CHAPTER 15

Turn Statistics into Substance

What You'll Learn

- Statistics are often reported or presented in ways that are misleading or unhelpful for decision making.
- If you think clearly about the question at hand, you can often translate a statistic into a piece of more substantively useful information.
- Quantitative evidence, on its own, can't tell you what to believe. Your beliefs depend on a combination of the new evidence and your prior beliefs. Bayes' rule tells us how to update beliefs in response to new information.
- Quantitative evidence, on its own, can't tell you what to do. For that, you must carefully combine your evidence-based beliefs with your values.

Introduction

Quantitative analysis should provide information that helps us make better decisions. The ideas we've emphasized thus far—how to establish whether a relationship exists, reversion to the mean, the difference between causation and correlation, using tools for estimating causal effects, and so on—are important inputs to that process. But they are not the end point.

Suppose you've estimated that some intervention has a positive effect on some outcome. Does that mean you should do the intervention? You can't know from a quantitative analysis alone. The decision also depends on your beliefs and values and on any trade-offs you might have to consider. To get from evidence to action, you need to translate statistical information into a substantive answer to your question.

People frequently get confused on this point. It's easy to stop thinking clearly once you've got some precise and authoritative-sounding quantitative finding, reaching incorrect conclusions even from correct information. In this chapter, we explore how to avoid such mistakes. The key is to turn statistics into substance, so as to make sure you are asking and answering the question you really care about.

What's the Right Scale?

There is more than one precise and accurate way to represent a piece of quantitative information. But they are not all equally helpful. How the information is presented can have an important effect on its perceived substantive meaning. For instance, changing scales can dramatically alter whether a relationship seems large or small, important or unimportant, a good reason for taking action or not. So it is important, when presented with such information, to think about whether the way that information is presented corresponds with the substantive question you are trying to answer, or whether reframing the relationship some other way might provide a better guide. To see what we mean, consider a couple examples.

Miles-per-Gallon versus Gallons-per-Mile

Suppose you work for the Environmental Protection Agency regulating automobile emissions. Your team brings you two proposed regulations to evaluate. One regulation will result in a 2-miles-per-gallon improvement in the fuel efficiency of small sedans. The other will result in a 2-miles-per-gallon improvement in the fuel efficiency of large SUVs. Suppose there are the same number of these two kinds of automobiles on the road and, on average, each gets driven 10,000 miles per year. The sedans get 30 miles-per-gallon (which the regulation would improve to 32 miles-per-gallon), while the SUVs get 10 miles-per-gallon (which the regulation would improve to 12 miles-per-gallon). The SUV regulation will cost a little bit more to implement. And since the two regulations each offer a 2-miles-per-gallon improvement for the same number of vehicles driven the same number of miles per year, your team recommends regulating the sedans. Does this make sense?

Let's start by remembering the substantive question you want to answer. Your job is to reduce automobile emissions by reducing gas consumption. Does improving fuel economy by 2 miles-per-gallon on sedans and SUVs translate into the same reduction in gas used? Let's turn the statistics into substance to check.

The SUVs get 10 miles-per-gallon, which means that, since the average driver drives 10,000 miles per year, on average SUVs use 1,000 gallons of gas per year ($\frac{10,000}{10}$). If you implement the regulation that improves fuel efficiency to 12 miles-per-gallon, then on average SUVs will use about 833 gallons per year ($\frac{10,000}{12}$). The 2-miles-per-gallon improvement saves 167 gallons of gas per SUV per year.

What about for the sedans? The sedans get 30 miles-per-gallon, which means that, since the average driver drives 10,000 miles per year, on average sedans use about 333 gallons of gas per year ($\frac{10,000}{30}$). A regulation that improves fuel efficiency to 32 miles-per-gallon results in sedans using about 313 gallons of gas per year ($\frac{10,000}{32}$). The 2-miles-per-gallon improvement saves only about 20 gallons of gas per sedan per year.

It wasn't obvious until we translated the statistics into substance, but now we can see that your team's recommendation looks wrong. The same 2-miles-per-gallon improvement in fuel economy has a much larger effect on gas consumption when applied to a gas-guzzling SUV than when applied to an already relatively fuel-efficient sedan. So you should regulate the SUVs unless doing so is much more expensive.

Figure 15.1 shows how much gas a vehicle that drives 10,000 miles a year uses as a function of the miles-per-gallon it gets. As expected, gas consumption decreases as miles-per-gallon increase. Less intuitive, but critical for understanding the results we just showed, is the fact that the slope of this curve is really steep for low values of miles-per-gallon (less efficient cars) and much less steep for high values of miles-per-gallon (more efficient cars). You get a lot more bang for your buck improving miles-per-gallon for inefficient cars than you do for efficient cars. This has interesting implications

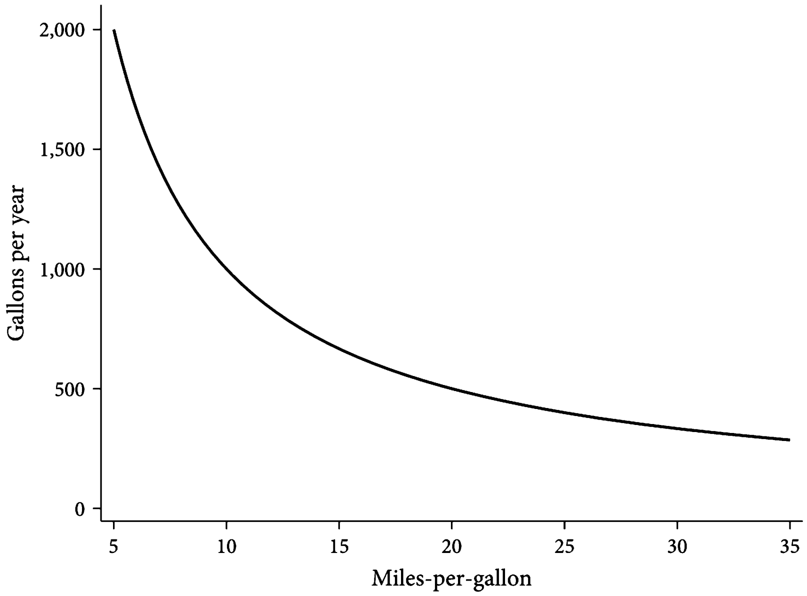

Figure 15.1. Gas consumed by driving 10,000 miles as a function of the miles-per-gallon.

beyond our example. In particular, moving people out of very fuel-inefficient vehicles into somewhat more fuel-efficient vehicles does a lot more to reduce emissions than moving people already in relatively fuel-efficient vehicles into very fuel-efficient vehicles like hybrids.

Returning to our example, there was nothing wrong with the quantitative information your team used to form its recommendations. Yet their recommendation was incorrect. Why? The problem came from the particular metric used to present the quantitative information. Miles-per-gallon is the most commonly reported metric for fuel efficiency in the United States. But it is not a particularly helpful statistic for substantive decision making.

The substantive question we care about is how much gasoline a car burns given how far it is driven. But miles-per-gallon tells you how far a car drives given how much gasoline is burned. That's backward for answering our question. You can do the math, as we did just now, to turn this statistic into substance. But most people won't. In fact, most people won't even notice the distinction. And therefore, consumers and regulators alike may be confused (or tricked) into making bad decisions.

If you wanted to provide more useful information, you would use a more substantively meaningful measure of fuel efficiency, something like gallons-per-hundred-miles, instead of miles-per-gallon. As we just saw, the same 2-miles-per-gallon improvement results in very different improvements in gallons-per-mile, depending on an automobile's baseline fuel efficiency. You'd have had no trouble making the right decision if your team had come to you with two regulations, one of which saved about 8.3 gallons-per-hundred-miles (the SUV regulation) and the other of which saved about 3.1 gallons-per-hundred-miles (the sedan regulation).

Indeed, in a 2008 study, Richard Larrick and Jack Soll show that the way statistics are reported can be quite consequential for important decisions. They quote an automotive expert who seems to think it's not worthwhile to try to make marginal improvements

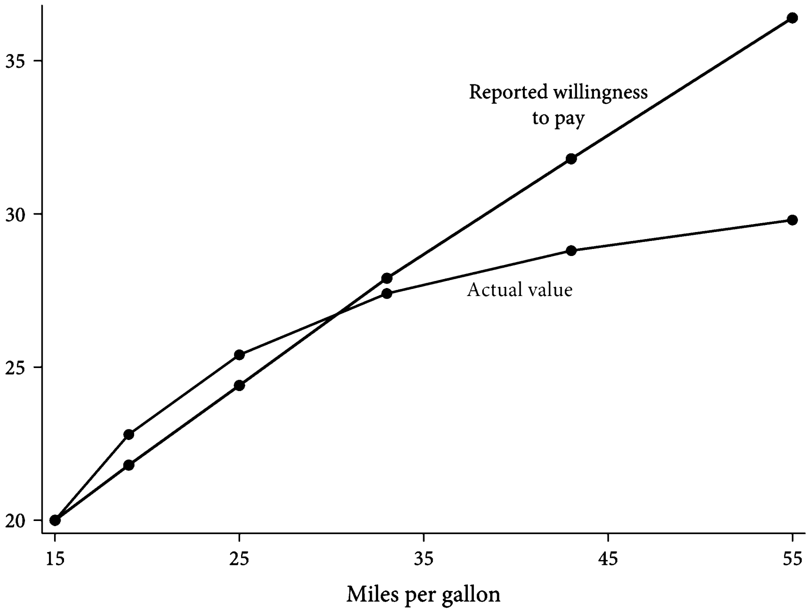

Figure 15.2. Willingness to pay versus the actual value of improvements in fuel efficiency (in thousands of dollars).

in the miles-per-gallon of large SUVs when that's, in fact, where engineers and policy makers would likely get the biggest bang for their buck in terms of mitigating emissions. Furthermore, they show that consumers are often misled by the statistics with which they are presented in ways that could have significant implications for purchasing decisions.

Specifically, Larrick and Soll asked college students to consider how much they would pay for a new car. Respondents were asked to imagine that they drive 10,000 miles per year. They were shown a car that gets 15 miles per gallon and asked to imagine that they value this car at $20,000. They were then shown alternative versions of that car that are reportedly identical to the baseline version in every way except that they get 19, 25, 33, 43, or 55 miles per gallon. How much would respondents be willing to pay for these more efficient cars?

Figure 15.2 shows the results. The black dots are the average willingness to pay (in thousands of dollars) reported by the survey respondents. Reported willingness to pay increases approximately linearly with miles-per-gallon. But it shouldn't! The gray dots in the figure show approximately how the respondents *should* have valued these cars (also in thousands of dollars), assuming that they'll keep the car for ten years and they have a 3 percent discount rate (i.e., a dollar tomorrow is worth ninety-seven cents today). As we saw in figure 15.1, a 1-mile-per-gallon increase in fuel efficiency is a lot more valuable if you're starting at a low level of efficiency than if you're starting at a high level. So the respondents in this study are making a big mistake in how they value these hypothetical cars.

Larrick and Soll go on to show that they can correct this mistake if they present fuel efficiency in terms of gallons-per-hundred-miles rather than miles-per-gallon. In other words, different ways of conveying the same information can be hugely consequential for decision making, so we need to think about the best way to present quantitative information so that decision makers can best translate their preferences into actions.

Percent versus Percentage Point

Often, in evaluating the substantive importance of some effect, we want to know how big the effect is. There are at least two ways the size of an effect might be reported: the percent change in the outcome it induces or the percentage point change in the outcome it induces. The *percentage point change* is the simple numerical difference between two percentages. The *percent change* is the ratio of the percentage point change to the initial value. So, for instance, moving from 20 percent to 22 percent is a 2 percentage point increase (22% − 20%) but a 10 percent increase ($\frac{2}{20}$)—which can lead to very different perceptions of the magnitude of an effect. So it is important to check your intuitions by translating back and forth and thinking clearly about which matters for your question. Here's an example.

The *Wall Street Journal* reported on a medical experiment showing that a new drug reduced the "risk of heart-related death, heart attacks, and other serious cardiac problems by 44%." A 44 percent reduction sounds big. This, coupled with the headline, "Cholesterol Drug Cuts Heart Risk in Healthy Patients," makes it sound very important that people have access to the treatment.

But let's stop to think clearly about the substantive question we are interested in when evaluating a quantitative result like this. To determine whether it is worthwhile to give a particular treatment to a large population, we'd like to know how much the treatment costs and how many people the treatment will save. Knowing that a treatment reduces heart attacks among otherwise healthy people by 44 percent doesn't actually tell you how many people it saves. To determine that, you also need to know how frequent heart attacks are in that population in the first place.

Later in the article we learn that 250 out of the 9,000 people randomly assigned to the control group, which received a placebo pill, had heart attacks over the course of the study. This suggests a baseline heart attack risk of about 2.8 percent ($\frac{250}{9000}$). A 44 percent decrease in heart attacks means going from about 2.8 percent of people having a heart attack to about 1.6 percent having a heart attack. Because heart attacks are so rare in this population, the 44 *percent* reduction in heart attacks translates into about a one *percentage point* reduction—not such a huge difference. Indeed, if the drug is expensive, you might well conclude that the treatment is not worthwhile.

Here, again, we see the value of translating statistics into substance. The article uses statistics that answer one question—Does the drug cause a large percent decrease in heart disease?—to which the answer is yes. But the headline makes it seem as though it is answering a much more important question—Will the drug save a lot of lives?—to which the answer is probably not. By translating the statistics (the percent reduction) into substance (number of heart attacks prevented per 100 people treated), we can easily spot the difference and answer the questions most relevant for decision making.

Visual Presentations of Data

One of the most common ways to present and consume quantitative information is through some sort of graph, figure, or visual display. Indeed, we have displayed data visually throughout this book.

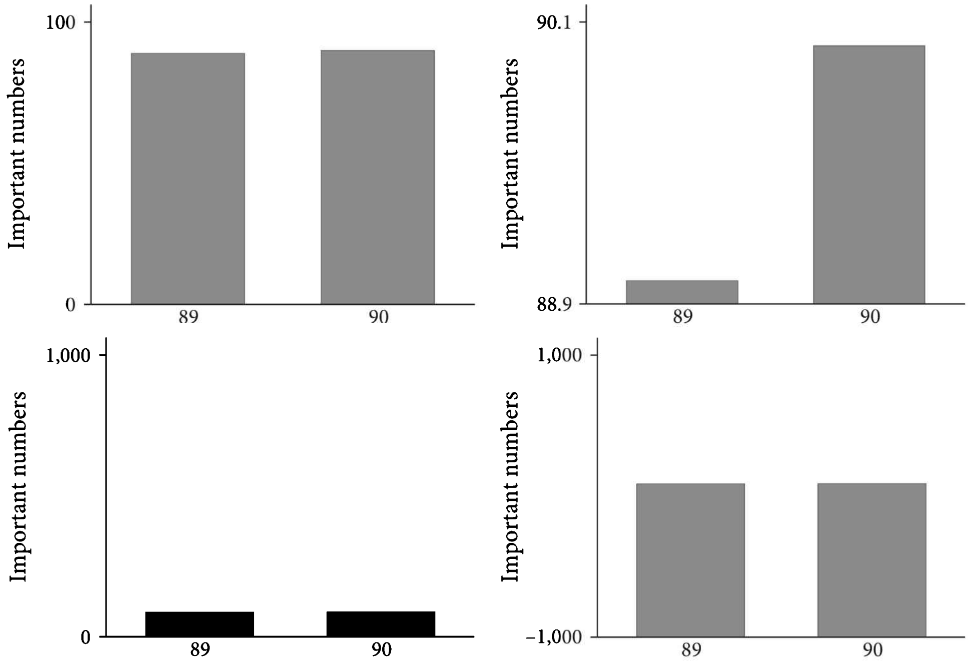

Figure 15.3. Four different ways to show the difference between 89 and 90 with a bar chart.

Displaying data accurately and informatively is part art and part science. So it is worth pausing briefly to reflect on some best practices. There are excellent books dedicated almost entirely to this topic (see the Readings and References section at the end of this chapter), so we won't belabor the discussion. But we want to hit on some essentials.

The most essential of all the essentials is this: no matter how beautiful, data visualizations are not a substitute for clear thinking. It's easy to be fooled by aesthetically pleasing, but misleading, graphics. So as a consumer of quantitative figures, you have to stay focused on thinking clearly about the substance. What are the underlying data and analyses that led to the figure? Are the underlying assumptions sound? Are there other statistics or representations of the data that would be more informative? Do the findings being presented answer the question being asked? Is the scale on which the data is presented appropriate, or was it chosen to hide or exaggerate the substantive magnitude of a relationship? Are there unnecessary, distracting features of the figure that could mislead you?

Choosing the scale on which to present data is one of the most consequential decisions in creating a data visualization. A seemingly innocuous change of scale can transform a graph from one that makes a relationship or finding look enormous to one that makes it look inconsequential, or vice versa.

To see what we mean, have a look at figure 15.3. That figure displays four different bar graphs, each of which is just a comparison of the number 89 to the number 90. But by altering the scale—here, by changing the range of the vertical axis—we change how much we zoom in or zoom out. The result is that we can make 89 and 90 appear to be hugely different from one another or nearly identical. And we can also make both numbers look very large or very small. Therefore, one of the simplest and most important ways to make sure you are thinking clearly about how to interpret a figure is to carefully read the axes and think about what the numbers mean substantively.

Importantly, there isn't a correct scale, separate from the question at hand. You should decide for yourself what constitutes a substantively meaningful difference in your particular context. There are some circumstances in which the differences between 89 and 90 is substantively large. For instance, if you chaperoned 90 school children on a field trip, there's a very big difference between 89 and 90 students returning home safely. Alternatively, it's not likely to be important whether the bus transporting the children home is 89 or 90 seconds late. The right scale for your graph depends on which kind of situation you are in.

If a graph is on a scale so big that you can't see substantively meaningful differences, you should worry that important information is being hidden. If a 1-point difference is substantively important, then a graph on a scale of 88.9 to 90.1 (upper-right panel of figure 15.3) appropriately reflects the important distinction between 89 and 90, while a graph on a scale of 0 to 1,000 (bottom-left panel) obscures that distinction.

And if a graph is on a scale so small that differences you shouldn't care about appear large, you should worry that findings are being exaggerated. For instance, if a 1-point difference is substantively negligible, then a graph on a scale of 88.9 to 90.1 inappropriately makes it look like an important difference, while a graph on the scale of 0 to 100 accurately reflects that the two numbers are essentially the same.

Concerns about the scale of a figure apply far more broadly than just these somewhat silly bar graphs (you could, after all, just report the numbers 89 and 90). By changing the scale of the axes, analysts can make correlations look strong or weak, they can make the slopes of regression lines appear large or small, and they can even make a linear relationship appear non-linear or vice versa (for example, by the choice of whether to show income or log-income). As we've discussed, there are plenty of good and bad reasons to transform a variable or carefully select the scale on which something is presented. An analyst should always think about how to present their data in the most informative way, and a consumer should turn what's being presented into the substance they care about most.

Policy Preferences and the Southern Realignment

Consider an example. In a 2016 book, Christopher Achen and Larry Bartels argue that voters' policy views have little relationship to their political behavior. That behavior, they claim, is driven by non-policy concerns. As one piece of supporting evidence, Achen and Bartels argue that policy views don't explain why white voters in the U.S. South shifted from supporting the Democratic to the Republican party during the so-called Southern realignment that occurred in the second half of the twentieth century. Their evidence for this claim is a visual representation of data, which we have attempted to reproduce as closely as possible in figure 15.4.

The figure separately plots the trend in party identification for white Southerners who opposed and did not oppose integration. The horizontal axis is years. The vertical axis shows the Democratic margin, measured as the percent of people who identify as Democratic minus the percent of people who identify as Republican. So the higher a data point is on the vertical axis, the more Democrats there are compared to Republicans.

The figure clearly shows the Southern realignment. In 1960, Southern whites were overwhelmingly Democrats. But that changed over time, so that by the end of the twentieth century they were overwhelmingly Republicans.

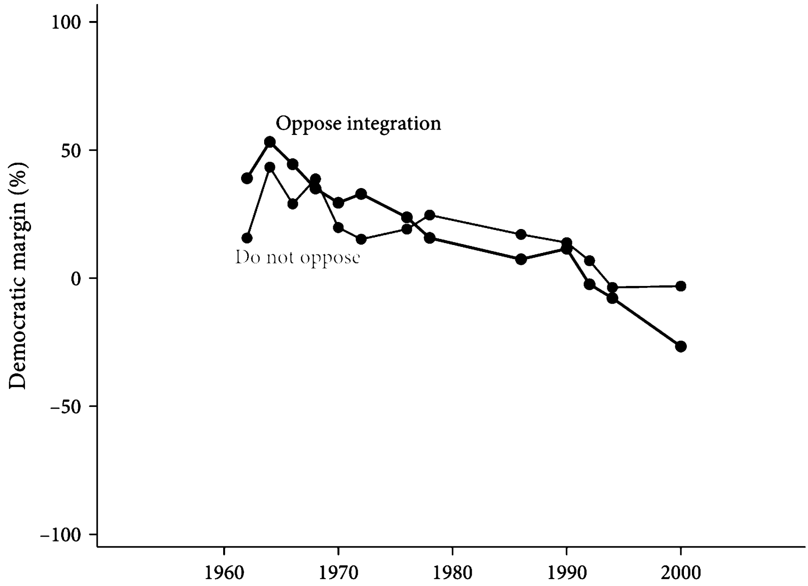

Figure 15.4. Trends in partisanship for white Southerners who opposed and did not oppose integration.

Achen and Bartels argue that the figure also shows that white voters' policy positions on integration do not affect changing party affiliation. That is, they claim these two trends are more or less the same. And this, they argue, suggests that voters' positions on even highly salient policy issues don't influence party affiliation.

What do you notice about figure 15.4? Is it obvious that the trends are more or less the same? First, we might want to look at the scale of the vertical axis. The measure of partisanship—the percent of individuals identifying as Democratic minus the percent of individuals identifying as Republican—in theory could range from −100 to 100. And that's the scale on which the figure is drawn. However, in practice, many people don't identify as either Democratic or Republican, so in almost any large population, we're probably not going to see a Democratic margin anywhere near the theoretical minimum or maximum. Because the range of the axis is so large, isn't it possible that there is a substantively meaningful difference that's difficult to see, much like in the bottom-right panel of figure 15.3?

Also, consider the horizontal axis. The figure only includes data from 1962 through 2000, but the graph is wide enough to include data from 1950 through 2010, leaving a bunch of empty, wasted space. There is no good reason to leave that space blank. But it does compress the data.

How would our substantive conclusions change if we removed some of that wasted space and redrew the same quantitative information on a scale that more accurately reflects the observed range of the data? We can see this in figure 15.5. We have also added linear regression lines, which we believe make it easier to visualize the average trends for the two different groups of voters.

The data visualization in figure 15.5 suggests an importantly different interpretation from the data visualization in figure 15.4. In particular, figure 15.5 shows that the trends in partisanship were actually quite different for people who opposed integration compared to people who did not oppose integration. Those who opposed

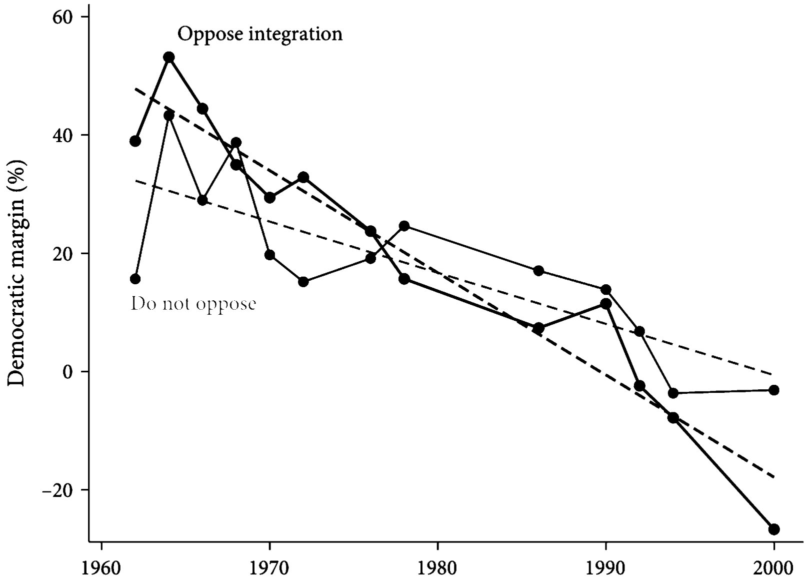

Figure 15.5. Trends in partisanship for white Southerners who opposed and did not oppose integration on a more appropriate scale and with regression lines.

integration were more Democratic in the 1960s than those who did not oppose integration. And they were more Republican by the end of the twentieth century. So their over-time trend was substantially steeper—people opposed to integration switched partisan affiliation at a faster rate than people who did not oppose integration. Perhaps policy views do help explain the shift in party identification during the Southern realignment.

Lest we fall for the trick in the top-right panel of figure 15.3, we'll want to interpret the numbers substantively to make sure what looks like a large difference is substantively meaningful. Most of us probably don't regularly think about percent margins of party identification, so perhaps there are better ways to convey this information. Let's give it some thought.

We see that, from 1962 through 2000, white Southerners who opposed integration went from something like a 48-point margin in favor of the Democratic Party to an 18-point margin in favor of the Republican Party. Those who did not oppose integration also became more Republican, but the change was more modest—from a 32-point Democratic margin to a 1-point Republican margin. So the shift for those who opposed integration was 33 percentage points larger, or twice as big, as the shift for those who did not oppose integration.

But is that a big or small difference? To provide a benchmark, if we look at 2020 data, 33 percentage points is approximately the difference in the Democratic margin between Massachusetts (a solidly blue state) and Idaho (a solidly red state). So we think it's safe to say that two trends that differ by 33 points on this scale are, in fact, meaningfully different, and the visualization in figure 15.4 was obscuring substantively important information.

We've illustrated one set of questions that you would want to ask about figure 15.4 and shown that they matter. But we've only scratched the surface of the questions worth asking as you attempt to turn statistics into substance in this case. For example, why is the right outcome for evaluating political behavior party identification rather than something more politically consequential, like voting behavior? Why start this analysis in 1962, when the Southern realignment is widely viewed as having started earlier? Is a single survey question about views on integration the best way to measure policy preferences in this context?

Some Rules of Thumb for Data Visualization

There is much more to think about in interpreting data visualizations. As we've said, we aren't going to try to provide a comprehensive overview. But here are some key principles that we think are important to keep in mind when creating or consuming graphical depictions of quantitative information.

- Keep it simple. If you don't need multiple colors, don't use colors. If you don't need fancy graphics, don't include them. If a third dimension doesn't add something crucial, use a two-dimensional plot. If you have complicated legends and labels, break things up.
- The focus should be on substance. You're trying to convey information in a transparent and easy- to- absorb manner. Make sure that the design choices you make are ones that advance the goal of conveying the answer to the question at hand.
- If you're just showing some simple numbers (like 89 and 90 or a regression coefficient), perhaps you can do away with the figure altogether and present the numbers in a table. Save figures for situations where a figure would convey more information than a table.
- Show the data. One of the great things about a figure is that you can show far more complicated relationships and far more detail than you might be able to do with a table. If the point of your figure is just to show the intercept and slope from a regression, you might as well just provide a table. But a figure can add a lot if you plot both a regression line and the data underlying the regression, so we can see whether the relationship is or isn't approximately linear. Think about figure 2.5 or 5.8. We learn a lot from the visualizations, relative to just reporting the correlation or regression coefficient, precisely because the underlying data are also displayed.
- When possible, convey uncertainty. Showing your data is a good way to do this. Instead of just showing means, consider showing distributions. If you're plotting estimates, also consider plotting standard errors or confidence intervals as we did in figure 12.4.

From Statistics to Beliefs: Bayes' Rule

The data never speaks for itself. Evidence is always interpreted in light of our existing ideas about how the world works, other related evidence we've seen earlier, and so on. So, in order to make use of quantitative information, it is important that we think clearly about how we should integrate that new information into our existing store of knowledge, so that we can translate statistics into beliefs. A key tool we have for doing so is called Bayes' rule. To get us thinking about Bayes' rule, how it works, and why we need it, let's start with an example.

In 1964 in Los Angeles, an elderly woman named Juanita Brooks was walking down an alley, pulling a basket of groceries with her purse resting on top, when she was pushed to the ground from behind and her purse was stolen. She didn't get a good look at the perpetrator. Around the same time, an eyewitness saw a woman run out of that same alley and enter a yellow car. The witness also didn't get a great look. But he did note that the woman running was white and had a blond ponytail and that the driver of the car was a Black man with a beard and mustache. On the basis of this eyewitness testimony, police later arrested Malcolm and Janet Collins and charged them with the robbery. Malcolm was a Black man with a beard and mustache. Janet was a white woman with a blond ponytail. And they drove a yellow car.

As Jonathan Koehler relates in an article, prosecutors brought in a mathematician to testify regarding the chances that, on the basis of this evidence alone, Malcolm and Janet were guilty of the robbery. The mathematician concluded that there was only about a 1 in 12 million chance that the couple was innocent. Here was the reasoning.

If we just arrested an innocent couple at random, it's very unlikely that the husband would be Black with a beard and mustache, that the wife would be white with a blond ponytail, and that they would drive a yellow car. Why is this?

The argument starts with some quantitative facts. If we just picked a man at random from the population, there's a 10 percent chance that he would be Black, because about 10 percent of the U.S. population is Black. Suppose that 10 percent of all men have beards, so there's also a 10 percent chance that he'd have a beard. Perhaps there's a 20 percent chance that he'd have a mustache. And there's only a 0.5 percent chance that he'd drive a yellow car, given the number of yellow cars on the road.

How do we take these numbers and turn them into an overall probability of a randomly selected innocent couple having this confluence of characteristics? Let's think about an analogy to a deck of cards. What is the probability that a randomly drawn card is the four of hearts? The probability that a randomly drawn card is a four is 1 in 13. And the probability that a randomly drawn card is a heart is 1 in 4. Since being a four and being a heart are independent (i.e., knowing that a card is a four doesn't tell you anything about how likely it is to be a heart and vice versa), the probability of the two characteristics occurring together is simply the product of the probability of each occurring individually. So if we draw a random card from a deck, the probability it is the four of hearts is $\frac{1}{13} \times \frac{1}{4} = \frac{1}{52}$. This makes sense. There are 52 cards in a deck. Only one of them is the four of hearts.

The prosecutor applied the same logic to the Collins couple. He argued that the chances that a randomly selected person would be Black, have a mustache and beard, and drive a yellow car is the product of the probabilities of the individual characteristics: $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{5} \times \frac{1}{200} = \frac{1}{100{,}000}$. He continued to add characteristics (being married, being an interracial couple, the woman having blond hair, and a ponytail, and so on), eventually arriving at a probability of 1 in 12 million. Indeed, as the prosecutor pointed out, even this was an under-estimate, since the couple had many other characteristics that had not been accounted for, so that the probability of innocence was probably more like 1 in 1 billion! A jury found the Collins couple guilty, and newspapers praised prosecutors for making such a quantitatively rigorous case.

What do you think: Does this example reflect clear thinking? We hope you said no, because indeed, there is so much wrong, it is hard to know where to start. But start we must.

So, first, these characteristics (unlike being a heart and being a four in a deck of playing cards) are not independent of one another. So you can't just multiply the probabilities of each individual characteristic together to get the probability of the confluence of characteristics. For instance, having a beard is positively correlated with having a mustache. As such, the probability of having a beard and a mustache is much higher than the probability of having a beard times the probability of having a mustache. That is, if 1 in 10 men have a beard and 1 in 5 have a mustache, since many of these are the same men, many more than 1 in 50 men have a beard and a mustache. Indeed, the probability is likely much closer to 1 in 10, since almost everyone with a beard also has a mustache. So, if we took into account all the relevant correlations, maybe we wouldn't conclude that the probability that a randomly selected couple fit the eyewitness description was 1 in 12 million. But we would still get a pretty low probability (maybe 1 in a million). That still seems like good evidence on which to convict, no?

No. It really isn't. We haven't even talked about the main thing that's gone wrong in the analysis, which is that it answers the wrong question entirely. If we think clearly about the right question, we reach a very different conclusion.

The jury has to decide whether or not to convict the Collins couple. They don't want to do so if the Collins couple is sufficiently likely to be innocent, and they do want to do so if the Collins couple is sufficiently likely to be guilty. So the right question for the jury is, How likely is the Collins couple to be innocent, given the evidence? The evidence is that the Collins couple matches the eyewitness description. So the right statistic to answer the jury's substantive question is the probability that the Collins couple is innocent given that they match that description. Write this as Pr(innocent | match). This is called a *conditional probability*, since it is the probability of one thing conditional on another. It is read in one of two ways. People either say "the probability they are innocent, conditional on them matching the evidence" or "the probability they are innocent, given that they match the evidence." Either one is fine. The probability they are guilty conditional on them matching the evidence is just Pr(guilty | match) = 1 − Pr(innocent | match).

The mathematical analysis we've discussed thus far has *not* told us this probability and, so, has not answered the right question. The analysis thus far tells us how likely it is that a randomly selected couple would match the eyewitness description. That is, it tells us Pr(match | innocent), which is read "probability a couple would match the evidence conditional on them being innocent." While this statistic may be useful for answering the jury's question, it is not itself the answer. The jury wants to know Pr(innocent | match). The prosecutor has told them Pr(match | innocent). But the jury (and the press) failed to notice the difference because they weren't thinking clearly.

Let's see why this matters. Suppose we agreed that Pr(match | innocent) is approximately equal to 1 in 1,000,000. We need to figure out Pr(innocent | match). Can we do it?

Table 15.1 will help. It categorizes couples in Los Angeles County according to two characteristics: whether they match the eyewitness description or not and whether they are guilty or not. We know that there is exactly one guilty couple and that couple matches the eyewitness description. So the guilty column is easy to fill in. The innocent column is a little trickier. We've agreed that the prosecutor's analysis, along with a bit of conjecture, suggests that the probability an innocent couple matches the eyewitness description is about 1 in 1,000,000. If we approximate that there were roughly 2 million innocent couples in LA County in 1964, we'll conclude that there were approximately 2 innocent couples who also matched the description. The remaining 1,999,998 couples in LA County fall into the last cell: innocent and don't match.

Table 15.1. LA couples by innocence or guilt and whether or not they match the evidence.

	Innocent	Guilty
Don't Match	1,999,998	0
Match	2	1

So, how likely is a couple to be innocent, given that they match the description—that is, what is Pr(innocent | match)? Well, there are three couples that match the description. Exactly one of them is guilty. So the true probability that a couple is innocent, given that they match the eyewitness description, is not anything like 1 in one million. It is 2 in 3. That means the probability the couple is guilty given that they match the eyewitness description is only 1 in 3. On the basis of the eyewitness evidence, the Collins couple was more likely to be innocent than guilty!

The discrepancy arises not because the mathematician, the prosecutor, and the press looked at incorrect quantitative information, but because they used the quantitative information to answer the wrong question. As the mathematician and prosecutor said, it is very unlikely that a randomly selected innocent couple would match the description of the criminals. But that doesn't mean it is very unlikely that a couple that matches the description of the criminals is innocent. Only one innocent couple in a million matches the description. But two couples out of three who match the description are innocent. If the jury had been able to think more clearly about the quantitative information, we suspect the Collins couple would not have been convicted. Few jurors want to send people to jail on the basis of there being a 1 in 3 chance that they committed a crime.[1]

Bayes' Rule

The analysis we've just done is an example of a general approach to figuring out what we should believe, given some evidence. A mathematical tool called *Bayes' rule* (or, sometimes Bayes' theorem or Bayes' law) gives us the formula for calculating this value. It is named after Thomas Bayes, an eighteenth-century philosopher and statistician.

Bayes' rule tells us the correct formula for how likely a claim is to be true, given the available evidence. It goes like this. Suppose we want to know the probability that a claim C is true, given evidence E. That is, we want to know $\Pr(C \mid E)$. In our example, the claim was that the Collins couple was innocent and the evidence was that they matched the eyewitness decision. Bayes' rule says

$$\Pr(C \mid E) = \frac{\Pr(E \mid C)\,\Pr(C)}{\Pr(E)}.$$

Let's go back to the Collins case to unpack this a bit. We want to know the probability the Collins couple is innocent, conditional on them matching the eyewitness description. In this case, Bayes' rule says

[1] Fun fact: Malcolm Collins appealed the guilty verdict on the grounds that the prosecutor had used a faulty mathematical argument to convict him. The California Supreme Court reversed the judgement, arguing for the importance of clear thinking. It wrote, "Mathematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search for truth, must not cast a spell over him."

$$\Pr(\text{Innocent} \mid \text{Match}) = \frac{\Pr(\text{Match} \mid \text{Innocent}) \Pr(\text{Innocent})}{\Pr(\text{Match})}.$$

We can use table 15.1 to find the values to plug in to see how this works.

What is Pr(Match | Innocent)? It is the probability a couple matches, given that they are innocent. There are 2 million innocent couples. Two of them match. So Pr(Match | Innocent) $= \frac{2}{2{,}000{,}000}$.

What is Pr(Innocent)? It is the overall probability that a random couple is innocent. There are 2,000,001 couples in LA County. Of them, 2,000,000 are innocent. So Pr(Innocent) $= \frac{2{,}000{,}000}{2{,}000{,}001}$.

Finally, what is Pr(Match)? Again, there are 2,000,001 couples, of which 3 match the eyewitness description. So Pr(Match) $= \frac{3}{2{,}000{,}001}$.

Putting these together we have

$$\begin{aligned}\Pr(\text{Innocent} \mid \text{Match}) &= \frac{\Pr(\text{Match} \mid \text{Innocent}) \Pr(\text{Innocent})}{\Pr(\text{Match})}\\ &= \frac{\frac{2}{2{,}000{,}000} \cdot \frac{2{,}000{,}000}{2{,}000{,}001}}{\frac{3}{2{,}000{,}001}}\\ &= \frac{2}{3}.\end{aligned}$$

Notice, we were able to figure this out earlier without knowing Bayes' rule, just by looking at the table. So there isn't much need to memorize the formula. But it is important to know how to calculate beliefs from evidence and to make sure you are thinking clearly about what question you want to ask and answer. Because it is really easy to convince yourself that Pr(Match | Innocent) is the same as Pr(Innocent | Match). But, as we've now seen, they can be really different.

Information, Beliefs, Priors, and Posteriors

Bayes' rule is useful anytime we receive new information and want to update our beliefs about how likely some claim is to be true. Before we get the new information, we have what we call a *prior belief* about the claim—that is, our belief about the probability that the claim is true, without knowing the new evidence. In the formula, this prior belief is represented by $\Pr(C)$—the probability the claim is true, without reference to the evidence. After we incorporate the new information, Bayes' rule gives us what we call a *posterior belief*: $\Pr(C \mid E)$.

In *People v. Collins*, the prior belief is the baseline probability that the Collins couple was innocent, before hearing about the eyewitness testimony. At that point, there was no reason to suspect them more than any other couple living in LA, so the prior belief was very close to 1—something like $\frac{2{,}000{,}000}{2{,}000{,}001}$, since all but one couple were innocent.

We learned that the Collins couple matches the description of the criminals. In fact, the chances that an innocent couple matches that description was only 1 in 1,000,000, which might make us think that they are almost certainly guilty. But Bayes' rule tells us to hold off before jumping to conclusions. On the one hand, the evidence seems pretty damning. It's extremely unlikely that an innocent couple would match the description. On the other hand, the prior belief pushes in the other direction. It's extremely unlikely that any given couple is guilty. To figure out how likely it is that the Collins couple is guilty, given both of these facts, we have to ask about the relative likelihood of each one. If we ignore either our prior belief or the new evidence, we arrive at the wrong conclusion. Incorporating both, we see that, while the Collins couple is way more likely to be guilty than a randomly selected couple, there's still a good chance that they are innocent.

One way of thinking about the problem with the prosecutor's argument is that he talked only about the new evidence, ignoring the prior. This is a common mistake that people make when they aren't thinking clearly about quantitative evidence.

Abe's Celiac Revisited

Way back in chapter 1, we told you the story of Ethan's son, Abe, being incorrectly diagnosed with celiac disease. In case you don't remember, here are the highlights of the story.

As a little kid, Abe was small for his age, which is an indicator for celiac. His pediatricians administered two blood tests. One came back positive (evidence that he had the disease), the other negative (evidence that he did not have the disease). The doctors concluded that Abe probably had celiac, because the positive test was "over 80 percent accurate."

The test on which Abe came up negative (let's call this Test 1) for celiac disease had quite low false negative and false positive rates, about 5 percent each. We can write this in our new notation. The false negative rate is the probability you get a negative test result given that you have the disease—that is, $\Pr(\text{Negative on Test 1} \mid \text{Celiac}) = .05$. The false positive rate is the probability you get a positive test result given that you don't have the disease—that is, $\Pr(\text{Positive on Test 1} \mid \text{No Celiac}) = .05$.

The test on which Abe came up positive (let's call this Test 2) for celiac disease had a false negative rate of about 20 percent—that is, $\Pr(\text{Negative on Test 2} \mid \text{Celiac}) = .2$. This, we suspect, is where the "80 percent accurate" claim came from. That test has a false positive rate of 50 percent—that is, $\Pr(\text{Positive on Test 2} \mid \text{Celiac}) = .5$.

Prior to the blood tests, a reasonable guess about the probability of Abe having celiac disease, given his small stature, was maybe 1 in 100. That is Ethan's prior: $\Pr(\text{Celiac}) = .01$.

Let's ignore Test 1 for a second, and just apply Bayes' rule to Test 2. Imagine a group of 10,000 kids, all of whom were similarly small in stature. Our prior tells us that, of those 10,000 kids, about 100 (1%) will have celiac. Test 2's false negative rate tells us that, of those 100 kids with celiac, about 20 (20%) will nonetheless test negative, while 80 will test positive. And Test 2's false positive rate tells us that, of the 9,900 kids without celiac, about 4,950 (50%) will nonetheless test positive and 4,950 will test negative. Table 15.2 provides a summary.

So what is the probability Abe has celiac, given that he was small in stature and tested positive on Test 2? Well, a total of $4{,}950 + 80 = 5{,}030$ kids test positive. Of those, 80 have celiac. So the probability that one of these kids has celiac given a positive result on Test 2 is $\frac{80}{5{,}030}$, or approximately 1.6 percent.

Notice, now that we know Bayes' rule, we could have done this without making the table:

$$\Pr(\text{Celiac} \mid \text{Positive on Test 2}) = \frac{\Pr(\text{Positive on Test 2} \mid \text{Celiac}) \Pr(\text{Celiac})}{\Pr(\text{Positive on Test 2})}.$$

Table 15.2. Outcomes of a celiac test on 10,000 kids.

	Celiac	No Celiac
Negative on Test 2	20	4,950
Positive on Test 2	80	4,950

We know enough to calculate each of these quantities. Pr(Positive on Test 2 | Celiac) is 1 minus the false negative rate, which is .8. Pr(Celiac) is our prior belief, which is .01.

Calculating Pr(Positive on Test 2) is a bit more involved. Here's how you do it. There are two kinds of people who test positive: kids with celiac who get a correct test result and kids without celiac who get a false positive. One percent of kids have celiac, and of these 80 percent get a positive test result. Ninety-nine percent of kids do not have celiac, and of these 50 percent get a positive test result. So,

$$\begin{aligned}\text{Pr(Positive on Test 2)} &= \text{Pr(Celiac) Pr(Positive on Test 2} \mid \text{Celiac)}\\ &\quad + \text{Pr(No Celiac) Pr(Positive on Test 2} \mid \text{No Celiac)}\\ &= .01 \times .8 + .99 \times .5\\ &= .503.\end{aligned}$$

Now we can calculate Ethan's posterior beliefs directly:

$$\begin{aligned}\text{Pr(Celiac} \mid \text{Positive on Test 2)} &= \frac{\text{Pr(Positive on Test 2} \mid \text{Celiac) Pr(Celiac)}}{\text{Pr(Positive on Test 2)}}\\ &= \frac{.8 \times .01}{.503}\\ &\approx .016\end{aligned}$$

Of course, Abe actually had two tests. What happens if we add in the fact that Abe tested negative on the more accurate Test 1? If we assume that false positives and false negatives on these two tests are independent, then we can just multiply to get the relevant quantities.

$$\begin{aligned}&\text{Pr(Celiac} \mid \text{Neg on Test 1 \& Pos on Test 2)}\\ &= \frac{\text{Pr(Neg on Test 1 \& Pos on Test 2} \mid \text{Celiac) Pr(Celiac)}}{\text{Pr(Neg on Test 1 \& Pos on Test 2)}}\end{aligned}$$

What is the probability that a kid with celiac gets a negative result on Test 1 and a positive result on Test 2? Well, Test 1 returns a negative for a kid with celiac (i.e., a false negative) only 5 percent of the time. Test 2 returns a positive for a kid with celiac

80 percent of the time. So, if the false negatives and false positives are independent across the two tests, then

$$\begin{aligned}\Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{Celiac}) &= .8 \times .05\\ &= .04.\end{aligned}$$

The prior belief, Pr(Celiac), remains the same, 1 percent. And, again, there are two kinds of kids who might get a negative on Test 1 and a positive on Test 2. First, the kid might have celiac (that's true of 1 percent of these kids). That kid would then need to get a false negative on Test 1 but a correct result on Test 2. As we've just seen, the probability of this is $.8 \times .05 = .04$. Second, the kid might not have celiac (that's true of 99% of these kids). That kid would then need to get a correct result on Test 1 and a false positive on Test 2. This happens with probability $.99 \times .5 = .495$. Now we can calculate the overall probability of these two test scores.

$$\begin{aligned}&\Pr(\text{Neg on Test 1 \& Pos on Test 2})\\ &\quad= \Pr(\text{Celiac}) \Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{Celiac})\\ &\qquad+ \Pr(\text{No Celiac}) \Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{No Celiac})\\ &\quad= .01 \times .04 + .99 \times .495\\ &\quad= .49045\end{aligned}$$

Plugging all of this into Bayes' rule, we get

$$\begin{aligned}&\Pr(\text{Celiac} \mid \text{Neg on Test 1 \& Pos on Test 2})\\ &\quad= \frac{\Pr(\text{Neg on Test 1 \& Pos on Test 2} \mid \text{Celiac}) \Pr(\text{Celiac})}{\Pr(\text{Neg on Test 1 \& Pos on Test 2})}\\ &\quad= \frac{.05 \times .01}{.49045}\\ &\quad\approx .001\end{aligned}$$

The probability that Abe had celiac given the two test results was approximately 1 in 1,000.[2]

Now that you know Bayes' rule, you can see that the doctors were not thinking very clearly about what the evidence really meant.

[2] We would have gotten the same answer if we had applied Bayes' rule iteratively. We could have started with the prior belief that Abe had celiac before seeing any evidence, shifted our beliefs according to the evidence from Test 1, treated this posterior belief as our new prior, and then shifted our beliefs again according to the evidence from Test 2. And the order in which we do this doesn't matter. We'd end up with the same beliefs in the end if we started with Test 2 and then went to Test 1. As a bonus exercise, you can try double-checking this yourself to make sure you understand how to apply Bayes' rule.

Finding Terrorists in an Airport

In the years following the terrorist attacks of September 11, 2001, the United States government poured resources into airport security. One of the major new programs was called Screening of Passengers by Observation Techniques (SPOT).

The idea of SPOT was to use behavioral cues to catch potential terrorists before they boarded a plane. Behavior Detection Officers watched people in the security line at airports, looking for indicators that a person was nervous or otherwise suspicious. Different kinds of suspicious behaviors were assigned different numbers of points. If a person exhibited a cluster of suspicious behaviors that rose above some point threshold, that person was targeted for additional questioning, searching, and screening.

By the year 2010, about 5 percent of the Transportation Security Administration's (TSA's) annual budget, hundreds of millions of dollars per year, went to fund the SPOT program. Let's use Bayes' rule to see why this wasn't a very good use of money.

The TSA needs to be able to answer questions like "Given a set of behaviors and characteristics, how likely is it that the person in question is a terrorist?" In other words, the TSA is trying to form a posterior belief about the probability that a traveler is a terrorist, given some evidence gleaned by observing the traveler's behavior. To form such posterior beliefs correctly on the basis of a program like SPOT, the TSA needs to know at least three pieces of information:

1. How likely is a random traveler to be a terrorist?
2. How likely is a terrorist to appear suspicious to a Behavior Detection Officer?
3. How likely is a non-terrorist to appear suspicious to a Behavior Detection Officer?

Unfortunately, according to the General Accountability Office (GAO)—an independent, non-partisan agency that works for Congress and is charged with investigating how the federal government spends taxpayer dollars—the TSA doesn't know the answers to any of these questions. No existing scientific research confirms, much less quantifies, the usefulness of behavioral observation for identifying terrorists. What we do know is that, even according to the TSA's own report intended to show the efficacy of the SPOT program, it seems that no terrorists have ever been caught by it. Indeed, the GAO reports that undocumented immigration status was by far the most common reason for detention of a person identified for additional screening by a Behavior Detection Officer.

So the government doesn't have the data that we need to calculate posterior beliefs on the basis of the evidence collected by the SPOT program. But we can see that this program was never going to work even without hard data. Let's ask how well the program would work in something like the best-case scenario. That is, we'll make up some data, being extremely generous to the SPOT program in all of our assumptions, and see whether the program would be a good idea under these assumptions. If the answer is no even under these generous assumptions, then we can be sure the answer is also no under more realistic assumptions.

First, according to the GAO, there are approximately 2 billion passenger trips through U.S. airports each year. For convenience, let's say there are 2 billion plus 100. Presumably the vast majority of those people are innocent travelers. Very few travelers are trying to hijack planes or engage in other forms of terrorism. Let's be generous to

Table 15.3. How many terrorists and non-terrorists appear suspicious.

	Not Terrorist	Terrorist
Not Suspicious	1,980,000,000	1
Suspicious	20,000,000	99

the government and suppose that each year, 100 would-be terrorists are in U.S. airports attempting to hijack airplanes. So that's our prior: Pr(Terrorist) $= \frac{100}{2{,}000{,}000{,}100}$.

Second, we need to know how likely these terrorists are to exhibit the suspicious behaviors that the Behavior Detection Officers are looking for. Of course, we have no idea. But all the scientific evidence suggests that these kinds of behavioral cues are quite unreliable. Again, let's be generous and stack the deck in favor of SPOT. Suppose that 99 percent of all terrorists exhibit the behavior that the TSA is looking for—that is, Pr(Suspicious | Terrorist) = .99. In reality, this number is surely much much lower.

Finally, we need to know how likely innocent travelers are to exhibit the suspicious behavior. As we've already said, these behaviors are unreliable indicators, so at least some innocent people will exhibit them. But we want to be generous to SPOT. So suppose that only 1 percent of innocent people exhibit suspicious behavior, such that Pr(Suspicious | Not Terrorist) = .01. Again, in reality, this number is surely much much higher. For this exercise, we are assuming SPOT is an incredibly accurate behavioral screening program.

How likely is a person who behaves suspiciously to be a terrorist? Even under these extremely generous assumptions, the answer is not very likely. Table 15.3 shows you the data you'd get based on our assumptions.

Of the 2,000,000,100 passenger trips, 100 involve terrorists. Ninety-nine of them will exhibit suspicious behavior. The remaining 2 billion trips involve innocent travelers. Just 1 percent of them will exhibit suspicious behavior. But this 1 percent amounts to 20 million people! A total of 20,000,099 people act suspiciously. Of them, 99 are terrorists. So the probability that someone is a terrorist given that they acted suspiciously is $\frac{99}{20{,}000{,}099}$. That is, approximately .000005—about 1 in 200,000.

We could have similarly calculated this directly from Bayes' rule.

$$\begin{aligned}
\text{Pr(Terrorist | Suspicious)} &= \frac{\text{Pr(Suspicious | Terrorist) Pr(Terrorist)}}{\text{Pr(Suspicious)}} \\
&= \frac{\frac{99}{100} \cdot \frac{100}{2{,}000{,}000{,}100}}{\frac{20{,}000{,}099}{2{,}000{,}000{,}100}} \\
&= \frac{99}{20,000,099}
\end{aligned}$$

Remember the numbers above come from assumptions that are extremely generous to the government. There is no way that terrorists actually exhibit the behavior the SPOT program looks for 99 percent of the time. And there is no way that innocent people actually exhibit the behavior the SPOT program looks for only 1 percent of the time. So the probability a suspicious person is a terrorist is actually much lower than 1 in 200,000. Indeed, if terrorists exhibit suspicious behavior 75 percent of the time and innocent people 10 percent of the time, the probability of being a terrorist given suspicious behavior becomes about 1 in 37 million:

$$\begin{aligned}
\Pr(\text{Terrorist} \mid \text{Suspicious}) &= \frac{\Pr(\text{Suspicious} \mid \text{Terrorist})\Pr(\text{Terrorist})}{\Pr(\text{Suspicious})} \\
&= \frac{\frac{75}{100} \cdot \frac{100}{2{,}000{,}000{,}100}}{\frac{200{,}000{,}075}{2{,}000{,}000{,}100}} \\
&= \frac{75}{200,000,075} \\
&\approx \frac{1}{37,000,000}
\end{aligned}$$

Remarkably, even with this number, we are still being too generous. According to a study by the National Academy of Sciences, screeners looking for just one facial characteristic (rather than the many things SPOT screeners are looking for) in perfect conditions get their assessment right only about 60 percent of the time. In more realistic conditions, they get their assessment right only about 30 percent of the time. With this level of accuracy and the tiny proportion of people who are terrorists, we think it is safe to say that the over $1 billion allocated to the SPOT program was not money well spent. This is easy to see when we ask the right questions.

Let us end this unpleasant tale with one more distressing tidbit that harkens back to the key lesson from chapter 4, correlation requires variation. The Government Accountability Office is a watchdog organization that is supposed to make sure that government agencies spend money appropriately. After investigating a program it may also provide the relevant government agency with advice about how to improve. This is precisely what the GAO did after evaluating the SPOT program.

One of the areas that the GAO was concerned about was the lack of a scientific basis for the behavioral characteristics that TSA had its SPOT screeners looking for. According to GAO, the TSA had no idea whether terrorists are actually more likely to exhibit the behaviors they are looking for or not. (As we just saw, even if they are, this program is a waste.) And so, here is what the GAO recommended to the TSA to improve accuracy:

> Studying airport video recordings of the behaviors exhibited by persons waiting in line and moving through airport checkpoints and who were later charged with or pleaded guilty to terrorism-related offenses could provide insights about behaviors that may be common among terrorists.

Suppose you watched these videos and found that, for example, all the people who turned out to be terrorists wore sunglasses and looked agitated waiting in the security line. Do you want to start arresting everyone who meets that description? We hope not. As we've known since chapter 4, correlation requires variation. If you want to get better at the (hopeless) task of identifying characteristics that predict whether or not a person is a terrorist, at the very least you must compare the characteristics of terrorists and non-terrorists. You can't just study the terrorists.

Bayes' Rule and Quantitative Analysis

One particularly interesting application of Bayes' rule is thinking about how confident we should be about the truth of some scientific hypothesis, in light of the evidence presented in a scientific study. Of course, we already discussed one approach to this issue in chapter 6. There, we learned that the p-value tells us how likely a given estimate is to have occurred by chance alone. But if you think about it clearly, that doesn't answer the right question. In fact, when an analyst finds a low p-value and concludes that the finding must be true, they've made the same mistake as the mathematician and the prosecutor in *People v. Collins*. They've calculated the probability they would have found a relationship in their data, even if there is no real relationship in the world—that is, Pr(result | relationship not real). But what they really want to know is how likely it is that there is no real relationship, given their result—that is, Pr(relationship not real | result). The probability there *is* a real relationship given the result is just 1 minus this.

Let's use Bayes' rule to think about this a little more clearly. Suppose we collect some data, test for a relationship, and obtain a statistically significant result at the .05 level (i.e., $p < .05$). What's the probability that the estimated relationship reflects a real relationship in the world (as opposed to appearing in the data due to noise)? Bayes' rule tells us.

$$\text{Pr(relationship real} \mid \text{result)} = \frac{\text{Pr(result} \mid \text{relationship real) Pr(relationship real)}}{\text{Pr(result)}}$$

And, like before, we can break down Pr(result) into two components. One way we might have found the result is that the relationship is real and the test correctly identified it. The probability of this is Pr(relationship real) × Pr(result | relationship real). The other way we could have found the result is that the relationship is not real but the test spuriously identifies it as real due to noise. The probability of this is Pr(relationship not real) × Pr(result | relationship not real). So we can write Bayes' rule as follows:

$$\text{Pr(relationship real} \mid \text{result)}$$
$$= \frac{\text{Pr(result} \mid \text{relationship real) Pr(relationship real)}}{\text{Pr(relationship real) Pr(result} \mid \text{relationship real)} + \text{Pr(relationship not real) Pr(result} \mid \text{relationship not real)}}$$

We know the Pr(result | relationship not real). This is just the significance level used in our hypothesis test. If we would declare a statistically significant result if $p < .05$, then Pr(result | relationship not real) $= .05$.

The other numbers are more complicated. The quantity Pr(relationship real) is our prior belief that a genuine relationship exists, before seeing any of our new evidence. The quantity Pr(result | relationship real)—that is, the probability you find a result in your data given that the relationship really exists in the world—is called the *statistical power* of the test. The statistical power is the answer to the following question: What is the probability we would find a statistically significant result in the data given that the relationship is real? There are ways of estimating the statistical power once we know more details about the data and test. For instance, one might conduct computer simulations to determine how likely it would be to statistically detect an effect of a certain magnitude.

Now we can rewrite the formula for Bayes' rule one more time in terms of these substantively interpretable quantities:

$$\begin{aligned}
&\Pr(\text{relationship real} \mid \text{result}) \\
&= \frac{\Pr(\text{result} \mid \text{relationship real})\Pr(\text{relationship real})}{\Pr(\text{relationship real})\Pr(\text{result} \mid \text{relationship real}) + \Pr(\text{relationship not real})\Pr(\text{result} \mid \text{relationship not real})} \\
&= \frac{\text{Power} \times \text{Prior}}{\text{Power} \times \text{Prior} + \text{Significance} \times (1 - \text{Prior})}
\end{aligned}$$

Let's put this formula to work to see what it implies about our posterior beliefs in light of new, statistically significant, scientific evidence. Suppose we have a hunch about some causal effect in the world. It's a bit of a long shot. We think there's a 5 percent chance that this effect exists (our prior belief is .05). So we run a randomized experiment. We want to be confident in the answer, so we get a big sample size, such that the statistical power of our test will be .8 (we'll have an 80 percent chance of detecting an effect if one really exists). And following convention, we use a .05 threshold for statistical significance. Now, we can ask, conditional on obtaining a statistically significant result, what should our posterior beliefs be about the probability that the effect is real?

Plugging these numbers into the above equation, we get

$$\begin{aligned}
\Pr(\text{effect real} \mid \text{result}) &= \frac{.8 \times .05}{.8 \times .05 + .05 \times .95} \\
&\approx .46.
\end{aligned}$$

What happened? Even conditional on getting a result that is statistically significant at the 95 percent level, there's still only a 46 percent chance that the effect we believe we are estimating exists at all! The logic is the same as that underlying the conclusion that the Collins couple was more likely to be innocent than guilty even though the probability a random couple matched the description was only 1 in a million. The p-value, just like that 1 in a million, is just one of the numbers we need in order to form our posterior beliefs. If the power is low or our prior beliefs are low, our posterior beliefs are likely to be low as well.

This kind of thinking also helps us to better understand the replication crisis in so many scientific disciplines that we described back in chapters 7 and 8. Remember the ESP study? What was your prior belief about humans having ESP before you saw the results from that study? Probably pretty low, right? So your correct posterior belief that the effect is real, even given the statistically significant evidence, isn't that high. Figure 15.6 gives you a sense of this. The vertical axis is the posterior probability that an observed relationship is real. The horizontal axis is the prior probability that it is real. The curve plots the correct posterior belief as a function of your prior belief, given that a study with statistical power of .8 and a significance threshold of .05 generated statistically significant evidence of the relationship.

Our prior beliefs are hugely important for our posterior beliefs. Indeed, if you have really low priors about ESP, like we do, then it might not even make sense to study ESP, because the results of the study will have virtually no effect on your beliefs.

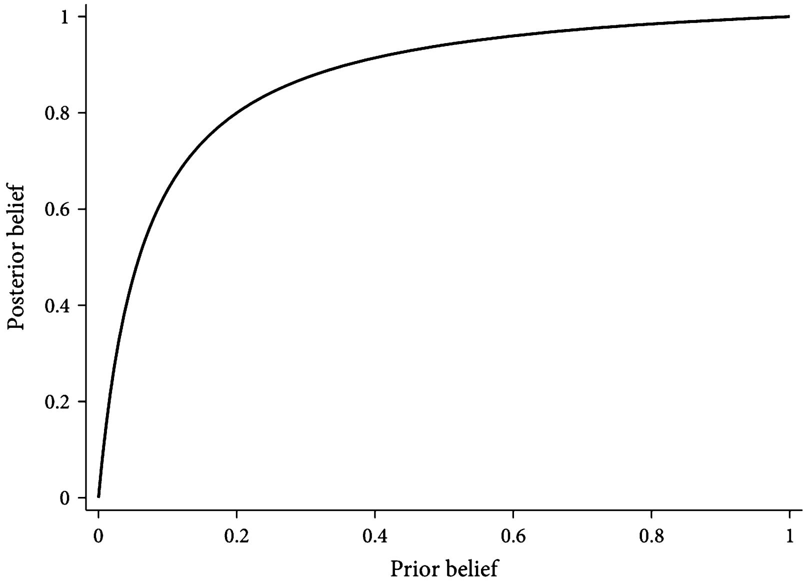

Figure 15.6. Posterior belief that an effect is real given statistically significant evidence, as a function of prior belief.

Figure 15.7 shows how the change in beliefs in response to new evidence relates to the prior. That is, it plots your posterior belief that a real relationship exists minus your prior belief that a real relationship exists, for different values of the prior belief, given that you saw statistically significant evidence in favor of the relationship. As you can see, if your prior belief is already very close to 0 or 1, it is very hard to move your beliefs. The effect of new evidence is largest for moderately surprising results (i.e., results where your prior belief was around .2).

Figure 15.7 also illustrates that two people can (and should) react quite differently to the same piece of information if they have different prior beliefs. Some people might see a piece of evidence about ESP, the consequences of global warming, or Russian interference in American elections and shift their beliefs dramatically, while others might see the same piece of evidence and barely shift their beliefs at all. When we experience this in our day-to-day lives, we often conclude that people who reacted differently than we did are unreasonable or irrational. But Bayes' rule tells us that it is perfectly understandable that different people react differently to the same information if, at the outset, they had different prior beliefs.

Some of this discussion might make you uncomfortable. As data analysts, aren't we supposed to let the data speak without imposing our own prejudices? And where do these priors come from, if not from data? These are tough questions. But there's no way around them. If you want to say something about the probability there is a genuine relationship in the world, given some piece of evidence, you need to have prior beliefs about the likelihood of that relationship. You can't just ignore your priors. Because, as we've seen, Pr(result | relationship not real) and Pr(relationship not real | result) can be very different.

Here's another wrinkle. Most of the time, we're not really interested in the probability that some phenomenon exists or doesn't exist (though we probably are in the ESP

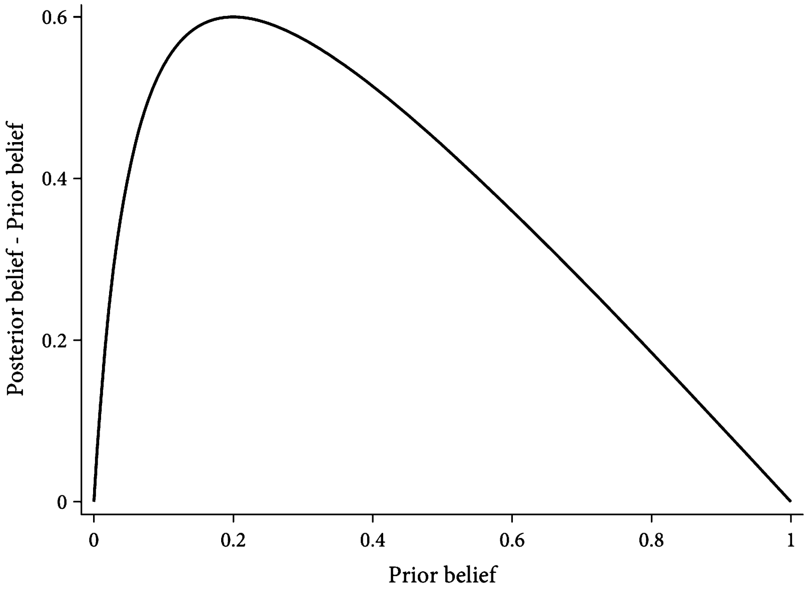

Figure 15.7. How much posterior beliefs change in response to new evidence, as a function of the prior belief.

example). Typically we want to know how substantively important or large an effect or relationship is, not just that it exists. That is, instead of just wanting to know if there is a real effect of, say, campaign strategy on vote share, we want to know the size of the effect of campaign strategy on vote share. How many supporters can a campaign turn out by launching a door-to-door canvassing campaign? Will turnout increase by 0.1, 1, or 10 percentage points? We can also incorporate Bayesian reasoning in such situations, but it's complicated. When thinking about the magnitude of a relationship, your prior belief is not just a single number, as it was when thinking about the probability that a relationship exists. Instead, it is a belief about how likely each possible relationship size is. And when you update your beliefs, you have to update your beliefs about each of these probabilities. Some analysts do this formally, specifying the whole prior distribution of beliefs about all the possible magnitudes and then doing complicated computations to estimate their posteriors. (This is called *Bayesian statistics*.) An alternative approach is to continue using conventional statistics like those described in chapter 6 (called *frequentist statistics*), while still trying to be careful when interpreting the results.

Expected Costs and Benefits

Your beliefs about effects are only one input to a decision. Even once you have made sure that things are on the right scale, you are answering the right question, and you have formed correct posterior beliefs based on the evidence and your prior beliefs, quantitative information still doesn't speak for itself. To use information and evidence to improve decision making, you have to combine your evidence-based beliefs with your values and goals to figure out how to act.

There is a sense in which this is obvious. Suppose a really well-designed series of studies convinces you that a certain kind of school intervention increases the likelihood that students will attend college by 30 percentage points. That's a big effect. But that alone doesn't tell you that the intervention is a good idea. To answer that question, at the very least, you have to know the value of college and how much the intervention costs.

It is easy, in the midst of forming beliefs based on sometimes complicated data analyses, to lose track of thinking about costs, benefits, values, and goals. A giant effect may seem compelling, just on its own. But it is important not to fall into this trap—because what may seem like obvious implications of a piece of evidence may turn out not to be so obvious. Let's see an example.

Screening Frequently or Accurately

As we write this section, a coronavirus pandemic is sweeping across the world. One of the central challenges in confronting the pandemic concerns testing—specifically, identifying infected people quickly enough that they can be isolated before they spread it to too many others.

As we've emphasized several times in this book, in thinking about the efficacy of a test for diagnosing a disease, both the false positive and the false negative rates matter. The lower each is, the more accurate the diagnosis. Not surprisingly, then, regulatory agencies like the Food and Drug Administration (FDA) demand tests that have low false positive and false negative rates, not allowing them on the market if they are too inaccurate on either front.

Much of the time, this is quite sensible. We don't want sick people concluding they are healthy (false negatives) or healthy people concluding they are sick (false positives). And we don't want to undermine testing by having people conclude that they can't trust tests in general.

In the early months of the coronavirus, medical scientists tried a variety of approaches to testing. Several of these had low false positive rates. But the nasal-swab-based polymerase chain reaction (PCR) tests had the additional virtue of low false negative rates. This is because they were able to detect the virus at quite low levels. Because they satisfied the FDA's requirements for low false positive and false negative rates, PCR tests were quickly approved and became the standard testing regimen.

A competing technology, tests that involved putting saliva on a paper strip, had a harder time getting approval. The reason was their higher false negative rate. The paper-strip tests could only detect the virus at higher levels of concentration. So they were more likely to miss someone who was infected, especially in the early days of an infection, when a person's viral load was still relatively low.

For many diseases, the FDA's position might make a lot of sense. If we are testing for celiac disease or cancer, it makes sense to only approve the most accurate tests. But the coronavirus case is, arguably, different in a bunch of ways that are worth thinking through.

In comparing the merits of two diagnostic tests, the false positive and false negative rates are important. But they aren't the only relevant criteria. One should also consider the relative costs of the two tests. And, especially in the case of a highly infectious disease like the coronavirus, one should also consider the speed of the test. It is one thing to wait a week or two for the results of a celiac test. It is another thing to wait a week or two for the results of a coronavirus test, during which time the person in question could spread the disease to many other people.

As it turns out, while the paper-strip tests have higher false negative rates than PCR tests, they are much cheaper, can be administered at home, and can deliver results in under an hour, as compared to the five to ten days people were waiting for PCR results. If we combine these additional pieces of information with the difference in the false negative rates, we might reach a quite different conclusion about whether the FDA did the right thing by delaying approval of the paper-strip tests.

To start to get a sense of the issues, think just about the difference in price. By some estimates, paper-strip tests cost \$1–\$5, while PCR tests cost \$50–\$100. So comparing one PCR test to one paper-strip test hardly seems fair. We could do at least ten paper-strip tests for every PCR test.

The main way you get a false negative on a test like this is if your viral load is too low to be detected by the test. The PCR test has a lower false negative rate because it can detect the virus in much lower concentration. But the coronavirus grows very quickly in a person. So scientists suspect it only takes a day or so to go from having the sort of viral load that can be detected by a PCR test to having the sort of viral load that can be detected by a paper-strip test.

If this is right, one way of thinking about the difference between the two tests is as follows. Suppose you can afford N paper-strip tests for each PCR test. So, to keep costs equal, let's imagine we do a paper-strip test every day or a PCR test once every N days. For the sake of argument, let's imagine $N = 10$ and let's ignore delays in getting test results back. You have to choose between taking the PCR test on day 1, 11, 21, and so on and taking a paper-strip test every day. Focus on days 1 through 10. Under the PCR regimen, if you have a low viral load on day 1, you detect the virus with the PCR test on that day, but you don't find out you are infected with the paper-strip test for another day or two. If your viral load is low on day 2, you don't detect you are sick with the PCR test until day 11, but you detect you are sick with the paper-strip test on day 3. The same is true for days 3 through 9. If your viral load is low on day 10, you find out you are sick with either test on day 11. So, all told, the probability you find out you are sick faster with PCR testing is 1 out of 10. The probability you find out you are sick faster with paper-strip testing is 8 out of 10. And the probability you find out you are sick at the same time under either regimen is 1 out of 10.

Of course, there may be other reasons that people get false negatives besides low viral loads. So here's another way to think about the comparison. Suppose, again just for the sake of argument, that both tests had a false positive rate of zero. So we are only worried about the false negative rate. Let p be the false negative rate of the PCR test and q be the false negative rate of the paper-strip test. The probability the PCR misses an infected person is p. How likely are ten paper-strip tests to miss this case? That depends on how correlated false negatives are across tests of the same person. If they are perfectly correlated (which surely isn't true, since a person's viral load is increasing over time), then if you get a false negative once, you will always get a false negative. In this case, the probability that ten paper-strip tests miss the case is the same as the probability that one paper-strip test misses the case, q. If, by contrast, false negatives are completely uncorrelated across cases (which also surely isn't true, since some people have lower viral loads than others and so their cases are harder to detect), then the probability that ten paper-strip tests miss an infected person is q^{10}. So, for instance, if the PCR test had a false negative rate of one-tenth of 1 percent and the paper-strip test had a false negative rate of 20 percent, ten paper-strip tests would be way more likely to catch an infected person than one PCR test ($.001 > .2^{10} \approx .0000001$). The truth, of course, lies somewhere in between.

There are still more factors to consider in evaluating these two approaches to testing. First, as we've already indicated, false negatives are more likely early in a person's infection, when the viral load is low. But this is also when people are less infectious. So, as it becomes more important to correctly diagnose people, the difference between the PCR and paper-strip test goes down.

Second, a test's speed is an incredibly important part of the cost-benefit calculation. The main benefit of testing is to keep people from infecting others once they are infected. The coronavirus grows rapidly in an infected person. So there are huge advantages to administering the test at home and getting results in less than an hour.

For all these reasons, studies that simulate disease spread under a variety of testing regimens find that differences in the frequency and rapidity of testing can be much more important than differences in false negative rates. As such, the FDA's sensible-sounding rule for approving diagnostic tests might not have been so sensible in this case.

One thing you might worry about is that, unlike the PCR tests, perhaps the paper-swab tests did not have false positive rates close to zero. If there are lots of false positives, then daily testing might lead to lots of costly and unnecessary self-quarantining. False positive rates are hard to study, but at least some evidence suggests that they were low, even for the paper-swab tests. But even if false positive rates were non-negligible, the combination of the two technologies suggests a reasonable solution. Paper-swab tests need not be treated as the final answer in order to be highly useful. If everyone did a paper-swab test every day, some people would get false positives. They could be asked to self-quarantine, while being immediately administered a more definitive PCR test. With the load on labs lightened by reduced PCR testing, turnaround times might even speed up. And, as a result, false positives could be corrected relatively quickly, with a minimum of inconvenience.

The point of this discussion is not to provide a definitive answer to this difficult policy problem, on which we are not experts. Rather, it is to illustrate the fact that we have to consider lots of different costs and benefits when we make decisions, and every person or society has to use their personal values to decide how to weigh those different costs and benefits. It's easy to fixate on one particular quantitative statistic like the false negative rate and make decisions accordingly, but that is typically a mistake. We'll return to these themes in the final two chapters.

Wrapping Up

Turning statistics into substance helps us think clearly about what exactly the evidence tells us about the questions we are trying to answer. Keeping those questions forefront in our minds is a key element of thinking clearly about how to use quantitative information. Indeed, we need to do so not only when interpreting the results of an analysis but when choosing how to measure, selecting the samples we study, and deciding which settings our results apply to. Those issues are the topic of chapter 16.

Key Words

- **Percentage point change:** The simple numerical difference between two percentages.
- **Percent change:** A way of measuring the degree of change. It is the difference between the initial value and the new value divided by the original value

(multiplied by 100). Unlike percentage point change, percent change is highly sensitive to the original value.

- **Conditional probability:** The probability of an event conditional on some other information. We write the probability of C conditional on E as $\Pr(C \mid E)$.
- **Prior belief:** Your belief about some thing before learning new evidence.
- **Posterior belief:** Your belief about some thing after incorporating new evidence.
- **Bayes' rule:** A formula for calculating your posterior belief conditional on new evidence and your prior belief. In particular: $\Pr(C \mid E) = \frac{\Pr(E|C)\Pr(C)}{\Pr(E)}$. Sometimes called Bayes' theorem or Bayes' law.
- **Statistical power:** The probability of finding a statistically significant result in the data given that the relationship really exists in the world.

Exercises

15.1 A newspaper reports, "Economic growth was 20 percent higher in Country A than in Country B last year."

The typical way that economists measure economic growth is the percent change in GDP from one year to the next. So we'd say economic growth was 3 percent in a particular country and year if the GDP was 3 percent higher at the end of the year than at the beginning.

(a) Suppose GDP growth in Country B was 10 percent. What was GDP growth in Country A?

(b) Suppose GDP growth in Country B was 0.1 percent. What was GDP growth in Country A?

(c) What's an alternative way to write the headline so that you don't misleadingly mask the difference between the scenarios described by (a) and (b)?

15.2 Now consider two other countries C and D. Suppose that growth in Country C is 1 percent while growth in Country D is 0.1 percent.

(a) What is the percent difference in growth? What is the percentage point difference?

(b) Write two headlines, each including a true statistical fact about the two countries. One should make the difference in their economic growth sound like a really big deal. The other should not.

(c) Now suppose that upon a statistical review, growth in Country D turns out to be just 0.001 instead of 0.1 percent. What now is the percent difference in growth between the two countries? What is the percentage point difference? Which of these two statistics better conveys the substantive significance of the shift from 0.1 percent to 0.001 percent? Why?

15.3 During the coronavirus pandemic, governments and private organizations around the world rushed to create diagnostic tests. Those tests varied in their accuracy. Let's think about one of those tests, which was reported to have a 1 percent false positive rate and a 10 percent false negative rate.

We don't know the underlying rate of coronavirus in the asymptomatic population. Suppose the probability an asymptomatic person has coronavirus is some number q—that is, the prior belief any given person is sick is $\Pr(\text{sick}) = q$.

(a) Using the information above about the false negative rate, what is the probability a person gets a positive result given that they really do have coronavirus (written, $\Pr(+ \mid \text{sick})$)? (Hint: You don't need Bayes' rule to answer this question.)

(b) There are two ways to get a positive test result. A person with coronavirus can get a correct test result. And a person who does not have coronavirus can get a false positive. Calculate the overall probability that an asymptomatic person gets a positive test:

$$\Pr(+) = \Pr(\text{sick}) \cdot \Pr(+ \mid \text{sick}) + \Pr(\text{not sick}) \cdot \Pr(+ \mid \text{not sick})$$

(Your answer will have q in it because it will depend on the prior belief that an asymptomatic person is sick.)

(c) Now use Bayes' rule to calculate $\Pr(\text{sick} \mid +)$—the probability that an asymptomatic person tests positive. (Your answer will, again, have q in it.)

(d) We don't actually know q. Let's think about different scenarios.

i. Calculate $\Pr(\text{sick} \mid +)$ if $q = .005$ (i.e., if half a percent of the asymptomatic population has coronavirus).
ii. Calculate $\Pr(\text{sick} \mid +)$ if $q = .01$ (i.e., if 1 percent of the asymptomatic population has coronavirus).
iii. Calculate $\Pr(\text{sick} \mid +)$ if $q = .05$ (i.e., if 5 percent of the asymptomatic population has coronavirus).
iv. Draw a figure with q on the horizontal axes (going from 0 to 1) that graphs $\Pr(\text{sick} \mid +)$.

15.4 Discrimination against certain groups in the job market is a major societal and policy concern. Many studies seek to bring quantitative evidence to bear on the extent of such discrimination.

Let's think through a very simple example. Imagine a society with two equally sized and equally qualified groups: the privileged and the unprivileged.

Using the conditional probability notation we developed earlier we will express the probability that a person gets a job given their group membership as $\Pr(\text{hired} \mid \text{group})$. Similarly, we will express the probability that a person is a member of a particular group given that the person got a job as $\Pr(\text{group} \mid \text{hired})$.

(a) Suppose you want to know whether, if they both apply for the same job, a member of the privileged group is more likely to be hired than a member of the unprivileged group. So you want to know whether, among those who apply for the job, the following is true:

$$\Pr(\text{hired} \mid \text{privileged \& applied}) > \Pr(\text{hired} \mid \text{unprivileged \& applied})$$

i. Use Bayes' rule to rewrite Pr(hired | privileged & applied) as a function of three terms, Pr(privileged | hired & applied), Pr(hired | applied), and Pr(privileged | applied).
ii. Use Bayes' rule to rewrite Pr(hired | unprivileged & applied) as a function of three terms, Pr(unprivileged | hired & applied), Pr(hired | applied), and Pr(unprivileged | applied).

(b) Suppose a study shows that people in a given job are equally likely to be privileged and unprivileged. Express that using our notation. Which two terms from your answer to (a) does that mean you know?
(c) Is the information in (b) sufficient to determine whether, if they both apply for the same job, a member of the privileged group is more likely to be hired than a member of the unprivileged group? Using your answer to part (a), what additional piece of information would you need to know?
(d) Suppose you learned that the same number of members of the two groups applied for the job. Now would you know the answer?

Readings and References

The study on the reporting of fuel-efficiency statistics is

Richard P. Larrick and Jack B. Soll. 2008. "The MPG Illusion." *Science* 320:1593–94.

The *Wall Street Journal* article on a cholesterol drug is

Ron Winslow. "Cholesterol Drug Cuts Heart Risk in Healthy Patients." *Wall Street Journal*, Nov. 10, 2008. https://www.wsj.com/articles/SB122623863454811545.

To learn more about how to create informative data visualizations and how to avoid being fooled by bad graphics, we recommend the following books.

Carl T. Bergstrom and Jevin D. West. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World.* Random House.

Kieran Healy. 2019. *Data Visualization: A Practical Introduction.* Princeton University Press.

Edward R. Tufte. 2001. *The Visual Display of Quantitative Information, 2nd Edition.* Graphics Press.

The figure on partisan trends in the U.S. South is from

Christopher H. Achen and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government.* Princeton University Press.

The story of the statistical errors made in the trial of the Collins couple is related in

Jonathan J. Koehler. 1995. "One in Millions, Billions, and Trillions: Lessons from People v. Collins (1968) for People v. Simpson (1995)." *Journal of Legal Education* 47(2): 214–23.

For more on the SPOT program have a look at two reports by the General Accountability Office:

The 2010 report: https://www.gao.gov/assets/310/304510.pdf.

The 2013 report: https://www.gao.gov/assets/660/658923.pdf.

An analysis of testing regimens for coronavirus can be found in

Daniel B. Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M. Burke, James A. Hay, Milind Tambe, Michael J. Mina, and Roy Parke. 2020. "Test Sensitivity Is Secondary to Frequency and Turnaround Time for COVID-19 Surveillance." https://www.medrxiv.org/content/10.1101/2020.06.22.20136309v3.

An early blog post on the idea is here:

Alex Tabarrok. "Frequent, Fast, and Cheap Is Better than Sensitive." Marginal Revolution. July 24, 2020. https://marginalrevolution.com/marginalrevolution/2020/07/frequent-fast-and-cheap-is-better-than-sensitive.html.