1
Explanations and Evidence

Key terms and concepts: variable, response, predictor, proxy, exogenous versus endogenous, clean variation versus dirty variation, confounding, directed graph, selection bias, cross-sectional study, longitudinal study, natural experiment, scatter plot

Cause and effect

WHY HAVE some nations become rich while others have remained poor? Do small class sizes improve student achievement? Does a high-salt diet lead to high blood pressure? Did an asteroid impact kill the dinosaurs? Does investing in "green" design improve the value of a commercial property? Do state-funded public-health programs lower the incidence of diabetes?

All of these questions involve a *predictor variable* and a *response variable*. Both terms are fairly intuitive: the predictor predicts the response! For example, take the first question posed above: why have some nations become rich while others have remained poor? We'll consider a specific hypothesis: that investment in education predicts economic growth. To express this idea in shorthand, we can put the predictor and response together in a *model expression*:

You may hear predictors called *independent variables*. But predictors will rarely be independent of one another in the formal sense. Therefore, know the term, but avoid it as far as possible, unless you relish the linguistic masochism of "dependent independent variables."

$$\text{Economic growth} \sim \text{Education spending}\,,$$

where the $\sim$ sign means "depends upon" or "is modeled by." To represent loose ideas like "growth" and "education," we'll have to use *proxy variables*: GDP growth rate, and percentage of GDP spent on education. Many important things in life cannot be measured, and often the use of a proxy—a firm measurable stand-in for the slippery non-measurable thing of interest—is the best we can do.

Models, of course, must answer to data. We try on different models the same way we try on different pairs of jeans. We keep going until we find one that fits the data well, and isn't too expensive—that is, complicated.

But not all data-based reasoning is created equal. In the left panel below, 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%.

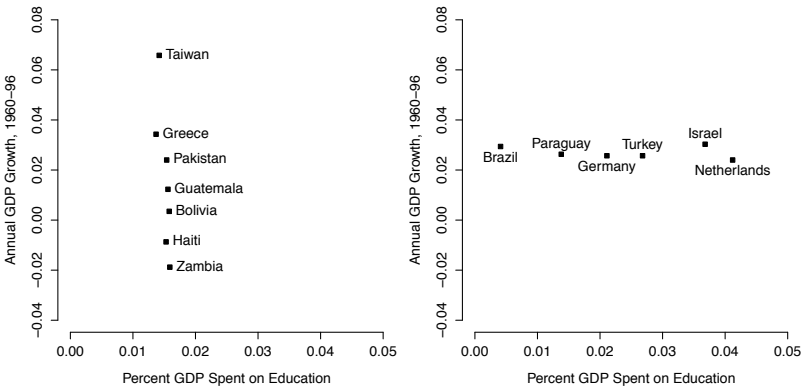


Figure 1.1: Two egregious examples of selective reporting.

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. Even good data can, if used poorly or dishonestly, be complicit in mushy cause-and-effect arguments.

The much bigger sample in Figure 1.2 provides a much more representative body of evidence—and a much more complicated story. This evidence takes the form of a *scatter plot* of GDP growth versus education spending for a sample of 79 countries worldwide. A scatter plot is an excellent way of visualizing the relationship between two variables. Each dot represents a single case. By convention the predictor gets plotted on the x axis, and the response on the y axis. The dot's position in two dimensions tells you how much the corresponding country spent on education (along the x axis), and its average GDP growth rate from 1960–1996 (along the y axis).

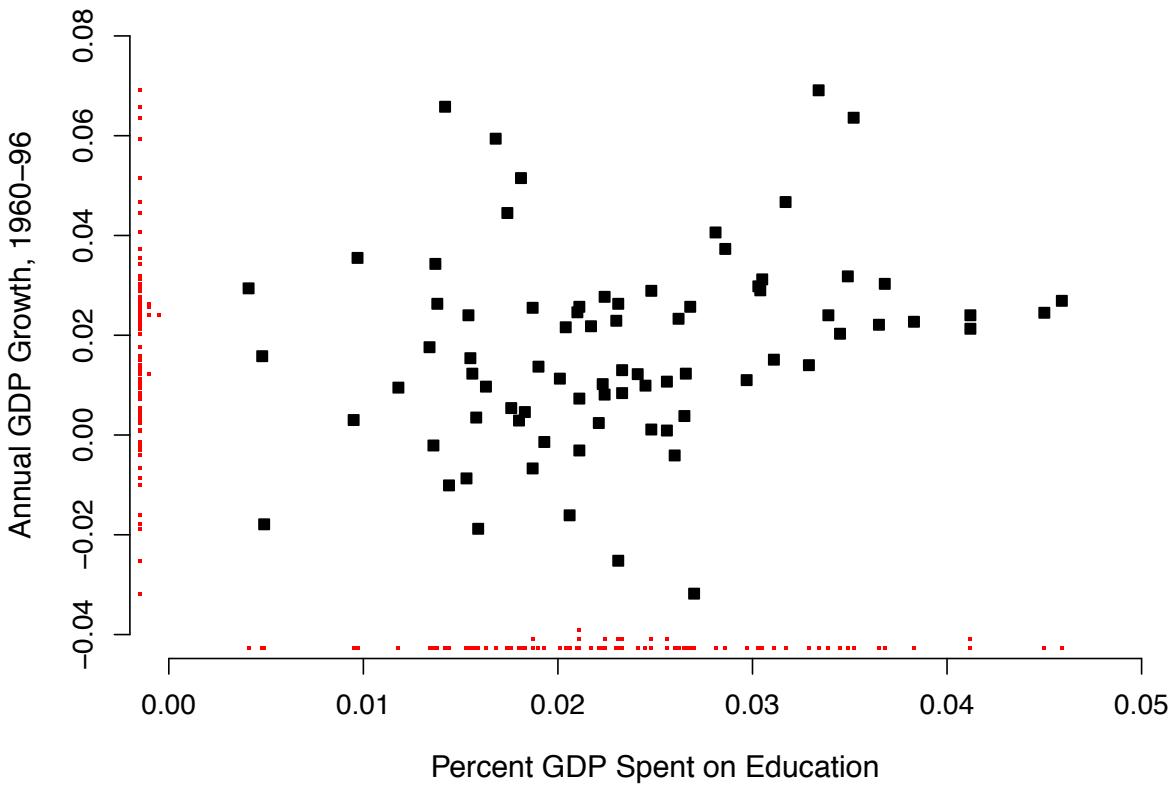


Figure 1.2: A scatter plot will show any systematic relationship between predictor and response. (Does the cloud of dots seem to ramp up as you scan left to right? Does it ramp down? Is it U-shaped? And so forth.) The tiny red dots clustered near the x and y axes are called *rug plots*. They are miniature histograms aligned with the axes of the predictor and the response.

What do the data above seem to say about the relationship between education spending and GDP growth? Does there seem to be an upward trend, a downward trend, or no trend at all? No two people will see exactly the same story in a scatter plot, but consider at least 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) Of the 18 countries that spent more than 3% of GDP on education, 16 fall above the median growth rate.

These two facts at least suggest, though by no means prove, that education spending might have some causal effect on future economic growth.

Yet how tempting just to cherry pick, and ignore the messy reality! Indeed, we're all used to seeing popular news stories that marshal highly selective evidence, often even worse than that of Figure 1.1, on behalf of some plausible "just-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 really quite a superficial analysis of the diverging economic fates of Boston and Detroit over the last thirty years.

Good evidence . . . and bad

Alas, most bad arguments are harder to detect than these last two howlers. Formal data-based reasoning is hard, and not at all a natural thing for humans to do when anecdotes, intuition, and bombast are so readily available.

If you've ever tried to self-diagnose the trouble with your golf or tennis swing, or any other highly choreographed motion, you will sympathize intensely with the difficulty of inferring explanations from evidence. It helps to keep in mind some of the different ways that variables can appear correlated. (Here, we're using "correlation" in the everyday-English sense of the word, to mean a mutual relationship between two things. Later, we'll provide a formal mathematical definition of this intuitive concept.)

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

(2) Two-way causality: flowers and honey bees prosper together. (Both A and B 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 A in causing B is hard to distinguish from the role of C, which we may not have observed.)

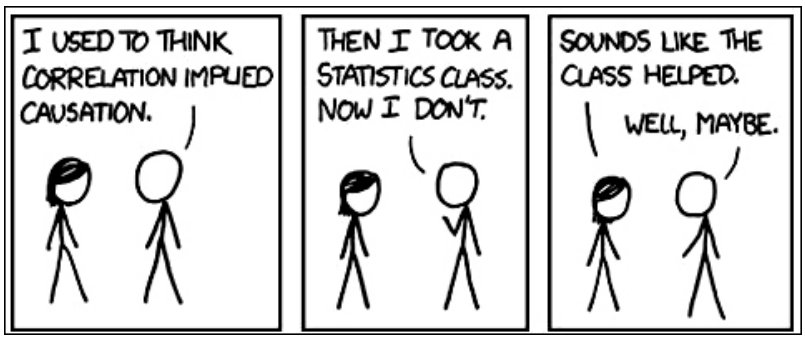


Figure 1.3: Originally published online at xkcd: http://xkcd.com/552/

(4) Common effect: either musical talent (A) or athletic talent (B) will help you get into Harvard (C). Among a population of Harvard freshmen, musical and athletic talent will thus appear negatively correlated, even if they are independent in the wider population. (A and B both contribute to some common outcome C, inducing a correlation among a subset of the population defined by C. 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—as if doing so bespoke deep intellectual seriousness—that "correlation does not imply causation." But if not to illuminate causes, what is the point of evidence? Of course correlation does not imply causality, or else playing professional basketball would make you freakishly tall. But that hasn't stopped humans from learning that smoking causes cancer, or that lightning causes thunder, on the basis of observed correlations.

What distinguishes these solid conclusions from the tenuous, or the downright absurd? How do we know that causation doesn't run the other way? These are deeply important questions, to which we'll return again and again. In facing them, the real mark of intellectual seriousness is not to repeat the platitude, but to understand *when*—that is, under what conditions—we can take observed correlation as good evidence of cause and effect.

Randomize and intervene

A very useful concept in evaluating the quality of evidence is *exogeneity*. Exogeneity defies easy definition, but means something like "outside the system." The full meaning of this term will unfold slowly, over the course of the entire book. But at least here in the beginning, it is best elaborated by example.

Let's start with the idea of a purposeful experimental manipulation, whose grade-school simplicity is as exogenous as it gets.

A clinical trial provides the archetypal example. 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 *nuisance variables*.

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 the old one. In fact, government regulators are so fastidious in their attention to possible biases that, in 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.

No, if you want to do things properly, follow 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 Zapaclot (the new drug), and everyone in the control group to take the old drug.

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.[2] 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.

[2] And if our sample is big enough—but that's for later.

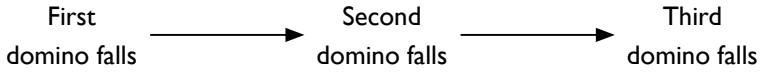
Take care in digesting that last sentence: it's not that diet, genes, and other lifestyle factors somehow stop affecting a patient's cholesterol level when we randomize and intervene. No, these predictors go on their merry causal way, just as before. They just do so equally in both the treatment and the control group! (Later on, we'll be able to make this notion precise.)

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. Therefore—and this is the key step—if we see a difference in patients' cholesterol levels, it must be because one drug is better than the other. Other than luck, what else could it be?

Together, randomization and intervention buy you *clean variation*. Here "variation" means the difference in cholesterol levels between the treatment and control groups. "Clean" means it is unsullied by the effects of some other *confounding variable*—that is, some systematic effect correlated with both the response and the predictors you care about. For this reason the problem of confounding is sometimes called the *omitted-variable bias*.

Directed graphs

A simple pictorial way of expressing complex relationships among variables is to use a *directed graph*, or simply a graph. A graph is like a set of marching orders for variables: it shows which ones lead, and which ones follow.

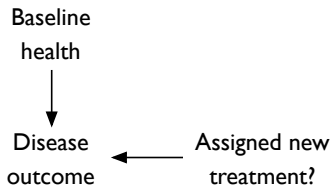


The little text blocks are called the *nodes* of the graph—you can put boxes or circles around the text if you wish—and the arrows show in which direction the assumed causation runs.

The key word here is "assumed." The graph is a *causal hypothesis* that describes a complete set of relationships among more than

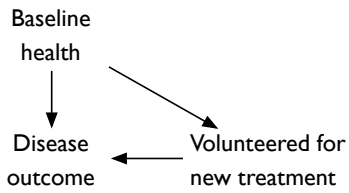
one quantity. Drawing one is easy: (1) specify all relevant variables and put each one in a node; and (2) specify all structural dependencies between nodes by drawing arrows pointed in the direction of assumed causation.

Directed graphs make excellent tools for cause-and-effect reasoning in multivariate situations, by the simple fact of getting everything there on paper in front of you. They also encourage honesty—or, at the very least, make mushy reasoning that much more obvious. Assumptions can't hide when they are right there in pictures.



Selection bias

Graphs can also help ferret out *selection bias.* Consider, for example, the two graphs at right. The top graph depicts a system in which patients are randomly assigned to receive a new drug. Both the drug and the patient's baseline level of health influence the ultimate outcome, but the treatment itself is exogenous. The bottom graph, on the other hand, depicts a system where each patient chooses on his or her own whether to take the new drug. Now the study is useless. What if, for example, only the healthiest, most proactive patients volunteer to receive the new drug? Even if we were to observe a correlation between the treatment and the disease outcome, we could not be sure whether the cause was the treatment itself, or what volunteering for the treatment said about the patient's health. The variable of interest is no longer exogenous, because it participates in another causal path influencing the response.



Here's another example of potential selection bias, from a story in *USA Today* on 19 November 2010:

> 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 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.[3]

Presumably this did not, even in California, involve a controlled experiment.

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

Of course, subjects in a study aren't the only ones who can introduce a selection bias. As one doctor 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.

Endogeneity

Yet many times a controlled experiment is impossible, impractical, unethical, or too expensive in time or money. We can imagine an experiment, for example, in which we assume dictatorial authority over every U.S. state, randomly perturb each state's health policies, and observe the pattern of changes in the incidence of diabetes. In doing so, we'd hope to test a model highly relevant to future policymaking, such as

$$\text{Diabetes rate} \sim \text{Health policy}\,.$$

But we could never perform such an experiment, or anything like it. We're left with an observational study, which is to say: we can only watch and listen.

In Figure 1.4 we see boxplots of a data set on state-by-state diabetes rates from 2005. The states are broken down by group, according to whether their state-level health policies were judged "weak" or "strong" by a team of researchers at the Harvard School of Public Health. (These judgments involved checking for policies like childhood nutrition programs, soda-machine bans in schools,

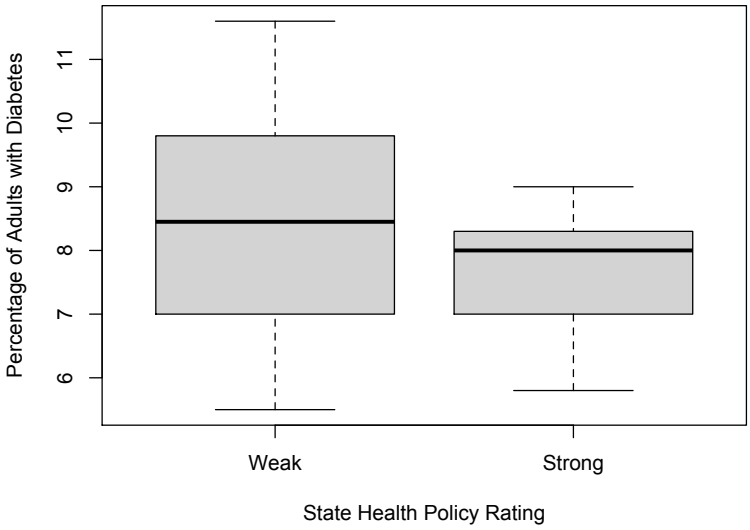


Figure 1.4: In each boxplot, the thick line shows the median; the box shows the middle 50% of cases; and the whiskers mark the upper and lower quartiles.

funding and availability of bike lanes, and so forth. The ratings are therefore subjective and imperfect, but far from useless.)

In light of the data, it appears that states with well funded, activist health policies may have slightly lower rates of diabetes than states that don't. But you should know better than to rely on this as good evidence: no randomization, no intervention, no basis for a causal judgment.

Broadly speaking, we describe variables as being *endogenous*, or exhibiting *dirty variation*, when they are caught up in some unknown, difficult-to-untangle knot of dependence between predictors and response. For example, consider the following two hypotheses, depicted in graphs at right:

(1) State health policies can change public attitudes toward health and lower the incidence of diabetes.

(2) The underlying attitudes toward health among the citizens of a state affect are the main causal factor in both the state-wide diabetes rate and the health policies adopted by the state's legislature.

Hypothesis 1 says that government policies are the cause, while hypothesis 2 says they are the effect—presumably the indirect effect of the decisions made by citizens when they go to the polls and choose their legislators. These two hypotheses are very different, and yet they can both result in the same observed patterns of

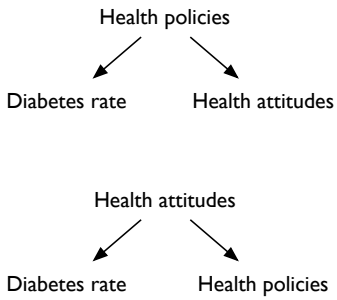


Figure 1.5: Directed graphs depicting two different causal hypotheses relating the same three variables.

correlation among the variables. Because of this possible common cause, health policies and diabetes rates are endogenous.

Some kinds of observational studies may still suggest cause and effect, even in the face of endogeneity. For example, you might consider one, or both, of the following study designs.

- Look at all 50 states, and see how the three variables—diabetes rate, health policies, and health attitudes—vary from one to the next. Do you see states where health attitudes are similar, but where policies and diabetes rates are different? This would be a *cross-sectional* study, since it involves taking a cross-section of subjects at a single time point and studying them all.
- Or, investigate how policies, attitudes, and health outcomes vary from year to year within states. This would be a *longitudinal study*, since it involves following the same group of subjects over time. A longitudinal analysis might uncover the fact that changes in policies tend to systematically lag, or lead, changes in diabetes rates.

To be sure, neither the cross-sectional nor the longitudinal analysis would yield solid proof that one causal hypothesis is true and the other false. But either might help tip the balance of evidence one way or another. (Which hypothesis would be supported if, for example, the longitudinal study showed that policy changes lagged changes in the diabetes rate?) Often, this kind of reasoning is the best we can hope for.

Natural experiments

Almost by definition, endogenous variables are not subject to experimental manipulation. In economics, business, and social science, these situations are the rule rather than the exception. Endogeneity means that we must be very careful to get a clean test of the effect we care about, for potential confounders are usually lurking around every corner when we aren't able to exert experimental control. But it does not mean that all is lost.

Endogeneity is less common in laboratory science, where designed experiments usually win the day. But then again, a careful understanding of endogeneity will make everything in experimental design seem easier.

Designed experiments get around the endogeneity problem, and provide good evidence of causation at work. But certain kinds of observed correlations, by themselves, may be almost as good. This brings us to the idea of a *natural experiment*.

A natural experiment is not something that you, as the investigator, design. Rather, it is an "experiment" where nature seems

to have done the randomization and intervention for you, thereby giving you exogenous variation for free. (Remember, without randomization, you're just back to plain old observation, and all the difficulties that entails.)

As with many ideas we've encountered so far, the idea of exogenous variation induced by a natural experiment is best understood by example. Suppose you want to study the effect of class size on student achievement. You reason that, in smaller classes, students can get more individual attention from the instructor, and that intructors will feel a greater sense of personal connection to their students. All else being equal, you believe that smaller class sizes will help students learn better.

A naïve way to study this question would be to compare the test scores of students in small classes to those of students in larger classes. Any of these confounders, however, would render such a study highly dubious: (1) students in need of remediation are sometimes put in very small classes; (2) highly gifted students are also sometimes put in very small classes; (3) richer school districts can afford both smaller classes and many other potential sources of instructional advantage; or (4) better, more experienced teachers lobby successfully to teach the smaller classes themselves.

An expensive, difficult way to study this question would be to design an experiment, in conjunction with a scientifically inclined school district, that randomly assigned both teachers and students to classes of varying size. This would guarantee exogenous variation in class sizes. In fact, a few school systems have done exactly this. A notable experiment is Project STAR in Tennessee—an expensive, lengthy experiment that studied the effect of primary-school class sizes on high-school achievement, and showed that reduced class sizes have a long-term positive impact both on test scores and drop-out rates.[5]

[5] The original study is described in Finn and Achilles (1990). "Answers and Questions about Class Size: a Statewide Experiment." *American Educational Research Journal* 28, pp. 557–77

But suppose you are neither naïve nor rich, and want to study the question yourself. If you're in search of a third way—one that's better than merely looking at correlations, yet cheaper than a full-fledged experiment—you might be interested to know the following fact about the Israeli school system.

> [I]n Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in a fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions, such as ability and family background, we can

Question	Problem	Natural experiment	Lingering issues
Does being rich make people happy?	Even if richer people are happier on average, maybe happiness and success are the common effect of a third factor. Or maybe the rich grade on a different curve than the rest of us.	Compare a group of lottery winners with a similar group of people who played the lottery but didn't win.	Lottery winners may play the lottery far more often than people who played the lottery but didn't win, which might correlate with other important differences.
Does smoking increase a person's risk for Type-II diabetes?	People who smoke may also engage in other unhealthy behaviors at systematically different rates than non-smokers.	Compare before-and-after rates of diabetes in cities that recently enacted bans on smoking in public places.	There is no control group; maybe the incidence of diabetes would have changed anyway.
Do bans on mobile phone use by drivers in school zones reduce the rate of traffic collisions?	Groups of citizens that enact such bans may differ systematically in their attitudes toward risk and behavior on the road.	Go to Texarkana, split by State Line Avenue. Observe what happens when Texas passes a ban and Arkansas doesn't.	There may still be systematic differences between the two halves of the city.

> think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."[6]

[6] Angrist and Pischke (2009). *Mostly Harmless Econometrics*, Princeton University Press, p. 21

This is a lovely example of a natural experiment—something you didn't design yourself, but that is almost as good as if you had.

Some natural experiments, of course, are better than others. Consider the examples in the table above. For each one, ask yourself two questions. (1) How good is the control group? (2) How good is the randomization? Think carefully about each one, and you'll begin to see clean and dirty variation as the black and white ends of a spectrum, with many shades of grey in between. Said concisely: only experimental variation is perfectly clean, but among other kinds of variation, some are cleaner than others.

Statistical adjustment: a look ahead

We are slowly building towards the idea of using models to adjust for the effects of confounders *statistically*, rather than experimentally. You may have heard this process described as "statistical control" or "statistical correction," both in the popular media and in scientific publications:

- "Schatz's numbers are unique in that they evaluate each play against the league average for plays of its type, adjust for the strength of the opponents' defense, and even try to divide credit for a given play among teammates."[7]

[7] "Pigskin Pythagoras: A guy from Framingham tries to remake the muddy field of football statistics." *Boston Globe*, February 1, 2004

- "The committee concluded that a statistical adjustment of the 1990 census leads to an improvement of the counts."[8]
- "Further adjustment for weight change and leukocyte count attenuated these risks substantially."[9]

Adjustment, control, correction. . . . The formal meaning of these terms is both specific and subtle, and happens to be one of the most important intellectual themes in statistical modeling. It will occupy us for the next several chapters.

[8] "Judge must decide on census adjustment." *Chicago Tribune*, 6//8/1992

[9] "Smoking, Smoking Cessation, and Risk for Type 2 Diabetes Mellitus: A Cohort Study." *Annals of Internal Medicine*, January 4, 2010