CHAPTER 7

The Importance of Data

"Garbage in, garbage out"

In the spring of 2012, researchers published a striking finding in the esteemed journal *Science*. According to this cutting-edge research, when male fruit flies are spurned repeatedly by female fruit flies, they drown their sorrows in alcohol. The *New York Times* described the study in a front page article: "They were young males on the make, and they struck out not once, not twice, but a dozen times with a group of attractive females hovering nearby. So they did what so many men do after being repeatedly rejected: they got drunk, using alcohol as a balm for unfulfilled desire."[1]

This research advances our understanding of the brain's reward system, which in turn can help us find new strategies for dealing with drug and alcohol dependence. A substance abuse expert described reading the study as "looking back in time, to see the very origins of the reward circuit that drives fundamental behaviors like sex, eating and sleeping."

Since I am not an expert in this field, I had two slightly different reactions upon reading about spurned fruit flies. First, it made me nostalgic for college. Second, my inner researcher got to wondering how fruit flies get drunk. Is there a miniature fruit fly bar, with assorted fruit based liquors and an empathetic fruit fly bartender? Is country western music playing in the background? Do fruit flies even like country western music?

It turns out that the design of the experiment was devilishly simple. One group of male fruit flies was allowed to mate freely with virgin females. Another group of males was released among female fruit flies that had already mated and were therefore indifferent to the males' amorous overtures. Both sets of male fruit flies were then offered feeding straws that offered a choice between standard fruit fly fare, yeast and sugar, and the "hard stuff": yeast, sugar, and 15 percent alcohol. The males who had spent days trying to mate with indifferent females were significantly more likely to hit the booze.

The levity notwithstanding, these results have important implications for humans. They suggest a connection between stress, chemical responses in the brain, and an appetite for alcohol. However, the results are not a triumph of statistics. They are a triumph of data, which made relatively basic statistical analysis possible. The genius of this study was figuring out a way to create a group of sexually satiated male fruit flies and a group of sexually frustrated male fruit flies—and then to find a way to compare their drinking habits. Once the researchers did that, the number crunching wasn't any more complicated than that of a typical high school science fair project.

Data are to statistics what a good offensive line is to a star quarterback. In front of every star quarterback is a good group of blockers. They usually don't get much credit. But without them, you won't ever see a star quarterback. Most statistics books assume that you are using good data, just as a cookbook assumes that you are not buying rancid meat and rotten vegetables. But even the finest recipe isn't going to salvage a meal that begins with spoiled ingredients. So it is with statistics; no amount of fancy analysis can make up for fundamentally flawed data. Hence the expression "garbage in, garbage out." Data deserve respect, just like offensive linemen.

We generally ask our data to do one of three things. First, we may demand a data sample that is representative of some larger group or population. If we are trying to gauge voters' attitudes toward a particular political candidate, we will need to interview a sample of prospective voters who are representative of all voters in the relevant political jurisdiction. (And

remember, we don't want a sample that is representative of everyone *living* in that jurisdiction; we want a sample of those *who are likely to vote.*) One of the most powerful findings in statistics, which will be explained in greater depth over the next two chapters, is that inferences made from reasonably large, properly drawn samples can be every bit as accurate as attempting to elicit the same information from the entire population.

The easiest way to gather a representative sample of a larger population is to select some subset of that population randomly. (Shockingly, this is known as a simple random sample.) The most attractive feature of this methodology is that each observation in the relevant population has an equal chance of being included in the sample. More specifically, if you plan to survey a simple random sample of 100 adults in a neighborhood with 4,328 adults, your methodology has to ensure that every possible sample of 100 adults has an equal chance of being surveyed, such as randomly dialing 100 phone numbers from a phonebook that includes the whole neighborhood. In turn, this random dialing (or a similar methodology) would ensure that each of those 4,328 residents has the same probability of ending up as one of the 100 adults who are surveyed. (For now, we'll assume unrealistically that all residents are equally likely to answer their phones.) Statistics books almost always illustrate this point by drawing colored marbles out of an urn. (In fact, it's about the only place where one sees the word "urn" used with any regularity.) If there are 60,000 blue marbles and 40,000 red marbles in a giant urn, then the most likely composition of a sample of 100 marbles drawn randomly from the urn would be 60 blue marbles and 40 red marbles. If we did this more than once, there would obviously be deviations from sample to sample—some might have 62 blue marbles and 38 red marbles, or 58 blue and 42 red. But the chances of drawing any random sample that deviates hugely from the composition of marbles in the urn are very, very low.

Now, admittedly, there are some practical challenges here. Most populations we care about tend to be more complicated than an urn full of marbles. How, exactly, would one select a random sample of the American adult population to be included in a telephone poll? Even a seemingly elegant solution like a telephone random dialer has potential flaws. Some individuals (particularly low-income persons) may not have a

telephone. Others (particularly high-income persons) may be more prone to screen calls and choose not to answer. Chapter 10 will outline some of the strategies that polling firms use to surmount these kinds of sampling challenges (most of which got even more complicated with the advent of cell phones). The key idea is that a properly drawn sample will look like the population from which it is drawn. In terms of intuition, you can envision sampling a pot of soup with a single spoonful. If you've stirred your soup adequately, a single spoonful can tell you how the whole pot tastes.

A statistics text will include far more detail on sampling methods. Polling firms and market research companies spend their days figuring out how to get good representative data from various populations in the most cost-effective way. For now, you should appreciate several important things: (1) A representative sample is a fabulously important thing, for it opens the door to some of the most powerful tools that statistics has to offer. (2) Getting a good sample is harder than it looks. (3) Many of the most egregious statistical assertions are caused by good statistical methods applied to bad samples, not the opposite. (4) Size matters, and bigger is better. The details will be explained in the coming chapters, but it should be intuitive that a larger sample will help to smooth away any freak variation. (A bowl of soup will be an even better test than a spoonful.) One crucial caveat is that a bigger sample will not make up for errors in its composition, or "bias." A bad sample is a bad sample. No supercomputer or fancy formula is going to rescue the validity of your *national* presidential poll if the respondents are drawn only from a telephone survey of Washington, D.C., residents. The residents of Washington, D.C., don't vote like the rest of America; calling 100,000 D.C. residents rather than 1,000 is not going to fix that fundamental problem with your poll. In fact, a large, biased sample is arguably worse than a small, biased sample because it will give a false sense of confidence regarding the results.

The second thing we often ask of data is that they provide some source of comparison. Is a new medicine more effective than the current treatment? Are ex-convicts who receive job training less likely to return to prison than ex-convicts who do not receive such training? Do students

who attend charter schools perform better than similar students who attend regular public schools?

In these cases, the goal is to find two groups of subjects who are broadly similar except for the application of whatever "treatment" we care about. In a social science context, the word "treatment" is broad enough to encompass anything from being a sexually frustrated fruit fly to receiving an income tax rebate. As with any other application of the scientific method, we are trying to isolate the impact of *one specific intervention or attribute*. This was the genius of the fruit fly experiment. The researchers figured out a way to create a control group (the males who mated) and a "treatment" group (the males who were shot down); the subsequent difference in their drinking behaviors can then be attributed to whether they were sexually spurned or not.

In the physical and biological sciences, creating treatment and control groups is relatively straightforward. Chemists can make small variations from test tube to test tube and then study the difference in outcomes. Biologists can do the same thing with their petri dishes. Even most animal testing is simpler than trying to get fruit flies to drink alcohol. We can have one group of rats exercise regularly on a treadmill and then compare their mental acuity in a maze with the performance of another group of rats that didn't exercise. But when humans become involved, things grow more complicated. Sound statistical analysis often requires a treatment and a control group, yet we cannot force people to do the things that we make laboratory rats do. (And many people do not like making even the lab rats do these things.) Do repeated concussions cause serious neurological problems later in life? This is a really important question. The future of football (and perhaps other sports) hangs on the answer. Yet it is a question that cannot be answered with experiments on humans. So unless and until we can teach fruit flies to wear helmets and run the spread offense, we have to find other ways to study the long-term impact of head trauma.

One recurring research challenge with human subjects is creating treatment and control groups that differ *only* in that one group is getting the treatment and the other is not. For this reason, the "gold standard" of research is randomization, a process by which human subjects (or

schools, or hospitals, or whatever we're studying) are randomly assigned to either the treatment or the control group. We do not assume that all the experimental subjects are identical. Instead, probability becomes our friend (once again), and we assume that randomization will evenly divide all relevant characteristics between the two groups—both the characteristics we can observe, like race or income, but also confounding characteristics that we cannot measure or had not considered, such as perseverance or faith.

The third reason we collect data is, to quote my teenage daughter, "Just because." We sometimes have no specific idea what we will do with the information—but we suspect it will come in handy at some point. This is similar to a crime scene detective who demands that all possible evidence be captured so that it can be sorted later for clues. Some of this evidence will prove useful, some will not. If we knew exactly what would be useful, we probably would not need to be doing the investigation in the first place.

You probably know that smoking and obesity are risk factors for heart disease. You probably don't know that a long-running study of the residents of Framingham, Massachusetts, helped to clarify those relationships. Framingham is a suburban town of some 67,000 people about twenty miles west of Boston. To nonresearchers, it is best known as a suburb of Boston with reasonably priced housing and convenient access to the impressive and upscale Natick Mall. To researchers, Framingham is best known as the home of the Framingham Heart Study, one of the most successful and influential longitudinal studies in the history of modern science.

A longitudinal study collects information on a large group of subjects at many different points in time, such as once every two years. The same participants may be interviewed periodically for ten, twenty, or even fifty years after they enter the study, creating a remarkably rich trove of information. In the case of the Framingham study, researchers gathered information on 5,209 adult residents of Framingham in 1948: height, weight, blood pressure, educational background, family structure, diet, smoking behavior, drug use, and so on. Most important, researchers have gath-

ered follow-up data *from the same participants* ever since (and also data on their offspring, to examine genetic factors related to heart disease). The Framingham data have been used to produce over two thousand academic articles since 1950, including nearly a thousand between 2000 and 2009.

These studies have produced findings crucial to our understanding of cardiovascular disease, many of which we now take for granted: cigarette smoking increases the risk of heart disease (1960); physical activity reduces the risk of heart disease and obesity increases it (1967); high blood pressure increases the risk of stroke (1970); high levels of HDL cholesterol (henceforth known as the "good cholesterol") reduce the risk of death (1988); individuals with parents and siblings who have cardiovascular disease are at significantly higher risk of the same (2004 and 2005).

Longitudinal data sets are the research equivalent of a Ferrari. The data are particularly valuable when it comes to exploring causal relationships that may take years or decades to unfold. For example, the Perry Preschool Study began in the late 1960s with a group of 123 African American three- and four-year-olds from poor families. The participating children were randomly assigned into a group that received an intensive preschool program and a comparison group that did not. Researchers then measured various outcomes for both groups *for the next forty years.* The results make a compelling case for the benefits of early childhood education. The students who received the intensive preschool experience had higher IQs at age five. They were more likely to graduate from high school. They had higher earnings at age forty. In contrast, the participants who did not receive the preschool program were significantly more likely to have been arrested five or more times by age forty.

Not surprisingly, we can't always have the Ferrari. The research equivalent of a Toyota is a cross-sectional data set, which is a collection of data gathered at a single point in time. For example, if epidemiologists are searching for the cause of a new disease (or an outbreak of an old one), they may gather data from all those afflicted in hopes of finding a pattern that leads to the source. What have they eaten? Where have they traveled? What else do they have in common? Researchers may also gather data from individuals who are not afflicted by the disease to highlight contrasts between the two groups.

In fact, all of this exciting cross-sectional data talk reminds me of the week before my wedding, when I became part of a data set. I was working in Kathmandu, Nepal, when I tested positive for a poorly understood stomach illness called "blue-green algae," which had been found in only two places in the world. Researchers had isolated the pathogen that caused the disease, but they were not yet sure what kind of organism it was, as it had never been identified before. When I called home to inform my fiancée about my diagnosis, I acknowledged that there was some bad news. The disease had no known means of transmission, no known cure, and could cause extreme fatigue and other unpleasant side effects for anywhere from a few days to many months.* With the wedding only one week away, yes, this could be a problem. Would I have complete control of my digestive system as I walked down the aisle? Maybe.

But then I really tried to focus on the good news. First, "blue-green algae" was thought to be nonfatal. And second, experts in tropical diseases from as far away as Bangkok had taken a personal interest in my case. *How cool is that?* (Also, I did a terrific job of repeatedly steering the discussion back to the wedding planning: "Enough about my incurable disease. Tell me more about the flowers.")

I spent my final hours in Kathmandu filling out a thirty-page survey describing every aspect of my life: Where did I eat? What did I eat? How did I cook? Did I go swimming? Where and how often? Everyone else who had been diagnosed with the disease was doing the same thing. Eventually the pathogen was identified as a water-borne form of cyanobacteria. (These bacteria are blue, and they are the only kind of bacteria that get their energy from photosynthesis; hence the original description of the disease as "blue-green algae.") The illness was found to respond to treatment with traditional antibiotics, but, curiously, not to some of the newer ones. All of these discoveries were too late to help me, but I was lucky enough to recover quickly anyway. I had near-perfect control of my digestive system by wedding day.

---

* At the time, the disease had a mean duration of forty-three days with a standard deviation of twenty-four days.

Behind every important study there are good data that made the analysis possible. And behind every bad study . . . well, read on. People often speak about "lying with statistics." I would argue that some of the most egregious statistical mistakes involve *lying with data*; the statistical analysis is fine, but the data on which the calculations are performed are bogus or inappropriate. Here are some common examples of "garbage in, garbage out."

Selection bias. Pauline Kael, the longtime film critic for *The New Yorker*, is alleged to have said after Richard Nixon's election as president, "Nixon couldn't have won. I don't know anyone who voted for him." The quotation is most likely apocryphal, but it's a lovely example of how a lousy sample (one's group of liberal friends) can offer a misleading snapshot of a larger population (voters from across America). And it introduces the question one should always ask: How have we chosen the sample or samples that we are evaluating? If each member of the relevant population does not have an equal chance of ending up in the sample, we are going to have a problem with whatever results emerge from that sample. One ritual of presidential politics is the Iowa straw poll, in which Republican candidates descend on Ames, Iowa, in August of the year before a presidential election to woo participants, each of whom pays $30 to cast a vote in the poll. The Iowa straw poll does not tell us that much about the future of Republican candidates. (The poll has predicted only three of the last five Republican nominees.) Why? Because Iowans who pay $30 to vote in the straw poll are different from other Iowa Republicans; and Iowa Republicans are different from Republican voters in the rest of the country.

Selection bias can be introduced in many other ways. A survey of consumers in an airport is going to be biased by the fact that people who fly are likely to be wealthier than the general public; a survey at a rest stop on Interstate 90 may have the opposite problem. Both surveys are likely to be biased by the fact that people who are willing to answer a survey in a public place are different from people who would prefer not to be bothered. If you ask 100 people in a public place to complete a short survey, and 60 are willing to answer your questions, *those 60 are likely to be different in significant ways from the 40 who walked by without making eye contact.*

One of the most famous statistical blunders of all time, the notorious *Literary Digest* poll of 1936, was caused by a biased sample. In that year, Kansas governor Alf Landon, a Republican, was running for president against incumbent Franklin Roosevelt, a Democrat. *Literary Digest*, an influential weekly news magazine at the time, mailed a poll to its subscribers and to automobile and telephone owners whose addresses could be culled from public records. All told, the *Literary Digest* poll included 10 million prospective voters, which is an astronomically large sample. As polls with good samples get larger, they get better, since the margin of error shrinks. As polls with bad samples get larger, the pile of garbage just gets bigger and smellier. *Literary Digest* predicted that Landon would beat Roosevelt with 57 percent of the popular vote. In fact, Roosevelt won in a landslide, with 60 percent of the popular vote and forty-six of forty-eight states in the electoral college. The *Literary Digest* sample was "garbage in": the magazine's subscribers were wealthier than average Americans, and therefore more likely to vote Republican, as were households with telephones and cars in 1936.[2]

We can end up with the same basic problem when we compare outcomes between a treatment and a control group if the mechanism for sorting individuals into one group or the other is not random. Consider a recent finding in the medical literature on the side effects of treatment for prostate cancer. There are three common treatments for prostate cancer: surgical removal of the prostate; radiation therapy; or brachytherapy (which involves implanting radioactive "seeds" near the cancer).[3] Impotence is a common side effect of prostate cancer treatment, so researchers have documented the sexual function of men who receive each of the three treatments. A study of 1,000 men found that two years after treatment, 35 percent of the men in the surgery group were able to have sexual intercourse, compared with 37 percent in the radiation group and 43 percent in the brachytherapy group.

Can one look at these data and assume that brachytherapy is least likely to damage a man's sexual function? No, no, no. The authors of the study explicitly warn that we cannot conclude that brachytherapy is better at preserving sexual function, since the men who receive this treatment are generally younger and fitter than men who receive the other treat-

ment. The purpose of the study was merely to document the degree of sexual side effects across all types of treatment.

A related source of bias, known as self-selection bias, will arise whenever individuals volunteer to be in a treatment group. For example, prisoners who volunteer for a drug treatment group are different from other prisoners *because they have volunteered to be in a drug treatment program.* If the participants in this program are more likely to stay out of prison after release than other prisoners, that's great—but it tells us absolutely nothing about the value of the drug treatment program. These former inmates may have changed their lives because the program helped them kick drugs. Or they may have changed their lives because of other factors that also happened to make them more likely to volunteer for a drug treatment program (such as having a really strong desire not to go back to prison). We cannot separate the causal impact of one (the drug treatment program) from the other (being the kind of person who volunteers for a drug treatment program).

Publication bias. Positive findings are more likely to be published than negative findings, which can skew the results that we see. Suppose you have just conducted a rigorous, longitudinal study in which you find conclusively that playing video games *does not* prevent colon cancer. You've followed a representative sample of 100,000 Americans for twenty years; those participants who spend hours playing video games have roughly the same incidence of colon cancer as the participants who do not play video games at all. We'll assume your methodology is impeccable. Which prestigious medical journal is going to publish your results?

None, for two reasons. First, there is no strong scientific reason to believe that playing video games has any impact on colon cancer, so it is not obvious why you were doing this study. Second, and more relevant here, the fact that something *does not* prevent cancer is not a particularly interesting finding. After all, most things *don't* prevent cancer. Negative findings are not especially sexy, in medicine or elsewhere.

The net effect is to distort the research that we see, or do not see. Suppose that one of your graduate school classmates has conducted a different longitudinal study. She finds that people who spend a lot of time

playing video games *do* have a lower incidence of colon cancer. *Now that is interesting!* That is exactly the kind of finding that would catch the attention of a medical journal, the popular press, bloggers, and video game makers (who would slap labels on their products extolling the health benefits of their products). It wouldn't be long before Tiger Moms all over the country were "protecting" their children from cancer by snatching books out of their hands and forcing them to play video games instead.

Of course, one important recurring idea in statistics is that unusual things happen every once in a while, just as a matter of chance. If you conduct 100 studies, one of them is likely to turn up results that are pure nonsense—like a statistical association between playing video games and a lower incidence of colon cancer. Here is the problem: The 99 studies that find no link between video games and colon cancer will not get published, because they are not very interesting. The one study that does find a statistical link will make it into print and get loads of follow-on attention. The source of the bias stems not from the studies themselves but from the skewed information that actually reaches the public. Someone reading the scientific literature on video games and cancer would find only a single study, and that single study will suggest that playing video games can prevent cancer. In fact, 99 studies out of 100 would have found no such link.

Yes, my example is absurd—but the problem is real and serious. Here is the first sentence of a *New York Times* article on the publication bias surrounding drugs for treating depression: "The makers of antidepressants like Prozac and Paxil never published the results of about a third of the drug trials that they conducted to win government approval, misleading doctors and consumers about the drugs' true effectiveness."[4] It turns out that 94 percent of studies with positive findings on the effectiveness of these drugs were published, while only 14 percent of the studies with nonpositive results were published. For patients dealing with depression, this is a big deal. When all the studies are included, the antidepressants are better than a placebo by only "a modest margin."

To combat this problem, medical journals now typically require that any study be registered at the beginning of the project if it is to be eligible for publication later on. This gives the editors some evidence on

the ratio of positive to nonpositive findings. If 100 studies are registered that propose to examine the effect of skateboarding on heart disease, and only one is ultimately submitted for publication with positive findings, the editors can infer that the other studies had nonpositive findings (or they can at least investigate this possibility).

Recall bias. Memory is a fascinating thing—though not always a great source of good data. We have a natural human impulse to understand the present as a logical consequence of things that happened in the past—cause and effect. The problem is that our memories turn out to be "systematically fragile" when we are trying to explain some particularly good or bad outcome in the present. Consider a study looking at the relationship between diet and cancer. In 1993, a Harvard researcher compiled a data set comprising a group of women with breast cancer and an age-matched group of women who had not been diagnosed with cancer. Women in both groups were asked about their dietary habits earlier in life. The study produced clear results: The women with breast cancer were significantly more likely to have had diets that were high in fat when they were younger.

Ah, but this wasn't actually a study of how diet affects the likelihood of getting cancer. *This was a study of how getting cancer affects a woman's memory of her diet earlier in life.* All of the women in the study had completed a dietary survey years earlier, before any of them had been diagnosed with cancer. The striking finding was that women with breast cancer recalled a diet that was much higher in fat than what they actually consumed; the women with no cancer did not. The *New York Times Magazine* described the insidious nature of this recall bias:

> The diagnosis of breast cancer had not just changed a woman's present and the future; it had altered her past. Women with breast cancer had (unconsciously) decided that a higher-fat diet was a likely predisposition for their disease and (unconsciously) recalled a high-fat diet. It was a pattern poignantly familiar to anyone who knows the history of this stigmatized illness: these women, like thousands of women before them, had searched their own memories for a cause and then summoned that cause into memory.[5]

Recall bias is one reason that longitudinal studies are often preferred to cross-sectional studies. In a longitudinal study the data are collected contemporaneously. At age five, a participant can be asked about his attitudes toward school. Then, thirteen years later, we can revisit that same participant and determine whether he has dropped out of high school. In a cross-sectional study, in which all the data are collected at one point in time, we must ask an eighteen-year-old high school dropout how he or she felt about school at age five, which is inherently less reliable.

Survivorship bias. Suppose a high school principal reports that test scores for a particular cohort of students has risen steadily for four years. The sophomore scores for this class were better than their freshman scores. The scores from junior year were better still, and the senior year scores were best of all. We'll stipulate that there is no cheating going on, and not even any creative use of descriptive statistics. Every year this cohort of students has done better than it did the preceding year, by every possible measure: mean, median, percentage of students at grade level, and so on.

Would you (a) nominate this school leader for "principal of the year" or (b) demand more data?

I say "b." I smell survivorship bias, which occurs when some or many of the observations are falling out of the sample, changing the composition of the observations that are left and therefore affecting the results of any analysis. Let's suppose that our principal is truly awful. The students in his school are learning nothing; each year half of them drop out. Well, that could do very nice things for the school's test scores—without any individual student testing better. If we make the reasonable assumption that the worst students (with the lowest test scores) are the most likely to drop out, then the average test scores of those students left behind will go up steadily as more and more students drop out. (If you have a room of people with varying heights, forcing the short people to leave will raise the average height in the room, but it doesn't make anyone taller.)

The mutual fund industry has aggressively (and insidiously) seized on survivorship bias to make its returns look better to investors than

they really are. Mutual funds typically gauge their performance against a key benchmark for stocks, the Standard & Poor's 500, which is an index of 500 leading public companies in America.* If the S&P 500 is up 5.3 percent for the year, a mutual fund is said to beat the index if it performs better than that, or trail the index if it does worse. One cheap and easy option for investors who don't want to pay a mutual fund manager is to buy an S&P 500 Index Fund, which is a mutual fund that simply buys shares in all 500 stocks in the index. Mutual fund managers like to believe that they are savvy investors, capable of using their knowledge to pick stocks that will perform better than a simple index fund. In fact, it turns out to be relatively hard to beat the S&P 500 for any consistent stretch of time. (The S&P 500 is essentially an average of all large stocks being traded, so just as a matter of math we would expect roughly half the actively managed mutual funds to outperform the S&P 500 in a given year and half to underperform.) Of course, it doesn't look very good to lose to a mindless index that simply buys 500 stocks and holds them. No analysis. No fancy macro forecasting. And, much to the delight of investors, no high management fees.

What is a traditional mutual fund company to do? Bogus data to the rescue! Here is how they can "beat the market" without beating the market. A large mutual company will open many new actively managed funds (meaning that experts are picking the stocks, often with a particular focus or strategy). For the sake of example, let's assume that a mutual fund company opens twenty new funds, each of which has roughly a 50 percent chance of beating the S&P 500 in a given year. (This assumption is consistent with long-term data.) Now, basic probability suggests that only ten of the firm's new funds will beat the S&P 500 the first year; five funds will beat it two years in a row; and two or three will beat it three years in a row.

* The S&P 500 is a nice example of what an index can and should do. The index is made up of the share prices of the 500 leading U.S. companies, each weighted by its market value (so that bigger companies have more weight in the index than smaller companies). The index is a simple and accurate gauge of what is happening to the share prices of the largest American companies at any given time.

Here comes the clever part. At that point, the new mutual funds with unimpressive returns relative to the S&P 500 are quietly closed. (Their assets are folded into other existing funds.) The company can then heavily advertise the two or three new funds that have "consistently outperformed the S&P 500"—even if that performance is the stock-picking equivalent of flipping three heads in a row. The subsequent performance of these funds is likely to revert to the mean, albeit after investors have piled in. The number of mutual funds or investment gurus who have consistently beaten the S&P 500 over a long period is shockingly small.*

Healthy user bias. People who take vitamins regularly are likely to be healthy—*because they are the kind of people who take vitamins regularly!* Whether the vitamins have any impact is a separate issue. Consider the following thought experiment. Suppose public health officials promulgate a theory that all new parents should put their children to bed only in purple pajamas, because that helps stimulate brain development. Twenty years later, longitudinal research confirms that having worn purple pajamas as a child does have an overwhelmingly large positive association with success in life. We find, for example, that 98 percent of entering Harvard freshmen wore purple pajamas as children (and many still do) compared with only 3 percent of inmates in the Massachusetts state prison system.

Of course, the purple pajamas do not matter; but having the kind of parents who put their children in purple pajamas *does matter*. Even when we try to control for factors like parental education, we are still going to be left with unobservable differences between those parents who obsess about putting their children in purple pajamas and those who don't. As *New York Times* health writer Gary Taubes explains, "At its simplest, the problem is that people who faithfully engage in activities that are good for them—taking a drug as prescribed, for instance, or eating what they believe is a healthy diet—are fundamentally different from those who

* For a very nice discussion of why you should probably buy index funds rather than trying to beat the market, read *A Random Walk Down Wall Street*, by my former professor Burton Malkiel.

don't."[6] This effect can potentially confound any study trying to evaluate the real effect of activities perceived to be healthful, such as exercising regularly or eating kale. We think we are comparing the health effects of two diets: kale versus no kale. In fact, if the treatment and control groups are not randomly assigned, we are comparing two diets that are being eaten by two different kinds of people. We have a treatment group that is different from the control group in two respects, rather than just one.

If statistics is detective work, then the data are the clues. My wife spent a year teaching high school students in rural New Hampshire. One of her students was arrested for breaking into a hardware store and stealing some tools. The police were able to crack the case because (1) it had just snowed and there were tracks in the snow leading from the hardware store to the student's home; and (2) the stolen tools were found inside. Good clues help.

Like good data. But first you have to get good data, and that is a lot harder than it seems.