CHAPTER 8

Reversion to the Mean

What You'll Learn

- Lots of things tend to revert toward the mean, meaning that extreme observations are often followed by less extreme observations.
- This phenomenon will arise for virtually any outcome that is a function of both signal (i.e., something real in the world) and noise.
- If you don't think clearly about reversion to the mean, it is easy to misinterpret evidence.
- We shouldn't expect reversion to the mean for things that reflect our beliefs about the future, like election projections or stock prices.

Introduction

As emphasized by our favorite equation, the world is noisy, and most quantitative measurements reflect both the thing we meant to measure and noise. This has lots of implications for the ways that we study and understand the world.

One of the most common yet least understood consequences of living in a noisy world is *reversion to the mean*. Loosely speaking, unusually large or small measurements tend to be followed (and preceded) by measurements that are closer to the mean.

Although reversion to the mean is not a standard topic in books on quantitative reasoning and data analysis, its pervasiveness means that you will often be misled by quantitative information if you don't understand this phenomenon. So we think it is important to devote some time to it.

Does the Truth Wear Off?

In chapter 7 we saw that, because of over-comparing and under-reporting, we should often be skeptical of new, surprising scientific findings. And so, when such findings are initially reported, often the first question that gets asked is "Does the result replicate?" That is, if we were to run a new, independent study designed similarly to the original study, would we find a similar effect? The instinct to ask this question is solid: we are worried that some reported results reflect the vagaries of chance, rather than real phenomena in the world. So, before hanging our hats on new findings, we want to see that they hold up in multiple studies. And, in fact, researchers often fail to replicate

hyped results in follow-up studies. Indeed, in some fields of study this is so common that people have started talking about a *replication crisis* undermining confidence in entire bodies of scientific enquiry.

Jonathan Schooler, a prominent psychologist at UC Santa Barbara, famously noticed such a pattern of replication failures in some of his own most influential studies. Interestingly, the effects Schooler estimated typically didn't disappear entirely when replicated, but they did get systematically smaller. He asked around and found that many colleagues had experienced the same thing; replicated results were often smaller than the original findings.

One potential explanation for this pattern is that once you've done a study and subjects are aware of the results, they change their behavior. This phenomenon, whereby subjects alter their behavior because they know they're being studied, is sometimes called the *Hawthorne effect*.[1] Another term—*demand effects*—refers to situations where the subjects behave differently because they know what the experimenters are looking for and are trying to please them.

Schooler and his colleagues quickly ruled out the Hawthorne effect and other similar explanations because they found the same pattern in studies of birds, who presumably have no idea what's being studied and don't care one way or the other about pleasing human researchers. So what else could explain the peculiar pattern of disappearing effects?

Schooler (perhaps jokingly) started referring to this phenomenon as *cosmic habituation*. He wondered whether there is some unknown force in the universe that causes effects to shrink every time they're studied. One analogy he gives is to the habituation of human sensory perception. When something first touches your arm, you are acutely aware of it. However, over time, you habituate, and your sensation of being touched diminishes. Maybe the universe is like that. The first time we observe some phenomenon, there is an acute effect. But, over time, the universe habituates to our studies and we observe the effect less and less. In other words, scientists actually alter reality every time they study it. Spooky.

Schooler's theory of cosmic habituation has received significant media attention, including an episode of the popular radio show and podcast *Radio Lab* and an article in the *New Yorker* entitled "The Truth Wears Off." But before we follow Schooler down the path of hypothesizing new cosmic forces, let's see if thinking a little more clearly can help us resolve the puzzle of shrinking effect sizes a bit less mystically.

Francis Galton and Regression to Mediocrity

As we described back in chapter 5, Francis Galton made a similarly eerie discovery in the 1860s. He collected data on the size of parents and their children. He did this for human height. He also did it for plants, collecting data on the size and weight of the seeds of parent and child sweet-peas.

Galton drew scatter plots of these kinds of data, putting the parents' size on the horizontal axis and the children's on the vertical axis. Then he plotted a regression line

[1] Fun fact: The term *Hawthorne effect* comes from a study of the relationship between working conditions and productivity at the Hawthorne Works factory outside Chicago. But it turns out that the data were analyzed badly. The economists Steven Levitt and John List reanalyzed the original data and showed that what looked to the original researchers like a Hawthorne effect was more likely attributable to other factors, such as differences across days of the week, rather than the subjects changing their behavior in response to being studied.

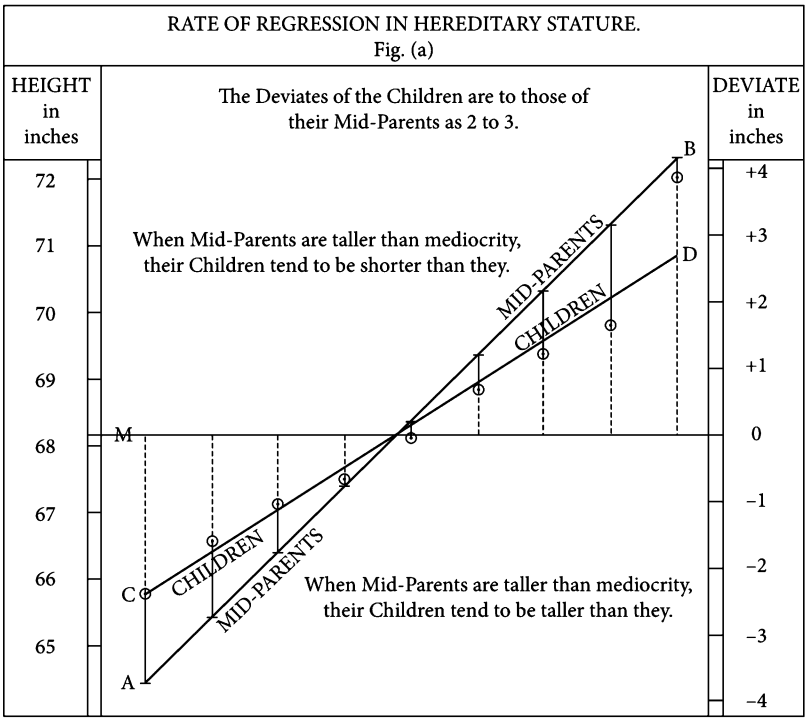


Figure 8.1. A reproduction of Galton's illustration of reversion to the mean.

through the data. You can see one of Galton's plots, for the heights of parents (adjusting for biological sex) and their children, in figure 8.1.

Originally, Galton expected the regression line would be a 45-degree line—that is, its intercept would be 0 and its slope 1. That seems a reasonable guess. It would be true if, on average, children are the same size as their parents (again, adjusting for biological sex).

As it turns out, however, this guess is not correct for humans or for sweet-peas. Let's look at Galton's figure. In that figure, the 45-degree line is labeled "Mid-Parents." The phrase refers to Galton's measure of the average height of a child's parents, after first adjusting to put female and male heights on the same scale. The (unlabeled) horizontal axis corresponds to parents' height. Then, the line tracing out parents' height must be the 45-degree line. The line labeled "Children" is the regression line running through the data that takes as its x-value the parents' height and as its y-value the children's height.

As you can see in the figure, Galton's regression shows a positive y-intercept—at the lowest value on the horizontal axis (parents' height), the regression line lies above the 45-degree line. And Galton's regression line has a slope that is positive but distinctly less than 1—the regression line is increasing more slowly than the 45-degree line.

What does this regression imply about the relationship between parents' height and children's height? The fact that the slope is positive means that, on average, the taller the parents, the taller their children. As you can see in the figure, the fact that the y-intercept is positive means that particularly short parents tend to have children who are taller than they are. On the left-hand side of the horizontal axis (where parents are short),

the regression line lies above the 45-degree line. But because its slope is less than 1, the regression line increases less quickly than the 45-degree line. And, indeed, the two lines cross in the middle. As such, on the right-hand side of the horizontal axis (where parents are tall), the regression line lies below the 45-degree line. Particularly tall parents tend to have children who are shorter than they are.

As we mentioned in chapter 5, Galton referred to this phenomenon as "regression to mediocrity." And it is because of this word choice that we now use the word *regression* to refer to the practice of fitting lines to data. It is also why some people refer to the phenomenon whereby things tend to revert toward the mean (which is the subject of this chapter) as *regression to the mean*. However, to avoid confusing the two concepts, we will refer to the latter by its other common name, *reversion to the mean* also sometimes referred to as *mean reversion*.

Galton's findings sound a lot like cosmic habituation. Perhaps there is some unseen force in the universe that pushes size within families toward the average. Maybe when the universe sees tall parents or big sweet-pea seeds, it restores order by making their offspring smaller. Or perhaps when Galton measured the heights of the parents and the diameter of sweet-pea seeds, he somehow made their offspring shrink! Galton was probably perplexed by his unexpected findings. But he wasn't so quick to jump to supernatural conclusions.

Eventually, Galton realized what was going on. Size is determined by many factors. For the purposes of seeing Galton's idea as clearly and simply as possible, let's think about sweet-pea seeds. And let's imagine that a sweet-pea seed's size is influenced by just two things: (1) the genes it inherits from its parent and (2) the amount of direct sunlight it gets while growing. Inheriting genes for largeness from its mother make a seed larger, all else equal. And getting more sunlight make a seed larger, all else equal.

Under this simple model, let's think about how a seed can end up especially large or especially small. Suppose you find a really large sweet-pea seed. It could be large because it got genes for really large size from its mother. It could also be large because it happened to grow in a year with uncommonly good sun. Or it could be some combination of the two. Odds are, if a seed is *really* large, it had both factors working in its favor: a parent with genes for large size and excellent sun.

So what should we expect if these really big seeds are planted and produce offspring of their own? They'll pass along their genes for larger-than-average size. But, most likely, the child seed won't experience the same outstanding sunlight as its parent did. On average, it will grow in average sun. So the child will be larger than average because of the genes it inherited. But the child will probably be smaller than its parent because its parent got particularly lucky with respect to sun exposure. The same holds for the children of very small seeds. They get their parents' genes for small size. But they are likely to experience better sunlight than their parents did and, thus, be larger than their parents.

So, if this simple model were correct, we would observe exactly Galton's pattern. Larger plants tend to have larger children (the regression line has a positive slope). But size reverts toward the mean—really small parents tend to have children who are smaller than average but larger than they are, and really large parents tend to have children who are larger than average but smaller than they are (the slope of the regression line is less than 1).

Obviously, seed size is more complicated than this. Many things influence it other than genes and sun. But the example makes the point. We should expect reversion to

the mean if size is partly determined by genes that are systematically transmitted from parent to child and partly determined by idiosyncratic or random factors that are uncorrelated across generations (like sun exposure). The same goes for human height, as we see in Galton's plot.

More generally, there will be reversion to the mean for any outcome that's partly a function of systematic factors (which we sometimes call *signal*) and partly a function of random or idiosyncratic factors (which we sometimes call *noise*). Imagine an outcome observed over and over, where with each observation, the outcome reflects a combination of a systematic signal (e.g., the genes) and random noise (e.g., sunlight). Extreme outcomes typically arise because of extreme values of both the signal and the noise. In other iterations, while the signal stays fixed, the noise takes a new, random value. And, in expectation, the value of the noise will be average. So extreme values in one iteration are expected to revert toward the mean in other iterations.

Many phenomena in the world have this signal and noise structure. So we should expect reversion to the mean to pop up a lot. Therefore, thinking clearly about evidence requires anticipating and accounting for reversion to the mean. In what follows, we first delve a little more deeply into the nature of reversion to the mean to make sure we are clear about exactly what is going on. We then consider a variety of different real-world settings to understand when we should and should not expect reversion to the mean to appear.

Reversion to the Mean Is Not a Gravitational Force

One common misconception about reversion to the mean is that it reflects something like a gravitational pull—that is, that the world is full of outliers and that ineluctably, over time, things are being pulled toward the mean. This isn't right.

To see if you are thinking clearly about reversion to the mean, try answering each of the following questions:

1. John Junior is exceptionally tall. If you had to guess, would you guess that John Junior's son, John III, is
   (a) shorter than John Junior?
   (b) the same height as John Junior?
   (c) taller than John Junior?

2. John Junior is exceptionally tall. If you had to guess, would you guess that John Junior's father, John Senior, is
   (a) shorter than John Junior?
   (b) the same height as John Junior?
   (c) taller than John Junior?

Before we turn to the answers, let's start by noting that you should have given the same answer to both questions. Understanding why is essential.

For many people, once they learn about reversion to the mean, question 1 is pretty intuitive. John III is probably shorter than John Junior. John Junior is particularly tall. So he probably has genes for tall height (signal). And he also probably had idiosyncratic things happen that led him to grow particularly tall (noise). His son, John III, will likely inherit his genes for tallness. But, if you had to guess, you'd guess the idiosyncratic other factors will likely be more average. This is the logic of reversion to the mean as explained by Galton.

But, in our experience, question 2 is often a bit more vexing. You may be inclined to reason as follows. John Junior is really tall. And there is reversion to the mean in the world. So John Junior's height is probably closer to the mean than was his father's.

Factoring in reversion to the mean, for John Junior to be so tall, his father must have been a virtual giant! Hence, one might reason, while the answer to question 1 is (a), the answer to question 2 must be (c).

It's okay if you thought something along those lines. The argument has a certain appeal. But it is wrong, and it's important that you see why. The answer to both questions 1 and 2 is (a). And the logic for John Junior's father is identical to the logic for John Junior's son: the logic of reversion to the mean. Here's how it goes, one more time.

Suppose you observe some outcome made up of signal and noise and that outcome is surprisingly large. (The argument, of course, works for surprisingly small too.) Then suppose you are going to observe another outcome, where the signal is the same as your first observation, but there will be a new, independent draw of the noise. Since the first observation is so large, it probably reflects a large value of the signal and a large value of the noise. The new observation again has a large signal value, but the value of the noise is likely to be smaller. So that new observation will likely be smaller.

Importantly, in making this argument, we said nothing about which outcome was determined first in time. We just talked about the order in which you observed them. Reversion to the mean is not a gravitational force pulling things toward the average over time. For the logic of reversion to the mean, it makes no difference which came first temporally. So, if John Junior is very tall, and his son has the same signal (genes) but independent noise, then his son is probably shorter than him. And, if John Junior is very tall, and his father has the same signal (genes) but independent noise, then his father is also probably shorter than him.

The easiest place to see this in the real world is in data from athletic competitions, where we observe the same competitor doing the same task over and over again. And, what we see in those settings is that reversion to the mean characterizes the data, moving forward and backward in time.

Figure 8.2 is a scatter plot of scores from the first two rounds of the 2019 U.S. Women's Open golf tournament. Players' scores from round 1 are on the horizontal axis and from round 2, on the vertical axis. What do you see here?

On average, there is a positive correlation between scores in the two rounds; the regression line is sloping upward. The players who did better (in golf, lower scores are better) in round 1 also tended to do better in round 2. That makes sense: some players are better than others (that's the signal). But the slope of the regression line is far less than 1; the regression line is shallower than the 45-degree line. If a player's round 1 score was worse than average, their round 2 score tended to be better than their round 1 score. And if a player's round 1 score was better than average, their round 2 score tended to be worse than their round 1 score. It's exactly the same as the pattern with the heights of parents and children or the size of mother sweet-peas and their offspring discovered by Galton. And we can assure you that we didn't cherry pick this example. It's a virtual guarantee that you'll see this same pattern for any golf tournament.

A golf commentator might look at this data and tell a story to explain the scores. Maybe the players who had a really good first round succumbed to the pressure. They choked. And that explains why they did worse in round 2. And maybe the players who had a bad first round realized they had to put in more practice, change strategy, or really focus. And they accordingly improved their scores.

That's possible. But it could also just be reversion to the mean. Golf scores are a function of both skill (signal) and luck (noise). The players with the best score in a given round are probably better than the average player in the field. But they probably also had some luck on their side. A few putts went in that could have just as easily lipped

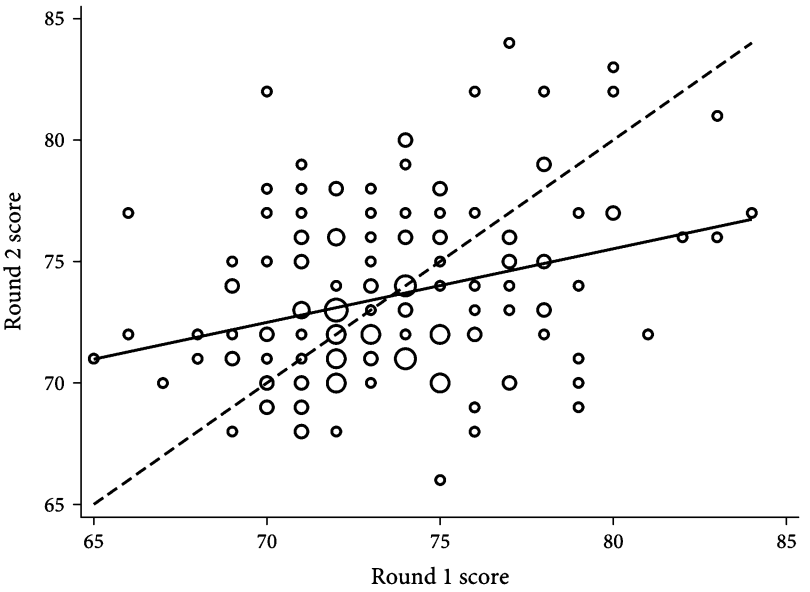


Figure 8.2. Scores across rounds in the 2019 U.S. Women's Open golf tournament. In cases where multiple players had identical scores, the size of the circle is increased to reflect the number of players. The 45-degree line is dashed, and the actual OLS regression line is solid.

out. Their few bad shots got fortunate bounces, keeping them out of trouble. And so on. In their other rounds, they're still better-than-average players, so we expect their scores to be better than average. But they probably won't have the same good luck, so their score will probably be worse than in that exceptional round.

Notice, nothing in the previous paragraph's logic depended on which round came first. That's because reversion to the mean isn't some gravitational force pulling things toward the mean over time. This realization allows us to probe which story—the commentator's or reversion to the mean—is more likely to be true.

Think about a player who has a particularly good score in round 2. Should we expect their round 1 score to be better or worse than their round 2 score? The old temptation was to think that because of reversion to the mean, in order for them to have a good score in round 2, they must have had a really good score in round 1, allowing them to still have a good round 2 score despite reversion to the mean. But we now know better.

The logic of reversion to the mean has nothing to do with time. The score in each round of golf is a combination of signal and noise. If a player had a particularly good round at some point, we should expect a different round by that player (with the same signal but different noise) to be worse, regardless of which round came first. And if a player had a particularly bad round at some point, we should expect a different round by that player to be better, regardless of which round came first.

Figure 8.3 shows the same graph you saw before, but with the axes flipped so that round 2 is on the horizontal axis and round 1 is on the vertical axis. The overall pattern is almost exactly the same. People who had particularly good scores in round 2 were better than average in round 1 but still worse in round 1 than in round 2. Just like with John Junior and John Senior, reversion to the mean works backward in time just as well as it works forward in time.

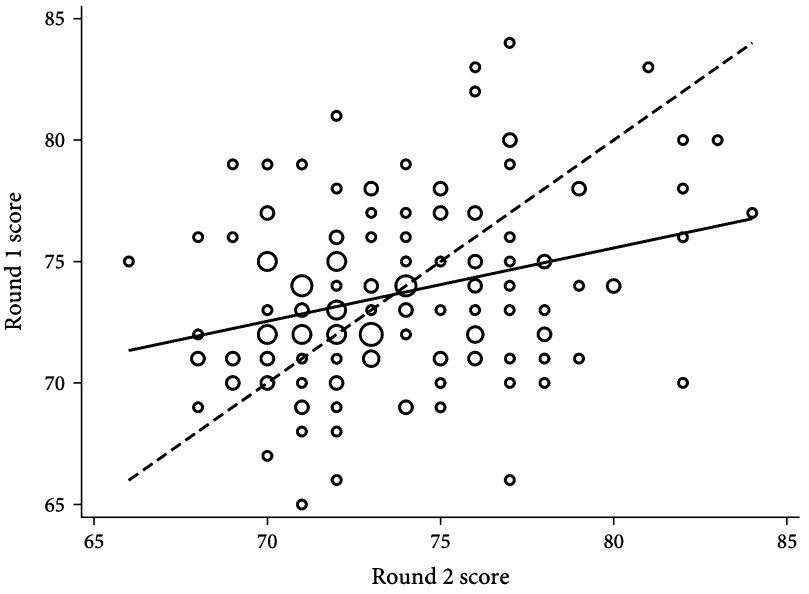


Figure 8.3. Scores across rounds in the 2019 U.S. Women's Open golf tournament, with flipped axes.

Seeing both versions of the graph demonstrates why the commentator's explanation is unlikely to be right. It is hard to believe that a good score in round 2 causes players to feel extra pressure, which somehow goes backward in time, making their round 1 score worse. And yet, we see the same pattern whether we look forward or backward in time. The explanation is reversion to the mean.

Seeking Help

Reversion to the mean can be particularly problematic for clear thinking in settings where we seek help when things have gone unexpectedly wrong (e.g., you suddenly become ill or you do poorly on an exam). Why is this?

If things have gone unexpectedly wrong, that suggests we have some underlying expectation of how things should be going—perhaps formed on the basis of long past experience—and that we have deviated from that expectation in a bad direction. We might think of the expectation of how things should be going as reflecting the signal. And deviations from that expectation might reflect noise.

Let's think about that in a couple of concrete settings.

Suppose you are a relatively healthy person. You might think of good health as reflecting the true underlying signal for you. Even most fundamentally healthy people feel unwell from time to time—because of the flu, a back ache, or what have you—for idiosyncratic reasons that perhaps do not reflect any fundamental change to their underlying state of healthiness. And most healthy people also have days when they feel particularly vigorous and well. Thus, we might think of a day when you feel fine as reflecting your true signal, with very little noise. Days when you feel sick can be thought of as days with particularly negative values for the noise. And days when you feel like you could go out and conquer the world can be thought of as days with particularly positive values for the noise.

Or maybe an example about school will speak more directly to our readers. You have some underlying level of academic skill that reflects how strong a student you are in any given area. That is your true signal. But some days you do way better than normal on a test—perhaps reflecting a particularly lucky draw of questions or a particularly good night's sleep. And some days you do way worse than normal on a test—perhaps reflecting an unlucky draw of questions or a late night. These idiosyncratic features constitute noise.

Now, what does this have to do with reversion to the mean? Ask yourself, On which days is a person likely to seek help from, say, a chiropractor? Probably on days when they wake up with a back ache that feels worse than normal—that is, on days when the noise is particularly negative. And ask yourself, Which students seek out help from a test prep company? Probably students whose performance on an important standardized test was worse than they had anticipated given their sense of their underlying ability. If this is right, then reversion to the mean tells us that, even if chiropractic adjustments or test prep strategies don't actually help at all, we should expect people who seek out this kind of assistance to see improvements. And, if they aren't thinking clearly about reversion to the mean, they are likely to give too much credit to the chiropractor or the test prep company. Reversion to the mean can be a good business model.

This kind of problem is pervasive. We've already seen an example in chapter 1, where we discussed broken-windows policing. Recall that, in New York City, when the police rolled out a new strategy, they targeted the highest crime precincts and found that, after the change of strategy, crime went down in those precincts. But this is just what we'd expect from reversion to the mean, even if broken-windows policing doesn't work at all. The highest-crime precincts will tend to get better and the lowest-crime precincts will tend to get worse, regardless of any policy change. Because of this reversion to the mean, policing strategies that target the highest-crime precincts will look to the naive observer as if they are really effective, even if they aren't.

And, actually, reversion to the mean was lurking underneath another of our examples in chapter 1. Remember when Ethan's son's doctors recommended that he try out a gluten-free diet because he was underweight? Their idea was that, if his weight percentile increased once he went on the gluten-free diet, that would be evidence of gluten intolerance. But reversion to the mean says that we should have expected Abe's weight percentile to increase even with no intervention. Month to month, a baby's weight is a function of both signal (e.g., health, genetics) and noise (e.g., random features of the environment, idiosyncrasies in the growth trajectory). If a kid has off-the-charts low weight in one month, he probably had extremely low values on the noise terms that month. Over time, we should expect more average values on the noise, so his weight percentile should increase. It would be a mistake to interpret this, on its own, as evidence that some change in behavior (e.g., going gluten-free) explains the weight increase.

Once you think about it clearly, you'll see that all kinds of interventions and treatments will look like they work even if they do nothing. People typically seek out help when things are at their worst. We'd expect things to improve, even without an intervention, because of reversion to the mean. Let's look at one particularly striking example that scientists have actually investigated to see whether this is going on.

Does Knee Surgery Work?

There are many expensive medical treatments for which the best available evidence is not so different from the evidence for broken-windows policing or SAT prep. For

instance, there are no randomized trials validating the efficacy of a variety of kinds of surgery. Consider a patient who goes to the surgeon with joint pain of some sort. The doctor recommends surgery. At the end of a recovery period, the patient says they feel better. Now, the doctor may have all sorts of knowledge about body mechanics and physiology that provides some good reasons for believing the surgery did help. But we should at least entertain the possibility that reversion to the mean is also at work—that is, that many patients would have experienced at least some significant improvement without surgery.

Indeed, once a randomized trial is run, researchers sometimes find that a common surgery does not in fact provide the hypothesized benefits. For example, in a 2002 study of arthroscopic surgery for osteoarthritis of the knee, researchers found that the commonly prescribed surgeries had no detectable effect on knee pain. Yes, patients reported less knee pain two weeks after surgery, but other subjects who simply received skin incisions and were told that they received the surgery reported the same reduction in knee pain. That's right: doctors actually gave sham surgeries to some of their patients, and those deceived patients were no worse off than those who received the real surgery. Why did all the patients seem to feel better? Presumably, you only go under the knife for knee pain when it's especially severe, so most of those patients might have felt better in a few weeks even without the surgery.

Reversion to the Mean, the Placebo Effect, and Cosmic Habituation

We've just seen that, if we fail to think clearly about reversion to the mean, we are likely to misinterpret the extent to which certain kinds of interventions, including medical interventions, are actually responsible for improved outcomes. But, it turns out, reversion to the mean can even create problems when we try to do careful scientific studies. Let's see why this is the case in a couple different settings. First, we will consider the much discussed *placebo effect*. Then we will revisit the problem of cosmic habituation with which we opened this chapter.

The Placebo Effect

Few phenomena in medicine are cited more often than the so-called *placebo effect*. Many people suspect that the belief that we've undergone treatment somehow activates the body's own healing powers, independent of the direct effects of the treatment itself. For this reason, medical researchers are careful to compare the effectiveness of new drugs or treatments to placebos. They want to account for the possibility that believing you're receiving a treatment will heal you all on its own. So they use things like sugar pills or fake surgeries, so that experimental subjects don't know whether they are getting a real treatment or not.

Why do medical researchers, and others, think that there's a placebo effect? One source of evidence comes from medical trials themselves. In such experiments there is a treatment group that gets the drug and a control group that gets a placebo pill. And often, in such studies, both groups' health improves. The improvement of the control group is taken as evidence for the placebo effect.

But now you can see that this kind of evidence for the placebo effect is unconvincing. The people in the control group (and the treatment group) entered the medical study because they were unwell. We might expect them to tend to get better even in

the absence of any treatment. This need not be because their minds are healing their bodies. It could just be reversion to the mean.

If you actually wanted to test for a placebo effect, you'd want to divide the pool of experimental subjects into a group that got a placebo pill and a group that got no treatment at all. (Of course, you could also have a group that got the real medicine. Let's not forget why we are here in the first place.) Few studies explicitly test for the effect of a placebo treatment relative to no treatment. The ones that do typically find no evidence of a placebo effect. Moreover, those studies that do find evidence for a placebo effect typically concern purely subjective outcomes. So people may perceive themselves to be feeling better after taking a placebo, even if they aren't objectively healthier.

For example, in 2011, a team of researchers from Harvard Medical School published a paper in the *New England Journal of Medicine* comparing the effects of real medical treatments, placebo treatments, and no treatment for asthmatic patients. Interestingly, both the real treatment (an albuterol inhaler) and the placebo treatments (a placebo inhaler or acupuncture) led patients to report that they felt better. But when the scientists actually measured the subjects' lung capacity, only the real treatment had an effect; the placebo treatments were no better than doing nothing. So to the extent that there's evidence for a placebo effect, it's evidence of mind over mind, not mind over matter.

In short, once we think clearly about reversion to the mean, we see that there's little compelling evidence of a placebo effect in medicine. Yet somehow, almost everyone believes in the placebo effect because they're bad at recognizing and correctly interpreting reversion to the mean. Even some of the greatest medical researchers have fallen victim to this confusion. Vitamin C is widely believed to provide significant health benefits, despite little evidence. (To be clear, a small amount of vitamin C is necessary to avoid scurvy, but for almost everyone in the developed world who naturally consumes some vitamin C, there is little evidence that additional vitamin C is beneficial.) An important source of this widely held superstition is Linus Pauling, a world-famous chemist and two-time Nobel Prize winner, who strongly advocated for vitamin C. Some argue that Pauling knew that vitamin C had little effect, but he believed in the placebo effect, and he thought that telling people that a vitamin C supplement, or a glass of orange juice, would heal them was a cheap and easy way to get the placebo effect working, inducing the human body to somehow cure itself.

Cosmic Habituation Explained

If you are paying particularly close attention, you might have noticed a connection between the talk of signal and noise in this chapter and our favorite equation. Recall, our favorite equation says that an estimate from data is made up of three things—the true estimand, bias, and noise:

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

Imagine a study that is really well designed to learn about an estimand so that there is no bias. The estimate generated by that study will be made up of the true estimand and noise from things like sampling variation. The true estimand stays constant across different studies of the same phenomenon. But the resulting estimates nonetheless vary from study to study because of the noise. So, if we imagine multiple studies of the same phenomenon, we can think of the true estimand as the signal. And we can think of the noise as, well, the noise.

Given this, we should expect repeated scientific studies of the same phenomenon to exhibit reversion to the mean. If the first study found a particularly large relationship, it is probably the case that the true relationship in the world is big, and it is also probably the case that the sampling variation happened to create noise in the positive direction in that study. In the next study, we should expect to find a smaller estimate because, while the true relationship (i.e., the estimand) is still probably large, the noise from the sampling variation will probably not be as large this time around.

With this realization, we are now, finally, ready to return to the idea of cosmic habituation and see why it probably isn't best explained by a mystical force whereby the universe accustoms itself to the activity of scientists here on earth.

If it isn't mystical forces, why do estimated effects tend to get smaller when replicated? Part of the answer is reversion to the mean. But that's not the whole answer. To really understand what is going on with cosmic habituation, you need to combine our new understanding of reversion to the mean with our discussion of publication bias from chapter 7. Let's see why.

Imagine several scientists, independently studying some phenomenon—say, whether giving people time to think leads them to make better decisions, which is one of Jonathan Schooler's areas of interest. Each scientist does a study. One finds that people who are given time to think make dramatically worse decisions. Another finds that people who are given time to think make slightly worse decisions. A third finds no relationship. And a fourth finds that people who are given time to think make slightly better decisions. Given the sample size of the various studies, only the large finding is statistically significant.

What generates these different estimates? Presumably, there is some true effect of giving people time to think on the quality of their decisions. This is the estimand in our favorite equation. We can also think of it as the *signal* that is common across each of these studies. But then there are lots of idiosyncratic features, the *noise*, that affect the observed relationship (the estimate) in any given study. For instance, even though these are experiments, in any one of them, by happenstance, it could occur that the people given time to think turn out to be intrinsically much worse decision makers than the people not given time to think. This large negative noise term would lead that study to find a particularly large negative effect. In another study the people given time to think might happen to be intrinsically slightly better decision makers than the people not given time to think. This positive noise term would lead that study to find a more favorable relationship between thinking time and decision making. So, we can see, the results of these studies are made up of both signal and noise. Thus, we should expect replications to experience reversion to the mean.

Now let's think about which of the findings are most likely to be replicated. From our earlier discussion of *p*-screening and publication bias, if we had to guess, we'd guess that only one of these studies is notable enough to the original researcher or the scientific community to warrant an independent replication—the one with the large, statistically significant negative relationship between time to think and the quality of decision making. This study has two things going for it over the other studies. First, it has a statistically significant finding, so it is more likely to be published. Second, that finding is pretty surprising—who would have thought that thinking things through leads to worse decisions?

Imagine that this is what happens. One large surprising finding gets published. It was found in a well-designed study, so people think it is probably true. But, because it is an important result, scientists will also want to see if it replicates. What should we

expect them to find? Well, we just saw that a finding is more likely to get published and warrant replication (both because of its surprise value and because it is more likely to pass the statistical significance threshold) when the estimated effect size is particularly large in magnitude. But we also know that particularly large estimates are probably the result of both large values of the signal *and* large values of the noise. So, because of reversion to the mean, when we go to replicate this study, we should expect to find a smaller (in magnitude) estimated effect size (as, indeed, the other three, unpublished studies had found). That is, because of a combination of publication bias and reversion to the mean, we should expect to see cosmic habituation!

Cosmic Habituation and Genetics

Figure 8.4 is our favorite illustration of the phenomenon of cosmic habituation resulting from publication bias and reversion to the mean. The figure describes the changing evidence on the link between particular genes and particular diseases. The different curves represent different hypothesized gene-disease linkages. Each data point shows the sign and size of the estimated relationship, taking into account all the available data at any given point in time. In this particular graph, a value of 1 on the vertical axis means that the evidence shows no relationship at all between the gene and the disease. A value below 1 means that the evidence shows a negative relationship between the gene and the disease. And a value above 1 means that the evidence shows a positive relationship between the gene and the disease. The farther from 1, the larger (in magnitude) the estimated relationship.

The data points on the far left of the plot are the estimated relationships between genes and diseases from the very first study published on the topic. Moving to the right, the next data point shows the estimated relationship taking into account the data from both the first and second published studies. This continues as we move to the right, until we get to the most recent published study.

What you can see in figure 8.4 looks like cosmic habituation. The first published study finds a large relationship between a gene and a disease. This is the kind of study that gets published in a prestigious journal and covered in the press. But as scientists engage in replication, reversion to the mean kicks in. The magnitude of the estimated effects is typically smaller in subsequent studies. As we add more and more information, we get closer to the truth, which is quite far from the over-estimate reported in the initial study. As we see, by the end, the evidence suggests at most a very weak relationship between the gene and disease in question. The follow-up newspaper story is rarely written.

Beliefs Don't Revert to the Mean

Once you understand reversion to the mean, you should start to worry about it a lot. Commentators, analysts, and casual observers constantly misunderstand it, introducing complex theories to explain patterns that reflect a simple statistical phenomenon. The preceding discussion makes it sound like we should expect reversion to the mean almost everywhere, and to a close approximation, that's right. We should see reversion to the mean for any variable that is influenced by signal and noise. So instead of listing all the instances of reversion to the mean, let's think about situations where we shouldn't expect it.

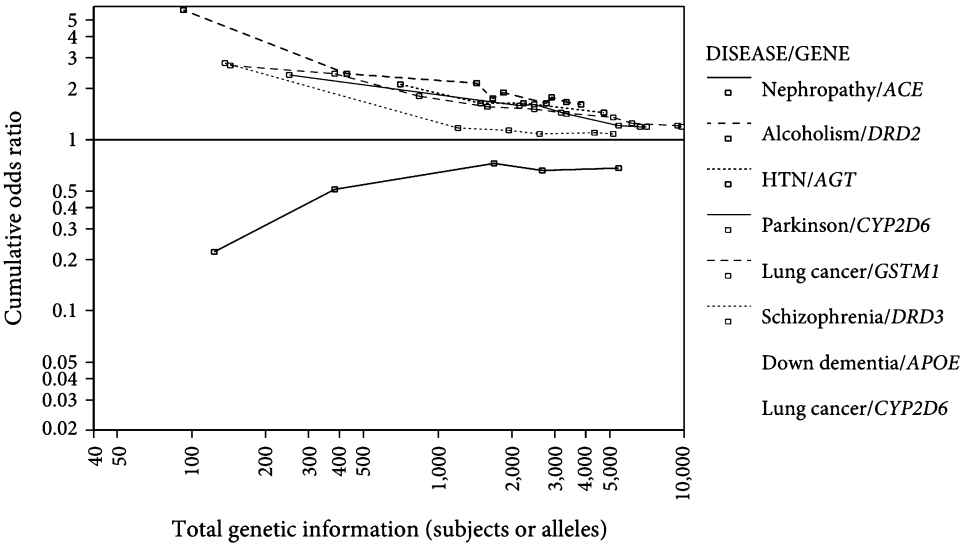


Figure 8.4. The estimated effect size shrinks as more data is accumulated in genetic studies.

First, if the signal is much greater than the noise, then we wouldn't expect to see much reversion to the mean. Suppose we repeated our earlier golf analysis, but instead of plotting scores from two different rounds within one tournament, we plotted average scores from two different seasons on the LPGA Tour. An average score from an entire season contains a lot more information about a player's ability. Much of the good and bad luck that constitutes the noise across rounds averages out, so there will be less reversion to the mean in that picture (but still some).

But there are some situations where we should expect no reversion to the mean at all. Let's think about the stock market. Should we expect to find reversion to the mean in stock prices? Maybe we could beat the market and become billionaires by exploiting reversion to the mean.

Reversion to the mean would seem to predict that, on average, companies with low stock prices should rebound in the future, and companies with high prices should drop. It also seems to predict that increases should be followed by decreases and vice versa. Could we exploit this information by buying the stocks that just dropped and selling the stocks that just increased?

The answer to this question is almost surely no. Suppose it were the case that there was reversion to the mean in stock prices. Clever investors would realize this, follow the strategy described above, and make a lot of money. But with enough investors following this strategy, the reversion to the mean would go away, because the prices of the low stocks would increase and the prices of the high stocks would decrease in response to the market's buying and selling decisions. The efficient-market hypothesis, discussed briefly in chapter 7, says that with enough traders looking for these kinds of opportunities, we shouldn't be able to predict changes in stock prices, and therefore, we shouldn't expect reversion to the mean.

Reversion to the mean is quite prevalent in the business world. We see it for corporate revenues and profits, despite the desire by startups and venture capitalists to project future revenues by making linear or sometimes exponential projections from

past revenues. So why don't we see reversion to the mean for stock prices? The main reason is that stock prices reflect beliefs about the future, while revenues don't. The price of the stock is driven by investors' beliefs about the long-term value of the firm. And if there were reversion to the mean, it would suggest that investors are making systematic mistakes in forming those beliefs. If there were a stock that we could expect to increase in price, investors would all buy it, driving the price up, erasing our expectation of a change.

The stock market is just one example of a general phenomenon. There should be no reversion to the mean when it comes to beliefs. It wouldn't make sense to say something like "Today, I believe that Republicans have a 60 percent chance of winning the House in the next election, but come Election Day, I expect my belief to be lower." That doesn't make sense because your belief is just your belief about the future—nothing else. If you expect that your belief will be 55, rather than 60, percent on Election Day, then your belief should be 55 percent today.

Wrapping Up

In part 2 we have learned how to quantify correlations and how to assess whether correlations found in data are likely to reflect real phenomena or just noise. We then turned to other challenges created by the presence of noise—over-comparing and under-reporting and reversion to the mean.

Importantly, our favorite equation told us that noise is not the only reason an estimate might not equal the estimand. We also have to worry about bias. For reasons that we started to learn about in chapter 3, bias is a particularly important concern when we are trying to learn about causal relationships—when we say correlation doesn't imply causation, what we mean is that the correlation between two features of the world may be a biased estimate of the causal relationship between them. In part 3, we will focus on causal relationships, first examining the sources of bias in more detail and then learning about strategies for estimating causal relationships in an unbiased way.

Key Words

- **Hawthorne effect:** The phenomenon whereby subjects change their behavior because they know they are being studied.
- **Demand effect:** A specific instance of a Hawthorne effect in which research subjects change their behavior to try to please the researcher.
- **Signal:** The systematic component of an outcome that is persistent across observations.
- **Noise:** Random components of an outcome that change from observation to observation.
- **Reversion to the mean:** The phenomenon whereby, if one observation of an outcome made up of signal and noise is particularly large (respectively, small) other observations will typically be smaller (respectively, larger).

Exercises

8.1 Early on in every baseball season, someone appears to be on pace to break the home-run record, but they almost never do. Let's think about why.

Suppose you hit a phenomenal number of home runs in the first twenty games of the season, and you're on pace to break the record.

(a) In the next twenty games, are you more likely to hit an above-average or below-average number of home runs? Why?
(b) In the next twenty games, are you likely to hit fewer home runs, the same number of home runs, or more home runs than you hit in the first twenty games? Why?
(c) A commentator notices that players on pace after the first twenty games almost never break the record. The commentator argues that this shows that players lose their nerve when they start thinking about setting the record. What data might you want to collect, and how might you want to analyze it in order to see if this interpretation is right?

8.2 Anthony once took a course from a famous econometrician who made the following argument: Paul's son John has a genius-level IQ. Therefore, because of mean reversion, Paul himself must have had a super-genius-level IQ.

(a) What's wrong with the econometrician's reasoning? If you had to guess, is Paul's IQ lower or higher than average? Is it lower or higher than John's?
(b) How would your answer change if we told you that the Paul in this example is Paul Samuelson, a Nobel Prize winner considered by many to be the foremost academic economist of the twentieth century?

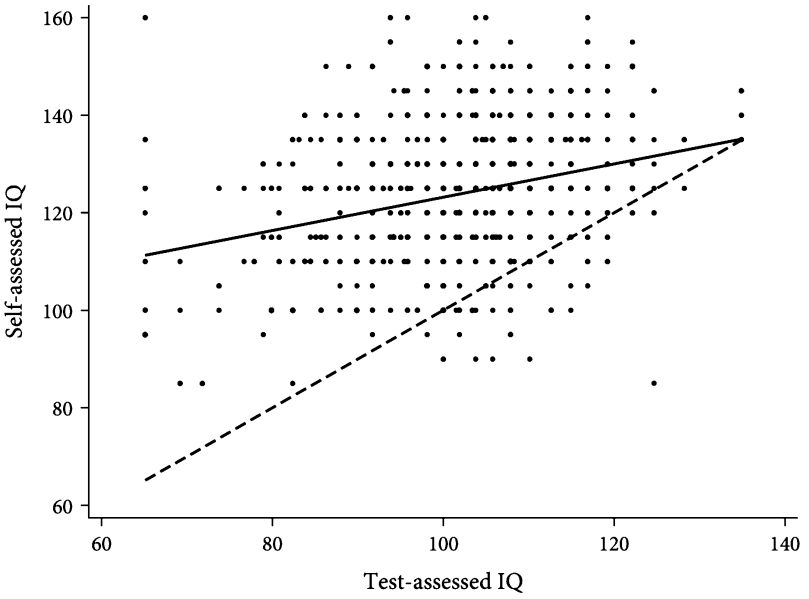
8.3 At the time of this writing, the stock price of Zoom (a company specializing in online video conferencing, with which many of us became all too familiar in the year 2020) has just fallen by about 18 percent in response to Joe Biden winning the 2020 U.S. presidential election and the release of promising results on COVID-19 vaccines. Because of mean reversion, your friend argues that now would be a great time to buy Zoom stock.

(a) Explain to your friend, in layperson's terms, what's wrong with their reasoning.
(b) Without knowing any additional details except that Zoom stock recently fell, would you expect your friend to make money, lose money, or break even on this investment?

8.4 Psychologists argue that the extent to which someone can accurately assess their own ability in some domain depends on their ability in that domain. Perhaps people who lack ability in a particular area don't even know enough to know how deficient they really are. Named after the researchers who first developed this hypothesis, this phenomenon is often called the Dunning-Kruger effect.

The typical evidence offered in favor of the Dunning-Kruger hypothesis is shown in the figure below. Subjects were first asked to predict their own IQ, and they later took an IQ test. The figure shows a scatterplot of these two scores for each subject. The regression line is shown in gray, and the dashed, 45-degree line is in black. People with low IQs (as measured by the test)

tended to over-estimate their score, and people with high IQs (as measured by the test) were, on average, correct.



(a) Can you come up with another explanation for this empirical phenomenon that wouldn't necessarily imply that high-IQ people are better at assessing their own IQs? You'll want to think about mean reversion, but that alone won't do the trick since the high-test-score people didn't under-estimate their IQ. Perhaps you'll also want to think about bias. Remember our favorite equation.

(b) To check your intuitions, simulate some data on your computer that generates a similar result even though the assessments of high-ability people are just as noisy as those of low-ability people. (Hint: It will help to remember that the test-assessed IQ is not a perfect measure of true intelligence.)

(c) Download "IQdata.csv" and the associated "README.txt," which describes the variables in this data set, at press.princeton.edu/thinking-clearly. These are the data used to produce the figure. (We obtained the data from a 2020 study by Gignac and Zajenkowski.) Let's think about how we can assess whether high-IQ people really are better at assessing their own IQs.

   i. First, compute the absolute value of the error for each subject (that is, how far off were their self-assessed IQs from their test-assessed IQs?).
   ii. Now, regress this absolute error on the test-assessed IQ and interpret the results.

(d) As you can see in the figure, people tend to over-estimate their own IQs.

   i. Estimate the average extent of this bias in this data.

ii. Subtract this estimate of the average bias from everyone's self-assessed IQ to get a bias-corrected self-assessment.
iii. Using this bias-corrected self-assessment, recompute the absolute value of the errors—that is, calculate how far off, on average, a person's bias-corrected self-assessment is from their test-assessed IQ.
iv. Finally, regress this new measure of error on the test-assessed IQ and interpret the results.

(e) Provide your final assessment of the Dunning-Kruger hypothesis on the basis of this data. Are high-intelligence people better at assessing their own intelligence?

8.5 Find a recent example where an analyst failed to consider mean reversion when they should have. Specifically, look for evidence that is presented in favor of a particular theory or phenomenon that could also easily be explained by mean reversion. Your example might come from a newspaper article, an academic study, a policy memo, or a statement by a politician, business leader, or sports commentator. Summarize the claim being made by the analyst and the evidence that purportedly supports the claim. Explain why the data is equally consistent with mean reversion. As a bonus, think about ways that you could potentially adjudicate between the analyst's claim and mean reversion.

Readings and References

The *New Yorker* article on cosmic habituation is:

Jonah Lehrer. 2010. "The Truth Wears Off: Is There Something Wrong with the Scientific Method?" *The New Yorker*. December 13.

The study we mentioned reanalyzing the evidence for the Hawthorne effect is

Steven D. Levitt and John A. List. 2011. "Was There Really a Hawthorne Effect at the Hawthorne Plant?: An Analysis of the Original Illumination Experiments." *American Economic Journal: Applied Economics* 3(1): 224–238.

The source for Figure 8.1 is Galton's original paper on the subject

Francis Galton. 1886. "Regression towards Mediocrity in Hereditary Stature." *Journal of the Anthropological Institute of Great Britain and Ireland* 15:246–263.

The study comparing real knee surgeries to sham surgeries is

J. Bruce Moseley, Kimberly O'Malley, Nancy J. Petersen, Terri J. Menke, Baruch A. Brody, David H. Kuykendall, John C. Hollingsworth, Carol M. Ashton, and Nelda P. Wray. 2002. "A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee." *New England Journal of Medicine* 347(2):81–88.

The article showing that, among asthma patients, the placebo effect appears to exist for subjective measures but not for objective measures is

Michael E. Wechsler, John M. Kelley, Ingrid O. E. Boyd, Stefanie Dutile, Gautham Marigowda, Irving Kirsch, Elliot Israel, and Ted J. Kaptchuk. 2011. "Active Albuterol

or Placebo, Sham Acupuncture, or No Intervention in Asthma." *New England Journal of Medicine* 365(2):119–126.

Figure 8.4 is taken from

John P. A. Ioannidis, Evangelia E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos-Ioannidis. 2001. "Replication Validity of Genetic Association Tests." *Nature Genetics* 29:306–309.

The exercise on the Dunning-Kruger effect is inspired by

Gilles E. Gignac and Marcin Zajenkowski. 2020. "The Dunning-Kruger Effect Is (Mostly) a Statistical Artefact: Valid Approaches to Testing the Hypothesis with Individual Differences Data." *Intelligence* 80:101449.