CHAPTER 6

Selection Bias

WE ARE BOTH SKIERS, AND WE HAVE BOTH SPENT OUR SHARE of time in the Wasatch mountains outside of Salt Lake City, Utah, enjoying some of the best snow on the planet. The perfect Utah powder even factored into the choice that one of us made about what college to attend. A number of ski resorts are situated in the Wasatch mountains, and each resort has its own personality. Snowbird rises out of Little Cottonwood Canyon, splayed out in steel and glass and concrete, its gondola soaring over sheer cliffs to a cirque from which terrifyingly steep chutes drop away. Farther up the canyon, Alta has equally challenging terrain, even better snow, and feels lost in time. Its lodges are simple wooden structures and its lifts are bare-bones; it is one of three remaining resorts in the country that won't allow snowboarders. Start talking to a fellow skier at a party or on an airplane or in a city bar, and the odds are that she will mention one or both of these resorts as among the best in North America.

In nearby Big Cottonwood Canyon, the resorts Brighton and Solitude have a very different feel. They are beautiful and well suited for the family and great fun to ski. But few would describe them as epic, and they're hardly destination resorts. Yet if you take a day off from Alta or Snowbird to ski at Solitude, something interesting happens. Riding the lifts with other skiers, the conversation inevitably turns to the relative merits of the local resorts. But at Solitude, unlike Alta or Snowbird—or anywhere else, really—people often mention Solitude as the best place to ski in the world. They cite the great snow, the mellow family atmosphere, the moderate runs, the absence of lift lines, the beauty of the surrounding mountains, and numerous other factors.

When Carl skied Solitude for the first time, at fifteen, he was so taken by this tendency that he mentioned it to his father as they inhaled burgers at the base lodge before taking the bus back to the city.

"I think maybe I underestimated Solitude," Carl told him. "I had a pretty good day here. There is some good tree skiing, and if you like groomed cruising . . .

"And I bet I didn't even find the best runs," Carl continued. "There must be some amazing lines here. Probably two-thirds of the people I talked to today like this place even more than Alta or Snowbird! That's huge praise."

Carl's father chuckled. "Why do you think they're skiing at Solitude?" he asked.

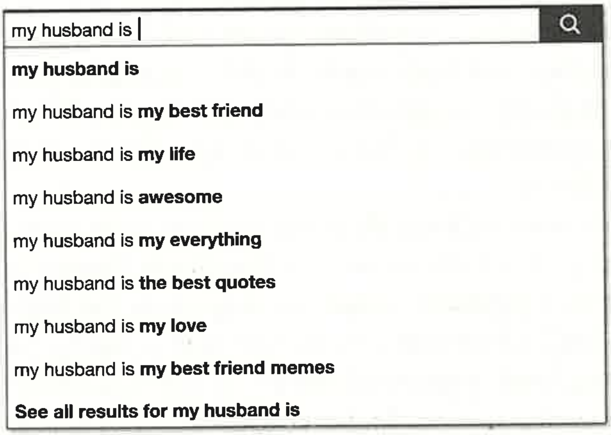
This was Carl's first exposure to the logic of selection effects. *Of course* when you ask people at Solitude where they like to ski, they will answer "Solitude." If they didn't like to ski at Solitude, they would be at Alta or Snowbird or Brighton instead. The superlatives he'd heard in praise of Solitude that day were not randomly sampled from among the community of US skiers. Skiers at Solitude are not a representative sample of US skiers; they are skiers who could just as easily be at Alta or Snowbird and choose not to be. Obvious as that may be in this example, this basic principle is a major source of confusion and misunderstanding in the analysis of data.

In the third chapter of this book, we introduced the notion of statistical tests or data science algorithms as *black boxes* that can serve to conceal bullshit of various types. We argued that one can usually see this bullshit for what it is without having to delve into the fine details of how the black box itself works. In this chapter, the black boxes we will be considering are statistical analyses, and we will consider some of the common problems that can arise with the data that is fed into these black boxes.

Often we want to learn about the individuals in some group. We might want to know the incomes of families in Tucson, the strength of bolts from a particular factory in Detroit, or the health status of American high school teachers. As nice as it would be to be able to look at every single member of the group, doing so would be expensive if not outright infeasible. In statistical analysis, we deal with this problem by investigating small samples of a larger group and using that information to make broader inferences. If we want to know how many eggs are laid by nesting bluebirds, we don't have to look in every bluebird nest in the country. We can look at a few dozen nests and make a pretty good estimate from what we find. If we want to know how people are going to vote on an upcoming ballot measure, we don't need to ask every registered voter what they are thinking; we can survey a sample of voters and use that information to predict the outcome of the election.

The problem with this approach is that *what you see depends on where you look*. To draw valid conclusions, we have to be careful to ensure that the group we look at is a random sample of the population. People who shop at organic markets are more likely to have liberal political leanings; gun show attendees are more likely to be conservative. If we conduct a survey of voters at an organic grocery store—or at a gun show—we are likely to get a misleading impression of sentiments citywide.

We also need to think about whether the results we get are influenced by the act of sampling itself. Individuals being interviewed for a psychology study may give different answers depending on whether the interviewer is a man or a woman, for example. We run into this effect if we try to use the voluminous data from the Internet to understand aspects of social life. Facebook's autocomplete feature provides a quick, if informal, way to get a sense of what people are talking about on the social media platform. How healthy is the institution of marriage in 2019? Let's try a Facebook search query:



my husband is

my husband is

my husband is **my best friend**

my husband is **my life**

my husband is **awesome**

my husband is **my everything**

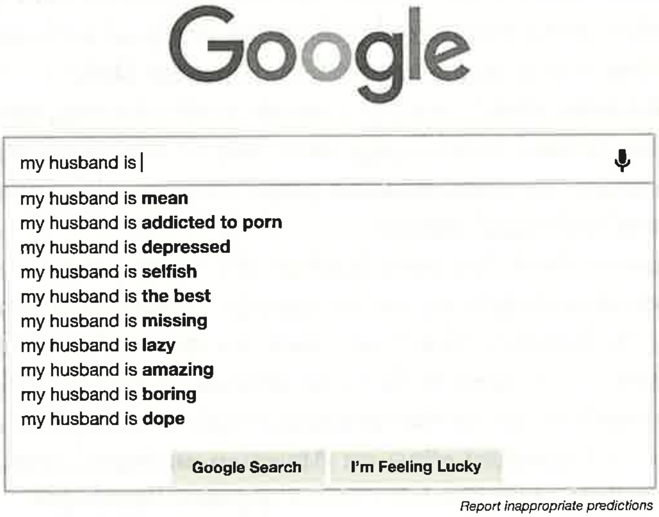
my husband is **the best quotes**

my husband is **my love**

my husband is **my best friend memes**

See all results for my husband is

This paints a happy—if saccharine—picture. But on Facebook people generally try to portray their lives in the best possible light. The people who post about their husbands on Facebook may not be a random sample of married people; they may be the ones with happy marriages. And what people write on Facebook may not be a reliable indicator of their happiness. If we type the same query into Google and let Google's autocomplete feature tell us about contemporary matrimony, we see something very different:



my husband is |

my husband is **mean**
my husband is **addicted to porn**
my husband is **depressed**
my husband is **selfish**
my husband is **the best**
my husband is **missing**
my husband is **lazy**
my husband is **amazing**
my husband is **boring**
my husband is **dope**

Yikes! At least "the best," "amazing," and "dope" (as opposed to "a dope") make the top ten. It seems that people turn to Google when looking for help, and turn to Facebook when boasting about their lives. What we find depends on where we look.

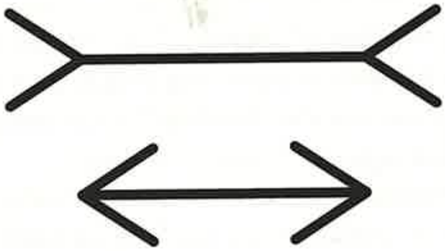
We should stress that a sample does not need to be completely random in order to be useful. It just needs to be random *with respect to whatever we are asking about*. Suppose we take an election poll based on only those voters whose names appear in the first ten pages of the phone book. This is a highly nonrandom sample of people. But unless having a name that begins with the letter *A* somehow correlates with political preference, our sample is random with respect to the question we are asking: How are you going to vote in the upcoming election?*

* Simply surveying people who are listed in the phone book can be a problem as well. Doing so excludes people who do not have landlines and people who have unlisted numbers. Moreover, those who take the time to answer a telephone survey may differ systematically from those who screen

Then there is the issue of how broadly we can expect a study's findings to apply. When can we extrapolate what we find from one population to other populations? One aim of social psychology is to uncover universals of human cognition, yet a vast majority of studies in social psychology are conducted on what Joe Henrich and colleagues have dubbed WEIRD populations: Western, Educated, Industrialized, Rich, and Democratic. Of these studies, most are conducted on the cheapest, most convenient population available: college students who have to serve as study subjects for course credit.

How far can we generalize based on the results of such studies? If we find that American college students are more likely to engage in violent behavior after listening to certain kinds of music, we need to be cautious about extrapolating this result to American retirees or German college students, let alone people in developing countries or members of traditional societies.

You might think that basic findings about something like visual perception should apply across demographic groups and cultures. Yet they do not. Members of different societies vary widely in their susceptibility to the famous Müller-Lyer illusion, in which the direction of arrowheads influences the apparent length of a line. The illusion has by far the strongest effect on American undergraduates.* Other groups see little or no difference in the apparent line length.



Again, where you look determines what you see.

their calls or hang up on the interviewer. As such, any form of sampling from a phone book may substantially undercount certain demographic groups including younger people and some minorities. This specific issue has become a major problem for telephone polling. Pollsters increasingly use cell phone interviews, but this approach comes with its own suite of problems.

* The reason that US undergraduates are most susceptible to the illusion is not well understood. One hypothesis notes that sharp, regular corners as in the arrowheads are relatively uncommon in nature. Perhaps the strength of the perceptual response to the arrowheads depends on the degree to which one has grown up in an environment with "carpentered corners." We are not sold, but we don't have a better explanation.

WHAT YOU SEE DEPENDS ON WHERE YOU LOOK

If you study one group and assume that your results apply to other groups, this is extrapolation. If you think you are studying one group, but do not manage to obtain a representative sample of that group, this is a different problem. It is a problem so important in statistics that it has a special name: *selection bias*. Selection bias arises when the individuals that you sample for your study differ systematically from the population of individuals eligible for your study.

For example, suppose we want to know how often students miss class sessions at the University of Washington. We survey the students in our Calling Bullshit class on a sunny Friday afternoon in May. The students' responses indicate that they miss, on average, two classes per semester. This seems implausibly low, given that our course is filled to capacity and yet only about two-thirds of the seats are occupied on any given day. So are our students lying to us? Not necessarily. The students who are answering our question are not a random sample of eligible individuals—all students in our class—*with respect to the question we are asking*. If they weren't particularly diligent in their attendance, the students in our sample wouldn't have been sitting there in the classroom while everyone else was outside soaking up the Friday afternoon sunshine.

Selection bias can create misleading impressions. Think about the ads you see for auto insurance. "New GEICO customers report average annual savings over $500" on car insurance. This sounds pretty impressive, and it would be easy to imagine that this means that *you* will save $500 per year by switching to GEICO.

But then if you look around, many other insurance agencies are running similar ads. Allstate advertisements proclaim that "drivers who switched to Allstate saved an average of $498 per year." Progressive claims that customers who switched to them saved over $500. Farmers claims that their insured who purchase multiple policies save an average of $502. Other insurance companies claim savings figures upward of $300. How can this be? How can all of the different agencies claim that switching to them saves a substantial amount of money? If some companies are cheaper than the competition, surely others must be more expensive.

The problem with thinking that you can save money by switching to GEICO (or Allstate, or Progressive, or Farmers) is that the people who switched to GEICO are nowhere near a random sample of customers in the market for automobile insurance. Think about it: What would it take to get you to go through the hassle of switching to GEICO (or any other agency)? You would have to save a substantial sum of money. People don't switch insurers in order to pay more!

Different insurance companies use different algorithms to determine your rates. Some weight your driving record more heavily; some put more emphasis on the number of miles you drive; some look at whether you store your car in a garage at night; others offer lower rates to students with good grades; some take into account the size of your engine; others offer a discount if you have antilock brakes and traction control. So when a driver shops around for insurance, she is looking for an insurer whose algorithms would lower her rates considerably. If she is already with the cheapest insurer for her personal situation, or if the other insurers are only a little cheaper, she is unlikely to switch. The only people who switch are those who will save big by doing so. And this is how all of the insurers can claim that those who switch to their policies save a substantial sum of money.

This is a classic example of selection bias. The people who switch to GEICO are not a random sample of insurance customers, but rather those who have the most to gain by switching. The ad copy could equivalently read, "Some people will see their insurance premiums go up if they switch to GEICO. Other people will see their premiums stay about the same. Yet others will see their premiums drop. Of these, a few people will see their premiums drop a lot. Of these who see a substantial drop, the average savings is $500." While accurate, you're unlikely to hear a talking reptile say something like this in a Super Bowl commercial.*

In all of these cases, the insurers presumably know that selection bias is responsible for the favorable numbers they are able to report. Smart consumers realize there is something misleading about the

* Your chances of finding a better deal on insurance appear not to be all that high. Only about one in three people shop for car insurance each year, and fewer than one-third of those shopping actually switch. Just because the average person who switches to Allstate saves $498, it doesn't mean that *you'll* save $498 by switching to Allstate. And just because almost every insurance company makes a similar claim, it doesn't mean that you'll save that kind of money if *you* switch. More likely, you'll end up being one of the majority of people who are unable to find a sufficiently compelling offer to justify a switch.

marketing, even if it isn't quite clear what that might be. But sometimes the insurance companies themselves can be caught unaware. An executive at a major insurance firm told us about one instance of selection bias that temporarily puzzled his team. Back in the 1990s, his employer was one of the first major agencies to sell insurance policies online. This seemed like a valuable market to enter early, but the firm's analytics team turned up a disturbing result about selling insurance to Internet-savvy customers. They discovered that individuals with email addresses were far more likely to have filed insurance claims than individuals without.

If the difference had been minor, it might have been tempting to assume it was a real pattern. One could even come up with any number of plausible *post hoc* explanations, e.g., that Internet users are more likely to be young males who drive more miles, more recklessly. But in this case, the difference in claim rates was huge. Our friend applied one of the most important rules for spotting bullshit: *If something seems too good or too bad to be true, it probably is.* He went back to the analytics team that found this pattern, told them it couldn't possibly be correct, and asked them to recheck their analysis. A week later, they reported back with a careful reanalysis that replicated the original result. Our friend still didn't believe it and sent them back to look yet again for an explanation.

This time they returned a bit sheepishly. The math was correct, they explained, but there was a problem with the data. The company did not solicit email addresses when initially selling a policy. The only time they asked for an email address was when someone was in the process of filing a claim. As a result, anyone who had an email address in the company's database had necessarily also filed a claim. People who used email were not more likely to file claims—but people who had filed claims were vastly more likely to have email addresses on file.

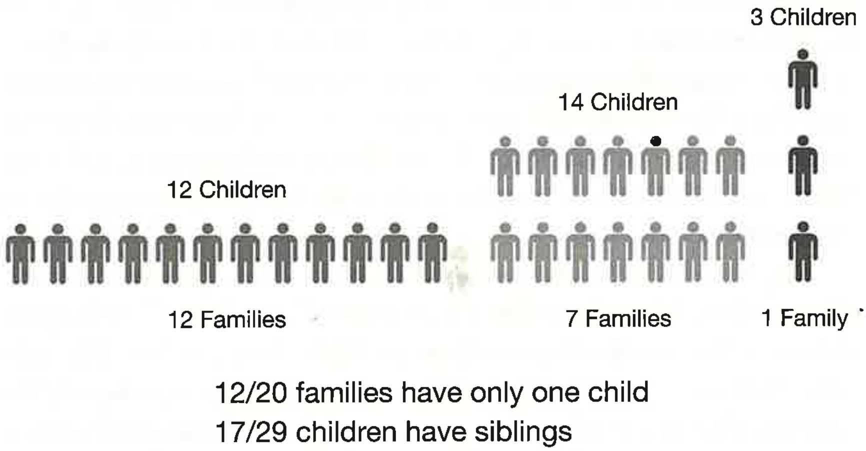
Selection effects appear everywhere, once you start looking for them. A psychiatrist friend of ours marveled at the asymmetry in how psychiatric disorders are manifested. "One in four Americans will suffer from excessive anxiety at some point," he explained, "but in my entire career I have only seen one patient who suffered from *too little* anxiety."

Of course! No one walks into their shrink's office and says "Doctor, you've got to help me. I lie awake night after night *not* worrying."

Most likely there are as many people with too little anxiety as there are with too much. It's just that they don't go in for treatment. Instead they end up in prison, or on Wall Street.

THE HIDDEN CAUSE OF MURPHY'S LAW

In Portugal, about 60 percent of families with children have only one child, but about 60 percent of children have siblings. This sounds impossible, but it's not. The picture below illustrates how this works. Out of twenty Portuguese families, we would expect to see about twelve with a single child, seven with two children, and one with three children. Thus most families are single-child, but because multi-child families each have multiple children, most children live in multi-child families.



What does this have to do with bullshit? Universities boast about having small class sizes, but students often find these statistics hard to believe: "An average class size of 18? That's bullshit! In three years I've only had two classes with fewer than 50 students!"

Both the students and the university are right. How? This difference in perception arises because, just as multi-child families have a disproportionately large number of children, large classes serve a disproportionately large number of students. Suppose that in one semester, the biology department offers 20 classes with 20 students in each, and 4 classes with 200 students in each. Look at it from an administrator's perspective. Only 1 class in 6 is a large class. The mean class size is $[(20 \times 20) + (4 \times 200)] / 24 = 50$. So far so good.

But now notice that 800 students are taking 200-student classes and only 400 are taking 20-student classes. Five classes in six are small, but only one student in three is taking one of those classes. So if you ask a group of random students how big their classes are, the average of their responses will be approximately $[(800 \times 200) + (400 \times 20)] / 1{,}200 = 140$. We will call this the *experienced mean class size,*[*] because it reflects the class sizes that students actually experience.

Because larger classes contain more students, the average student is enrolled in a class that is larger than the average class. Institutions can exploit this distinction to promote their own agendas. A university recruiting pamphlet might proclaim, "The average biology class size is 50 students." The student government, lobbying for reduced class sizes, might report that "the average biology student is taking a class of 140." Neither is false, but they tell very different stories.

This principle explains why faculty are likely to have a different notion of class sizes than students do. If you wanted to know how big classes are, you might think you could ask either students or teachers and get the same answer. So long as everyone tells the truth, it shouldn't matter. But it does—a lot. Large or small, each class has one instructor. So if you sample instructors at random, you are likely to observe a large class or a small class in proportion to the frequency of such classes on campus. In our example above, there are more teachers teaching small classes. But large classes have many students and small classes have few, so if you sample students at random, students are more likely to be in large classes.[†]

Recall from chapter 5 Goodhart's law: "When a measure becomes

[*] Technically this is a weighted mean, albeit a curious one in which each value (class size) is weighed by that value itself.

[†] It is difficult to obtain exhaustive data on universities' course sizes, but Marquette University in Milwaukee provides a relatively fine-grained view, listing the number of courses of size 2–9, 10–19, 20–29, 30–39, 40–49, 50–99, and 100 or greater. We can use these data to estimate the gulf between average class size and experienced average class size at a midsize US institution. If we approximate the class sizes by the midpoints of the bins (and use 150 for the midpoint of the 100+ bin) we have the following class sizes (footnote continues on next page):

a target, it ceases to be a good measure." Class sizes provide an example. Every autumn, college and university administrators wait anxiously to learn their position in the *U.S. News & World Report* university rankings. A higher ranking improves the reputation of a school, which in turn draws applications from top students, increases alumni donations, and ultimately boosts revenue and reputation alike. It turns out that class size is a major ingredient in this ranking process, with a strong premium placed on small classes.

> Schools receive the most credit in this index for their proportions of undergraduate classes with fewer than 20 students. Classes with 20 to 29 students score second highest, 30 to 39 students third highest and 40 to 49 students fourth highest. Classes that are 50 or more students receive no credit.

By summing up points across classes, the *U.S. News & World Report* ranking is rewarding schools that maximize the number of small classes they offer, rather than minimizing the experienced mean class size. Since the student experience is what matters, this may be a mistake. Consider the numerical example we provided earlier. The biology department in that example has 24 instructors and 1,200 students enrolled. Classes could be restructured so that each one has 50 students. In this case the experienced mean class size would plummet from 140 to 50, but the department would go from a good score to a bad score according to the *U.S. News & World Report* criterion.* To eliminate this perverse incentive for universities to pack most of their

Class size	5	15	25	35	45	75	150
Number of classes	101	318	197	59	66	28	22

Given this distribution of class sizes, the mean class size is about 26. But not from the students' perspective. Only 505 students are in 5-person classes, whereas 3,300 students are in 150-student classes. The experienced mean class size turns out to be approximately 51—nearly twice the mean class size.

* Or consider the data from Marquette University, as presented in the previous footnote. Marquette *could* improve its *U.S. News & World Report* ranking, without hiring additional instructors, by offering 726 "small" classes with 15 students each, and assigning the remaining students to 65 classes of 150 students each. In doing so it would go from its current impressive score of 61 percent small classes to an extraordinary 92 percent of small classes. Sounds good, right? But if Marquette chased the rankings in this way, they would substantially decrease the quality of the student experience, because a large number of students would be in 150-student classes and the experienced mean class size would climb from 51 to 73.

students into large classes, we suggest that the *U.S. News* ranking use experienced class size, rather than class size, in their calculations.

The same mathematical principles explain the curious fact that most likely, the majority of your friends have more friends than you do. This is not true merely because you are the kind of person who reads a book about bullshit for fun; it's true of anyone, and it's known as the friendship paradox. Understanding the friendship paradox is a bit more difficult than understanding the class size issue that we just treated, but a basic grasp of the problem should not be beyond reach. The sociologist Scott Feld, who first outlined this paradoxical result, explains it as follows. Suppose people have ten friends, on average. (We say ten rather than five hundred because Feld's paper was written in 1991, back when "friends" were people you actually had met in person—and often even liked.) Now suppose in your circle there is an introvert with five friends and a socialite with fifteen friends. Taken together they average ten friends each. But the socialite is friends with fifteen people and thus makes fifteen people feel insecure about how few friends they have, whereas the introvert is friends with only five and thus makes only five people feel better about themselves.

It's well and good to make intuitive arguments, but is the friendship paradox actually true in real life? Feld looked at a diagram of friendships among 146 adolescent girls, painstakingly collected three decades prior. He found while many of these girls had fewer friends than their friends, relatively few had more friends than their friends.

But this is just one sample of one group in one small town. We would like to address this question on a far broader scale, and in the social media age, researchers can do so. One team looked at 69 billion friendships among 720 million users on Facebook. They find, indeed, that most users have fewer friends than their friends. In fact, this is the case for 93 percent of Facebook users! Mind twisting, right? These researchers found that the Facebook users have, on average, 190 friends, but their friends have, on average, about 635 friends.

Subsequent studies have distinguished between a weak form and a strong form of the friendship paradox. The weak form pertains to the mean (average) number of friends that your friends have. The weak form is maybe not so surprising: Suppose you follow Rihanna and 499 other people on Twitter. Rihanna has over ninety million followers, so the 500 people you follow will average at the very least

90,000,000 / 500 = 180,000 followers—far more than you have. The strong form is more striking. It states that most people have fewer friends than their median friend has. In other words, order your friends by the number of friends they have. Pick the friend that is right in the middle. That friend likely has more friends than you do. This phenomenon can't be attributed to a single ultrapopular friend. The same research team found that the strong form holds on Facebook as well: 84 percent of Facebook users have fewer friends than the median friend count of their friends. Unless you are Kim Kardashian or someone of similar ilk, you are likely to be in the same situation.

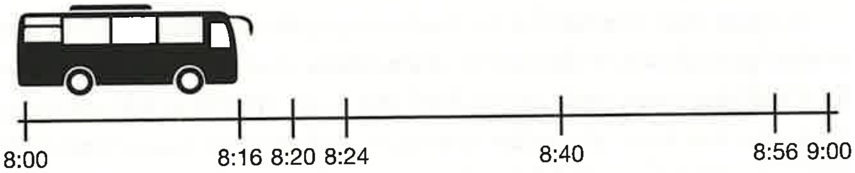
You may find it disconcerting to realize that the same logic applies to your past sexual history. Chances are, the majority of your partners have slept with more people than you have.

Okay. Forget we mentioned that. Back to statistics. Selection effects like this one are sometimes known as *observation selection effects* because they are driven by an association between the very presence of the observer and the variable that the observer reports. In the class size example, if we ask *students* about the size of their classes, there is an association between the presence of the student observer and the class size. If instead we ask *teachers* about the sizes of their classes, there are no observation selection effects because each class has only one teacher—and therefore there is no association between the presence of a teacher in the classroom and the size of the class.

Observation selection effects explain some of what we typically attribute to bad luck. If you commute by bus on a regular basis, you have probably noticed that you often have to wait a surprisingly long time for the next bus to arrive. But what is considered a "long wait"? To answer this question, we need to compare your wait to the average waiting time. Suppose that buses leave a bus stop at regular ten minute intervals. If you arrive at an arbitrary time, how long do you expect to wait, on average? The answer: five minutes. Since you might arrive anywhere in the ten-minute window, a nine-minute wait is just as likely as a one-minute wait, an eight-minute wait is just as likely as a two-minute wait, and so on. Each pair averages out to five minutes. In general, when the buses run some number of minutes apart, your average waiting time will be half of that interval.

What happens if the city operates the same number of buses, so that buses leave every ten minutes *on average*—but traffic forces the buses to run somewhat irregularly? Sometimes the time between buses is quite short; other times it may extend for fifteen minutes or more. Now how long do you expect to wait? Five minutes might seem like a good guess again. After all, the same number of buses are running and the average time between buses is still ten minutes.

But the actual average time that you will wait is longer. If you were equally likely to arrive during any interval between buses, the differences in between-bus gaps would average out and your average waiting time would be five minutes, as before. But you are not equally likely to arrive during any interval. You are more likely to arrive during one of the long intervals than during one of the short intervals. As a result, you end up waiting longer than five minutes, on average.



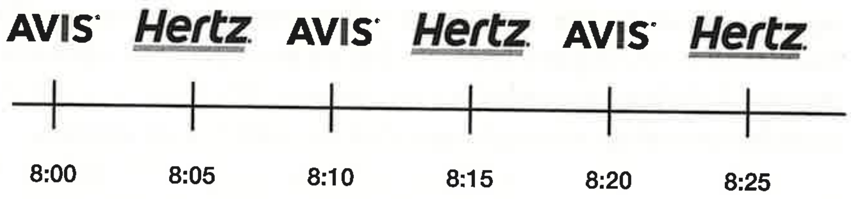
In the picture above, buses run every 10 minutes, on average, but they are clumped together so that some intervals are 16 minutes long and others are only 4 minutes long. You have an 80 percent chance of arriving during one of the long intervals, in which case you will wait 8 minutes on average. Only 20 percent of the time will you arrive during one of the short intervals and wait 2 minutes on average. Overall, your average wait time will be (0.8 × 8) + (0.2 × 2) = 6.8 minutes, substantially longer than the 5 minutes you would wait on average if the buses were evenly spaced.

So while it seems as though you tend to get unlucky with waiting times and wait longer than expected, it's not bad luck at all. It is just an observation selection effect. You are more likely to be present during a long between-bus interval, so you end up waiting and waiting.

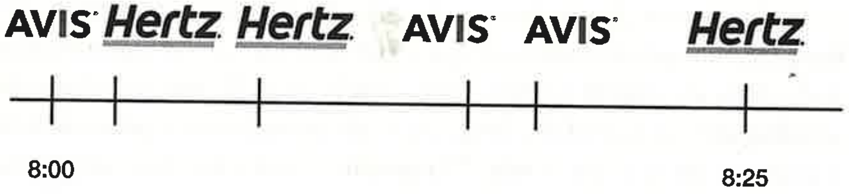
Something similar happens at the airport when you are waiting for a hotel van, an airport parking bus, or a rental car shuttle. A few days after writing the section above, Carl and his daughter flew into Los

Angeles and waited to catch a rental car shuttle. They stood and watched as multiple shuttles from several other car rental companies drove past, but no sign of their shuttle. After a while, Carl's daughter complained about their bad luck—but it wasn't particularly bad luck. It was an observation selection effect. Here's the explanation he gave her on the LAX curb.

To keep things simple, let's just think about two rental car companies, yours and its main competitor.

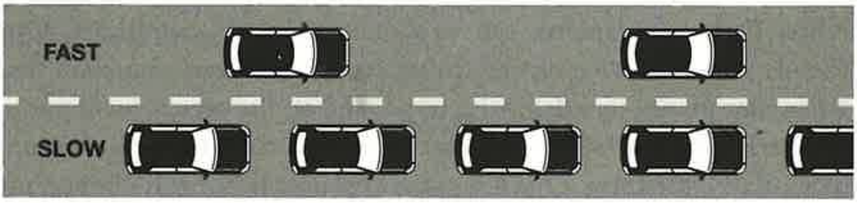


Suppose that shuttles for the two companies were evenly spaced as in the figure above. If you arrive at a random time, your bus will come first half the time, and second half the time. You'll never see more than one bus from the other company before your bus arrives. On average, you will see 0.5 buses from the other company.



But now suppose that the buses are not spaced out regularly, and instead arrive at random times. Half the time your bus will arrive first, just as when the buses were staggered. But half the time, the other bus will arrive first. The key thing to notice is that if it does, you're back where you started right when you first arrived, with a one-half chance that your bus comes next and a one-half chance that *another* bus from the other company comes next. If the same number of buses are running but they arrive at completely random times, you will see an average of 1 bus from the other company instead of 0.5 buses. The same thing happens if there are many different rental car companies running shuttle buses. If *n* different companies run buses at the same rate, your bus will on average be the *n*th to arrive.* This feels like bad luck. "There are only eight rental car companies at this airport, and my bus was the eighth to arrive!" If your intuitions are based on the case where buses are spaced out evenly, this is twice as long a wait as you expect. As a result, you'll get the impression that your company runs only half as many buses as its competitors—no matter which company you choose.

Observation selection effects can also explain why, when driving on a congested four-lane highway, you so often seem to pick the slow lane. You might think that with two lanes to pick from, you ought to be able to pick the faster lane half of the time. Not so! On the highway, the faster cars are traveling, the greater the distance between them. If traffic is moving at different speeds in two adjacent lanes, the cars in the slower-moving lane are more tightly packed. When two lanes are going in the same direction but at different speeds, the majority of the cars on the road will be in the slower lane and the majority of drivers will be cursing their lane-picking luck.



* In slightly more technical language, we are referring to the case where buses for each company arrive according to the Poisson arrival process with the same arrival rate. The waiting time from when you arrive at the curb to when your bus comes is then exponentially distributed; the arrival times of the other companies' buses are also exponentially distributed with the same distribution of waiting times. The story about being right back where you started if the other bus arrives first can be expressed mathematically using *first-step analysis*. Suppose there are *n* different rental car companies, all running shuttles at the same frequency, and let *s* be the expected number of shuttles that pass before yours arrives. With probability $1/n$ your shuttle arrives first and then you see zero other shuttles. With probability $n-1\ /\ n$ some other shuttle passes first, you've seen one shuttle so far, and then you are back right where you started. Therefore we can write $s = 0 \times (1/n) + (1+s) \times (n-1)\ /\ n$. Solving for *s*, we get $s = n-1$. If there are *n* shuttle companies, on average your bus will be the *n*th to arrive.

HOT GUYS AND TOP CODERS

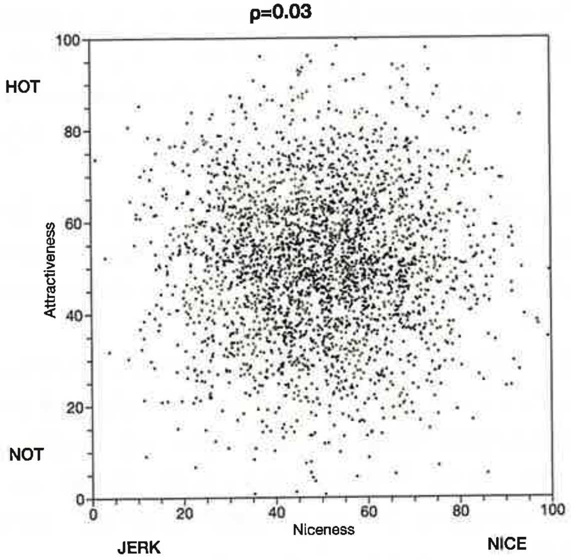
Five years ago, engineers at Google applied machine learning techniques to their own hiring process, in an effort to identify the most productive employees. They found a surprising result: Job performance was negatively correlated with prior success in programming contests. This could be because people who are good at programming contests have other attributes that make them less than ideal as employees. Google research director Peter Norvig conjectured that perhaps contest winners are used to working quickly, whereas a slower pace is more effective on the job. But we should not rush to that conclusion. Instead, this negative correlation between programming contest results and job performance could have something to do with the fact that these employees are far from a random sample of the population at large. They have already been selected strongly for programming ability and other skills during Google's hiring process. But to understand how that could generate a negative correlation, we will first turn to a question of broader interest: Why are hot guys such jerks?

Mathematician Jordan Ellenberg has suggested that a phenomenon called Berkson's paradox can explain a common complaint. Our friends who are active on the dating scene sometimes complain that when they go out on a date with someone hot, the person turns out to be a jerk, whereas when they meet someone nice, the person is not particularly attractive. One common explanation for this disappointing observation is that attractive people can afford to be jerks because they're afforded high status and are highly sought as partners. But there is another possible explanation.

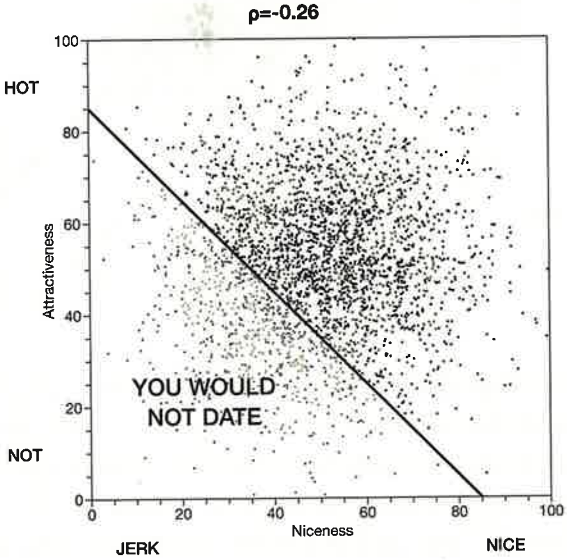
Ellenberg asks us to imagine placing potential partners on a two-dimensional plot, with how nice they are on the horizontal access and how attractive they are on the vertical axis. That might give you a plot of individuals that looks something like the next chart.

In this image, attractiveness and niceness are basically uncorrelated. Guys who are hot are no more or less likely to be jerks than guys who are not. So far, we have no reason to think that nice guys are unattractive and attractive guys are not nice.

But now let's look at what happens once you consider whom you'd



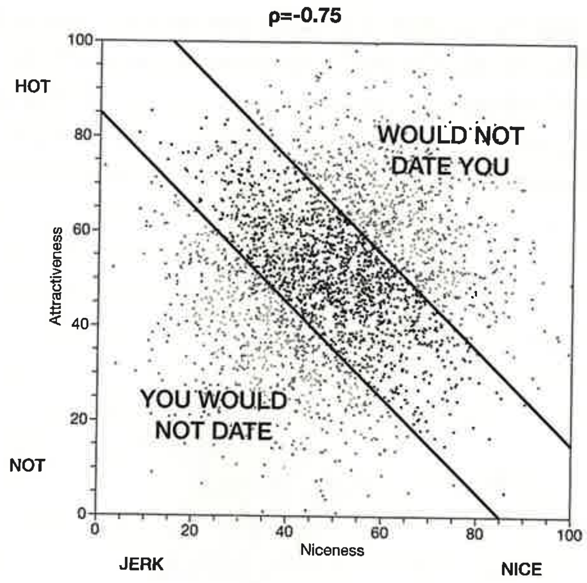
actually be willing to date. Certainly not anyone who is both unattractive and a jerk. Maybe you'd put up with some deficit in the looks department if a fellow was a really great person, or vice versa. So in the space of guys, the people you might date would be in the region above the diagonal:



Among the guys you would date, there is now a modest negative correlation between attractiveness and niceness. A really hot guy is less likely to be nice than the average guy that you would date, and a really nice guy is less likely to be hot than the average guy you would date.

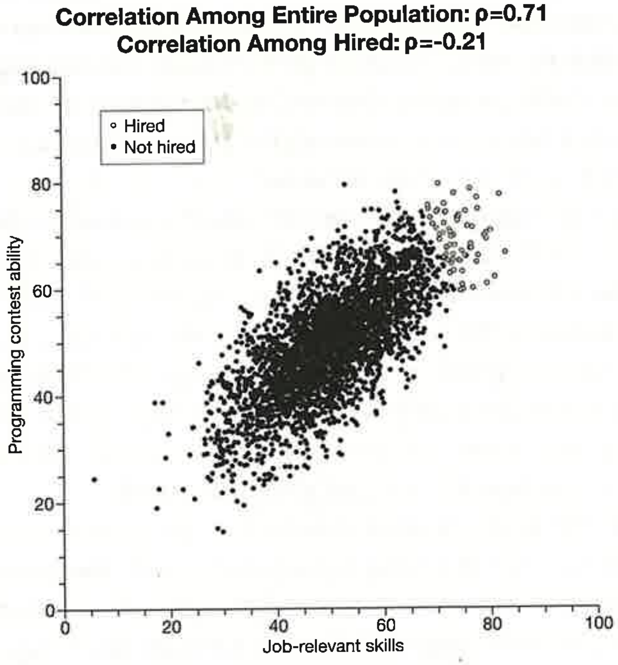
This is Berkson's paradox in operation. By selecting for both niceness and attractiveness, you have created a negative correlation between niceness and attractiveness among the people whom you would be willing to date.

An analogous process happens in the opposite direction. Not only do you pick whom you are willing to date, people pick whether they are willing to date you. We apologize for putting this bluntly, but unless you are John Legend, there will be some people who will have better options. So this rules out another set of possible people that you might date:



Who is left? Your dating pool is now restricted to a narrow diagonal band across the space of potential partners. Within this band, there is a strong negative correlation between niceness and attractiveness. Two separate instances of Berkson's paradox—one involving whom you would date and the other involving who would date you—have created this negative correlation among your potential partners, even though there is no negative trend in the population at large.

Let's return to the negative correlation between the ability to win programming contests and job performance at Google. In the population at large, we have every reason to expect a strong positive correlation between ability in programming contests and on-the-job ability at Google. The majority of US employees don't know how to program at all, but programming skill is a prerequisite for a job as a Google engineer. Google employees have been highly selected from the larger pool of potential employees, presumably for skills related to programming. Indeed, the assessments that hiring managers use during the hiring process may look a lot like the sorts of tasks or challenges that are posed in programming contests. Perhaps the hiring process puts a bit too much weight on programming contest ability at the expense of other qualities that make one effective at work. If this were the case, one could end up with a negative correlation between contest aptitude and job performance. The figure below illustrates:

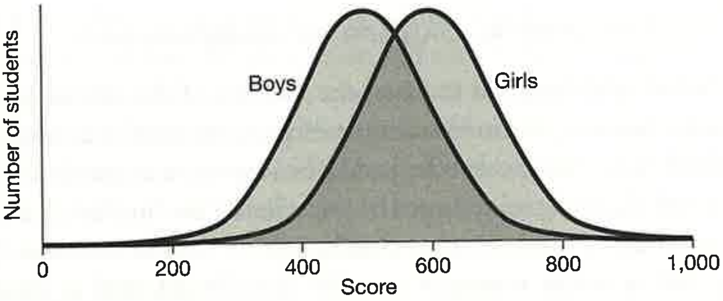


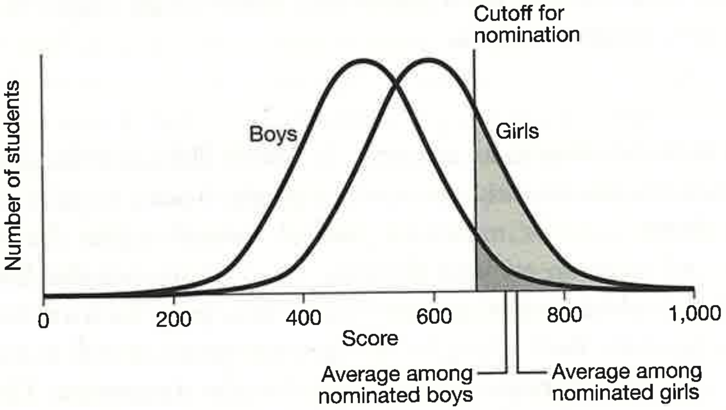
Once you start thinking about Berkson's paradox, you'll see it all over the place. Why do the best fielders in Major League Baseball tend to be mediocre hitters, and the best hitters tend to be mediocre fielders? Answer: There are plenty of players who are mediocre both in the field and at the plate, but they never make it to The Show. If you play *Dungeons & Dragons,* why do your best fighters have such low charisma and why are your best sorcerers so weak? Answer: because you literally trade ability in one domain for ability in another when your character reaches a new level and you allocate bonus power points. Why are some of the best songwriters cursed with Dylanesque voices and some of the best singers incapable of writing their own songs? Answer: There are plenty of musicians who sing like Dylan and write lyrics fit for a hair metal band, but thankfully you'll never hear them on the radio.

Even when selection happens right out in the open, the statistical consequences can seem counterintuitive. Suppose that a school can nominate students for a national scholarship competition based upon their grade point averages (GPAs). After the nominations are announced, we look at the two hundred students who were nominated. We find that the nominated girls have a GPA of 3.84 whereas the nominated boys have a GPA of 3.72. There are so many students in the sample that the differences are not merely due to chance. Can we conclude that the school must be preferentially nominating boys despite lower GPAs? It seems this way at first glance. If the school is using the same nomination standards for girls and boys, why wouldn't their GPAs be approximately the same?

A likely explanation is that the GPA distribution of the boys is different from the GPA distribution of the girls. If so, the GPA distributions of the nominated boys and the nominated girls will still differ even after selecting the best students, using the very same nomination threshold for both genders. The figure on page 125 illustrates. In this figure, the top-scoring girls have extremely high scores whereas the top-scoring boys more narrowly exceed the cutoff. This drives the difference in averages for students above the cutoff.

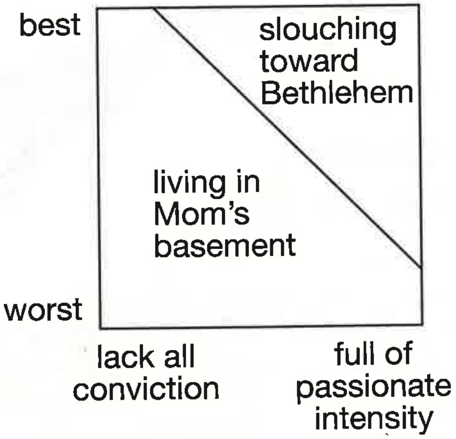
This isn't Berkson's paradox exactly, but the point is that selection can have all sorts of interesting consequences, and when trying to understand patterns from data, it is worth thinking about whether there has been any selection bias or deliberate selection operating, and if so, how those factors affect the patterns you observe.





Finally, we want to note that Berkson's paradox provides a hitherto unrecognized explanation for William Butler Yeats's dark observation in his classic modernist poem "The Second Coming":

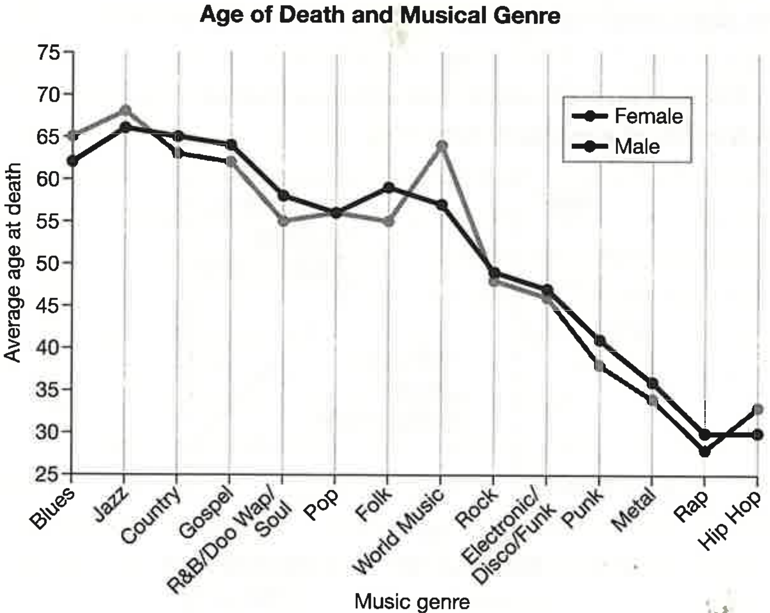
> The best lack all conviction, while the worst
> Are full of passionate intensity.



THE MORTAL DANGER OF MUSICIANSHIP

In a clinical trial designed to assess the severity of the side effects of a certain medication, the initial sample of patients may be random, but individuals who suffer side effects may be disproportionately likely to drop out of the trial and thus not be included in the final analysis. This is *data censoring,* a phenomenon closely related to selection bias. Censoring occurs when a sample may be initially selected at random, without selection bias, but a nonrandom subset of the sample doesn't figure into the final analysis.

LET'S DIVE RIGHT INTO an example. In March 2015, a striking graph made the rounds on social media. The graph, from a popular article about death rates for musicians, looked somewhat like the figure below and seems to reveal a shocking trend. It appears that being a musician in older musical genres—blues, jazz, gospel—is a relatively safe occupation. Performing in new genres—punk, metal, and especially rap and hip-hop—looks extraordinarily dangerous. The researcher who conducted the study told *The Washington Post* that "it's a cautionary tale to some degree. People who go into rap music or hip



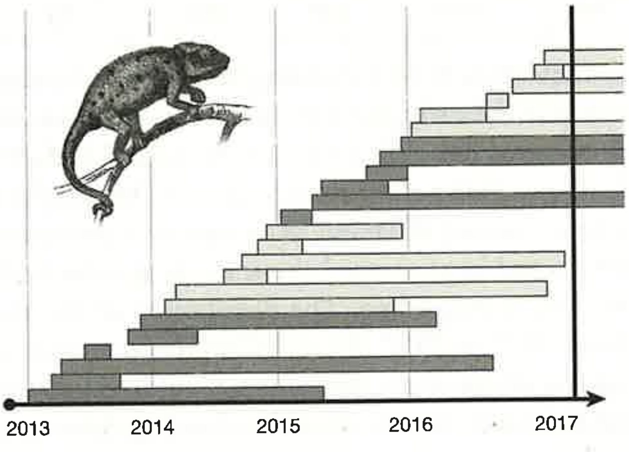
hop or punk, they're in a much more occupational hazard [sic] profession compared to war. We don't lose half our army in a battle."

This graph became a meme, and spread widely on social media. Not only does it provide quantitative data about an interesting topic, but these data reinforce anecdotal impressions that many of us may already hold about the hard living that goes along with a life in the music business.

Looking at the graph for a moment, however, you can see something isn't right. Again, if something seems too good or too bad to be true, it probably is. This certainly seems too bad to be true. What aroused our suspicion about this graph was the implausibly large difference in ages at death. We wouldn't be particularly skeptical if the study found 5 to 10 percent declines for musicians in some genres. But look at the graph. Rap and hip-hop musicians purportedly die at about thirty years of age—half the age of performers in some other genres.

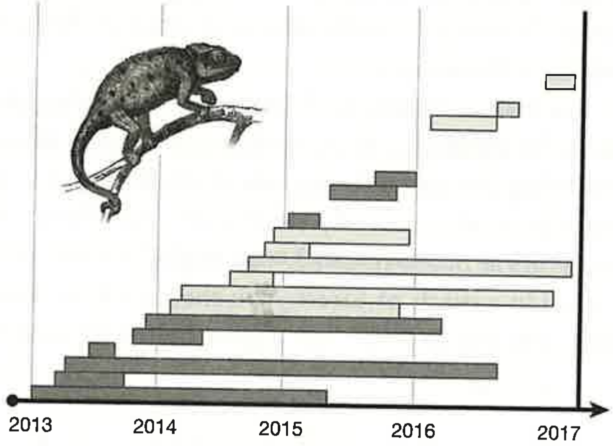
So what's going on? The data are misleading because they are *right-censored*—individuals who are still alive at the end of the study period are removed from the study.

Let's first look at an example of how right-censoring works, and then return to the musicians. Imagine that you are part of a conservation team studying the life cycle of a rare chameleon species on Madagascar. Chameleons are curiously short-lived; for them a full life is two to three years at most. In 2013, you begin banding every newly born individual in a patch of forest; you then track its survival until your funding runs out in 2017. The figure below illustrates the lon-



gevity of the individuals that you have banded. Each bar corresponds to one chameleon. At the bottom of the figure are the first individuals banded, back in 2013. Some die early, usually due to predation, but others live full chameleon lives. You record the date of death and indicate that in the graph. The same thing holds for the individuals born in 2014. But of the individuals born in 2015 and 2016, not all are deceased when the study ends in 2017. This is indicated by the bars that overrun the vertical line representing the end of the study.

So how do you record your data? If you were to simply count those individuals as dying at the study's end, you'd be introducing a strong bias to your data. They didn't really die; you just went home. So you decide that maybe the safest thing to do is to throw out those individuals from your data set entirely. In doing so, you are right-censoring your data: You are throwing out the data that run off the right side of the graph. The right-censored data are shown below.



Look at what seems to be happening here. Some chameleons born in 2013 and 2014 die early, but others live two or three years. From these right-censored data, all of those born in 2015 and 2016 appear to die early. If you didn't know that the data had been right-censored, you might conclude that 2015 and 2016 were very dangerous years to be a chameleon and be concerned about the long-term health of this population. This is misleading. The distribution of life spans is unchanged from 2013 to 2017. By throwing out the individuals who live past the end of the study, you are looking at all individuals born in 2013 and 2014, but only the short-lived among those that are born in 2015 and 2016. By right-censoring the data, you've created a misleading impression of mortality patterns.

The right-censoring problem that we see here can be viewed as a form of selection bias. The individuals in our sample from 2015 and 2016 are not a random sample of chameleons born in those years. Rather, they are precisely those individuals with the shortest life spans.

The same problem arises in the graph of average age at death for musicians. Rap and hip-hop are new genres, maybe forty years old at the most, and popular musicians tend to launch their careers in their teens or twenties. As a consequence, most rap and hip-hop stars are still alive today, and thus omitted from the study. The only rap and hip-hop musicians who have died already are those who have died prematurely. Jazz, blues, country, and gospel, by comparison, have been around for a century or more. Many deceased performers of these styles lived into their eighties or longer. It's not that rap stars will likely die young; it's that the rap stars who have died must have died young, because rap hasn't been around long enough for it to be otherwise.

To be fair to the study's author, she acknowledges the right-censoring issue in her article. The problem is that readers may think that the striking differences shown in the graph are due to differing mortality rates by musical genre, whereas we suspect the majority of the pattern is a result of right-censoring. We don't like being presented with graphs in which some or even most of the pattern is due to statistical artifact, and then being told by the author, "But some of the pattern is real, trust me."

Another problem is that there is no caveat about the right-censoring issue in the data graphic itself. In a social media environment, we have to be prepared for data graphics—at least those from pieces directed to a popular audience—to be shared without the accompanying text. In our view, these data never should have been plotted in the way that they were. The graph here tells a story that is not consistent with the conclusions that a careful analysis would suggest.

DISARMING SELECTION BIAS

We have looked at a number of ways in which selection bias can arise. We conclude by considering ways to deal with the problem.

Selection biases often arise in clinical trials, because physicians, insurance companies, and the individual patients in the clinical trial have a say in what treatments are assigned to whom. As a consequence, the treatment group who receive the intervention may differ in important ways from the control group who do not receive the intervention. Randomizing which individuals receive a particular treatment provides a way to minimize selection biases.

A recent study of employer wellness programs shows us just how important this can be. If you work for a large company, you may be a participant in such a program already. The exact structures of corporate wellness programs vary, but the approach is grounded in preventative medicine. Wellness programs often involve disease screening, health education, fitness activities, nutritional advice, weight loss, and stress management. Many wellness programs track employees' activity and other aspects of their health. Some even require employees to wear fitness trackers that provide fine-scale detail on individual activity levels. A majority of them offer incentives for engaging in healthy behaviors. Some reward employees for participating in activities or reaching certain fitness milestones. Others penalize unhealthy behaviors, charging higher premiums to those who smoke, are overweight, and so forth.

Wellness programs raise ethical questions about employers having this level of control and ownership over employees' bodies. But there is also a fundamental question: Do they work? To answer that, we have to agree about what wellness programs are supposed to do. Employers say that they offer these programs because they care about their employees and want to improve their quality of life. That's mostly bullshit. A sand volleyball court on the company lawn may have some recruiting benefits. But the primary rationale for implementing a wellness program is that by improving the health of its employees, a company can lower insurance costs, decrease absenteeism, and perhaps even reduce the rate at which employees leave the company. All of these elements contribute to the company's bottom line.

Companies are getting on board. By 2017, the workplace wellness business had ballooned into an eight-billion-dollar industry in the US alone. One report suggests that half of firms with more than fifty employees offer wellness programs of some sort, with average costs running well over $500 per employee per year.

Meta-analyses—studies that aggregate the results of previous studies—seem encouraging. Such studies typically conclude that wellness programs reduce medical costs and absenteeism, generating considerable savings for employers. But the problem with almost all of these studies is that they allow selection biases to creep in. They compare employees within the same company who did take part in wellness activities with those who did not. Employers cannot force employees to participate, however, and the people who choose to participate may differ in important ways from the people who choose not to. In particular, those who opt in may already be healthier and leading healthier lifestyles than those who opt out.

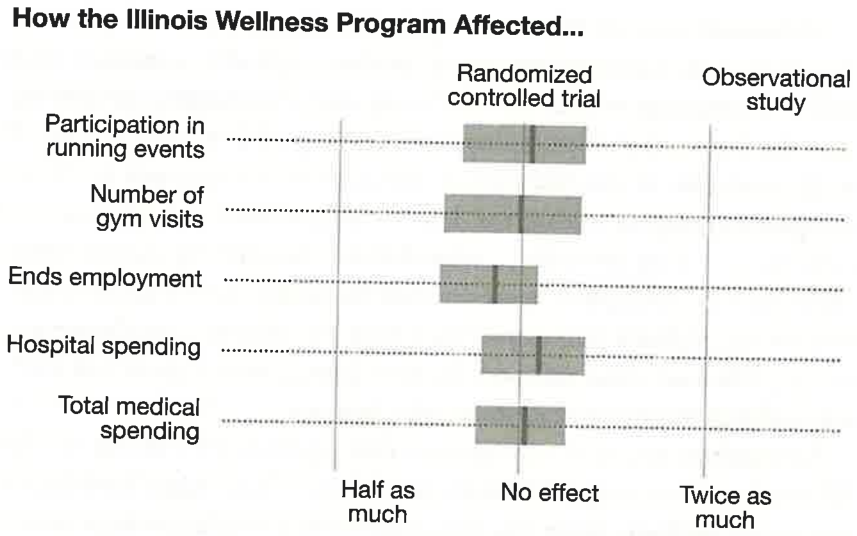
Researchers found a way around this problem in a recent study. When the University of Illinois at Urbana-Champaign launched a workplace wellness program, they randomized employees into either a treatment group or a control group. Members of the treatment group had the option of participating but were not required to do so. Members of the control group were not even offered an opportunity to take part. This design provided the researchers with three categories: people who chose to participate, people who chose not to, and people who were not given the option to participate in the first place.* The authors obtained health data for all participants from the previous thirteen months, affording a baseline for comparing health before and after taking part in the study.

Unlike previous observational studies, this study found that offering a wellness program had no statistically significant effect on healthcare costs or absenteeism, nor did it increase gym visits and similar types of physical activity. (The one useful thing that the wellness program did was to increase the fraction of participants who underwent a health screening.) The figure on page 132 summarizes these results.

The randomized controlled trial found that being offered the wellness program had no effect on fitness activities, employee retention, or medical costs.

But why? Previous studies largely found beneficial effects. Was there something particularly ineffective about how the Illinois well-

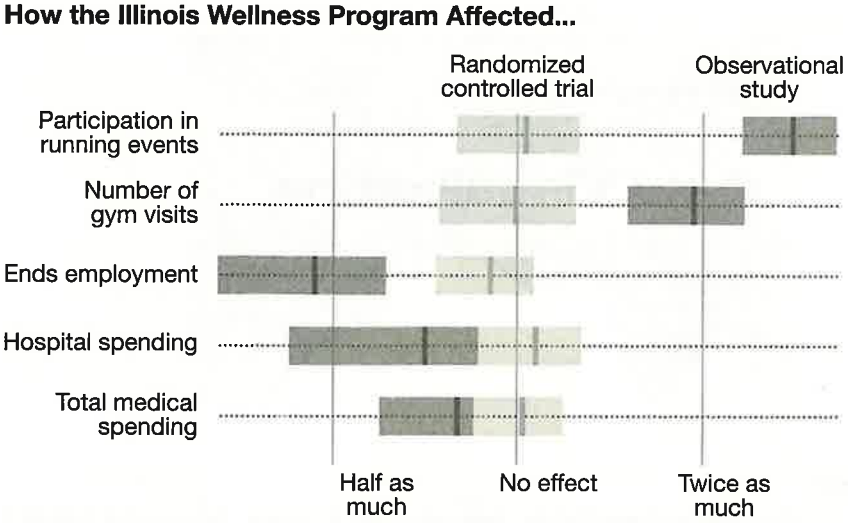
* Those offered the option to participate in the workplace wellness program were actually divided into six groups that received different payments, etc., for participation, but to keep the story simple we will not go into those details here.



ness program was designed? Or did this point to selection effects in the previous studies? In order to find out, the investigators ran a second analysis in which they set aside their control group, and looked only at the employees who were offered the chance to participate in the wellness program. Comparing those who were active in that program with those who were not, they found strong disparities in activity, retention, and medical costs between those who opted to participate and those who opted not to. These differences remained even when the authors tried to control for differences between those who participated and those who declined, such as age, gender, weight, and other characteristics. If these authors had done an observational study like previous researchers, they would also have observed that employees in workplace wellness programs were healthier and less likely to leave the firm.

This is a classic example of a selection effect. People who are in good health are more likely to participate in wellness programs. It's not that wellness programs cause good health, it's that being in good health causes people to participate in wellness programs.*

* The authors note that if a workplace wellness program differentially recruits or retains individuals in the best health, it can confer financial benefits to the company by screening for the healthiest employees. Even though the program itself doesn't make anyone healthier, it does improve the average health of individuals who choose to join or to stay with the company.



In this chapter, we looked at one of the most important ways that statistical analyses fail and mislead—namely, sampling nonrandomly. Another way to tell data stories is with visualizations and figures. Just like statistical analyses, visual presentations of data are subject to misuse as well. This will be our subject for the next chapter.