1 Intro

What is causal inference

How can we define causal inference? Causal inference is the art and science of estimating causal relationships. Some date the beginning of the modern causal inference to Ronald Fisher (1935), Trygve Haavelmo (1943), or Donald Rubin (1974). Some connect it to the work of early pioneers like John Snow (1855). We should give a lot of credit to labor economists for their groundbreaking work from the late 1970s to late 1990s. In fact, labor economics and causal inference is tightly linked.

For example, Nobel prize in economics in 2021 was given to David Card (*"for his empirical contributions to labour economics"*) and to Josh Angrist + Guido Imbens (*"for their methodological contributions to the analysis of causal relationships"*).

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

© Nobel Prize Outreach. Photo: Paul Kennedy

David Card

Prize share: 1/2

© Nobel Prize Outreach. Photo: Risdon Photography

Joshua D. Angrist

Prize share: 1/4

© Nobel Prize Outreach. Photo: Paul Kennedy

Guido W. Imbens

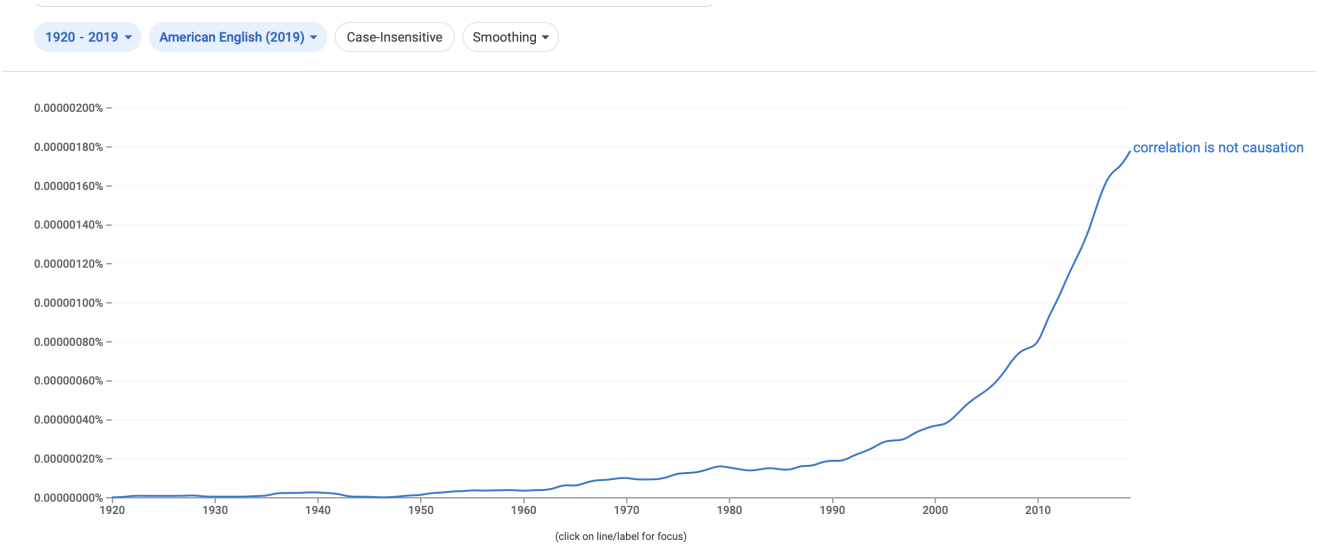
Prize share: 1/4

A classic textbook on causal inference, "Mostly Harmless Econometrics" was written by Joshua Angrist and Jörn-Steffen Pischke, who are labor economists. But of course, causal inference is not limited to labor economics. It applies throughout the field of empirical economics: political economy, health economics, environmental economics.

Causal inference has now matured into a distinct field. It is sometimes reviewed in a lengthy chapter on “program evaluation” in econometrics textbooks (Wooldridge 2010), or given entire book-length treatments. To name just a few textbooks, there’s Angrist and Pischke (2009) "Mostly Harmless Econometrics", Morgan and Winship (2014) "Counterfactuals and Causal Inference," Imbens and Rubin (2015) "Causal Inference for Statistics, Social, and Biomedical Sciences." There are also lengthy treatments of specific strategies, such as those found in Angrist and Krueger (2001) on Instrumental Variables, Imbens and Lemieux (2008) on Regression Discontinuity.

Correlation is not causation

You probably heard the phrase "correlation does not imply causation" or "correlation is not causation"



What does it mean? Here are a few possible definitions:

> Just because two variables are correlated does not necessarily mean that one causes the other

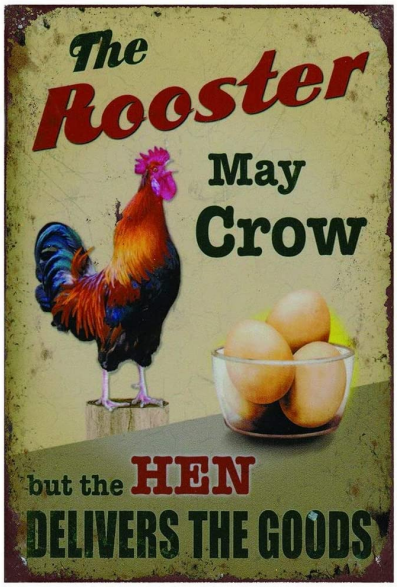
Or

> The inability to legitimately deduce a cause-and-effect relationship between two variables on the basis of an observed association between them

The purpose of this course is to explain why correlations, particularly in observational data, are unlikely to be reflective of causal relationships and how to uncover true causal relationships.

Here are a few examples of "correlation does not imply causation".

> When the rooster crows, the sun soon after rises, but does it mean that the rooster causes the sun to rise?



The sun does not rise because of the rooster. Why do roosters crow, by the way?

Here is another example

> We can observe that the number of people who wear shorts is much higher on days when people eat ice cream. Shorts-wearing can be a great predictor for ice-cream buying, but does it mean that shorts-wearing causes people to buy ice-cream?

Photo by Choi Mo on Unsplash

Shorts-wearing isn't *why* people buy ice cream. They buy ice cream and wear shorts because it's hot.

Sometimes there are causal relationships between variables and yet *no observable correlation* between them. Consider a central bank that expects a recession.

Seeing evidence that a recession is emerging, the bank enters into open-market operations, buying bonds and pumping liquidity into the economy. Insofar as these actions are done optimally, these open-market operations will show no relationship whatsoever with actual output. In fact, in the ideal, banks may engage in aggressive trading in order to stop a recession, and we would be unable to see any evidence that it was working *even though it was*! However, had the central bank *not* intervened, the recession would have occurred. We would need a **potential outcome** to deduce whether the actions of the central bank have a causal effect on the GDP.

Why do we worry so much about causality, anyway? Because many research questions we are interested in are *causal* in nature. We don't want to know if countries with higher minimum wages have less poverty, we want to know if *raising the minimum wage reduces poverty*. We don't want to know if people who take a popular common-cold-shortening medicine get better, we want to know if *the medicine made them get better more quickly*. We don't want to know if the central bank raising interest rates was shortly followed by a recession, we want to know if *the interest rate increase caused the recession*.

Defining causality can be a bit tricky. But a useful way to think about it is this: We can say that *X* causes *Y* if, were we to intervene and change the value of *X*, then the distribution of *Y* would also change as a result.

This definition lets us distinguish between correlation and causation.

If we bring a fox and put it next to the rooster, they fox would probably eat the rooster. Would it mean that the sun would no longer rise? Definitely not. Hence, it was a non-causal relationship.

If we were to swap out someone's pants for shorts, would it make them more likely to eat ice cream? Probably not! So this is a non-causal relationship.

If we were to stop the central bank and not let it engage in the open-market operations, would it lead to a change in GDP? Good question.

In some trivial cases, whether a relationship is causal or not is obvious. But in many cases it is not obvious, and the goal of researchers to figure that out.

For example, let's say that our research question is "does adding an additional highway lane reduce traffic?" How might we go about answering this question? Our first pass might be to just compare traffic patterns on, say, three-lane highways and on two-way highways. Seems reasonable. But then you do it, and it turns out that more lanes have more traffic. However, why do those highways have more lanes in the first place? It might be that the busiest routes tend to be the ones that get expanded, and so it's no surprise that more lanes are associated with more traffic.

Notice that we aren't *actually* interested in how much traffic there is on three-lane highways vs. two-lane highways. We are probably interested in whether we can make traffic go down by turning a two-lane highway into a three-lane highway. But as much as we want them to, the numbers we have don't actually tell us that right away. All we have are two-lane highways and three-lane highways. We typically don't have a "what if" highway that would tell us how much traffic there *would have been* if we'd made *that* two-lane highway one lane wider. The goal of good research design is to find the best possible proxy to that "what if" highway.

Consider another research question: Do hospitals make people healthier? To make it more realistic, imagine we are studying a poor elderly population that uses hospital emergency rooms for primary care. Some of these patients are admitted to the hospital. This sort of care is expensive, crowds hospital facilities, and is, perhaps, not very effective. In fact, exposure to other sick patients by those who are themselves vulnerable might have a net negative impact on their health.

Since those admitted to the hospital get many valuable services, the answer to the hospital-effectiveness question still seems likely to be "yes". But will the data back this up? The naive approach would be to compare the health status of those who have been to the hospital to the health of those who have not. The National Health Interview Survey (NHIS) contains the information needed to make this comparison. Specifically, it includes a question "During the past 12 months, was the respondent a patient in a hospital overnight?" which we can use to identify recent hospital visitors. The NHIS also asks "Would you say your health in general is excellent, very good, good, fair, poor?" The following table displays the mean health status (assigning a 1 to excellent health and a 5 to poor health) among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

The difference in the means is a 0.71, a large and highly significant contrast in favor of the non-hospitalized, with a t-statistic of 58.9.

Taken at face value, this result suggests that going to the hospital makes people sicker. It's not impossible this is the right answer: hospitals are full of other sick people

who might infect us, and dangerous machines and chemicals that might hurt us. Still, it's easy to see why this comparison should not be taken at face value: people who go to the hospital are probably less healthy to begin with. Moreover, even after hospitalization people who have sought medical care are not as healthy, on average, as those who never get hospitalized in the first place, though they may well be better than they otherwise would have been.

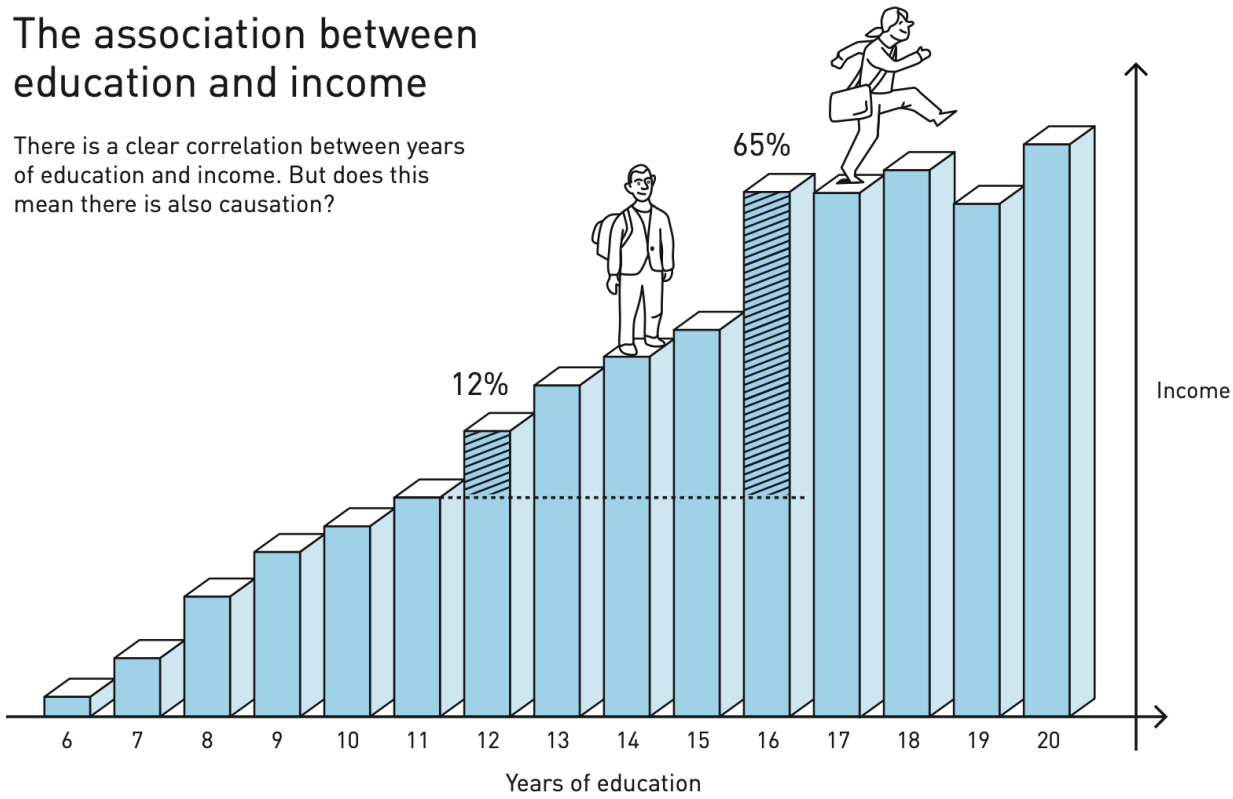
An iconic example from the field of labor economics is the evaluation of government-subsidized training programs. These are programs that provide a combination of classroom instruction and on-the-job training for groups of disadvantaged workers such as the long-term unemployed, drug addicts, and ex-offenders. The idea is to increase employment and earnings. Paradoxically, studies based on non-experimental comparisons of participants and non-participants often show that after training, the trainees earn less than plausible comparison groups (see, e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Lalonde 1995). Here too, selection bias is a natural concern since subsidized training programs are meant to serve men and women with low earnings potential. Not surprisingly, therefore, simple comparisons of program participants with non-participants often show lower earnings for the participants. In contrast, evidence from randomized evaluations of training programs generate mostly positive effects (see, e.g., Lalonde, 1986; Orr, et al, 1996).

Let's take another example from the field of education. Some studies suggest there is little or no link between a class size and student learning. So perhaps school systems can save money by hiring fewer teachers with no consequent reduction in achievement. The observed relation between a class size and student achievement should not be taken at face value, however, since weaker students are often deliberately grouped into smaller classes.

Consider another example from the literature on human capital. The theory of human capital says that education has a causal effect on the subsequent labor market earnings of individuals. Educational training provides skills that increase the potential productivity of workers. Because productivity is prized in the labor market, firms are willing to pay educated workers more.

Now, suppose we collect some data and find that people with more education earn higher wages than people with less education. Should we take this result at face

value?



The complication is that ability enhances productivity as well. Thus, because those with relatively high ability are assumed to be more likely to obtain higher educational degrees, the highly educated are presumed to have higher innate ability and higher natural rates of productivity. As a result, some portion of the uncovered effect of education on earnings may instead reflect innate ability rather than any productivity-enhancing skills provided by educational institutions. The degree of "ability bias" in standard estimates of the causal effect of education on earnings has remained one of the largest causal controversies in the social sciences since the 1970s (Card 1999).

Types of data

Broadly speaking, there are two types of data. There's experimental data and non-experimental, or observational, data. Our focus in this course is observational data.

Here is the crucial distinction between the two types of data (Cox and Reid 2000)

> The word *experiment* is used... to mean an investigation where the system under study is under the control of the investigator. This means that the individuals or material investigated, the nature of the treatments or manipulations under study and the measurement procedures used are all selected, in their important features at least, by the investigator.
> By contrast in an observational study some of these features, and in particular the allocation of individuals to treatment groups, are outside the investigator's control.

Experiments have their origins in the work of statistician Ronald A. Fisher during the 1920s, which then diffused throughout various research communities via his widely

read 1935 book, "The Design of Experiments." Experimental data come in a variety of different flavors. One use taxonomy of economic experiments is given by Harrison and List (2004):

- a conventional *lab experiment* is one a that employs standard subject pool of students, an abstract framing, and an imposed" set of rules
- an *artefactual field experiment* is the same as a conventional lab experiment but with a nonstandard subject pool
- a *framed field experiment* is the same as an artefactual field experiment but with field context in either the commodity, task, or information set that the subjects can use
- a *natural field experiment* is the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know 13 that they are in an experiment.

The most credible and influential research designs use experimental data. For example, take the Perry preschool project, a 1962 randomized experiment designed to assess the effects of an early-intervention program involving 123 Black preschoolers in Ypsilanti (Michigan). The Perry treatment group was randomly assigned to an intensive intervention that included preschool education and home visits. The experiment generated follow-up data through 1993 on the participants at age 27. Dozens of academic studies cite or use the Perry findings (see, e.g., Barnett, 1992). Most importantly, the Perry project provided the intellectual basis for the massive Head Start pre-school program, begun in 1964, which ultimately served (and continues to serve) millions of American children.

Another well-known experiment in the field of education is the Tennessee STAR experiment, it points to a strong and lasting payoff to smaller classes (see Finn and Achilles, 1990, for the original study, and Krueger, 1999, for an econometric analysis of the STAR data). The STAR experiment was unusually ambitious and influential, and therefore worth describing in some detail. It cost about $12 million and was implemented for a cohort of kindergartners in 1985/86. The study ran for four years, i.e. until the original cohort of kindergartners was in the third grade, and involved about 11,600 children. The average class size in regular Tennessee classes in 1985/86 was about 22.3. The experiment assigned students to one of three treatments: small classes with 13-17 children, regular classes with 22-25 children and a part-time teacher's aide, or regular classes with a full time teacher's aide. The estimates show a positive effect of small classes relative to regular-size classes of about 5 to 6 percentile points, which is about 0.2 s.d.

Observational data are collected through surveys, by observing people's behavior in the naturally occurring settings, or as the by-product of some other business activity. In many observational studies, you collect data about what happened previously. The key difference from the experimental data is that the researcher is a passive actor in

the processes creating the data itself. She observes actions and results but is not in a position to interfere with the environment in which the units under consideration exist.

Good empirical research using observational data is most valuable when it uses data to answer specific causal questions, as if in a randomized controlled trial. This view shapes our approach to all research questions. In the absence of a real experiment, we look for well-controlled comparisons and/or natural quasi-experiments.

The quasi-experimental study of class size by Angrist and Lavy (1999) illustrates the manner in which non-experimental data can be analyzed in an experimental spirit. The Angrist and Lavy study relies on the fact that in 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 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 think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."

The Angrist-Lavy study compares students in grades with enrollments above and below the class-size cutoffs to construct well-controlled estimates of the effects of a sharp change in class size without the benefit of a real experiment. As in Tennessee STAR, the Angrist and Lavy (1999) results point to a strong link between class size and achievement. This is in marked contrast with naive analyses, also reported by Angrist and Lavy, based on simple comparisons between those enrolled in larger and smaller classes. These comparisons show students in smaller classes doing worse on standardized tests.

Research design FAQs

The FAQs ask about:

- the relationship of interest,
- the ideal experiment,
- the identification strategy,
- the mode of inference.

In the beginning, we should ask: What is the causal relationship of interest? A causal relationship is useful for making predictions about the consequences of changing circumstances or policies; it tells us what would happen in alternative (or "counterfactual") worlds. For example ,consider the research studying the causal effect of schooling on wages (Card, 1999). The causal effect of schooling on wages is the increment to wages an individual would receive if he or she got more schooling. A range of studies suggest the causal effect of a college degree is about 40 percent higher wages on average, quite a payoff. The causal effect of schooling on wages is useful for predicting the earnings consequences of, say, changing the costs of

attending college, or strengthening compulsory attendance laws. This relation is also of theoretical interest since it can be derived from an economic model.

Causal questions can be asked about individuals, firms, countries. An example of the latter is Acemoglu, Johnson, and Robinson's (2001) research on the effect of colonial institutions on economic growth. This study is concerned with whether countries that inherited more democratic institutions from their colonial rulers later enjoyed higher economic growth as a consequence. The answer to this question has implications for our understanding of history and for the consequences of contemporary development policy. Today, for example, we might wonder whether newly forming democratic institutions are important for economic development in Iraq and Afghanistan. The case for democracy is far from clear-cut; at the moment, China is enjoying robust growth without the benefit of complete political freedom, while much of Latin America has democratized without a big growth payoff.

The second research FAQ is concerned with the experiment that could ideally be used to capture the causal effect of interest. In the case of schooling and wages, for example, we can imagine offering potential dropouts a reward for finishing school, and then studying the consequences (Angrist and Lavy 2007). In the case of political institutions, we might like to go back in time and randomly assign different government structures to former colonies on their Independence Days. Ideal experiments are most often hypothetical. Still, hypothetical experiments are worth contemplating because they help us pick fruitful research topics. Imagine yourself as a researcher with no budget constraint and no IRB.

> A design of experiments ... is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although-unfortunately-most economists do not describe their design of experiments explicitly. If they did, they would see that the experiments they have in mind may be grouped into two different classes, namely, (1) experiments that we should like to make to see if certain real economic phenomena when artificially isolated from "other influences"-would verify certain hypotheses, and (2) the stream of experiments that Nature is steadily turning out from her own enormous laboratory, and which we merely watch as passive observers. In both cases the aim of the theory is the same, to become master of the happenings of real life. (Haavelmo 1944).

A good research question is a question *that can be answered*, and for which having that answer will *improve your understanding of how the world works*. What does it mean to have a question *that can be answered*? It means that it's possible for there to be some set of evidence in the world that, if you found that evidence, your question would have a believable answer. But does it *improve our understanding of how the world works?* What this means is that the research question, once answered, should tell you about something broader than itself. It should inform *theory* in some way. Theory doesn't have to be something as important as the theory of gravity or the

theory of evolution. Theory just means that there's a *why* or a *because* lurking around somewhere.

A good research question *takes us from theory to hypothesis*, where a hypothesis is a specific statement about what we will observe in the world, like “people who wash their hands will get sick less often.” That is, a research question should be something that, if you answer it, helps improve your *why* explanation. Great research questions often come from the theory themselves - the line of thinking being “if this is my explanation of how the world works, then what should I observe in the world? Do I observe it?”

The third and fourth research FAQs are concerned with the nuts-and-bolts elements that produce a specific study. Question Number 3 asks: what is your identification strategy? Angrist and Krueger (1999) used the term *identification strategy* to describe the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment. Again, returning to the schooling example, Angrist and Krueger (1991) used the interaction between compulsory attendance laws in American schools and students' season of birth as a natural experiment to estimate the effects of finishing high school on wages (season of birth affects the degree to which high school students are constrained by laws allowing them to drop out on their birthdays).

The fourth research FAQ borrows language from Rubin (1991): what is your mode of statistical inference? The answer to this question describes the population to be studied, the sample to be used, and the assumptions made when constructing standard errors. Sometimes inference is straightforward, as when you use Census micro-data samples to study the American population. Often inference is more complex, however, especially with data that are clustered or grouped.

Identification, DGP, context

When dealing with any kind of data, an important question to ask is what is the *data-generating process*. These are the laws that work behind the scenes, doing what they do whether we know about them or not. We can’t see them directly, but we do see the *data* that result from them. In addition to thinking that there *are* underlying laws like this, scientists also believe that we can learn what these laws are from empirical observation.

Thinking about the data-generating process gives us two ideas. The first is the idea of *looking for variation*. The data generating process shows us all the different processes working behind the scenes that give us our data. But we are often only interested in part of that variation. How can we find the variation we need and focus just on that?

The task of figuring out how to answer our research question is really the task of figuring out *where your variation is*. It’s unlikely that the variation in the raw data

answers the question you're really interested in (correlation ≠ causation). So where is the variation that *does* answer your question? How can you find it and dig it out? What variation needs to be removed to unearth the good stuff underneath?

The second is the idea of *identification*. How can we use the data generating process to be sure that the variation we are digging out is the right variation? That process - finding where the variation you're interested in is lurking and isolating *just that part* so you know that you're answering your research question - is called identification. It's called identification because we've ensured that our calculation *identifies* a single theoretical mechanism of interest. In other words, we've gotten past all the misleading clues and identified our culprit.

> It is useful to separate the inferential problem into statistical and identification components. Studies of identification seek to characterize the conclusions that could be drawn if one could use the sampling process to obtain an unlimited number of observations. Identification problems cannot be solved by gathering more of the same kind of data. The study of identification logically comes first. Negative identification findings imply that statistical inference is fruitless: it makes no sense to try to use a sample of finite size to infer something that could not be learned even if a sample of infinite size were available. Positive identification findings imply that one should go on to study the feasibility of statistical inference. (Manski 1995)

A research question takes us from theory to hypothesis, making sure that the hypothesis we're testing will actually tell us something about the theory. Identification takes us from hypothesis to the data, making sure that we have a way of testing that hypothesis in the data, and not accidentally testing some other hypothesis instead.

Identification requires statistical procedures in order to properly get rid of the kinds of variation we don't want. But just as important, it relies on *theory* and *assumptions* about the way that the world works in order to figure out what those undesirable explanations are, and which statistical procedures are necessary. Specifically, it relies on theory and assumptions about what the DGP looks like. We need to make a claim about what we *already know* in order to have any hopes of learning something new.

So then that's our goal. If we want to identify the part of our data that gives the answer to our research question, we must:

1. Using theory, paint the most accurate picture possible of what the data generating process looks like
2. Use that data generating process to figure out the reasons our data might look the way it does that *don't* answer our research question
3. Find ways to block out those alternate reasons and so dig out the variation we need

This process is a lot more difficult than just "look at the data and see what it says." But if we don't go the extra mile of following these steps, we can end up with confusing, inconsistent, or just plain wrong results. Let's see what happens when we don't take identification quite seriously enough.

R Demo: Human Capital and Ability Bias

Let's generate some data. A good way to think about how data generating processes (DGPs) can help with research is to cheat a little and make some data where we know the data generating process for sure.

Let's play around with simulating our own data generating process and looking at what happens when we try to estimate effects naively. Imagine the following DGP, inspired by the human capital theory, that describes the relationship between ability, education, and earnings.

1. Ability is uniformly distributed on [0, 1]
2. Each unit of ability increases your income by 20 thousand euros per year
3. To decide whether you go to college or not, you first flip a coin (heads = go to college, tails = do not go to college). However, if it is tails, you then check your ability. If it is higher than a threshold of 0.5, you go to college anyway.
4. College degree increases your income by 10 thousand euros per year
5. The base income (no college, zero ability) is normally distributed with a mean of 50 thousand euros per year and a s.d. of 20 thousand
6. We observe a sample of 5000 individuals

`R Demo: educ_abil_inc.Rmd`