CHAPTER 7

Over-Comparing, Under-Reporting

What You'll Learn

- If analysts make lots of comparisons but report only the statistically significant ones, there will be lots of false positive results and over-estimates.
- These false positives can be the result of nefarious researcher behavior (p-hacking). But they can also arise in a community of entirely honest researchers (p-screening).
- There's no easy solution, but analysts and consumers have some tools at their disposal to reduce the risk of being misled.

Introduction

Although statistical hypothesis testing is a useful tool, it's far from foolproof. To understand why scientific studies and quantitative data analyses so often produce misleading or unreliable results, we're going to start in an unlikely place—the story of a seemingly impressive sea creature.

Can an Octopus Be a Soccer Expert?

In 2008 and 2010, Paul the Octopus made headlines for his apparent prowess in predicting the outcome of soccer matches. Before matches between two national teams, Paul's keepers would present him with two boxes of food, each marked with the flag of one of the competitor countries. The keepers interpreted the box Paul ate from first as his forecast for who would win the match. Paul was surprisingly accurate, and journalists and gamblers eagerly awaited his predictions.

Paul was the subject of much fascination and some scorn. According to Nick Collins of *The Telegraph*, an Argentinian chef was so angry after Paul correctly forecast Germany's defeat of Argentina that "he threatened to cook Paul in retribution." Gamblers were betting on the accuracy of Paul's predictions before he had even made them. Collins reported that "William Hill, the bookmaker, says it has taken so many bets on whether Paul will call the final between Spain and Holland correctly that it had to cut odds from evens to 10/11."

A skeptic might point out that, although octopuses are impressively intelligent, there's no way that Paul could actually have had special insight into the outcome of

Table 7.1. Ways to flip a coin 3 times.

	Three Heads	**Two Heads**	**One Head**	**Zero Heads**
Ways It Can Happen	HHH	HHT HTH THH	HTT THT TTH	TTT

soccer matches. Even experts have a hard time calling games in a sport in which, as far as we Americans can tell, basically no one ever scores. And Paul presumably knew nothing about the teams playing or even about soccer in general. Was Paul's success dumb luck?

As discussed in chapter 6, we have tools for assessing whether an observed pattern is likely to be the result of dumb luck or, more scientifically, noise. We can conduct a hypothesis test and calculate a p-value.

How does Paul fare in such a hypothesis test? Paul made 14 predictions over the course of his career, and he was correct in 12 of those 14 games. That's pretty good. Suppose the null hypothesis that this was dumb luck is true—that is, Paul was picking in a completely random fashion, with each box equally likely to be selected. To figure out whether it is plausible that Paul's record emerged from dumb luck alone, we want to know how likely it is that Paul would have guessed correctly at least 12 times if he was in fact just guessing at random.

This problem is simple enough that you can compute the p-value by hand. The basic idea is this. Assume Paul is guessing at random. Calculate how likely it is that he'd get exactly 12 correct, how likely it is that he'd get exactly 13 correct, and how likely it is that he'd get exactly 14 correct. The sum of those three probabilities is the probability Paul would have done at least as well as he did by dumb luck.

Before we go on with the Paul story, let's pause and learn how to calculate these probabilities. Doing so will help make sure we are all thinking clearly about what dumb luck really means.

It will help to start with a simplification of the problem. Our null hypothesis is that Paul the Octopus is guessing at random—that is, Paul predicting the winner of a game is analogous to a person flipping a coin and having it land on heads. So let's think about flipping a coin. Suppose you flip a coin 3 times (we'll get to Paul's 14 in a bit). Table 7.1 shows all the things that could happen.

Equivalently, if Paul forecast three games, then zero, one, two, or three of his predictions could be correct.

What is the probability that you get, say, exactly two heads? Well, there are eight total things that could happen, and if we're flipping a fair coin, they're all equally likely. Of those eight, three involve getting two heads. So the probability you get exactly two heads when you flip three coins is $\frac{3}{8}$. Analogously, if Paul was forecasting three games at random, the probability he would get exactly two correct is $\frac{3}{8}$.

But that isn't quite the quantity we want to know. We want to know the probability you get *at least* two heads or that Paul correctly forecasts at least two games.

Well, in addition to getting two heads, you could also get three heads. There is one way for this to happen, so the probability you get three heads is $\frac{1}{8}$. Hence, the probability you get at least two heads is $\frac{3}{8} + \frac{1}{8} = \frac{1}{2}$. Analogously, if Paul had called three games, and he was just guessing randomly, the probability he'd get at least two right is one-half.

But Paul didn't just forecast three games. He forecast 14. Making a table for fourteen coin flips would be pretty boring. So let's think about how to analyze this problem a little more generally.

Suppose you flipped a coin n times. How likely is it that you get exactly k heads? Let's start by calculating the probability that each of the first k coin flips lands on heads and the remainder land on tails. The probability that the first k flips land heads is $\frac{1}{2}^k$. The probability that the remainder land tails is $\frac{1}{2}^{n-k}$. So the probability that the first k flips land heads and the remaining $n-k$ flips land tails is $\frac{1}{2}^k \times \frac{1}{2}^{n-k}$.

Of course, that's only one way to get exactly k heads. The k heads don't have to be the first k flips. They can be any group of k out of the n flips. There are $\frac{n!}{k!(n-k)!}$ different ways to get exactly k heads when flipping a coin n times. So the overall probability of getting exactly k heads out of n coin flips is

$$\frac{1}{2}^k \times \frac{1}{2}^{n-k} \times \frac{n!}{k!(n-k)!}.$$

The exclamation points above mean *factorial*. We refer to the expression $n!$ as *n factorial* and it's defined as the product of n and every positive whole number less than n. So, for example, $3! = 3 \times 2 \times 1 = 6$.

Let's see if this confirms our finding from before in our three-coin-flip example. If we flip a coin three times, what is the probability we get exactly two heads? Since $n = 3$ and $k = 2$, we calculate the probability as follows:

$$\frac{1}{2}^2 \times \frac{1}{2}^{3-2} \times \frac{3!}{2!(3-2)!} = \frac{1}{4} \times \frac{1}{2} \times \frac{3 \times 2 \times 1}{2 \times 1 \times 1} = \frac{3}{8}$$

And now we can calculate the probability that Paul the Octopus would correctly predict 12 or more games out of 14, if he was picking at random. The probability he gets exactly 12 right is

$$\frac{1}{2}^{12} \times \frac{1}{2}^{14-12} \times \frac{14!}{12!(14-12)!} \approx .00555.$$

The probability he gets exactly 13 right is

$$\frac{1}{2}^{13} \times \frac{1}{2}^{14-13} \times \frac{14!}{13!(14-13)!} \approx .00085.$$

The probability he gets all 14 right is

$$\frac{1}{2}^{14} \times \frac{1}{2}^{14-14} \times \frac{14!}{14!(14-14)!} \approx .00006.$$

So the probability Paul calls twelve or more games correctly is approximately $.00555 + .00085 + .00006 \approx .0065$, around 1 in 155. In other words, if Paul had no special insight into soccer, it's highly unlikely that he would have been as accurate as he was. That's, of course, precisely why everyone was obsessed with Paul. And it seems like perhaps they were right to be. Using the standard statistical hypothesis testing approach that we introduced in chapter 6, we can reject the null hypothesis that Paul is just guessing at random and conclude that we have statistically significant evidence that Paul is indeed an expert soccer forecaster.

The analysis above is pretty similar to what two mathematicians, Chris Budd and David Spiegelharter, did when they were interviewed about Paul back in 2010. But if we look at Paul's games a little more closely, we can see that this analysis may be overly generous to Paul's psychic powers.

Paul lived in Germany, and he was primarily asked to predict the outcome of games in which Germany was competing. In fact, 13 of the 14 games involved Germany. Furthermore, Paul had a strong tendency to pick Germany. Maybe he liked that flag because he'd been seeing it for years. Maybe the German box happened to be his favorite box for reasons unbeknownst to us. Maybe Paul's handlers subconsciously trained Paul to pick Germany. Who knows? It also turns out that Germany is good at soccer—they win most of the time. So maybe Paul's success isn't so shocking. Let's redo the analysis above with this information in mind.

Paul predicted the outcome in 13 games involving Germany and he picked Germany to win 11 of those games. Germany in fact won 9 of them. Our null hypothesis is again that Paul's predictions were dumb luck, in the sense of having no special insight into soccer. But this time, instead of imagining that he was equally likely to choose either box, imagine he was predisposed to pick the German box. Let's assume his predisposition meant that his probability of picking Germany was $\frac{11}{13}$ in any game Germany played, since that's the frequency with which Paul in fact selected Germany. If Germany won 9 of 13 games and Paul selected Germany with a probability of $\frac{11}{13}$ each time, how likely is it that he would have been correct 11 or more times just by pure chance? This p-value could be computed by hand, but it is complicated to do so. So instead we ran a simple simulation on our computer to approximate it. With these tweaked assumptions, the chances that Paul gets 11 or more games right out of 13 is about .03 or 1 in 33—still unlikely, but much more likely than 1 in 155.

Now, what do we think? It still looks pretty unlikely that Paul's success is attributable purely to dumb luck. Even if he was predisposed to predict the strong German team, there was only a 3 percent chance that Paul would be as successful as he was. Therefore, a traditional hypothesis test with a .05 threshold would still lead us to reject the null. We continue to have statistically significant evidence that Paul is good at predicting soccer matches.

You won't be surprised to learn that we're still skeptical. But why? Paul is not the only octopus out there. What if there were actually ten octopuses scattered around Germany, each trying to predict the outcomes of soccer matches? The world, of course, would only ever hear about the most successful one. If this is right, then we still haven't tested the right hypothesis to figure out how likely it is that Paul's accuracy was due to dumb luck. If there really were ten octopuses trying to predict soccer matches, and if Paul just happened to be the one who did well and therefore became famous, instead of asking how likely it is that Paul would be so accurate by luck, we have to ask how likely it is that any one of the ten octopuses would be that accurate by luck. Because if it had turned out that Paulina the Octopus was right 12 out of 14 times instead of Paul, then we'd be talking about Paulina and we'd never have heard of Paul.

Figuring out how likely it is that some octopus out of the ten would have been as accurate as Paul is relatively straightforward. But to do the calculation, we need to understand one more fact about p-values. Recall that the p-value is the probability of observing an outcome at least as extreme as the one you observe if the null hypothesis is true. So, if the null hypothesis is true, how often will you observe an outcome as extreme as an outcome with a p-value of .05? Exactly 5 percent of the time. And if the null hypothesis is true, how often will you observe an outcome as extreme as an outcome with a p-value of .2? Exactly 20 percent of the time. And so on for each and every p-value. This is just a restatement of the definition of the p-value.

But from this fact, we learn something important. When the null hypothesis is true, we should observe a p-value less than or equal to .05 in 5 percent of cases, p-values less than or equal to .2 in 20 percent of cases, p-values less than or equal to .5 in 50 percent of cases, and so on. Hence, it must be that, when the null hypothesis is true, you are equally likely to find each p-value. (The technical jargon for this is that p-values are *uniformly distributed under the null.*)

So, what's the probability that at least one of our German octopuses would generate a record of prediction with a p-value at least as good as Paul's record of .03 by dumb luck alone? We just saw that the probability that any one octopus generates a p-value of .03 or lower by dumb luck alone is .03. Therefore, the probability that any one octopus generates a p-value higher than .03 is .97. If there are two octopuses and they're making their guesses independently, the probability that neither of them generates a p-value better than .03 is therefore $.97^2$. So the probability that at least one of them generates a p-value of .03 or better is $1 - .97^2$ (i.e., one minus the probability that both generate a p-value worse than .03). If ten octopuses are taking random guesses, the probability that at least one generates a p-value as good as Paul's is $1 - .97^{10} \approx .26$. In other words, if ten German octopuses went through the same ridiculous prediction exercise as Paul, there's about a 1 in 4 chance that at least one of them would have accumulated a record of predictive accuracy at least as glorious as Paul's, even if none of the octopuses were in fact soccer experts. This should make us much more skeptical of Paul's abilities.

We don't know how many German octopuses were in the soccer forecasting business. But we do know that lots of other animals got in on the action. No joke, Leon the Porcupine, Petty the Hippopotamus, Anton the Tamarin, and Mani the Parakeet all forecast the winners of soccer matches around the same time as Paul. And those are just the ones that made the news. Presumably, there were dozens more that we never heard about. And this discussion only includes soccer. What about all the other sports? And what about all the non-athletic things there are to predict? If Judy the Badger were good at predicting the winners of college football games, Steve the Cat were good at predicting the winners of congressional elections, and Fran the Otter were good at predicting stock market shifts, they would be celebrities too. But their predictions turned out to be no better than chance, so we never heard about them.

Budd and Spiegelharter, the mathematicians, were quick to point this out. Spiegelharter notes that "if someone flips a coin and gets the same result 9 or 10 times, it is not remarkable in itself, but it will seem remarkable to the person flipping the coin." In other words, if enough people flip coins, one of them is bound to flip a bunch of heads in a row. And despite the fact that somebody was bound to have a lucky streak of heads by chance, that person might wrongly conclude that they have an unfair coin or that they are a particularly skilled coin flipper. Unfortunately, as we'll see, this problem applies to much more serious situations than coin flipping and soccer forecasting, with far-reaching implications.

Publication Bias

Statistical hypothesis testing and *p*-values are clearly useful. When we find patterns in data, we want to know if they reflect genuine phenomena or if they could have easily been produced by random chance.

But there is a problem, which the story of Paul the Octopus highlights. Neither the public nor the broader scientific community gets to see all the hypothesis tests that were (or could have been) conducted. Often, the only results reported and published are the statistically significant ones. It's just not that interesting to write about Mary the Octopus, who is about as good as a coin flip at predicting soccer matches. But if there are twenty different animals out there making soccer predictions, we'd expect one of them to have a *p*-value less than .05 by pure luck, even if none of them actually have any special insight into soccer. Only that 1 in 20 case will be written up or reported in the news. So if we base our beliefs purely on what gets reported, we will have systematically misguided beliefs.

Making a lot of comparisons, *over-comparing*, while selectively reporting only the interesting or statistically significant ones, *under-reporting*, is a dangerous combination. But it is widespread. And because of it, when we hear about a new, exciting scientific result, it is much harder to know how likely it is to reflect a genuine phenomenon than the simple logic of hypothesis testing suggests.

This problem of over-comparing and under-reporting doesn't just affect how confident you should be that one particular finding is genuine. It also affects our ability to accumulate knowledge over time in a field. We know that any one estimate, even if unbiased, may be far from the true estimand because of noise. The hope is that as a field accumulates estimates, the noise averages out, so that the average of a large number of unbiased estimates gets very close to the true estimand. Over-comparing and under-reporting means that this may not work for the collection of published estimates, a troubling phenomenon called *publication bias*. To see why, let's go back to our favorite equation:

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

Suppose there are a large number of studies, all on the same question. Each study is really well designed, producing an unbiased estimate of the phenomenon under consideration. So the only reason the estimates differ from one another or from the true quantity of interest in the world (the estimand) is because of noise.

But let's also suppose, in the spirit of under-reporting (i.e., not reporting every result), that we only hear about the results of studies in which the evidence is strong enough to reject the null hypothesis that the true estimand is zero (i.e., that our estimate was the result of noise). For a result to be statistically distinguishable from zero, the estimated relationship must be sufficiently large, relative to the standard error. So if we only hear about statistically significant results, we are only hearing about the estimates that were sufficiently large in magnitude.

This means that, for any given true estimand, the estimates that end up being reported will be those for which the noise happened to be large enough in magnitude to push the magnitude of the estimate far enough away from zero to make it statistically significant. So, as a result of over-comparing and under-reporting, not only will our

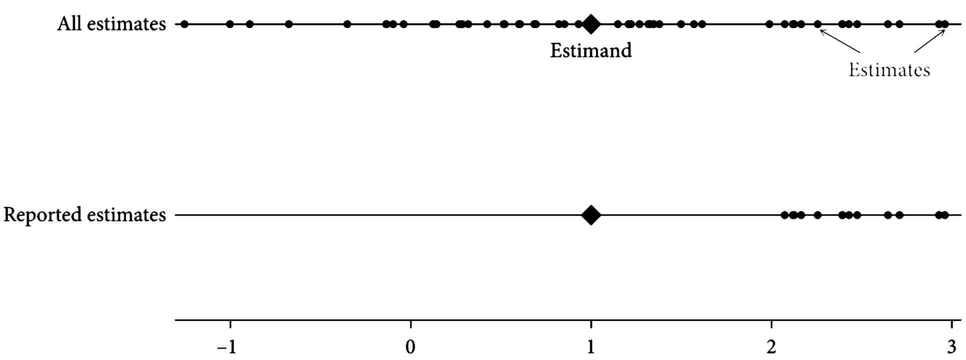

Figure 7.1. Only reporting statistically significant estimates creates publication bias.

p-values be wrong, but the collection of estimates that we learn about from published studies will systematically over-estimate the magnitude of the true estimand.

Distressingly, even though we started by assuming no bias in our estimates, we've learned that the process of over-comparing and under-reporting introduces bias, not into any one estimate but into the overall distribution of estimates reported in a scientific literature. So, when we average all the estimates, we do not get close to the true estimand, even if the number of estimates is very large. That is, we have what is called *publication bias*.

Figure 7.1 illustrates how this works. In the top figure, we see fifty unbiased but imprecise estimates. The true estimand is 1, and since our estimates are unbiased, the average of all estimates is also 1.

We calculate the standard error and find that the 95% confidence interval is from −2 to 2. That is, the probability an estimate would have arisen even if there was no relationship (i.e., the true estimand was 0) is less than 5 percent only if that estimate is greater than 2 in magnitude. So only the estimates larger than 2 are deemed statistically significant (we don't have any estimates less than −2). The statistically significant estimates are in the lower figure.

Suppose that only these statistically significant estimates are ever reported. Now, of course, the reported estimates are systematically greater than the estimand. So if we based our beliefs about the true value of the estimand on these published estimates, we'd be systematically over-estimating the truth. This is publication bias.

Over-comparing and under-reporting that results in publication bias can arise in a variety of ways. Let's consider a couple.

p-Hacking

One way we might end up with over-comparing and under-reporting is through bad or dishonest behavior by individual analysts. The scientific community calls the behavior of playing around with the data or tests until a p-value below a particular threshold emerges *p-hacking*. For instance, suppose a scientist does an experiment and doesn't quite get statistically significant evidence for the expected or desired result. That scientist might reason that something probably wasn't quite right in the initial attempt, and so try a slight tweak on the experiment. Indeed, the scientist can keep trying similar experiments until one comes up statistically significant. Because of noise, if they try the experiment enough times, eventually they will get a result, even if there is no real phenomenon being studied. That's the problem of over-comparing. And, of course, if an unscrupulous scientist only reports the results from the one experiment that yielded a statistically significant result, we have the problem of under-reporting and, thus, publication bias.

Or maybe the analyst has some flexibility in how a particular statistical test is implemented. Suppose you are asked to do an analysis at work about the relationship between productivity and having a standing desk. Should you pool the entire workforce together or run separate regressions for different age groups? Should you include higher-order polynomials of age? Should you separate women and men in the analysis? How about people with different job responsibilities, medical histories, and so on? As you can see, there are lots of reasonable ways to think about doing the analysis. If you keep trying out different ones, you'll eventually find a statistically significant result, even if there is in fact no real relationship between productivity and standing. Searching over specifications is, thus, another way of over-comparing.

Yet another way of over-comparing is by trying out lots of different outcomes. Suppose you want to evaluate the efficacy of some new pill for heart disease. You might run an excellent clinical trial that generates no bias at all. But perhaps you collected data on lots of outcomes for the experimental subjects: mortality, heart attacks, strokes, cholesterol, days of hospitalization, ability to exercise, subjective sense of well-being, and so on. You can then test whether the pill has a statistically significant effect on each of these outcomes. If you have enough different outcomes, you're likely to find a statistically significant result on one of them just due to noise—that is, the people given the pill and the people given the placebo will happen to differ on some outcome, even if the pill doesn't actually do anything. And, if you lack the proper ethics, you might just report the results for that one outcome in the hope of convincing doctors to prescribe your new pill.

As we've seen, *p*-hacking can take many different forms, and you have to work hard to avoid it as a quantitative analyst and to detect it as a consumer of quantitative evidence.[1]

p-Screening

Of course, *p*-hacking is a big concern. But it need not be the case that any individual is acting in a dishonest or negligent way for the problem of over-comparing and under-reporting to occur. It can happen even if absolutely everyone behaves in a completely honest and responsible manner!

Imagine that twenty scientists around the country all have the same scientific hunch. Let's suppose it's about the efficacy of a potential new cancer drug. In truth, the hunch is false—the drug doesn't work. But there is no way for the scientists to know this at the outset. So, as scientists do, they design studies to test the drug. Indeed, all twenty of their labs, independently and unaware of the others, run the same high-quality experiment, but on different samples. They each recruit a large sample of patients with the relevant type of cancer. At random, they assign half of them to receive the new drug. The other half receives a placebo sugar pill. At the end of the study period, they assess whether the

[1] Fun fact: The term *p*-hacking was coined by Joseph Simmons, Leif Nelson, and Uri Simonsohn in a clever study in which they showed, among other things, that by using standard methods in social science, they could provide statistically significant evidence that listening to the song "When I'm Sixty-Four" by the Beatles makes subjects younger!

group that received the drug was more likely to go into remission than the group that received the placebo.

Nineteen out of the twenty labs find no statistically significant evidence—the remission rates among those who received the drug and those who received the placebo are indistinguishable—and conclude that the drug doesn't work. Such null results aren't considered very exciting. "Another drug doesn't cure cancer" isn't a great headline. So scientific journals are reluctant to accept papers reporting null results for publication. As a consequence, these labs may not bother to write up their results, instead just moving on to more promising lines of research. This is sometimes called the *file drawer problem* because statistically insignificant results get locked away in a file drawer. Even if the labs do write up their findings, they might have trouble finding a journal interested in publishing them. In either case we get under-reporting, and the scientific community and the public fail to learn about these nineteen null results.

One (un)lucky lab out of the twenty finds statistically significant evidence suggesting the drug works. We know the drug doesn't work (though the scientists don't), so we know this is sheer chance. It just so happens that, for reasons having nothing to do with the drug, the people assigned to receive the drug in this experiment also had higher remission rates than the people assigned to receive the placebo. These things happen. The estimate can differ from the estimand, even absent bias, because of noise.

Since the other studies were either never written up or never accepted for publication if they were written up, as far as the scientist in charge of this one lab is aware, all existing evidence points toward this new drug working. So this one lab, appropriately, writes a scientific paper on their findings. Because the result is surprising and noteworthy, it is likely to be well published and reported on by the scientific press. And, indeed, if you look at this one study, it looks great. The lab ran a good, unbiased experiment. They made only the one appropriate comparison about the one appropriate outcome. There was no p-hacking. And the data support their hypothesis. So everyone believes this result even though it is in fact completely wrong and, if we had access to all the data (i.e., from the nineteen "failed" experiments), we'd see that the preponderance of the evidence points in exactly the opposite direction. That is, we end up with publication bias.

There isn't a term in common usage that describes both scientists not bothering to write up results that find small or no effects because they'll be hard to publish (which is the file drawer problem) and journals being reluctant to publish such findings even if they are written up. But we think these two phenomena are usefully thought of together, since they both give rise to publication bias despite no individual acting inappropriately. So, by analogy to p-hacking, we call this problem *p-screening*. The issue here isn't that some individual researcher is p-hacking their way to a statistically significant result. The issue is that the scientific community, through its publication practices, screens out studies with p-values that are above some threshold. Under p-hacking we don't see null results because dishonest researchers hide them. Under p-screening we don't see null results because honest researchers can't publish such results. Either way the outcome is the same. The results we do see suffer from publication bias due to lots of comparisons being made but only the statistically significant ones being reported.

Yikes! Because of p-screening, the scientific record (and our knowledge in a lot of other areas) can be unreliable even when everyone behaves just as they should. This should make you worry that it is going on all the time. In fact, stop and ask yourself, For how many things I believe might this story characterize the state of knowledge? Once you start thinking clearly about the problem, you'll see its potential everywhere.

Are Most Scientific "Facts" False?

As we've seen, over-comparing and under-reporting gives rise to publication bias. And these practices are pretty deeply entrenched in a lot of scientific practice and culture. This realization has led to something of an existential crisis in many scientific fields, with practitioners wondering whether many widely accepted scientific facts are actually false, the artifacts of over-comparing and under-reporting.

This is a real concern. It is surely the case that many things that we believe are true are in fact false because of publication bias. But certainly not everything. And analysts have started to think more clearly about how we might diagnose when a scientific consensus or literature is or is not likely to suffer from severe publication bias. To see how, we are going to talk through a couple examples of the problem and various attempts to address it. We'll even discuss some tips on how to detect *p*-hacking.

ESP

In 2010, Cornell psychologist Daryl Bem made news by publishing a study in the *Journal of Personality and Social Psychology*, a prestigious academic psychology journal, claiming that human beings have extrasensory perception (ESP). Often, academic researchers and quantitative analysts are the ones debunking claims about the paranormal, but in this case, a respected, tenured Ivy League professor was the source of the outlandish claim.

In Bem's experiment, students were asked to predict which virtual curtain on their computer screen (left or right) had an object of interest hiding behind it. Bem reported statistically significant evidence that his subjects were better than chance at predicting the future and identifying the correct curtain.

This is a very exciting finding if you are a journal editor who cares about notoriety or a science journalist who cares about readership. The result is cool. The scientist in question works at a reputable university. The article is published in a major scholarly journal. There is no reason to think the data are faked. The study provides scientific evidence of, to say the least, a surprising phenomenon. What journalist with blood running through their veins could resist the story?

This study and all the media attention it received notwithstanding, we're fairly confident that people don't have ESP. So what is going on?

There are, of course, the normal concerns with statistical hypothesis testing. If an analyst uses a significance threshold of .05, there's a 5 percent chance of finding support for a result (i.e., rejecting the null), even if the result is false (i.e., the null is true). And as we'll see in part 4, if you already have good reasons to believe that people do not have ESP, then you shouldn't shift those beliefs much in response to this one study.

But we have other concerns based on the themes of this chapter. This is a case where many researchers might be conducting experiments, but only the one with statistically significant evidence of an unlikely phenomenon gets published. Presumably, nobody is going to publish a paper that reports that people are no better than chance at guessing the correct curtain. That's what we all already believe. So we should worry a lot about publication bias due to *p*-screening.

We should probably also consider the possibility that the results were *p*-hacked. Bem reported the results of nine different experiments carried out over the course of ten years. These experiments are relatively inexpensive to conduct. Since Bem was, by all accounts, committed to the study of ESP, it doesn't seem far-fetched to imagine that he might have conducted more ESP experiments over this ten-year period. And, if so, the nine experiments that were reported on might well have been the ones with the strongest confirmatory evidence of ESP.

There are also some signs of over-comparing and under-reporting within the study itself. For example, Bem doesn't find evidence of ESP in general; he only finds it when the object behind the curtain is erotic in nature. For other kinds of objects, he finds no evidence of the paranormal. Of course! Wouldn't it make sense for humans to have evolved ESP that allows us to detect erotic activity around the corner but not other kinds of activity? Furthermore, in some tests, he only finds effects for women, not men; in others, he finds results only for those who are easily bored. Given all of the different tests conducted by Bem, it would be surprising if he *hadn't* stumbled upon a few statistically significant results, just by chance.

Reassuringly, the community of psychologists remained skeptical and responded quite quickly to Bem's paper. Several follow-up studies tried and failed to replicate the findings. Disappointingly, however, the *Journal of Personality and Social Psychology* initially refused to publish the replication studies debunking Bem's claim. The editor justified this decision on the grounds that the journal has a long-standing policy of refusing to publish *mere replication*. Fortunately, the journal eventually had a change of heart and published a meta-analysis of replication attempts, which strongly suggested that the original result was unreliable. This illustrates one important corrective to the problem of over-comparing and under-reporting: a vigilant commitment to investigating whether findings replicate within a scientific community. We will discuss replication in more detail later in this chapter.

Get Out the Vote

Political campaigns engage in lots of activities—phone calls, direct mail, door-to-door canvassing—to try to get out the vote. Since the 1990s, scholars have teamed up with campaigns to conduct experiments to learn about the efficacy of these efforts. In such studies, some people are randomly assigned to treatment (e.g., a direct mailing with information about the date of the election or the location of their polling place) and other people are randomly assigned to control (e.g., not getting any extra information). We can learn about the average effect of get-out-the-vote efforts on voter turnout by comparing the turnout rates in the two groups.

In the published record, the average estimated effect of a get-out-the-vote intervention is about a 3.5 percentage point increase in voter turnout. Moreover, almost no published paper reports an effect of less than 1 percentage point. So, if a campaign consulted the published record, it would conclude that get-out-the-vote efforts are quite effective.

But there have been many more get-out-the-vote experiments than there are scientific papers published on the topic—which means that some of those experiments did not result in publication. Why not?

If our fears about over-comparing and under-reporting are right, we might expect that the answer is *p*-screening—experiments that yielded no statistically significant evidence of an effect did not result in a published paper. If this is true, then there is publication bias. So we should expect that the true average effect of get-out-the-vote interventions is smaller than what is suggested by the published findings.

Three political scientists, Don Green, Mary McGrath, and Peter Aronow, decided to investigate this possibility quantitatively. They managed to get their hands on the data from over two hundred experiments done by a variety of scholars over many years. Some of those experiments had resulted in published papers. Others had not. They did an analysis to find the average effect of get-out-the-vote interventions across all two hundred of these interventions. The result: half a percentage point! Dramatically less than the 3.5 percentage point average effect shown in the published record. The unpublished record is, indeed, much less supportive of the efficacy of get-out-the-vote efforts than is the published record.

The efficacy of get-out-the-vote efforts is one of the most rigorously studied topics in the social sciences. And so, candidates or campaigns wanting to figure out the best way to allocate scarce resources might naturally turn to published studies to inform their decision. However, we now know that doing this would lead them to over-estimate how effective get-out-the-vote efforts are by a factor of 7, demonstrating again that *p*-screening can have serious consequences.

p-Hacking Forensics

It's always hard to know for sure if *p*-hacking took place in an individual study. And it is a good practice to be charitable, assuming that most people are attempting to behave in an above-board and honest manner most of the time. That said, clear thinking can help us get a sense of how widespread the *p*-hacking problem is. The best evidence comes from observing the *p*-values in actual published scientific literature. Doing so doesn't tell us whether any individual study is *p*-hacked. But it can help us sniff out whether an overall literature looks like it has a bunch of *p*-hacking going on.

Here's how it works. You start by thinking about what the distribution of *p*-values in a literature would look like in four different possible states of the world:

1. If there is no real relationship in the world and there is no *p*-hacking;
2. If there is a real relationship in the world and there is no *p*-hacking;
3. If there is no real relationship in the world and there is *p*-hacking; or
4. If there is a real relationship in the world and there is *p*-hacking.

You are then going to compare the actual distribution of reported *p*-values in a scientific literature to the distributions you would get in each of these four states of the world to try to figure out which state you are most likely to be in. For what follows we are going to assume there is still *p*-screening going on (so there are no *p*-values greater than .05). We just want to figure out if the literature is also *p*-hacked. But everything we are going to say is true even without *p*-screening.

The logic of the cases can be understood with reference to figure 7.2, which is adapted from a 2014 study by Simonsohn, Nelson, and Simmons, who first proposed examining the distribution of *p*-values in order to assess *p*-hacking.

Start with case 1: there is no real relationship in the world and there is no *p*-hacking. If there is no real relationship out there in the world, that means the null hypothesis is true. And, as we've already discussed, if the null hypothesis is true, then all *p*-values are equally likely to emerge in any given study. So, if there is no *p*-hacking, the observed distribution of *p*-values in published studies should look approximately uniform—that is, between 0 and .05, different *p*-values should appear with approximately equal frequency. This is illustrated by the light-gray line in figure 7.2.

There are two reasons we might see a deviation from this uniformity. The first is that there is a real relationship in the world. The second is that there is *p*-hacking.

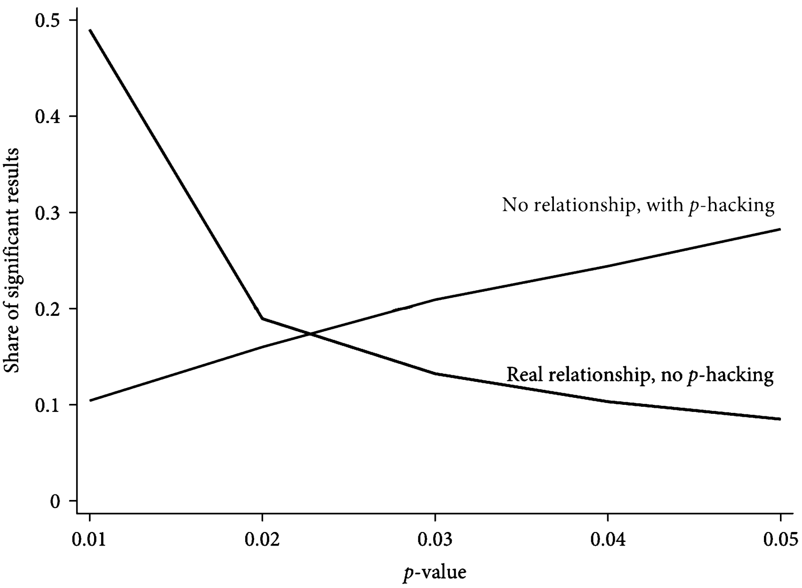

Figure 7.2. *p*-hacking distorts the distribution of *p*-values in a literature.

This brings us to case 2: there is a real relationship in the world (i.e., the null hypothesis is false) and there is no *p*-hacking. If we are actually studying a real relationship in the world, we are more likely to detect a statistically significant relationship than if there is no real relationship in the world. So, in case 2, where there is a real relationship and no *p*-hacking, we expect a distribution of *p*-values in the published record that is skewed such that there are more low *p*-values. That is, reflecting the fact that we are detecting a real relationship, there should be more low *p*-values in case 2 than in case 1. So, if we see a distribution with more low *p*-values, that is suggestive evidence that the literature is detecting a real relationship in the world. This is illustrated by the dark curve in figure 7.2.

The other reason we might see a deviation from case 1 is because of *p*-hacking. This takes us to case 3: there is no real relationship in the world and there is *p*-hacking. As we've already discussed, when there is no real relationship in the world, every *p*-value is equally likely. But, what happens in the presence of *p*-hacking? Well, suppose a researcher finds a *p*-value below .05. They can just report that statistically significant result. But suppose they find a *p*-value close to, but above, .05. They might be tempted to *p*-hack, playing around with specifications, subgroups, and so on until they find a *p*-value below .05 that they can report as statistically significant. The consequence of this *p*-hacking will be a whole bunch of reported *p*-values close to, but just below, .05. So, unlike in case 2, where we saw more low *p*-values than high *p*-values among statistically significant results, in case 3, we should expect more high *p*-values than low *p*-values among statistically significant results. This is illustrated by the medium-gray curve in figure 7.2.

Case 4 combines cases 2 and 3. If there is a true relationship, that skews things toward low *p*-values. If we also *p*-hack, that skews things back toward high *p*-values. So it is hard to know what to expect in this case. But, nonetheless, just with the distinctions between cases 1, 2, and 3, we can make some progress diagnosing *p*-hacking in a literature.

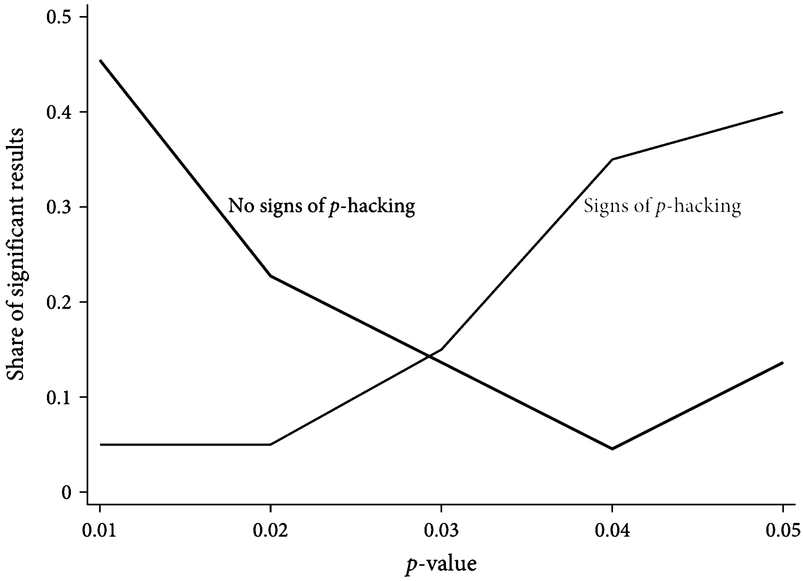

Figure 7.3. Using the distribution of p-values to diagnose p-hacking.

Sadly, many academic literatures exhibit a distribution of p-values consistent with case 3. Simonsohn, Nelson, and Simmons examined papers in a prominent psychology journal to see if there were any red flags that might indicate p-hacking. They identified certain words that might be a sign of over-comparing. For instance, one of their words of concern is *excluded*, as in "I excluded this variable (or group, or outcome) from my analysis because it didn't give the result I wanted." Another is *transformed* as in "I transformed age into age^2, age^3, age^4, and so on until the results supported my hypothesis." The darker curve in figure 7.3 shows the distribution of p-values for studies that don't use words that are signs of p-hacking. Reassuringly, for these studies, we see more low p-values, indicating that they are identifying genuine relationships in the world (case 2). The lighter curve in figure 7.3 shows the distribution of p-values for studies that do use words that are signs of p-hacking. Disturbingly, for these studies, there is reason to suspect p-hacking; we see more high than low p-values (case 3). So, while these forensics don't tell us exactly which papers are p-hacked, they allow us to look at the distribution of p-values in a literature and ask how worried we should be that any scientific consensus based on those studies is biased by p-hacking.

Potential Solutions

Publication bias is an insidious problem for science. And so scientists have started thinking about how they might change scientific practice to reduce the problem of over-comparing and under-reporting.

Reduce the Significance Threshold

Maybe we can solve the problem of publication bias by using a more stringent significance threshold for p-values. Maybe the conventional .05 threshold makes it too easy to hunt around until you find a statistically significant result. In 2017, seventy-two researchers across various fields published a letter in *Nature Human Behaviour* urging the scientific community to adopt a dramatically lower significance threshold of $p < .005$.

On the one hand, a lower significance threshold would certainly make it harder to conjure up a statistically significant result by over-comparing. On the other hand, lowering the significance threshold would likely increase incentives for p-hacking by making statistically significant results rarer and, thus, more valuable. It might even increase complacency about these issues, leading us all to think a little less critically. And, while a threshold of .005 means fewer false positives (i.e., rejecting the null hypothesis when it is true), that comes at the cost of more false negatives (i.e., failing to reject the null hypothesis when it is false). It's not obvious where we should draw the line to balance that trade-off. The answer probably depends on the particular question.

Adjust p-Values for Multiple Testing

The p-value is supposed to tell us the probability of obtaining a result at least as strong as your result if the null hypothesis is true. As we've seen, if researchers engage in over-comparing and under-reporting, the p-value doesn't reflect this probability. It is too low.

If we know how many tests were run, we can try to correct the p-value. As we discussed in the case of Paul the Octopus, if researchers conduct ten independent tests but only report their lowest p-value of .03, the true p-value is more like $1 - (1 - .03)^{10} \approx .263$. Correcting p-values in this way to account for the number of tests run is a good way for researchers to be transparent and for consumers of quantitative information to better assess the state of the evidence. Unfortunately, this is also not a panacea. The kind of simple calculation done above only works if the tests are truly independent. If the tests are closely related to each other—for example, if we're testing the same hypothesis with the same data but using slightly different variables in a regression or focusing on slightly different subgroups of observations—if may be much less clear how to adjust the p-values correctly.

Don't Obsess Over Statistical Significance

The threshold .05 is just an arbitrary number. Substantively important effects may be statistically insignificant, and statistically significant results may be substantively unimportant. Statistical hypothesis testing is a useful tool for quantifying uncertainty, but it can be abused. We should use p-values when appropriate. But they're not the end-all-be-all for assessing the believability of a quantitative finding. You can't just calculate; you have to keep thinking clearly. In part 4, we'll talk about how to incorporate quantitative evidence with other knowledge in order to think clearly about what our beliefs should be after seeing some new evidence.

Pre-Registration

At least in some settings—such as when the researchers are creating the data themselves with a new survey or experiment—researchers can pre-commit to exactly the tests they are going to do before they ever see the data. To do so, they pre-register their study, writing down and publishing exactly what they plan to test for and how they plan to do so, before actually doing the study. As long as they pre-register a reasonable number of tests, this prevents them from over-comparing. It also makes it harder to under-report.

People will be suspicious if a scientific paper only includes the results from 3 out of 10 promised tests. Moreover, some scientific journals are now willing to accept scientific studies for publication based only on the pre-registered plan—committing to report the results regardless of what the researcher finds, which also helps with under-reporting.

Pre-registration is a useful tool for mitigating the problems of over-comparing and under-reporting. Let's see an example of it in action, so we can get a sense of its merits and its limits.

Requiring pre-registration in drug trials

The problem of over-comparing and under-reporting is an important one in clinical trials for new drugs—a company that has invested a lot in a new drug might be tempted to search over specifications, subgroups, or outcomes until they find some result that suggests the drug is efficacious for some outcome on some group of people. The National Heart, Lung, and Blood Institute (NHLBI) has funded many clinical trials of new drugs and dietary supplements since 1970, and in 2000, they came up with a clever way to use pre-registration to combat this problem. They required the developers of a drug or supplement to announce beforehand the goals of the product. Under these new rules, a clinical drug trial is only declared a success if the researchers show a statistically significant effect on the pre-registered outcome of interest.

A 2015 study by Robert Kaplan and Veronica Irvin shows that after the NHLBI started requiring pre-registration, the rate of successful trials dropped from 57 percent to 8 percent. This suggests that many of the "successful" trials prior to pre-registration were the result of over-comparing rather than of any genuine effect of the drug or supplement. This was a big success for pre-registration.

Importantly, even if pre-registration is working to curtail over-comparing and under-reporting, we still have to worry about all the other problems of statistical inference. Think about the 8 percent success rate after pre-registration. Kaplan and Irvin use a significance threshold of .05, which means that even if none of the drugs or supplements were effective, we'd expect 5 percent to generate statistically significant evidence of efficacy just by chance. So the 8 percent success rate is not much higher than what we'd expect even if none of the drugs work. That means that even after we observe a successful, pre-registered trial, we still should not be that confident that the drug is effective. In fact, it looks like there's still a 5 in 8 chance that a positive, pre-registered result is a false positive.

Replication

One way to assess whether an estimated effect is genuine is to replicate it on new, independently generated data. Replication isn't foolproof. But, as we saw with the ESP studies, it can help provide some evidence of whether an estimated effect is genuine or just the result of over-comparing and under-reporting.

Suppose we do just one comparison and use a significance threshold of .05. The likelihood of finding statistically significant evidence of a relationship, even if none exists, is .05. But if we run the study twice, using independent data each time, the probability of finding statistically significant evidence in *both* studies if no real relationship exists is $.05 \times .05 = .0025$. If we do it a third time, the probability we find evidence for the nonexistent relationship in all three studies is $.05^3 = .000125$, pretty unlikely. By replicating, we reduce the chance that we reach a spurious positive conclusion. This is especially true if the replication is done by independent teams who don't have a vested interest in validating the initial finding.

Of course, replication is not a panacea, and we have to keep thinking clearly. Failure to reject the null hypothesis is not proof that the null is true. So, if we only believe results if they replicate multiple times, we might sometimes wrongly reject real effects, especially if we conduct replications on sparse or noisy data, where effects are hard to detect. Ideally, we'd get lots of data and replicate on large samples.

Sometimes such replication is feasible and sometimes it's not. If researchers conducting a drug trial for a cholesterol drug happen to collect weight data and find an unexpected effect of the drug on weight loss, researchers can recruit a new pool of subjects and see if the new treatment group shows similar weight loss relative to the new control group. But if we discover a phenomenon regarding twentieth-century gubernatorial elections, the behavior of the moons of Venus, or leadership strategies in the run-up to World War I, there's just one sample of them. There's no way to go collect more data. In that case, we may want to think less literally and more conceptually about replication. We do this not by directly replicating the existing findings but by asking something like, "If this phenomenon is genuine, what are some other hypotheses that should also be true?" Let's see an example.

Football and elections

Anthony has a paper with Pablo Montagnes that illustrates this approach. The paper re-examines a prominent study by Andrew Healy, Neil Malhotra, and Cecilia Mo, published in the *Proceedings of the National Academy of Sciences*, that claims that the outcome of college football games affects who wins elections. Specifically, the incumbent party reportedly performs better in congressional and gubernatorial elections in the home counties of teams that won prior to the election. This kind of finding makes some people worry about democracy. (Not us, but that is a topic for another day.)

Anyway, the Healy et al. study is in many ways good. Football wins and losses seem pretty random, so there isn't lots of reason to be concerned about bias. But this is exactly the kind of setting where you might worry about a false positive resulting from over-comparing and under-reporting. For instance, there is almost certainly a *p*-screening problem—would a prestigious scientific journal have published a paper showing that college football games do *not* appear to have any influence on elections? Moreover, there are lots of sports that might have been used to predict incumbent success: other research teams, studying the effects of basketball or curling losses on elections, may have found no relationship and, thus, not published papers. So we shouldn't leap to conclusions just because one published paper presents evidence supporting the claim that losses in one particular sport are associated with election outcomes.

Anthony and Pablo couldn't conduct a purely independent replication because there's no way to re-run decades of college football games and elections. Instead, they thought about independent theoretical predictions—additional hypotheses that seem like they should also hold if football losses really do affect elections. For instance, if football losses affect elections, you might expect the relationship to be particularly strong in places that care a lot about college football. If voters blame incumbent politicians for bad football outcomes, you might expect the impact of a football loss on the incumbent's party to be bigger when the incumbent is actually seeking reelection relative to when some new candidate from the same party is running. And so on. Testing such hypotheses, which speak to the underlying mechanism, is a way to probe whether some estimated effect is likely to reflect a real relationship in the world or is the result of noise (i.e., a false positive).

Here are some examples of what Anthony and Pablo found. It turns out that the estimated effect of football games is smaller in counties where more people follow college football than in counties where fewer people follow college football, no greater when an incumbent actually runs for reelection as opposed to just the incumbent's party running a candidate, and just as strong outside the home county of the team as inside the home county of the team. Furthermore, they found no evidence of a relationship between football losses and electoral outcomes for NFL games, despite the fact that NFL teams have the same kind of regional support as college football teams and NFL games are roughly ten times more popular than college football games.

Anthony and Pablo tested multiple, independent theoretical predictions that seemed like they should have held if the relationship between football and elections was real, but none of them received support in the data. From this, they concluded that it seems unlikely that college football games really do influence elections. This isn't a classic replication. But it shows how looking at the evidence for additional hypotheses that are related to the mechanism underlying the original hypothesis can help shed light on the strength of the evidence for surprising results.

Related to the idea of independent replication is the use of hold-out samples, which we discussed in chapter 5 when we talked about overfitting. Suppose you have a large sample of data and want to explore it for relationships. It might be a smart idea to hold out a randomly selected subset of that data from the exploration. For example, you could randomly select half the observations and use them as an exploratory data set. Then, after you've found a few interesting relationships, you could test whether those relationships also appear in the hold-out sample of data that you have not yet analyzed. If over-comparing produced a false positive in your initial analysis, we would expect that the same relationship is unlikely to appear in the hold-out sample. But if you've identified a real phenomenon, then we should expect it to persist in the hold-out data.

Test Important and Plausible Hypotheses

If you read a study that would have never been published had the researchers found the opposite of what they found (e.g., failed to reject the null), you should be particularly worried about over-comparing and under-reporting. But if a study answers a question for which we care about the answer intrinsically, regardless of what that answer turns out to be, a lot of the problems of over-comparing and under-reporting disappear. In particular, if the findings can be published regardless of the result, we can worry less about *p*-screening and we might think the researcher has less incentive to engage in *p*-hacking.

Happily, many important scientific questions fall into this latter category. If a study is testing a serious theoretical hypothesis, exploring a medical treatment that there are good reasons to think might work, or evaluating a real policy intervention, the answer is interesting, whatever it turns out to be.

By contrast, a lot of fun-sounding questions with surprising answers fall into the former category. And, unfortunately, such studies appear irresistable to much of the science press. Think of the ESP study. No one would be interested in a paper that found no evidence for ESP. So there were concerns both about incentives to *p*-hack and about the file drawer problem. An example will illustrate the point.

The power pose

A famous study by Amy Cuddy, Dana Carney, and Andy Yap purportedly shows the remarkable efficacy of adopting a simple power pose. Our attitudes, the argument goes, often cause our behaviors, rather than the other way around. And small changes to the way you hold yourself can change your attitude. In particular, by standing in a posture that you associate with being powerful, you will inspire feelings of assertiveness and will then behave accordingly.

Though the underlying science is strongly disputed, its promoters continue to argue that striking the right pose causes people to experience feelings of power and leads to physiological changes, including increased testosterone and cortisol levels. There were no pre-existing good reasons to think this might be true. And it is hard to imagine a major journal publishing a study showing that adopting a power pose had no effect on anything. So readers should have been skeptical from the outset. Nonetheless, because the findings were fun, surprising, and optimistic, the study got enormous attention. It was published in a prestigious scientific journal and written up in major media outlets, and Cuddy was invited to give what turned out to be a wildly popular TED Talk.

Not surprisingly, the result turns out to be wrong. Multiple attempts at replication fail to find similar effects. And one of the coauthors, Dana Carney, eventually went so far as to disavow the work, documenting the many ways in which the finding was the result of p-hacking.

Beyond Science

We have focused on the ways in which over-comparing and under-reporting create deep challenges for the scientific community. But, as the story of Paul the Octopus illustrates, the problem is broader than that. Indeed, we suspect you run into it on a regular basis, often without noticing.

Suppose someone is trying to sell you something—perhaps a car, financial advice, or a subscription to a dating app. The salesperson might tell you, "This car was rated number one in customer satisfaction five years after purchase!" Sounds great. But you might want to ask yourself how many measures of satisfaction they looked at before finding the one on which this car was rated number one. Did they look at reliability, repair record, safety, longevity, gas mileage, and resale value in addition to customer satisfaction? Did they also look one year from purchase, two years from purchase, three years from purchase, and so on? If so, they made a lot of comparisons and told you about the one that puts the car in the best possible light. This is not an unbiased estimate of the car's quality; it reflects the salesperson's equivalent of p-hacking.

Similarly, your financial advisor might tell you, "This mutual fund outperformed the S&P 500 for seven of the last eight years." That sounds good. But how did it do relative to the Dow Jones or a broad market index? How did it do over the last nine years, ten years, fifteen years? Did the advisor choose to compare to the S&P over the last eight years because that was the natural comparison or because it was the comparison that made the mutual fund look best?

In general, then, you need to think clearly about the problems of over-comparing and under-reporting not just when formally thinking about hypothesis testing and statistical significance. Whenever you are offered a piece of evidence, you should be asking yourself whether this particular comparison is the natural one or the first one you would have thought to look at. If not, you might pause to contemplate how many plausible comparisons there are and, thus, how difficult it would have been to come up with some comparison that made whatever point the speaker was trying to make.

In the spirit of appreciating the fact that this problem is, indeed, everywhere, let us leave you with one final story that illustrates yet another way in which over-comparing and under-reporting frequently rears its head—identifying superstars.

Superstars

We like to admire and study people who are really successful. We've already seen one reason why this can produce misleading inferences: correlation requires variation. Another reason that we shouldn't be so quick to admire and study superstars is that there may not be anything particularly special about them beyond good luck.

Bill Miller majored in economics in college, served as a military intelligence officer, dabbled in a doctoral program in philosophy, and worked as treasurer for a steel and cement company before taking a position at Legg Mason Capital Management as director of research in 1981, at the age of thirty-one. Miller was clearly a smart guy with a promising career ahead of him. The next year, he began running the Legg Mason Value Trust mutual fund. For the first decade or so, the fund's performance was mediocre, slightly underperforming the market average. But Miller eventually hit his stride, scoring some big returns in the late 1990s and early 2000s. By 2006, fellow investors and reporters noticed that Legg Mason Value Trust had outperformed the market for fifteen years running, an unprecedented streak that launched Bill Miller into the upper echelons of finance stardom.

Inevitably, everyone wanted to know Miller's secret. What made him such a successful investor? Perhaps surprisingly, Miller didn't achieve his success by developing intimate knowledge of niche industries or technical trading algorithms. His fund primarily invested in a small number of already well-known companies like Google, Amazon, eBay, J.P. Morgan, and Aetna. When describing his investment philosophy in a 2006 letter to investors, Miller reported that he simply looks for the "best value." He further speculated on what separates his fund from so many competitors: "We differ from many value investors in being willing to analyze stocks that look expensive to see if they really are. Most, in fact, are, but some are not."

Miller makes it sound easy. He just invests in companies that are undervalued. But before we conclude that Bill Miller is a genius investor, let's consider the possibility that Miller was simply lucky, like Paul the Octopus.

There is an idea in finance called the efficient-market hypothesis. It more or less says that no fund or investment strategy should be able to systematically beat the market average over the long run. Loosely, the logic goes like this. If some genius investor came up with an investment strategy that predictably beat the market, other investors would mimic that strategy. This would change the prices of the assets traded under that strategy. Investors would keep doing this until that strategy no longer beat the market. For instance, if a company's stock price fully reflects all available information about the value of that company, which we'd expect in a large market with lots of people trading on the best available information, there should be no way to systematically predict whether the stock price will go up or down without insider knowledge.

If the efficient-market hypothesis is right, then Miller and the other fund managers and stock pickers are just doing the equivalent of flipping coins. And we know that if enough people flip coins, a few of them will flip a long string of heads by sheer luck. So, to assess whether Miller is indeed a genius, we need to ask how likely it is that he just happens to be the one guy who hit a long string of heads by luck.

To get started, let's imagine that beating the market is really just like flipping heads with a fair coin. This is our null hypothesis. Then we want to ask, If our null hypothesis is true, how likely is it that someone would flip 15 heads in a row?

The chances that a given investor in a given 15-year period beats the market every single year by chance is really low. The probability that some investor beats the market by luck in one year is $\frac{1}{2}$. The probability that an investor beats the market by luck two years in a row is $\frac{1}{2} \times \frac{1}{2}$. Extending this logic, the probability that a given investor beats the market 15 years in a row by sheer luck is $\frac{1}{2}^{15}$, or about 1 in 30,000. So maybe Miller is a genius; if he was just flipping coins, there's only a 1 in 30,000 chance that he would be so successful. But maybe not. Let's make sure we are thinking clearly about a few things.

There are lots of investors out there, and if any one of them beat the market 15 years in a row, they would have been just as famous as Miller, and we'd be discussing them instead of him. Therefore, the relevant question is not how likely it is that one particular fund manager, Bill Miller, would beat the market 15 years in a row by chance. The relevant question is how likely it is that some fund manger would beat the market 15 years in a row by chance.

Notice, this is just like publication bias or the problem of Paul the Octopus. With publication bias, we only hear about the few studies out of many with statistically significant results. With Paul, we only heard about the one animal out of many who correctly forecast a lot of soccer games. And, similarly, we only hear about the few investors who have really long winning streaks. In all three cases, if we only think about the studies, animals, or investors we get to hear about, we over-estimate how likely it is that their success reflects a real phenomenon in the world.

In any given year, there are at least 24,000 professional funds trading, and presumably each will continue trading if it beats the market. So let's assume (as our null hypothesis) that there are 24,000 fund managers, none with any special insight. That means that whether each of them beats the market in any given year is a 50-50 proposition. So figuring out how likely it is that one of them beats the market 15 years in a row is just like figuring out how likely someone is to flip 15 heads in a row if 24,000 people each flips 15 fair coins.

Doing the same kind of calculation we did for Paul the Octopus, the answer is about .52 or 1 in 2.[2] It is very unlikely that any particular investor will beat the market 15 years in a row by sheer luck. But when you consider the thousands of investors out there, it's actually more likely than not that one of them will beat the market 15 years in a row, even if none of them has any special insight and they're all just flipping coins.

These calculations look even worse for Miller if we consider that a 15-year streak would have seemed just as impressive had it started in another year. Once we consider all the funds and all the possible 15-year periods, it seems extremely likely that some fund manager would have such a streak at some point just by chance. These calculations, combined with our knowledge of the efficient-market hypothesis, should make us skeptical of anyone who claims to know the secrets to beating the market. The sheer

[2] The probability that one investor gets 15 years in a row right is $.5^{15}$. So the probability that one investor doesn't gets 15 years in a row right is $1 - .5^{15}$. So the probability that none of 24,000 gets 15 years in a row right is $(1 - .5^{15})^{24,000}$. So the probability that at least one investor does get 15 years in a row right is $1 - \left((1 - .5^{15})^{24,000}\right)$, which is approximately .52.

number of traders and funds means that there are bound to be some exceptionally good track records. And those are the ones we hear about. So, before handing over your life savings to an investment manager, you should ask whether you would invest the same money betting on Paul the Octopus's soccer picks. If not, let us recommend that you consider low-cost index funds.

What do you think happened to Bill Miller after his flurry of press coverage in the mid-2000s? The streak ended in 2006. His fund lost 55 percent of its value during the 2008 financial crisis, the fund continued to trail the market for several more years, and he eventually stepped down from his post in 2012. Looking across the full thirty-year period in which Miller managed Legg Mason Value Trust, it actually *underperformed* the market. Alas, his historic winning streak still earns him regular appearances on cable news programs, where he pontificates on market conditions and hot stock picks. In 2017, his new fund, Miller Opportunity Trust, was once again making news for impressive returns in 2017. The secret? A big bet on Apple, the most highly valued company in the world.

Wrapping Up

Over-comparing and under-reporting can happen because of nefarious researcher behavior (p-hacking) or in a community of entirely honest researchers (p-screening). In either case, it results in publication bias—the phenomenon whereby published results are systematically misleading because there is a bias toward publishing statistically significant findings. There is no simple solution to the problem of over-comparing and under-reporting. But once we learn to think clearly about it, we can get a better sense for when it is likely to be occurring and come up with some practices that at least mitigate the problem.

In chapter 8, we turn to another challenge created by the presence of noise: reversion to the mean. Once we learn to think clearly about reversion to the mean, we will see that it, in combination with over-comparing and under-reporting, helps to explain what appears to be a truly puzzling phenomenon—the tendency of scientific estimates to shrink over time.

Key Terms

- **Publication bias:** The phenomenon whereby published results are systematically over-estimates because there is a bias toward publishing statistically significant results.
- ***p*-hacking:** Searching over lots of different ways to run an experiment, make a comparison, or specify a statistical model until you find one that yields a statistically significant result and then only reporting that one.
- ***p*-screening:** A social process whereby a community of researchers, through its publication standards, screens out studies with p-values above some threshold, giving rise to publication bias.

Exercises

7.1 Briefly return to the question from chapter 6 about Pete's roulette wheel. Would you revise any of your advice or conclusions in light of the lessons from this chapter?

7.2 In late April 2020, the National Institutes of Health announced the results of a study on the use of the drug remdesivir to treat COVID-19. Some COVID-19 patients were randomly given remdesivir; others were given a placebo. The study found statistically significant evidence that treatment with remdesivir reduced recovery time, as measured by the number of days it took for a patient to be discharged from the hospital after being put on the drug. The study was double blind (neither the patients nor doctors knew whether a subject had been put on the real drug or the placebo). The study size was reasonably large (hundreds of patients). And treatment assignment was random.

(a) On the basis of the lessons of this chapter, identify two more pieces of information that would help you assess how confident you should be in the efficacy of remdesivir.
(b) It turns out the study was pre-registered. The pre-registration plan identified twenty-eight outcomes that the scientists were going to measure. How does this change your beliefs about whether the findings identify a real effect in the world? Why?
(c) The pre-registration plan also identified one outcome as the primary outcome of interest. That primary outcome of interest was the one reported: how long it took for a patient to be discharged from the hospital. Does this affect your answer to the previous question? Why or why not?
(d) But wait, there is one final twist. The pre-registration plan was actually revised during the course of the study. It turns out that the length of hospitalization was not listed as the main outcome of interest until a revision on April 16, 2020. Prior to that, the primary outcome of interest was listed as a patient's score on an eight-point scale measuring disease severity. This is reflected in the April 2, 2020, version of the plan. The researchers, in a statement, explained that they had not seen the data coming out of the study prior to changing their primary outcome of interest. Reflect on how all of this affects your views on the study's findings.

7.3 Download "VoterSurveyData2016.csv" and the associated "README.txt," which describes the variables in this data set, at press.princeton.edu/thinking-clearly. Suppose we want to know whether prior exposure to Donald Trump before he was a politician affected political behavior in the 2016 U.S. presidential election. To proxy for exposure to Trump, a survey asked people whether they watched *The Apprentice*, a television show starring Trump, and whether they watched *Home Alone 2*, a movie featuring a cameo by Trump.

(a) Using the data available, try to find at least three interesting, statistically significant relationships suggesting that prior exposure to Trump corresponded to political behaviors in the 2016 presidential election.

If you're struggling to find statistically significant relationships, think about all the different things you can test for. You could use having seen *The Apprentice*, *Home Alone 2*, or both as your measure of prior Trump exposure. You can use support for Trump, support for Hillary Clinton, or voter turnout in 2016 as your outcome of interest. You can subset the data to look specifically at voter subgroups of interest (e.g., women, Blacks, Southerners, rich, young, and so on).

(b) Once you've found three statistically significant relationships, interpret them substantively and think about what they mean. Did you learn something interesting about American electoral behavior?

(c) The data you just analyzed is real survey data from the 2016 Cooperative Congressional Election Study. We randomly selected one thousand respondents and shared a subset of their responses with you. However, we lied above when we said that respondents were asked whether they watched *The Apprentice* or *Home Alone 2*. We made those variables up. (Sorry, we won't lie to you again.) Furthermore, the values for those variables were generated completely at random. Explain why you were nonetheless able to find a relationship between those variables and political behavior. Would you expect that relationship to continue to hold if we provided data on another thousand respondents and again randomly generated the exposure data?

7.4 Find a recently published academic study for which you are worried about the problems of over-comparing and under-reporting. Explain your concerns. Without collecting additional data, is there anything the authors could do to address or mitigate your concerns? Is there additional information you'd like the authors to disclose? Are there additional analyses you'd like them to conduct?

Readings and References

For more on *p*-hacking, see

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–66.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. 2014. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology* 143(2):534–47.

To see the ESP study and to read about some of the failed replications, see

Daryl J. Bem. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influence on Cognition and Affect." *Journal of Personality and Social Psychology* 100(3):407–25.

Jeff Galak, Robyn A. LeBoeuf, Leif D. Nelson, and Joseph P. Simmons. 2012. "Correcting the Past: Failures to Replicate Psi." *Journal of Personality and Social Psychology* 130(6):933–48.

For more on *p*-screening in the context of get-out-the-vote studies, see

Donald P. Green, Mary C. McGrath, and Peter M. Aronow. 2013. "Field Experiments and the Study of Voter Turnout." *Journal of Elections, Public Opinion and Parties* 23(1): 27–48.

The essay by seventy-two researchers on lowering our threshold for statistical significance is

Benjamin, Daniel J., et al. 2017. "Redefine Statistical Significance." *Nature Human Behavior* 2:6–10.

The study on the frequency of null results in NHLBI studies before and after preregistration is

Robert M. Kaplan and Veronica L. Irvin. 2015. "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time." *PLoS One* 10(8).

To read more about whether college football outcomes influence elections, see

Andrew J. Healy, Neil Malhotra, and Cecilia Hyunjung Mo. 2010. "Irrelevant Events Affect Voters' Evaluations of Government Performance." *Proceedings of the National Academy of Sciences* 107(29):12804–09.

Anthony Fowler and B. Pablo Montagnes. 2015. "College Football, Elections, and False-Positive Results in Observational Research." *Proceedings of the National Academy of Sciences* 112(45):13800–04.

There is a good discussion of the whole power pose episode on Andrew Gelman's blog, which frequently covers issues related to publication bias: https://statmodeling.stat.columbia.edu/2017/10/18/beyond-power-pose-using-replication-failures-better-understanding-data-collection-analysis-better-science/. The original study on power poses and the first replication attempt are

Dana R. Carney, Amy J. C. Cuddy, and Andy J. Yap. 2010. "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance." *Psychological Science* 21(10):1363–68.

Eva Ranehill, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. 2015. "Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women." *Psychological Science* 26(5):653–56.

Carney's disavowal can be found at http://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf.

If you would like to see the full history of revisions of the pre-analysis plan for the remdesivir study mentioned in exercise 2, you can find it at https://clinicaltrials.gov/ct2/history/NCT04280705.