11
Experiments

Statistical questions versus causal questions

WHY have some nations become rich while others have remained poor? Do small class sizes improve student achievement? Does following a Mediterranean diet rich in vegetables and olive oil reduce your risk of a heart attack? Does a "green" certification (like LEED, for Leadership in Energy and Environmental Design) improve the value of a commercial property?

Questions of cause and effect like these are, fundamentally, questions about *counterfactual statements.* A counterfactual is an if–then statement about something that has not actually occurred. For example: "If Colt McCoy had not been injured early in the 2010 National Championship football game, then the Texas Longhorns would have beaten Alabama." If you judge this counterfactual statement to be true—and who but the most hopelessly blinkered Crimson Tide fan doesn't?—then you might say that Colt McCoy's injury caused the Longhorns' defeat.

Statistical questions, on the other hand, are about correlations. This makes them fundamentally different from causal questions.

- Causal: "If we invested more money in our school system, how much faster would our economy grow?" Statistical: "In looking at data on a lot of countries, how are education spending and economic growth related?"

- Causal: "If I ate more vegetables than I do now, how much longer would I live?" Statistical: "Do people who eat a lot of vegetables live longer, on average, than people who don't?"

- Causal: "If we hire extra teachers at our school and reduce our class sizes, will our students' test scores improve?" Statistical: "Do students in smaller classes tend to have higher test scores?"

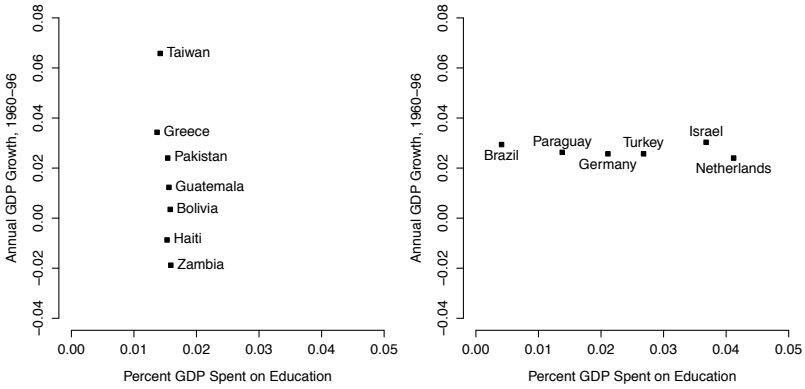

Figure 11.1: Two egregious examples of selective reporting.

Causal questions all invoke some kind of hypothetical intervention, where one thing is changed and everything else is held equal. In such a hypothetical intervention, there is no competing explanation for what might be causing the change we expect to see—in our economy, our lifespan, our students' test scores, a football game, or whatever outcome we're interested in.

Statistical questions, on the other hand, are about the patterns we observe in the real world. And the real world is rarely so simple as the hypothetical interventions we imagine. For example, people who eat more vegetables live longer—that's a clear pattern. But those same people also tend to exercise more, live in better housing, and have higher-status jobs. These other factors are *confounders*. A confounder is a competing explanation—some other factor correlated with both the "treatment" assignment (whether someone eats vegetables) and the response (lifespan). So in light of these confounders, how do we know it's the vegetables, rather than all that other stuff, that's making veggie-eaters live longer?

This is just a specific version of the general question we'll address in this chapter: under what circumstances can causal questions be answered using statistics?

Good evidence . . . and bad

Most of the cause-and-effect reasoning that you'll see out there in the real world is of depressingly poor quality. A common flaw is *cherry picking*: that is, pointing to data that seems to confirm some argument, while ignoring contradictory data.

Here's an example. In the left panel in Figure 11.1 we see a

group of seven countries that all spend around 1.5% of their GDP on education, but with very different rates of economic growth for the 37 years spanning 1960 to 1996. In the right panel, we see another group of six countries with very different levels of spending on education, but similar growth rates of 2–3%.

Both highly selective samples make it seem as though education and economic growth are barely related. If presented with the left panel alone, you'd be apt to conclude that the differences in growth rates must have been caused by something other than differences in education spending (of which there are none). Likewise, if presented with the right panel alone, you'd be apt to conclude that the large observed differences in education spending don't seem to have produced any difference in growth rates. The problem here isn't with the data—it's with the biased, highly selective *use* of that data.

This point seems almost obvious. Yet how tempting it is just to cherry pick and ignore the messy reality. Perhaps without even realizing it, we're all accustomed to seeing news stories that marshal highly selective evidence—usually even worse than that of Figure 11.1—on behalf of some plausible because-I-said-so story:

> [H]igher levels of education are critical to economic growth. . . . Boston, where there is a high proportion of college graduates, is the perfect example. Well-educated people can react more quickly to technological changes and learn new skills more readily. Even without the climate advantages of a city like San Jose, California, Boston evolved into what we now think of as an "information city." By comparison, Detroit, with lower levels of education, languished.[1]

[1] "Economic Scene." *New York Times* (Business section); August 5, 2004

And this from a reporter who presumably has no hidden agenda. Notice how the selective reporting of evidence—one causal hypothesis, two data points—lends an air of such graceful inevitability to what is a startlingly superficial analysis of the diverging economic fates of Boston and Detroit over the last half century.

Of course, most bad arguments are harder to detect than this howler from the New York Times. After all, using data to understand cause-and-effect relationships is hard. For example, consider the following summary of a recent neuroscience study:

> A study presented at the Society for Neuroscience meeting, in San Diego last week, shows people who start using marijuana at a young age have more cognitive shortfalls. Also, the more marijuana a person used in adolescence, the more trouble they had with focus and attention. "Early onset smokers

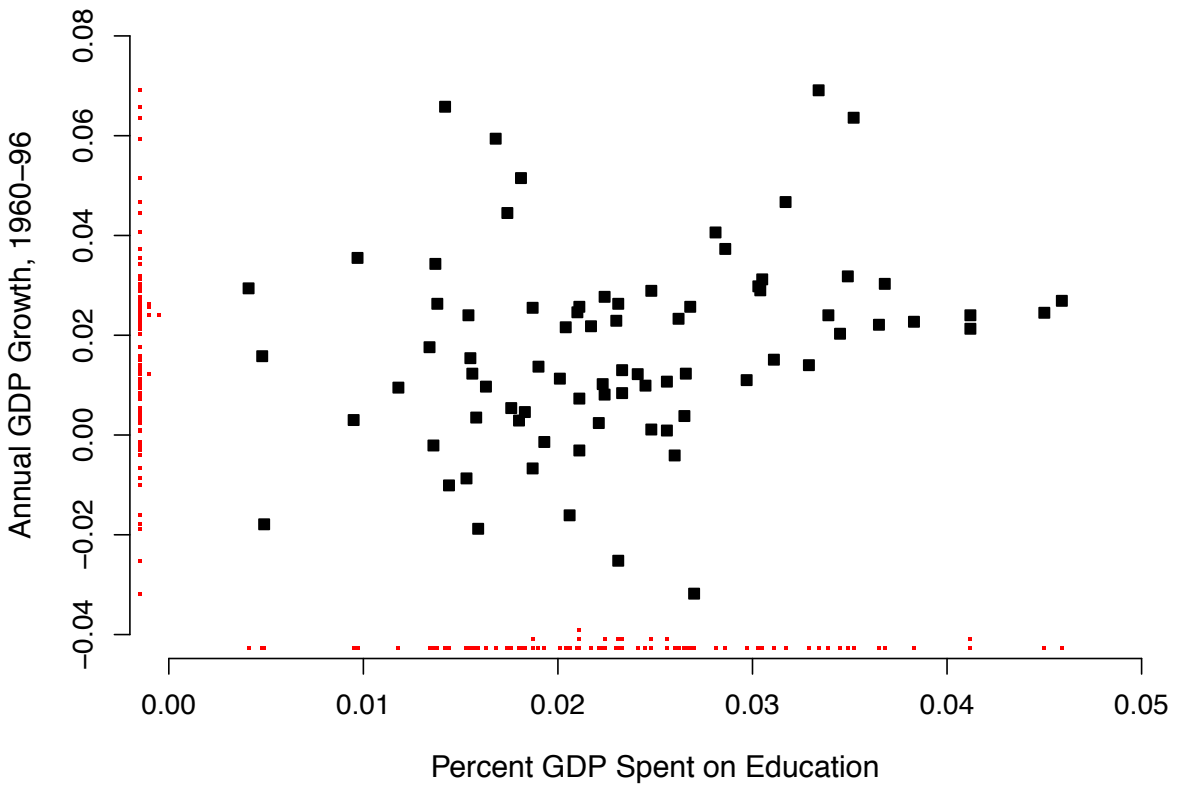

Figure 11.2: A scatter plot of GDP growth versus education spending for 79 countries. The tiny red dots clustered near the x and y axes are called *rug plots*. They are miniature histograms aligned with the axes of the predictor and the response.

have a different pattern of brain activity, plus got far fewer correct answers in a row and made way more errors on certain cognitive tests," says study author Staci Gruber.[2]

Did the marijuana smokers get less smart, or were the less-smart kids more likely to pick up a marijuana habit in the first place? It's an important question to consider in making drug policy, especially for states and countries where marijuana is legal. But can we know the answer on the basis of a study like this?

For another example, consider the bigger sample of countries in Figure 11.2, which provides a much more representative body of evidence on the GDP-versus-education story. This evidence takes the form of a scatter plot of GDP growth versus education spending for a sample of 79 countries worldwide. Notice the following two facts:

(1) Of the 29 countries that spent less than 2% of GDP on education, 18 fall below the median growth rate (1.58%).

[2] www.usatoday.com/yourlife/health/medical/pediatrics/2010-11-20-teendrugs22_ST_N.htm

(2) Of the 18 countries that spent more than 3% of GDP on education, 16 fall above the median growth rate.

These two facts, together with the upward trend in the scatter plot, suggest that economic growth and education spending are correlated. But this does not settle the causal question. For example, it might be that countries spend a lot on education because they are rich, rather the other way around.

The generic difficulty is that there are many different ways that two variables X and Y can appear correlated.

(1) One-way causality: the first domino falls, then the second; the rain falls, and the grass gets wet. (X causes Y directly.)

(2) Two-way causality: flowers and honey bees prosper together. (Both X and Y play a role in causing each other.)

(3) Common cause: People who go to college tend to get higher-paying jobs than those who don't. Does education directly lead to better economic outcomes? Or are a good education and a good job both just markers of a person's underlying qualities? (The role of X in causing Y is hard to distinguish from the role of C, which we may not have observed.)

(4) Common effect: either musical talent (X) or athletic talent (Y) will help you get into Harvard (Z). Among a population of Harvard freshmen, musical and athletic talent will thus appear negatively correlated, even if they are independent in the wider population. (X and Y both contribute to some common outcome C, inducing a correlation among a subset of the population defined by Z. This is often called Berkson's paradox; it is subtle, and we'll encounter it again.)

(5) Luck: the observed correlation is a coincidence.

This is the point where most books remind you that "correlation does not imply causation." Obviously. But if not to illuminate causes, what is the point of looking for correlations? Of course correlation does not imply causality, or else playing professional basketball would make you tall. But that hasn't stopped humans from learning that smoking causes cancer, or that lightning causes thunder, on the basis of observed correlations. The important question is: what distinguishes the good evidence-based arguments from the bad?

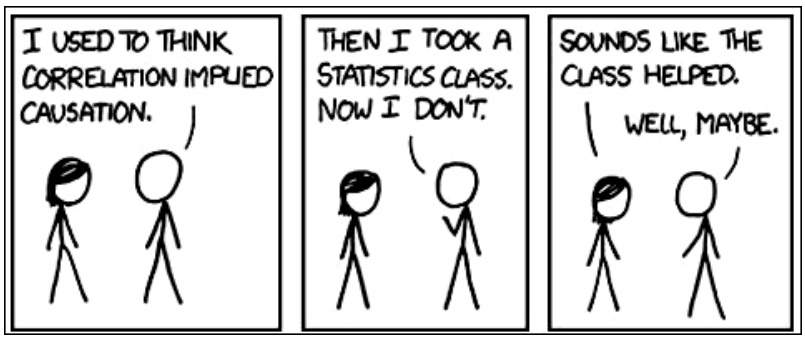

Figure 11.3: Originally published online at xkcd: http://xkcd.com/552/

Four common identification strategies

The key principle in using evidence to draw causal conclusions is that of a *balanced comparison*. To make things simple, we'll imagine that our predictor X is binary (i.e. has two groups), and we'll borrow the lingo of a clinical trial by referring to the two groups as the "treatment" and "control." To reach the conclusion that X causes Y, you must do two things: (1) *compare cases* in the treatment and control groups, to see how their Y values differ; and (2) *ensure balance*, by removing all other systematic differences between the cases in the treatment and control groups. Balance is crucial; it's what allows us to conclude that the differences in X (and not something else) cause the differences we observe in Y.

In general, there are four common ways to make a balanced comparison. These are often called *identification strategies*, in the sense that they are strategies for identifying a causal effect.

(1) Run a real experiment, randomizing subjects to the treatment and control groups. The randomization will ensure that, on average, there are no systematic differences between the two groups, other than the treatment.

(2) Find a natural experiment: that is, find a situation where the way that cases fall naturally into the treatment and control groups plausibly resembles a random assignment.

(3) Matching: artificially construct a balanced data set from an unbalanced data set, by explicitly matching treated cases with similar control cases, and discarding the cases without a good match. This will correct for lack of balance between control and treatment groups.

(4) Modeling: use regression modeling to adjust for confounders and isolate a partial relationship between the response and the treatment of interest. This an extension of the idea of statistical adjustment or "making fair comparisons" idea that we encountered in fitting simple straight lines.

In this chapter, we'll focus on the first option: run an experiment. In later chapters, we'll talk about the other ideas.

The power of experiment

The idea of an experiment is simple. If you want to know what would happen if you intervened in some system, then you should intervene, and measure what happens. There is simply no better way to establish that one thing causes another.

Indeed, one kind of experiment—the randomized, controlled clinical trial—is one of the most important medical innovations in history. Suppose we want to establish whether a brand new cholesterol drug—we'll call it Zapaclot—works better than the old drug. Also suppose that we've successfully recruited a large cohort of patients with high cholesterol. We know that diet and genes play a role here, but that drugs can help, too. We express this as

$$\text{Cholesterol} \sim \text{Diet} + \text{Genes} + \text{Drugs}\,.$$

Interpret the plus sign as the word "and," not like formal addition: we're assuming that cholesterol depends upon diet, genes, and drugs, although we haven't said how. Of course, it's that third predictor in the model we care about; the first two, in addition to some others that we haven't listed, are potential confounders.

First, what not to do: don't proceed by giving Zapaclot to all the men and the old drug to all the women, or Zapaclot to all the marathon runners and the old drug to the couch potatoes. These highly non-random assignments would obviously bias any judgment about the relative effect of the new drug compared to the old one. We refer to this sort of thing as *selection bias*: that is, any bias in the selection of cases that receive the treatment. Moreover, you shouldn't just give the new drug to whomever wants it, or can afford it. The people with more engagement, more knowledge, more money, or more trust in the medical system would probably sign up in greater numbers—and if those people

have systematic differences in diet or genes from the people who don't sign up, then you've just created a hidden selection bias.

Instead, you should two simple steps.

Randomize: randomly split the cohort into two groups, denoted the treatment group and the control group.

Intervene: allocate everyone in the treatment group to take the treatment (e.g. Zapaclot, the new drug), and everyone in the control group to take something else (e.g. the old drug or a placebo).[3]

[3] Everyone in the control group should be taking the *same* something else, whether it's the old drug or a placebo.

Randomize and intervene: a simple prescription, but the surest way to establish causality. The intervention allows you to pick up a difference between the new and old drug, if there's one to be found. The randomization ensures that other factors—even unknown factors, in addition to known ones like diet and lifestyle—do not lead us astray in our causal reasoning. The Latin phrase *ceteris paribus*, which translates roughly as "everything else being equal," is often used to describe such a situation. By randomizing and intervening, we have ensured that the only *systematic* difference between the groups is the treatment itself. The randomization gives us a balanced comparison.

This last point is crucial. It's not that diet, genes, and other lifestyle factors somehow stop affecting a patient's cholesterol level when we randomize and intervene. It's just that diet, genes, and lifestyle factors aren't correlated with the treatment assignment, and so they're balanced between the two groups, on average.

The need to avoid selection bias sounds obvious. But if selection bias in medical trials were not rigorously policed, then it would be easy for doctors to cherry pick healthy patients for newly proposed treatments. After all, a doctor who invents a new, seemingly effective form of treatment will almost surely become both rich and famous. As one physician reminisces:

> One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction. At the end of the lecture, a young student at the back of the room timidly asked, "Do you have any controls?" Well, the great surgeon drew himself up to his full height, hit the desk, and said, "Do you mean did I not operate on half of the patients?" The hall grew very quiet then. The voice at the back of the room very hesitantly replied, "Yes, that's what I had in mind." Then the

> visitor's fist really came down as he thundered, "Of course not. That would have doomed half of them to their death." God, it was quiet then, and one could scarcely hear the small voice ask, "Which half?"[4]

[4] Dr. E. Peacock, University of Arizona. Originally quoted in *Medical World News* (September 1, 1972). Reprinted pg. 144 of *Beautiful Evidence*, Edward Tufte (Graphics Press, 2006).

These last two words—"Which half?"—should echo in your mind whenever you are asked to judge the quality of evidence offered in support of a causal hypothesis. There is simply no substitute for a controlled experiment: not a booming authoritative voice, not even fancy statistics.

In fact, government regulators are so fastidious in their attention to possible selection biases that, in most real clinical trials, neither the doctors nor the patients are allowed to know which drug each person receives. Such a "double-blind" experiment avoids the possibility that patients might simply imagine that the the latest miracle drug has made them feel better, in a feat of unconscious self-deception called the placebo effect.

A placebo, from the Latin *placere* ("to please"), is a fake treatment designed to simulate the real one.

Some history

The notion of a controlled experiment was certainly around in pre-Christian times. The first chapter of the book of Daniel relates the tale of one such experiment. Daniel and his three friends Hananiah, Mishael, and Azariah arrive in the court of Nebuchadnezzar, the King of Babylon. They enroll in a Babylonian school, and are offered a traditional Babylonian diet. But Daniel wishes not to "defile himself with the portion of the king's meat, nor with the wine which he drank." He goes to Melzar, the prince of the eunuchs, who is in charge of the school. Daniel asks not to be made to eat the meat or drink the wine. But Melzar responds that he fears for Daniel's health if he were to let them follow some crank new-age diet. More to the point, Melzar observes, if the new students were to fall ill, "then shall ye make me endanger my head to the king."

So Daniel proposes a trial straight out of a statistics textbook:

> Prove thy servants, I beseech thee, ten days; and let them give us pulse to eat, and water to drink.
> Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the king's meat: and as thou seest, deal with thy servants.[5]

[5] King James Bible, Daniel 1:12–13.

The King agreed. When Daniel and his friends were inspected ten days later, "their countenances appeared fairer and fatter in flesh" than all those who had eaten meat and drank wine. Suitably

impressed, Nebuchadnezzar brings Daniel and his friends in for an audience, and he finds that "in all matters of wisdom and understanding," they were "ten times better than all the magicians and astrologers that were in all his realm."

As for a placebo-controlled trial, in which some of the patients are intentionally given a useless treatment (the "placebo"): that came much later.[6] The first such trial seems to have taken place in 1784. It was directed by none other than Benjamin Franklin, the American ambassador to the court of King Louis XVI of France. A German doctor by the name of Franz Mesmer had gained some degree of notoriety in Europe for his claim to have discovered a new force of nature that he called "magnétisme animal," and which was said to have magical healing powers. The demand for Dr. Mesmer's services soon took off among the ladies of Parisian high society, whom he would "Mesmerize" using a wild contraption involving ropes and magnetized iron rods.

[6] See "The Power of Nothing" in the December 12, 2011 edition of *The New Yorker* (pp. 30–6).

Much to the king's dismay, his own wife, Marie Antoinette, was one of Mesmer's keenest followers. The king found the whole Mesmerizing thing frankly a bit dubious, and presumably wished for his wife to have nothing to do with the Herr Doctor's magnétisme animal. So he convened several members of the French Academy of Sciences to investigate whether Dr. Mesmer had indeed discovered a new force of nature. The panel included Antoine Lavoisier, the father of modern chemistry, along with Joseph Guillotin, whose own wild contraption was soon to put the King's difficulties with Mesmer into perspective. Under Ben Franklin's supervision, the scientists set up an experiment to replicate some of Dr. Mesmer's prescribed treatments, substituting non-magnetic materials—history's first placebo—for half of the patients. In many cases, even the patients in the control group would flail about and start talking in tongues anyway. The panel concluded that the doctor's method produced no effect other than in the patients' own minds. Mesmer was denounced as a charlatan, although he continues to exact his revenge via the dictionary.

A more recent and especially striking example of a placebo comes from Thomas Freeman, director of the neural reconstruction unit at Tampa General Hospital in Florida. Dr. Freeman performs placebo brain surgery. (You read that correctly.) According to the British Medical Journal,

> In the placebo surgery that he performs, Dr Freeman bores into a patient's skull, but does not implant any of the fetal

> nerve cells being studied as a treatment for Parkinson's disease. The theory is that such cells can regenerate brain cells in patients with the disease. Some colleagues decry the experimental method, however, saying that it is too risky and unethical, even though patients are told before the operation that they may or may not receive the actual treatment.[7]

[7] BMJ. 1999 October 9; 319(7215): 942

"There has been a virtual taboo of putting a patient through an imitation surgery," Dr. Freeman said. (Imagine that.) "This is the way to start the discussion." Freeman has performed 106 real and placebo cell transplant operations since 1992. Dr. Freeman argues that the medical history is littered with examples of unsafe and ineffective surgical procedures—think of that small voice at the back of the room, asking "which half?"—that were not tested against a placebo and resulted in needless deaths, year after year, before doctors abandoned them.

Experimental evidence is the best kind of evidence

Let's practice here, by comparing two causal hypotheses arising from two different data sets. The first comes from a clinical trial in the 1980's on a then-new form of adjuvant chemotherapy for treating colorectal cancer, a dreadful disease that, as of 2015, has a five-year survival rate of only 60-70% in the developed world.

The trial followed a simple protocol. After surgical removal of their tumors, patients were randomly assigned to different treatment regimes. Some patients were treated with fluorouracil (the chemotherapy drug, also called 5-FU), while others received no follow-up therapy. The researchers followed the patients for many years afterwards and tracked which ones suffered from a recurrence of colorectal cancer.

The outcome of the trial are in Table 11.1, below. Among the patients who received chemotherapy, 39% (119/304) had relapsed by the end of the study period, compared with 57% of patients (177/315) in the group who received no therapy:

Chemotherapy?		Yes	No
Recurrence?	Yes	119	177
	No	185	138

Table 11.1: Data from: J. A. Laurie et. al. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. J. Clinical Oncology, 7:1447–56, 1989. There was also a third treatment arm of the study in which patient received a drug called levamisole, which isn't discussed here. Survival statistics on colorectal cancer from Cunningham et. al (2010). "Colorectal cancer." Lancet 375 (9719): 1030–47.

The evidence strongly suggests that the chemotherapy reduced the risk of recurrence by a substantial amount: the relative risk of

a relapse under the treatment group is 0.7, with a 95% confidence interval of $(0.59, 0.83)$.

We can be confident that this evidence reflects causality, and not merely correlation, because patients were randomly assigned to the treatment and control groups. Randomization ensures *balance*: that is, it ensures that there are no systematic differences between the two groups with respect to any confounding factors that might be correlated with the patients' survival chances. This would obviously not be true if we had non-randomly assigned all the healthiest patients to the treatment group, and all the sickest patients to the control group.

It's worth emphasizing a key fact here. Randomization ensures balance both for the possible confounders that we can measure (like a patient's age or baseline health status), as well as for the ones we might *not* be able to measure (like a patient's will to live). This is what makes randomization so powerful, and randomized experiments so compelling. We don't even have to know what the possible confounding variables are in order for the experiment to give us reliable information about the causal effect of the treatment. *Randomization balances everything*, at least on average.

Next, let's examine data from a study from the 1990's conducted in sub-Saharan Africa about HIV, another dreadful disease which, at the time, was spreading across the continent with alarming speed. Several studies in Kenya had found that men who were uncircumcised seemed to contract HIV in greater numbers. This set off a debate among medical experts about the extent to which this apparent association had a plausible biological explanation.

Circumcised?		Yes	No
HIV positive?	Yes	105	85
	No	527	93

Table 11.2: Data from Tyndall et. al. Increased risk of infection with human immunodeficiency virus type 1 among uncircumcised men presenting with genital ulcer disease in Kenya. Clin. Infect. Dis. 1996 Sep; 23(3):449–53.

Table 11.2, above, shows some data from one of these studies, which found that among those recruited for the survey, 48% of uncircumcised men were HIV-positive, versus only 17% of circumcised men. The evidence seems to suggest that circumcision reduced a Kenyan man's chance of contracting HIV by a factor of 3.

Evaluating the evidence. If you suffer from colon cancer, should you get chemotherapy? Almost certainly: the researchers in the

first study randomized and intervened, giving chemotherapy only to a random subset of patients. Unless you believe that the chemotherapy patients in this trial just happened to be much luckier than their peers, this result establishes that the reduction in recurrence must have been caused by the treatment.

But should all Kenyan men head straight to a surgeon? In this case we can't really be sure. The researchers in the second study neither randomized nor performed any snipping themselves. They merely asked whether each man was circumcised. It is therefore possible that they've been fooled by a confounder. To give one plausible example, a man's religious affiliation might affect both the likelihood that he is circumcised and the chances that he contracts HIV from unprotected sex. If that were true, the observed correlation between circumcisions and HIV rates might be simply a byproduct of an imbalanced, unfair comparison, rather than a causal relationship.[8]

To summarize: a randomized, controlled experiment is the gold standard of evidence for a causal hypothesis. Yet many times an experiment is impossible, impractical, unethical, or too expensive in time or money. In future chapters, we'll consider some alternative strategies for understanding cause and effect that don't involve running an experiment.

[8] The authors of the study were obviously aware of these possible confounders. They used a technique called logistic regression to attempt to account for some them and isolate the putative effect of circumcision on HIV infection. This is like our fourth method for making balanced comparisons: use a model to adjust for confounders statistically. See the original paper for details.