand* is the true proportion of Republicans in the population (q)—it is the unobserved quantity that we are trying to learn about with our data analysis. The process of sampling 100 voters and calculating the proportion who support the Republican is called the *estimator*—it is the procedure we apply to generate a numerical result. The proportion of Republicans in our sample ($\hat{q}$) is our *estimate*—it is the numerical result arising from the application of our estimator, which we hope approximates the estimand.

By understanding the distinction between estimates and estimands, we can take a first step toward making sense of our favorite equation:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

The quantity we are interested in is the estimand. The quantity we observe in the data is the estimate. In an ideal world, the estimate would equal the estimand, so that our estimator would reveal to us the true quantity of interest. But our favorite equation says this isn't the case. Estimates differ from estimands because of bias and noise. To understand why, we need to learn more about those two troublesome quantities.

Why Do Estimates Differ from Estimands?

Bias and noise are both important to understand. But they differ, and their difference is often lost on people, leading to unclear thinking. So we'll take them in turn. But before discussing bias and noise in detail, an analogy might help.

Anthony likes to play the Scottish game of curling. In curling, two teams take turns sliding heavy granite stones down a long sheet of ice. As a stone slides, other team members sweep in front of the stone like crazy while running along the ice. We recommend watching a video; it's pretty fun. Anyway, the team with the stone closest to the center of a target on the far end of the sheet of ice (called the button) scores points.

Anthony's quite good at curling. He can more or less get his stones to go where he wants. But, despite his skill, sometimes his stones miss the button (you're not always trying to "draw to the button" in curling, but for the purposes of this discussion, we'll assume that this is your goal). Why is this? Well, there are all sorts of factors outside the control of the thrower that affect how a stone slides. Maybe there was some debris on the sheet, causing a well-aimed stone to divert off course. Or maybe Ethan slipped on the ice while trying to sweep and that accidentally "burned the stone." Anyway, for all of these reasons, Anthony's well-aimed stones might miss the button.

Ethan, by contrast, is terrible at curling. So when he goes curling with Anthony, his stones frequently miss the target, typically to the left (let's not even talk about distance). He'd like to claim it is because of idiosyncratic factors, like with Anthony's misses. But if that were true, he wouldn't be more likely to miss left than right. No, the truth is, Ethan's technique is poor, so his stones are systematically improperly aimed.

There is, we think, a useful analogy between curling and data analysis. Think of the button as the estimand: it is the truth you are aiming at. Think of your estimator as the act of sliding a stone down the ice. And think of the outcome of one stone throw as an estimate arising from one iteration of the estimator.

Your stone (estimate) might miss the button (estimand) for two reasons. First, like Anthony, you may have aimed well, but random factors may have moved the stone this way or that. These random factors are like noise. Since these factors are random, on average, they don't make Anthony miss more to the left or to the right. Indeed, on average, his stone's location is on the button. But that doesn't mean every individual stone finishes on the button; his misses just average each other out. This is what noise does—estimates can equal the estimand on average, but because of noise, any given estimate may not equal the estimand.

Second, like Ethan, you may systematically aim too far to the left. There is still noise, so you might sometimes miss to the right. But, on average, your stone finishes to the left of the button. These systematic errors are like bias. Unlike Anthony, Ethan's average stone misses the button. This is what bias does—when there is bias, even the average estimate doesn't equal the estimand, let alone any given estimate.

Okay, now that we have an analogy to help us see the difference between bias and noise, let's talk about them in a bit more detail.

Bias

One reason an estimator might give you an estimate that differs from the estimand is because it is *biased*. Imagine applying your estimator over and over again an infinite number of times, each time to a new, independent sample of data. Doing so would generate an infinite number of estimates. Because of noise, some of those estimates will be

bigger than the estimand (i.e., you'll get a larger share of Republicans in some of your samples than there are in the population) and some of those estimates will be smaller than the estimand (i.e., you'll get a smaller share of Republicans in some of your samples than there are in the population). But you would like the average of that infinite number of estimates to be equal to the estimand. That is, you don't want to predictably (or systematically) over-estimate or under-estimate the number of Republicans. You want to be aimed at the truth. We say that an estimator is *unbiased* if the average value of the estimates it generates would equal the estimand if we repeatedly applied the estimator to new, independent samples an infinite number of times.

We also sometimes talk about the average value of a variable over an infinite number of draws in terms of *expectations*. So, we might say that an unbiased estimator equals the estimand, *in expectation*. Or we might say that the *expected value* of an unbiased estimator is the estimand.

There are lots of reasons a political poll might be biased. Suppose voters systematically lie to pollsters. Perhaps voters believe that pollsters are themselves likely to be Democrats and the voters want to please the pollsters, so some Republican voters report supporting the Democrat. Then our estimator will be biased in favor of Democrats—on average, reporting more voter support for Democrats than there really is. Or suppose Democrats are more likely to turn out to vote than Republicans, but equally likely to answer polls. Then poll respondents will differ from voters, and the estimates from polls will be biased in favor of Republicans—on average, reporting more voter support for Republicans than there actually is. Finally, what if pollsters contact people by phone and phone owners are systematically different in their political leanings than the population as a whole? This will also lead to bias. For lots of reasons, if we ran the poll an infinite number of times and averaged the estimates, that average might not equal the true proportion of Republicans in the population of voters, which is our estimand. Thus, the poll could be biased for any of these reasons.

In subsequent chapters, we will be very concerned with thinking about sources of bias. For the remainder of this chapter, however, we are going to ignore these potential sources of bias to focus on the second potential problem with estimators, noise.

Noise

When you take a sample of the population, you inevitably introduce some noise into your estimate. When you ask 100 randomly selected people out of 100 million their opinions on political candidates, sometimes by chance you happen to talk to a disproportionate share of Republicans and sometimes you happen to talk to a disproportionate share of Democrats. As a result, even without bias, any individual estimate need not equal the estimand. Suppose your estimand is unbiased. If you applied it an infinite number of times, you would not over-estimate Republican or Democrat support, on average. But each individual estimate would likely differ somewhat from the estimand because of noise—that is, natural variability that results from sampling. This natural variability is sometimes referred to as *sampling variation*, a common source of noise.

We have ways of quantifying the amount of noise associated with an estimator. Think about repeatedly applying an estimator with new, independent samples of data an infinite number of times. The closer the various estimates would be to each other, the more *precise* is the estimator. Thus, a more precise estimator is one with less noise.

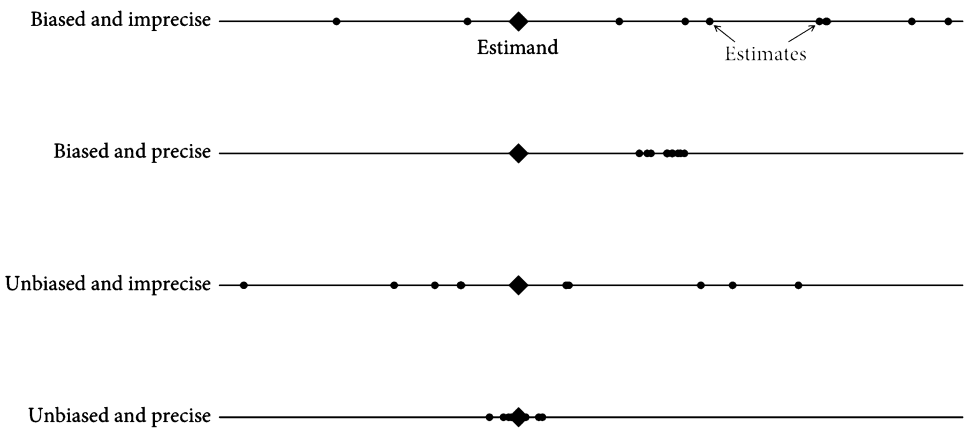

Figure 6.1. Understanding the difference between *unbiased* and *precise*.

What Makes for a Good Estimator?

In the end, we are trying to learn the true value of the estimand. Since our estimate can differ from the estimand because of bias or noise, what we really want is an estimator that is both unbiased and precise.

If our estimator is unbiased but imprecise, our estimates will typically differ from the estimand because there is so much noise. For instance, in our polling example, if we talk to just one voter at random, their opinion is an unbiased estimate of the average opinion in the electorate (if you did this an infinite number of times, q of those times you'd get a Republican, and $1 - q$ of those times you'd get a Democrat). But the sampling variation associated with estimating voter opinion by asking just one person's opinion is huge—we will always estimate either 100 percent Republicans or 100 percent Democrats.

If our estimator is biased but precise, our estimates will typically differ from the estimand because they are very precisely estimating the wrong quantity. For instance, if we sample ten thousand voters, but only do so in Republican neighborhoods, we will get answers clustered very tightly around each other, but they will systematically over-estimate the number of Republicans.

Figure 6.1 illustrates that estimators can be unbiased, precise, neither, or both. The black diamonds represent the estimand—the true value in the world we are interested in. The gray dots show various estimates that arise from repeated applications of a given estimator, each time with an independent sample of data. If the gray dots are symmetrically distributed around the diamond (like Anthony's curling stones around the button), the estimator is unbiased. That is, the estimates it provides are right on average. If the gray dots are clustered tightly together, the estimator is precise. That is, there is very little noise, so the estimator yields similar estimates with each iteration. All else equal, we would obviously like to have our estimator be less biased *and* more precise. However, sometimes there are trade-offs between these goals, and we have to decide how much bias we're willing to accept for a certain gain in precision.

For a concrete example of the possible trade-offs between bias and precision, let's return to the topic of polling. Suppose you have $2,000 and you'd like to conduct a reliable poll to understand how popular a political candidate, policy proposal, product, or potential advertising campaign would be. You could post the survey online, pay

people twenty cents per response, and obtain ten thousand responses. Or you could pay a professional polling firm to obtain a random, representative sample at a cost of $20 per response, meaning you'll be able to afford only one hundred responses.

The online convenience sample is much larger, so your estimates of public opinion will be more precise, but they'll also likely be biased. The kinds of people who voluntarily take surveys for modest compensation are not likely to be representative of the general population. The professionally conducted survey will likely give you less biased estimates, but the sample size will be smaller, so your estimates will be less precise.

This kind of trade-off between bias and precision is quite common for data analysts, and we'll see more examples in part 3. The right way to make this trade-off will depend on your goals, the costs of different kinds of errors, and the particular question you are hoping to answer.

If an estimator is unbiased, we'd also like it to be as precise as possible. And as we've discussed, we might even allow for a little bias in exchange for a big gain in precision. But if an estimator is really biased, it's no longer obvious that precision is a good thing. For one thing, a precise biased estimator will never be anywhere close to the truth. Whereas with less precision, you might sometimes make good predictions despite the bias, albeit by accident. (Ethan would probably be better off if there was an earthquake just as he released his curling stone because at least sometimes his stone would stay in play.) Furthermore, precision might give you a false sense of confidence. Beware a precise estimate with an unknown bias.

Quantifying Precision

Remember the motivating question for this chapter: When we estimate something from a sample of data, how confident should we be in drawing inferences about the larger population? As we've seen, if our estimator is biased, we should certainly be worried. But even if our estimator isn't biased, we still have to be worried that our estimates do not reflect the true relationship in the larger population (the estimand) because of noise. In order to know how worried we should be about this possibility, we need to quantify the precision of an estimator. We do so through a statistic called the standard error, which we can then use to construct confidence intervals.

Standard Errors

In chapter 2, we talked about the standard deviation as one way to measure how spread out a variable's distribution is (or, equivalently, how variable the variable is). Well, imagine that we repeated our estimator an infinite number of times, each time with a newly drawn sample of data. In that thought experiment, we could think of the estimate itself as a variable. Each time we draw the data and run our estimator, we get a different value of the estimate because of the noise. So, we can imagine the distribution of estimates we would get after repeating our estimator an infinite number of times. That imagined distribution is called the *sampling distribution*. The standard deviation of that sampling distribution is called the *standard error*. The standard error, if we knew what it was, would give us a sense of how far any given estimate will be from the average estimate, since it measures how variable our estimates will be. If the estimator is unbiased, the average estimate equals the estimand. So, for an unbiased estimator, the standard error tells us approximately how far a typical estimate is from the estimand, which is the true value we are trying to learn about.

If the standard error is large, then the estimates would be very spread out and the estimator is relatively imprecise (i.e., there is a lot of sampling variation). If the standard error is small, then the estimates would be very close together and the estimator is relatively precise (i.e., there is little sampling variation). Look back at figure 6.1. The third row shows an example of some estimates from repeated runs of an estimator with a relatively large standard error—as a consequence, the estimates we see are quite spread out. (Of course, we aren't seeing the full sampling distribution since we don't have infinite estimates.) The fourth row shows an example of some estimates from repeated runs of an estimator with a relatively small standard error—as a consequence, the estimates we see are tightly clustered.

We can provide some insight into what makes an estimator precise or imprecise (i.e., when the standard error will be small or large). In our polling example, the standard error is approximately equal to $\sqrt{\frac{q(1-q)}{N}}$, where N refers to the sample size (the number of people polled). While we aren't going to show you how to derive this formula (a topic for a different book), we can learn some things about what makes an estimator more or less precise by thinking about the formula.

Let's start with understanding the numerator, $q(1-q)$. Notice this term is maximized at $q = \frac{1}{2}$ and decreases as q gets larger or smaller. So, suppose the true proportion of Republicans in the population, q, is either very large (close to 1) or very small (close to 0). This makes $q(1-q)$ very small and, therefore, makes the standard error small. Why? When q is very large or very small, there is little possibility of sampling error. If 99 percent of voters are Republicans, when you collect your sample of, say, one thousand voters, it will be very unlikely that you find many Democrats. By contrast, if q is close to one-half, the standard error is large. This reflects the fact that there is lots of room for sampling error. You could easily find a 55-45 or 45-55 split in your sample of data drawn from a 50-50 population. The closer q is to one-half, the more natural variation there is in our outcome of interest, making the standard error larger.

Now consider the denominator. It tells us that as the size of our sample increases, our standard error goes down. This makes sense. The problem of imprecision comes from the fact that our sample might not accurately reflect the whole population. When the sample is large, it will more closely approximate the population. We can more precisely estimate the opinions of a million people by talking to ten thousand people than by talking to ten people.

The formula for the standard error actually tells us something a little more subtle than just that small sample sizes lead to imprecision. It tells us that the standard error shrinks in proportion to $\sqrt{N}$. Suppose the true proportion of Republicans in the population is $q = .5$. Then, if we took a poll of 1000 voters, we'd have a standard error of $\sqrt{\frac{.5 \cdot .5}{1{,}000}} \approx .016$. Suppose we conducted a much larger poll of 10,000 people. Then our standard error is $\sqrt{\frac{.5 \cdot .5}{10{,}000}} = .005$. So increasing the sample size by a factor of 10 only improves the precision of the poll by approximately threefold. If we further increased the sample size to 100,000, we'd get another roughly threefold improvement in precision, ($\sqrt{\frac{.5 \cdot .5}{100{,}000}} = .0016$). In other words, there are diminishing returns to bigger and bigger sample sizes. The standard error of a survey with 10,000 respondents is already tiny, and adding more respondents doesn't meaningfully improve precision.

One tricky thing you might have noticed is that we need to know q in order to calculate the standard error. But we don't know q; that's why we're doing the poll to begin

with. In practice, we approximate the standard error by substituting $\hat{q}$, our estimate of q, into the formula. Of course, this approximation would run into problems if you had a really small N or a value of q that was really close to 0 or 1. Suppose you talked to five people and found that none of them were Republicans. Thoughtlessly applying the procedures above, you would wrongly conclude that nobody is a Republican and that your standard error is 0. Of course that's wrong, and that's because with small samples and with extreme values of q, your approximation using $\hat{q}$ is misleading.

We should also point out that although there is a nice formula for approximating the standard error in our polling example, this won't always be the case. Fortunately, our computers can often produce reasonably reliable approximations of standard errors, even in more complicated circumstances.

Small Samples and Extreme Observations

It is worth pausing to note that the fact that small samples lead to imprecision explains a common phenomenon that you may have noticed out there in the world. If you look up data on the towns with the highest or lowest cancer rates or the highest or lowest average income, you will find a list of towns with a pretty small number of residents. Similarly, if you look up the schools with the highest or lowest average test scores, you will find a list of schools with a small number of students. Why is this?

Think of the average cancer rate or income in a town as an estimate of the national cancer rate or income, just like the average support for Republicans in a polling sample is an estimate of the average support for Republicans in the whole population. When the number of residents in a town is small, that is equivalent to having a small sample size. That leads to less precision (more noise) in your estimate. That means it is more likely your estimate will have an extreme value in either direction. Small towns tend to dominate the list of places with extreme cancer rates or average incomes, not because they are necessarily on average more or less cancer prone or more or less wealthy, but because their cancer rates and average incomes are more variable than places with more people to average over.

To see this in the extreme, imagine a town with just one resident. That town either has a 100 percent cancer rate or a 0 percent cancer rate. But a town with one hundred thousand residents is going to have a cancer rate somewhere in the middle, much closer to the national average.

Figure 6.2, inspired by a similar graph in Howard Wainer's *Picturing the Uncertain World*, illustrates the point. The figure plots data from California middle schools in 2012. We observe students' average academic performance (the Academic Performance Index, which is largely determined by standardized test scores) and the size of the student body for each school. The hollow data points represent the very worst performing schools (bottom 5 percent on academic performance), and the solid-black data points represent the very best performing schools (top 5 percent on academic performance). As you can see from the regression line, there is actually a positive correlation in this data between school size and academic performance—on average, larger schools perform better than smaller schools. But, more importantly for us, small schools are overrepresented in both groups.

Understanding that small sample sizes lead to imprecision is important for many reasons. One is that, as Wainer points out, failing to think clearly about the issue can lead to bad decision making. The Bill and Melinda Gates Foundation spent billions of dollars on an ultimately ineffective small schools initiative. The evidence that led

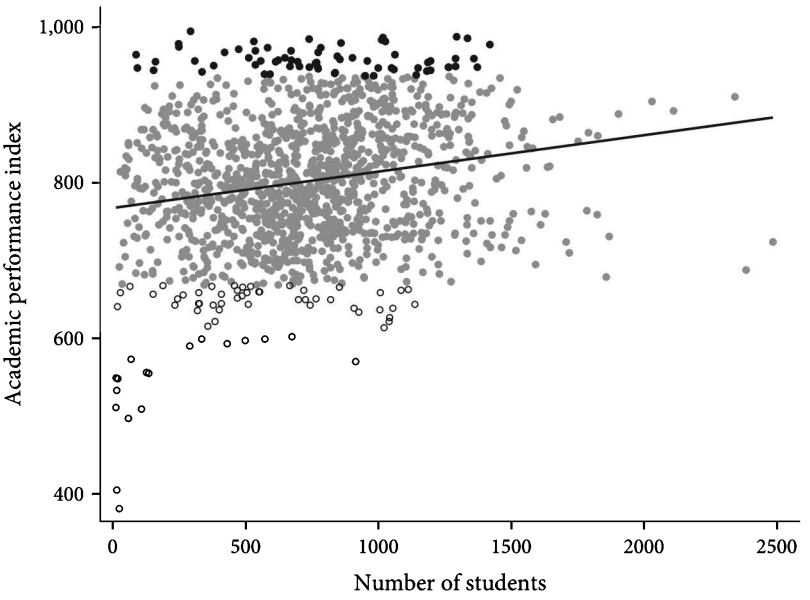

Figure 6.2. A scatter plot and regression line showing a slight positive correlation between average academic performance and school size for California middle schools in 2012.

them to make this misguided investment was the observation that schools with a small number of pupils were over-represented on lists of schools with the best test scores. Had they thought a little more clearly, they would have also checked the lists of schools with the worst test scores and found small schools over-represented on those lists too.

Confidence Intervals

Another way we often quantify precision is through a confidence interval.

An important mathematical fact, called the Law of Large Numbers, tells us that as our sample size gets really big, the noise will essentially disappear. But how big is big enough?

Another important mathematical fact, called the Central Limit Theorem, tells us that if our poll is indeed unbiased, then if we were to repeatedly run our poll, approximately 95 percent of our estimates ($\hat{q}$, Republicans in our sample) will end up being within approximately 2 standard errors of our estimand (q, Republicans in the population). Therefore, pollsters will often report what they call the *margin of error*, which is simply twice the standard error.

Researchers and pollsters also sometimes report what they call the *95% confidence interval*. This is the interval that ranges from the estimate ($\hat{q}$) minus two times the standard error up to the estimate plus two times the standard error.

The 95% confidence interval is a source of some confusion. Often, people will casually say that we're 95 percent confident that the true value lies within the 95% confidence interval. But that's not quite right. The correct statement is a lot clunkier. Technically, we can say that if there is no bias and if we repeated our estimator an infinite number of times, the true estimand would be inside the 95% confidence interval 95 percent of the time.

To get a picture in our head of how confidence intervals work, let's go back to our curling analogy. Suppose Anthony pushes an infinite number of stones down an infinite number of ice sheets. Think of the spot on the ice where the exact center of his stone comes to rest as the estimate. That estimate is extremely unlikely to be sitting exactly on the button—your estimate is almost never exactly equal to the true value of the estimand. But the stone is wider than that one spot. So we might ask how often the button will be touching some part of the stone. That will depend on how wide the stone is. (Of course, there is a regulation width in curling, but allow us a little poetic license. We aren't Scottish. And this isn't the Olympics.) We could find the exact width of a stone such that the button would be touching some part of the stone on 95 percent of Anthony's throws. This is like the 95% confidence interval. It isn't that, on any one throw, we're 95 percent confident that the button is covered by some part of the stone. It is that, on 95 percent of throws, the button is covered by some part of the stone.

This analogy can also help us think about confidence intervals other than the 95% confidence interval. Sometimes we want to be more confident than 95 percent. So we might be interested in the 99% confidence interval. Ask yourself whether the 99% confidence interval is wider or narrower than the 95% confidence interval. If Anthony wants 99 percent of his throws to end with some part of the stone touching the button, he is going to need to use a wider stone. So the 99% confidence interval is wider than the 95% confidence interval. We have to admit a larger range of possible Republicans if we want to be sure our estimate is within that range 99 percent of the time rather than just 95 percent of the time.

Statistical Inference and Hypothesis Testing

Now we can finally turn to this chapter's motivating question: How do we make inferences about populations using estimates from samples? Let's stick with our polling example. As we emphasized, when we conduct a poll, even if we think it is unbiased, we also want it to be precise because we want to know that it is giving us an estimate that's close to the truth. How can we assess this? How well does a sample of, say, one thousand voters estimate the views of the 140 million Americans that will determine who wins an upcoming presidential election? Let's see.

Hypothesis Testing

Often, we want to assess some particular hypothesis. For instance, we might want to know whether it is reasonable to believe that the estimand is greater than, less than, or different from some particular reference point. For that, we need to think about *hypothesis testing*.

In the example of an electoral poll, we might like to know which candidate is going to win the election. Suppose we conducted an unbiased poll of one thousand voters and this yielded an estimate of the Republican candidate's vote share of $\hat{q} = .532$ or 53.2 percent. How confident should we be that the Republican is actually going to win the election—that is, how confident should we be that more than 50 percent of voters will vote for the Republican or, put differently, that $q > .5$? Hypothesis testing provides a way for us to say something about that.

One way of thinking about this question is as follows. We have some evidence from our poll that the Republican candidate is more popular than the Democrat. But we want to know how good that evidence is. That is, we want to know how likely it is that

we could have observed such evidence *even if the Republican is not more popular than the Democrat*. So we test how likely it is that we would have observed the evidence we observe if the two candidates were actually equally popular. This *no relationship* benchmark is typically referred to as the *null hypothesis*.

To understand the hypothesis test, start by assuming the null hypothesis is true—that is, the two candidates are exactly equally popular, so $q = .5$. Now ask how likely would it be for us to obtain a poll result at least as favorable for the Republican as the one we actually found, $\hat{q} = .532$.

We already have the information we need to answer this question. With a true vote share of $q = .5$ and a poll of one thousand voters, the standard error of our estimator is approximately 1.6 percentage points ($\sqrt{\frac{.5 \cdot .5}{1{,}000}} \approx .016$). Our estimate of .532 is 2 standard errors above the null hypothesis ($.5 + 2 \cdot .016 = .532$). (We didn't choose these numbers by accident.)

As we said earlier, the Central Limit Theorem tells us that 95 percent of estimates from an unbiased poll of the sort we ran will fall within 2 standard errors of the truth, meaning that only 5 percent of estimates will fall more than 2 standard errors from the truth. Moreover, in half of those unfortunate cases, the estimate will be 2 standard errors *below* the truth (showing the Democrat to be notably ahead). So if the null is true, the probability that we'd get a poll result as favorable for the Republican as the one we got is about 2.5 percent, or 1 in 40.

We obviously picked numbers that would make this calculation straightforward, but your computer can do this calculation for any poll result. Statisticians are in the business of developing methods for conducting these calculations. In statistics lingo, the analysis we just did is called a *one-sided z-test*. You don't need to know about z-tests to understand the rest of this book, but if you want to learn about them, you can consult virtually any statistics book. (Wikipedia is pretty reliable for such material as well.) More generally, the important thing is that hypothesis testing is a strategy for assessing the probability of getting a result as extreme as yours under the assumption that the null hypothesis is true.

Statistical Significance

We just saw that if the null is true, the probability we would have gotten a result as favorable to the Republican as what we found is only .025. This probability is called a *p-value*. If our p-value is really low, then we might conclude that the null is unlikely to be true. Thus, we have some statistically compelling evidence that the Republican is indeed favored by voters—if the true vote share were evenly split, it's quite unlikely that the poll result would be this favorable to the Republican (and even more unlikely if the true vote share favored the Democrat).

A common strategy is to pre-specify a particular threshold (most commonly, .05), and if the p-value is below that threshold, then we say we reject the null hypothesis and conclude that we have *statistically significant* evidence for the hypothesis we were testing against the null hypothesis.

Of course, hypothesis testing does not provide certain conclusions. With a significance threshold of .05, there's a 5 percent chance of obtaining a statistically significant result even if the null hypothesis is true. But hypothesis testing provides one way of thinking quantitatively about whether a pattern or result you have detected in your data set is likely to reflect a genuine phenomenon rather than simply being the product of noise.

One common error is to assume that the p-value tells you the probability that the null hypothesis is true. It doesn't. It tells you the probability of getting an estimate as extreme as the one you got if the null is true. Those two numbers are typically different. Indeed, to calculate the former quantity, you'd have to have a lot more information (e.g., how likely you thought it was that the null was true before you saw the evidence). We'll discuss these issues in part 4.

Statistical Inference about Relationships

So far, we've developed our ideas about bias, noise, and hypothesis testing in a simple setting where we are just trying to learn about the share of voters who support the Republican candidate. But all these concepts and tools of statistical inference can be applied to much more interesting problems, including estimating relationships like correlations.

Suppose we ran a regression to estimate the relationship between some outcome variable and some explanatory variable. The previous chapter discussed how we can utilize linear regression to find coefficients that describe the relationship between two variables in a data set. But now let's think about this in terms of estimation and statistical inference.

Suppose our data set consists of information on the income and education of a random sample of one thousand workers, but we are actually interested in the average relationship between income and education in the population of all workers. So we are trying to make inferences about a correlation in the population (our estimand) based on a correlation in the data (our estimate). How do we do this?

Start with the following equation, which describes the relationship between income and education in the population.

$$\text{Income}_i = \alpha^{OLS} + \beta^{OLS} \cdot \text{Education}_i + \text{error}_i$$

This equation is just like the one that we studied in chapter 5. Income_i is person i's income, Education_i is person i's years of education, and error_i is the difference between person i's income and the income predicted by the OLS regression line for a person with person i's education. The parameters α^{OLS} and β^{OLS} take whatever values minimize the sum of squared errors across the population. For example, β^{OLS} is the average extent to which income increases with each additional year of education in the population. These parameters α^{OLS} and β^{OLS} are features of the world. We don't know them. But, since we would like to know, on average, how income changes with education, β^{OLS} is our estimand.

We don't know β^{OLS} because we don't observe the income and education of every single person in the population. But we can estimate it by applying linear regression to our data on one thousand workers. Following our convention of labeling estimates with hats, let's call the estimate from that regression $\hat{\beta}^{OLS}$. It is the regression coefficient, and it reflects the correlation between education and income in our sample. (Often, people drop the OLS superscript and just talk about their estimate as $\hat{\beta}$, which is fine as long as it is clear what you are up to.)

Unfortunately, β^{OLS} and $\hat{\beta}^{OLS}$ are not the same thing. The former is an estimand and the latter is an estimate, which, as we know from our favorite equation, may differ from the estimand because of both bias and noise. Let's assume that our sample of workers

was randomly selected from the population, so there is no bias. (We'll talk more about random sampling and unbiasedness in chapter 11). But there is still noise. So if we want to know how close $\hat{\beta}^{OLS}$ is likely to be to the true β^{OLS}, we're going to need to think about the standard error.

Just as $\hat{q}$, our estimate of the proportion of Republicans in the population, had a standard error, so too does our estimate of the relationship between income and education, $\hat{\beta}^{OLS}$. The standard error gives us a sense of how far, on average, our estimate would be from the truth if we repeated our estimator an infinite number of times with independent samples of data. Just as with the poll result, there are formulas for calculating this standard error. For now, you don't need to worry about the formula because a computer will calculate the standard error for you. Thinking more technically about standard errors is a topic for a different book.

Once you've estimated the standard error associated with a regression coefficient, you can do the same kinds of things that you did with the standard error of the poll result. For instance, you can construct a 95% confidence interval. You can also conduct hypothesis tests and compute p-values. All of this can help you assess how precise your estimate of the true relationship is.

One common question people are interested in is whether there is compelling evidence that there is any true relationship at all. That is, suppose you find a positive $\hat{\beta}^{OLS}$; income and education are positively correlated in your sample. Should you be confident, on the basis of that evidence, that they are positively correlated in the larger population?

You can start to answer that question by testing the null hypothesis that the true relationship between income and education is in fact zero. To do so, you ask how likely it is that you would have gotten an estimate as big as $\hat{\beta}^{OLS}$ if there was in fact no correlation between income and education in the population (i.e., $\beta^{OLS} = 0$). If you obtain a small p-value and reject the null hypothesis, then you have statistically significant evidence that there is a relationship between income and education in the population.

One reason statistical inference of this sort is so useful is that we're bound, from time to time, to find relationships in our data that do not reflect genuine relationships in the larger population. This is the nature of noisy data. So we need to check whether we have good reason to believe our findings aren't just the result of noise.

What If We Have Data for the Whole Population?

Sometimes we have data for an entire population of interest. For instance, suppose we want to know the correlation between participation in varsity athletics and GPA for University of Chicago students. It's conceivable that we could convince the university to give us the relevant data for every single student, in which case we wouldn't need to estimate the correlation in the population by calculating it in a sample. We could perfectly measure the true estimand, the correlation between athletics and GPA for the whole population of University of Chicago students.

Here's a tricky question. Does it still make sense to think about standard errors, confidence intervals, and statistical significance when we have data for the entire population? One argument is that these tools are irrelevant because there was no sampling. So there is no noise. The estimate is the estimand. As such, there's no need to think about statistical inference.

But we still think there are good reasons to pay attention to the concept of noise and the associated measures of uncertainty even if we have data for an entire population. Let's talk about why.

Suppose we found a small positive correlation between playing a varsity sport and GPA. It still seems reasonable to ask whether that difference arose for a reason or whether it arose just by coincidence.

What would it mean for a correlation to arise by coincidence? Suppose that there's no good reason to think there should be an on-average difference between the GPAs of athletes and non-athletes: the admissions standards are the same for both kinds of students, athletic participation has no effect on GPA, academic performance has no effect on athletic participation, and so on. Nonetheless, there are all kinds of idiosyncratic differences between students that lead their GPAs to be different from one another. And with any finite number of students, there's bound to be at least a slight difference between the GPAs of athletes and non-athletes, even if there's no good reason for the difference.

To assess whether this observed correlation arose by coincidence or for a reason, we can't collect more data. We already have all the data there is to have on University of Chicago students. However, it turns out that the tools of statistical inference and hypothesis testing still provide a useful way to think about whether an observed pattern was a coincidence or not.

One way to think about the problem is that, although we have data on all the actual students at the university, these actual students are just a small sample of a much larger hypothetical population of students that could have been at the university. We can start with the null hypothesis that the true correlation in that larger, hypothetical population is zero, and we can ask how likely it is that we would have observed a correlation at least as large as the one we observed among the actual students by chance. Of course, doing this requires a metaphysical leap from actual populations to hypothetical populations. But engaging in a little bit of metaphysics is a price probably worth paying to preserve our ability to think about whether some observed relationship reflects a genuine, predictable pattern or was just a fluke.

Substantive versus Statistical Significance

Statistical hypothesis testing is often helpful and informative because we want to know whether an observed phenomenon is likely to have arisen purely by chance (e.g., due to sampling variation). However, *statistical* significance (a low p-value indicating that a result was unlikely to have arisen by chance) is not the same thing as *substantive* significance, and we must be careful not to conflate these two concepts. Often, we don't want to know just whether some phenomenon *exists* or not, which is the question of statistical significance; we want to know how big or small the phenomenon is because that will tell us whether it is *important* or not, which is the question of substantive significance. For example, executives at Coca-Cola probably already know that their marketing has some effect on sales. But that doesn't tell them how much to spend on marketing. For that, they want to know how big the effects of marketing on sales are. Let us give you examples that illustrate the two ways quantitative analysts can be led astray by emphasizing statistical significance over substantive significance.

Social Media and Voting

In 2012, six researchers published a study in *Nature* showing that people were more likely to vote in the 2010 U.S. midterm elections if their Facebook pages displayed a banner indicating which of their friends voted. The study was notable for several reasons.

Facebook allowed the researchers to randomize the experience of sixty-one million voting-age Facebook users in the United States on Election Day. And indeed, the experimental intervention appears to have increased turnout—the estimated effect of seeing that a close friend voted is highly statistically significant ($p = .02$). The researchers concluded that "strong ties are instrumental for spreading both online and real-world behaviour in human social networks." And the study received significant press coverage for demonstrating how important social pressure is for voting.

What most observers failed to notice was that the estimated effect of the Facebook banners on voter turnout was less than 0.4 percentage points. This is a substantively small effect, arguably of little relevance for campaigning or understanding elections. The fact that 0.4 percent of eligible voters can be persuaded to vote through a Facebook banner reporting their friends' voting choices does not tell us that strong ties are instrumental for spreading behavior. Of course, with a sample size of sixty-one million, almost any non-zero estimate will be statistically significant. That's not a bad thing. Big sample sizes mean that our estimates are quite precise, so we will more reliably detect genuine relationships. However, we can't just assume that any statistically significant result is also substantively significant.

We've seen that statistically significant results can be substantively insignificant. Now let's see that the opposite is also true.

The Second Reform Act

In a 2011 article in the *Quarterly Journal of Political Science*, Samuel Berlinski and Torun Dewan estimate the effects of the Second Reform Act of 1867 on elections in the United Kingdom. Despite the fact that the Second Reform Act roughly doubled the size of the eligible electorate and brought working-class voters to the polls for the first time, the authors report that there was little effect on election outcomes: "There is no evidence relating Liberal [one of the major British parties] electoral support to changes in the franchise rules."

But is this the right way to interpret the evidence? When the authors say there is no evidence, what they mean is that their estimates of the effect of the Reform Act are not statistically significant. So they can't say that those results were unlikely to have arisen by chance. But, while not statistically significant, the evidence from the study actually suggests that the Second Reform Act had important consequences. The numerical estimates indicate that the Reform Act's doubling of the electorate increased the Liberal Party's vote share by 8 percentage points, a substantively large effect implying that the new, working-class voters enfranchised by the reform were much more likely to support the Liberal Party than were wealthier, previously enfranchised voters. However, although the estimate is substantively large, it is also imprecise and therefore not statistically significant. Focusing on this statistical insignificance, Berlinski and Dewan conclude that the Second Reform Act had little effect. But the evidence actually indicates that our best guess is it had a big effect. It's just that that guess is uncertain.

Although statistical significance is useful and informative, it is often misused and misunderstood. A theme throughout this book is that clear thinking and data are complements, not substitutes. Just because we're doing statistics doesn't mean we can stop thinking substantively about the questions we really want to answer. We should utilize statistical inference when possible. But we should also always remind ourselves to make substantive inferences from the evidence.

Wrapping Up

Estimates can differ from estimands for two reason: bias and noise. Bias will be a major focus of chapter 9. In this chapter, we focused on noise—differences between the estimate and the estimand that arise because of idiosyncratic features of our sample. Because noise is idiosyncratic, it would average out to zero if we were to follow our estimation procedure over and over again an infinite number of times, each time on an independent sample of data. But in any one sample, noise can be quite important.

The presence of noise means that we are always at least somewhat uncertain whether a relationship in a sample of data (the estimate) in fact reflects a real relationship in the larger population of interest (the estimand). We have discussed techniques for quantifying this uncertainty and for testing the hypothesis that an estimated relationship is real against the null hypothesis that the estimated relationship was the result of noise alone.

The presence of noise creates challenges beyond uncertainty. For instance, in chapter 7 we will consider the problem that, if the same study is run over and over again, some iterations will yield statistically significant results because of noise, even if the relationship under investigation isn't real. If only those statistically significant findings get reported, then the scientific enterprise may lead to systematically incorrect conclusions. In chapter 8 we will examine how the presence of noise creates the puzzling phenomenon of reversion to the mean—extreme observations tend to be followed by less extreme observations—which, if we don't think clearly, can lead to all sorts of misinterpretations of evidence.

Key Terms

- **Population:** The units in the world we are trying to learn about.
- **Sample:** A subset of the population for which we have data.
- **Estimand:** The unobserved quantity we are trying to learn about with our data analysis.
- **Estimator:** The procedure we apply to data to generate a numerical result.
- **Estimate:** The numerical result arising from the application of our estimator to a specific set of data.
- **Bias:** Differences between our estimand and our estimate that arise for *systematic* reasons—that is, for reasons that will persist on average over many different samples of data.
- **Noise:** Differences between our estimand and our estimate that arise due to *idiosyncratic* facts about our sample.
- **Unbiasedness:** An estimate/estimator is unbiased if by repeating our estimation procedure over and over again an infinite number of times, the average value of our estimates would equal the estimand.
- **Expectation or Expected value:** The average value of an infinite number of draws of a variable is the variable's expected value or its value in expectation.
- **Precision:** An estimate/estimator is precise if by repeating our estimation procedure over and over again, the various estimates would be close to each other. The more similar the hypothetical estimates from repeating the estimator, the more precise the estimate.

- **Sampling distribution:** The distribution of estimates that we would get if we repeated our estimator an infinite number of times, each time with a new sample of data.
- **Standard error:** The standard deviation of the sampling distribution. If the estimator is unbiased, the standard error gives us a sense of how far, on average, our estimate would be from the estimand if we repeated our procedure over and over with independent samples of data.
- **Margin of error:** Pollsters often multiply the standard error by 2 and report this as the margin of error.
- **95% confidence interval:** If we applied the estimator an infinite number of times, each time on a new sample of data, the estimand would be contained in the 95% confidence interval (newly calculated each time) 95 percent of the time. Importantly, it is not true that we are 95 percent confident that the true estimand lies in the 95% confidence interval.
- **Hypothesis testing:** Statistical techniques for assessing how confident we should be that some feature of the data reflects a real feature of the world rather than arising from noise.
- **Null hypothesis:** The hypothesis that some feature of the data is entirely the result of noise.
- **Statistical significance:** We say that we have statistically significant evidence for some hypothesis when we can reject the null hypothesis at some pre-specified level of confidence (typically, 95% confidence).
- ***p*-value:** The probability of finding a relationship as strong as or stronger than the relationship found in the data if the null hypothesis is true. We use p-values to assess statistical significance. For instance, if the p-value is less than .05, then we have statistically significant evidence (at the 95% confidence level) that the relationship is real. Importantly, the p-value is not equal to the probability that the null hypothesis is true.

Exercises

6.1 Consider the following strategies for conducting a political poll to predict the vote shares in an upcoming election. Discuss the likely extent of bias and precision for each one.

(a) Fox News asks their viewers to call in and tell them who they are supporting in the election. They get more than one hundred thousand responses.

(b) Nailbiter Polling (a new firm on the scene) conducts polls, and then, regardless of the answers, they always report that the race is a dead heat: 50 percent in favor of candidate A, and 50 percent in favor of candidate B.

(c) Surprising News Polls (another new player) conducts large, representative polls, computes the average support for each candidate, and then flips a coin. If the coin is heads, they add 10 percent to candidate A's support, and if the coin is tails, they subtract 10 percent from candidate A's support.

(d) Middle America Polling obtains a physical copy of the voter file (the list of registered voters), they flip to the middle page, and they

contact and interview the ten individuals in the middle of that middle page.

6.2 Anthony's father, Pete, recently purchased a roulette wheel to run an underground casino in his garage. In case you're not familiar with roulette, the wheel is spun and a ball is dropped seemingly randomly into one of thirty-eight pockets on the wheel, each of which corresponds to a number and a color. On this wheel, there are eighteen red pockets, eighteen black pockets, and two green pockets. A gambler might bet on red, in which case they will double their money if the ball falls into a red pocket but lose their money otherwise. If the wheel is indeed fair, meaning that the ball is equally likely to fall in any pocket, Pete expects to make money on these bets because the gambler wins 18 out of 38 times, while Pete wins the other 20 out of 38 times. Of course, if the wheel is not fair, Pete could have just made a terrible investment. To test the wheel, Pete conducted three practice spins (with no gambling), and much to his dismay, the ball fell into a red pocket all three times. Given the information available to us thus far, what can we say from a statistical standpoint about whether the table is likely to be biased toward the red pockets?

(a) What's the null hypothesis?
(b) What's the p-value?
(c) Provide a substantive interpretation for the p-value and, importantly, explain what the p-value is *not*.
(d) Ignoring the legality of garage roulette, what additional advice would you give to Pete to help him figure out if his table is fair?

6.3 Let's return to the analysis of schooling and earnings from last chapter's exercises. When you regress earnings on schooling, in addition to giving you estimated coefficients, your computer also probably gave you some other numbers that you didn't understand until you read this chapter. For the coefficient associated with years of schooling, you should have obtained an estimate of 1.16, indicating that each additional year of schooling corresponds with increased earnings of about $1,160. What are the estimated standard error, p-value, and 95% confidence interval associated with that coefficient? Provide a substantive interpretation for each one.

Readings and References

If you are interested in the history of how scientists and statisticians came to settle on the widely accepted 5 percent threshold for statistical significance, which is surprisingly interesting, see

Michael Cowles and Caroline Davis. 1982. "On the Origins of the .05 Level of Statistical Significance." *American Psychologist* 37(5):553–58.

If you are generally interested in the history of probability and statistics (which you should be), have a look at

Ian Hacking. 2006. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability and Statistical Inference, 2nd Edition.* Cambridge University Press.

Ian Hacking. 1990. *The Taming of Chance.* Cambridge University Press.

For a fascinating history of statistical hypothesis testing and the problems that arise when scientists conflate statistical and substantive significance, we recommend

> Stephen T. Ziliak and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Errors Cost Us Jobs, Justice, and Lives.* University of Michigan Press.

Our discussion of small sample sizes, small schools, and the Gates Foundation drew on material from

> Howard Wainer. 2009. *Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display.* Princeton University Press.

The data on academic performance and enrollment in California schools is from https://www.cde.ca.gov/re/pr/reclayout12b.asp.

The paper on Facebook and voter turnout that we mention is

> Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489(7415):295–98.

The paper on the British Second Reform Act is

> Samuel Berlinski and Torun Dewan. 2011. "The Political Consequences of Franchise Extension: Evidence from the Second Reform Act." *Quarterly Journal of Political Science* 6(34):329–76.