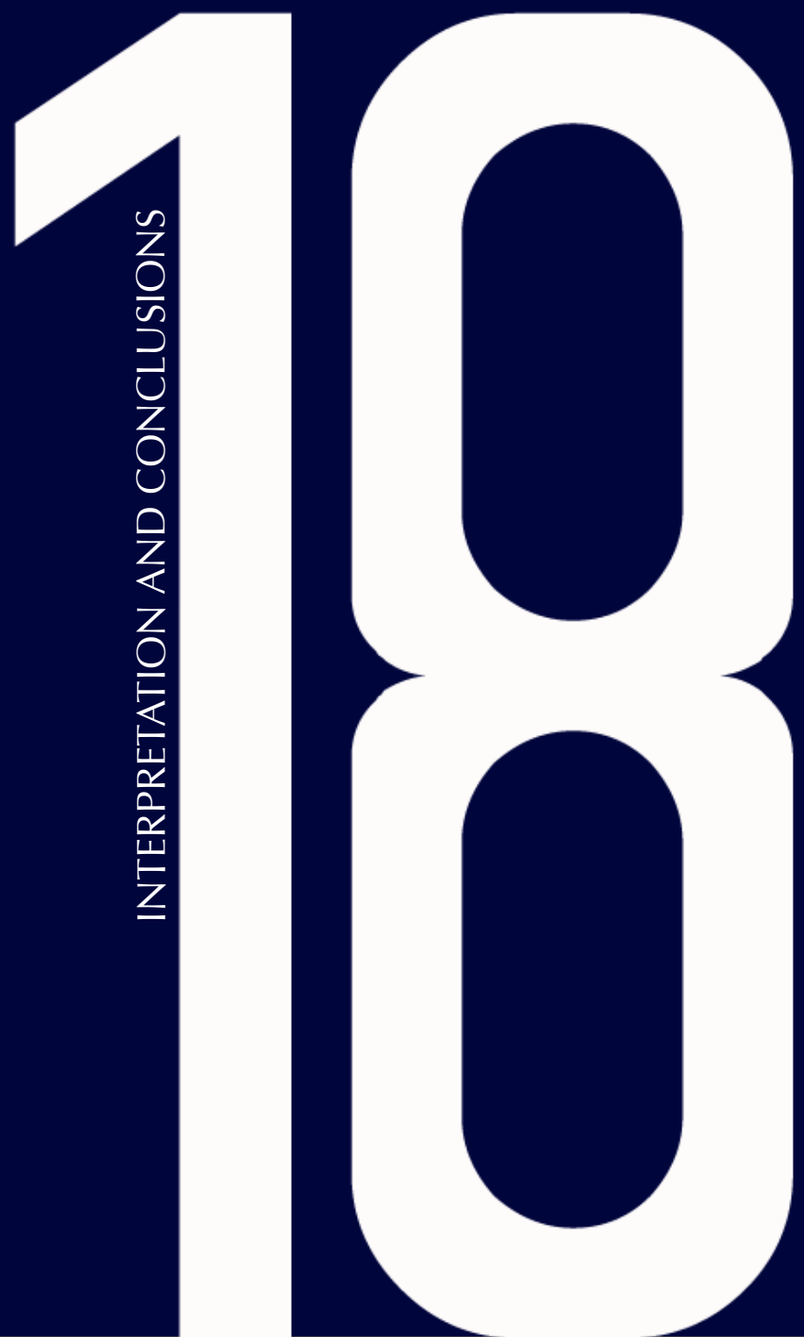
CHAPTER 18: CORRELATIONS ARE HARD TO INTERPRET



In his essay The Danger of Lying in Bed, Mark Twain made folly of people who bought travel insurance. He pointed out that far more people died in bed than on public transportation, so the REAL danger came from lying down.

SECTION 1

Introduction

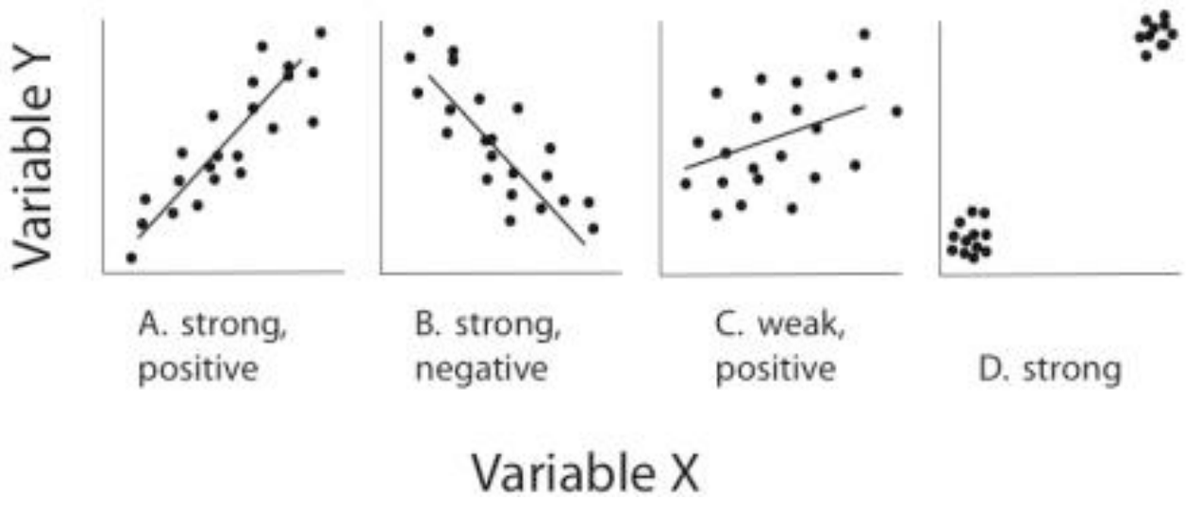
In most scientific inquiries, we seek the cause of something. We want to know what causes cancer, what drugs cause us to recover from disease or to feel less pain, what cultural practices cause environmental problems, what business practices lead to (cause) increased profits, what kind of sales pitch increases sales, what kind of resume is the most effective in getting a job, and so on. In these cases, we are testing causal models. Not everything we've discussed so far requires evaluation of a causal model: in DNA and drug testing, we are merely trying to measure properties of an individual (a drug level, a DNA bar code). But these exceptions notwithstanding, the most common kind of evaluation everyone encounters is testing of a causal model. "What can we change in our lives or our world to cause a certain outcome?" is the essence of what we want in a causal model.

Causal models are typically evaluated, at least initially, with data that describe an association or correlation between variables. If smoking causes lung cancer, then cancer rates should be higher (associated) with smokers. If some patterns of investment lead to higher profits, then companies which practice those kinds of investment ought to be associated with greater returns to their investors. If alcohol causes reckless driving, then a higher rate of accidents should be associated with drunk driving. The catch is this. Although a causal relationship between 2 sets of data leads to an association between them (drinking and driver accidents), an association may occur even when there is no causation. How then do we decide if the causal model is supported or refuted? This chapter is about associations among variables -- correlations -- and how and when we can tease out causation.

What Are Correlations?

Correlations are associations between variables. The first question to answer in understanding a correlation is therefore "What are variables?" Variables are things we measure that can differ from one observation to the next, such as height, weight, behavior, fat intake, life-span, grade-point average, and income. With these variables we can easily assign a number to represent the value of the variable. Perhaps less obviously, we can also treat sex (gender), country of origin, and political preference as variables, even though we don't know how to assign a number to represent each category. In general, a variable is a measure of something that can take on more than one value. It is somewhat arbitrary how we define a variable, but in general, you must be able to put the different values a variable can take onto a single axis of a graph. If you are wonder whether something you have defined is a variable and it would require two axes, then you are likely dealing with a couple of variables combined.

When an association exists between two variables, it means that the average value of one variable changes as we change the value of the other variable (Fig. 18.1). A correlation is the simplest type of association -- linear. When a correlation is weak (e.g., Model C), it means that the average value of one variable changes only slightly (only occasionally) in response to changes in the other variable. In some cases, the correlation may be positive (Models A, C), or it may be negative (Model B). If the points in such a graph pretty much fall inside a circle or horizontal ellipse such that the "trend-line" through them is horizontal, then a correlation does not exist (the same as a zero or no correlation). When either or both variables cannot be assigned numbers (e.g., political party or country of origin), a correlation may still exist but we no longer apply the terms positive and negative (e.g., Model D, depending on the nature of the variables). Since a correlation is an association among variables, a correlation cannot exist (is not defined) with just one variable; "undefined" is not the same as a zero correlation or no correlation. A graph of points with only one variable would have all points on a perfectly horizontal line or a perfectly vertical line (with no scatter around the line).



Different kinds of correlations:

The horizontal axis represents one variable (X) and the vertical axis represents a different variable (Y), with values of X and Y increasing according to the distance from the origin. Models A, B & C show correlations for continuous variables which can take on a range of values (e.g., height, weight), whereas Model D reveals a correlation for discrete variables (variable X might be gender, variable Y presence or absence of the Y chromosome). Model A reveals a strong positive correlation, Model B a strong negative correlation, and Model C a weak positive correlation. The correlation in Model D would be regarded as positive if values could be assigned to X and Y, but if values cannot be assigned (e.g., gender and presence of Y chromosome), we would not refer to the correlation as being positive or negative.

Correlations are common in Business . Businesses often obtain large quantities of correlational data as they go about their activities. An insurance company in the course of doing business obtains data about which types of customers are more often involved in accidents. These data are purely observational - the company can't force a 68 year old grandmother to drive a pickup if she doesn't want to. The data consist of driver age, sex, make and model of car, zip code, street address and so forth. In addition, the company knows how many and the type of accidents for each customer. These correlations are clearly quite useful in predicting what customers will have more accidents.

Correlations are used to manipulate us. Most advertisements, sales pitches, and political speeches invoke correlations to influence our behavior. A company tends to display its product in favorable settings to build an imaginary correlation between its product and the desirable surroundings (e.g., beer commercials using attractive members of the opposite sex, 4WD autos being pictured with a backdrop of remote, montane scenery). Negative campaigning usually involves describing some unfavorable outcome that occurred during an opponent's tenure in office to develop a correlation in the viewer's mind between the candidate and bad consequences of their election to office.

The reason that correlations are used so often in commercials is that they work-- people make the causal extrapolation from correlations. We tend to blame our current president for many social problems, even though the president has little control over many of them. In a well-known but unfortunate psychological experiment of some decades ago, a child was encouraged to develop a close attachment to a white rat, whereupon the experimenters intentionally frightened the child with the rat. Thereafter, the child avoided white objects -- a rather surprising correlate of the rat. Other studies have shown that people respond differently to an item of clothing according to what they are told about an imaginary person who wore it: the response is more favorable if the supposed previous wearer is famous than if the person is infamous. The information thus established a correlation between the clothing and a desirable or undesirable person, and the subjects mentally extrapolated that correlation to some kind of causation of good or bad from wearing the object. And some of our responses to correlations are very powerful. The experience of getting overly drunk on one kind of alcoholic beverage is often enough to cause a person to avoid that beverage years into the future but not to avoid other kinds of alcoholic beverages.

Negative uses. A more negative context for the application of correlation to influence behavior is the practice known as character assassination. A person can be denigrated in one aspect of their life by identifying an unfavorable characteristic in some other (and perhaps trivial) aspect of their life We automatically extrapolate the negative correlation to them as a whole.

How to identify a correlation

Correlations are not necessarily easy to recognize. For a correlation to possibly exist, you must either

i) have measures of the same two (variable) characteristics on several individuals, or

ii) have measures of the same (variable) characteristic on two populations.

Then, a correlation exists (is not zero) if the two variables change together on average, or if the two populations differ.

The possibly tricky part is that it may not be obvious how many variables and populations there are, and also because there are often multiple ways to interpret the same data. For example, if 94% of University of Texas students have cell phones (a statistic invented here for illustration), there is one population and one variable. So there is no possibility of a correlation because we either need another variable or another population. If we add that 90% of Texas A&M students have cell phones, now we have two populations and one variable. A correlation is now possible and indeed exists, because 94% differs from 90%. We could also interpret that data as one population with two variables: university attended is one variable and cell phone ownership is the other. The correlation still exists under this alternative interpretation of variables and populations, as it must.

If the numbers for UT and A&M were both 90% (or were both 94%), we still have enough data for a correlation to possibly exist, but the correlation is zero because the two numbers are the same.

Likewise, we may measure the right numbers of characteristics but they needn't satisfy the criteria for a correlation to exist. Thus, if we look within UT students for a correlation between owning a computer and being assigned a social security number, presumably all students would have a social security number but only some students would own a computer. If all students have a social security number, that characteristic is not variable. In this case, we have one population but only have one variable (computer ownership), so a correlation could not possibly exist.

Correlation versus Causation

A correlation is merely an association: certain values of X tend to be associated with specific values of Y. The effect may be subtle and only evident statistically. A correlation does not tell us how or why the association occurs. In contrast, causal models tell us the why of an association.

correlation: association without a reason

causation: a reason for an association

We usually want causal models because those tell us how to change our lives to achieve specific goals, such as avoid cancer or improve society.

As examples of the difference,

'Studying improves exam scores' gives causation

'People who study have higher exam scores than those who don't' is a correlation

'Talking on cell phones causes drivers to have higher accident rates' is causation

'Drivers talking on cell phones have higher accident rates than drivers not talking on cell phones' is merely correlation.

Correlations are everywhere. And they have often been used to infer causation, though not always correctly. For example, malaria is an infectious disease transmitted by mosquitoes. The word translates as 'bad air,' due to an early mistaken idea that it came from the air. The first guesses about the cause

of AIDS were drugs, because the men whose immune systems were collapsing were using a variety of illegal drugs that might have been imagined to harm an immune system. Mistaken correlations abound in the field of diet and health, and anyone who has been an adult for more than a decade is aware of new dietary advice overturning old dietary advice.

The Problem with Inferring Causation from Correlations: Hidden Variables

The problem that underlies evaluation of correlations is extremely common in science. We observe an association, or correlation, between two or more variables. In the nuclear power plant example, there is a correlation between residential proximity to a nuclear plant and cancer, because people near power plants are more likely to get cancer than those who live away from power plants. And we try to infer the causation from that correlation: does the plant actually cause cancer?. Time and again, science has learned the hard way that we cannot infer causation from correlation: **correlation does not imply causation.**

What does this mean? Say that you observe a correlation between smoking and lung cancer. To infer that smoking CAUSES lung cancer, you would argue that people should stop smoking to lower their lung cancer rates. If smoking does not cause lung cancer, however, then stopping smoking would actually have no effect on lung cancer rates (we are very confident, however, that smoking causes lung cancer).

How can a correlation not reflect causation? A correlation will not reflect causation when a confounding (3rd or hidden) variable is the cause. If X is correlated with Y, we have 3 possible models (in simple cases):

X causes Y

Y causes X

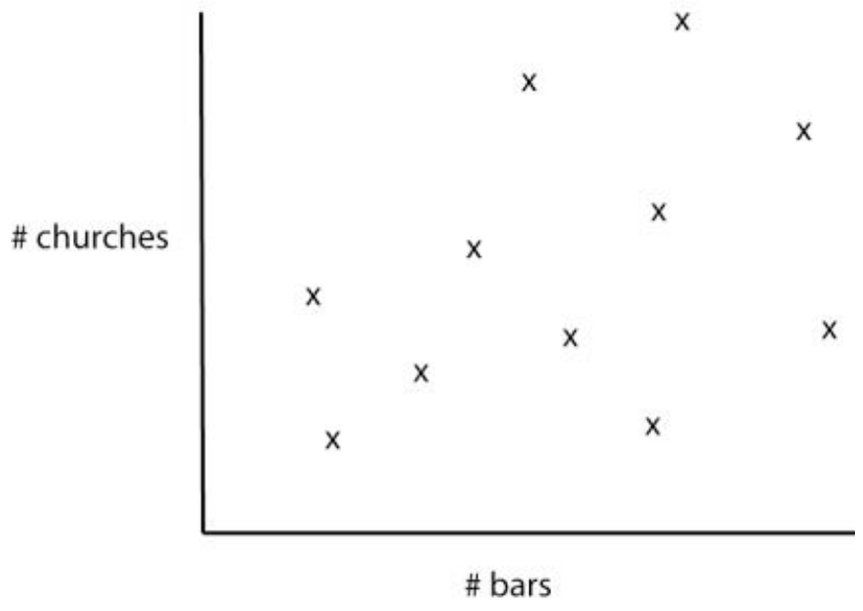
the variable Z, which we have not measured, causes X and/or Y.

All three models may not be feasible in specific cases. For example, if we consider smoking rates and lung cancer, one model would be that smoking causes cancer; the converse would be that cancer causes people to smoke. The latter might be feasible. But now suppose we observe a (negative) correlation between magnesium in municipal water supplies and tooth decay rates. We could imagine that magnesium might lower tooth decay rates, but we could not imagine that tooth decay rates influence the magnesium in the municipal water supply (which comes out of the ground or a river). In this case, the likely important 3rd variable would be fluoride in the water. It is well established that fluoride reduces tooth decay by hardening enamel. It would not be surprising if magnesium and other minerals in the natural water were correlated with fluoride levels (although fluoride is now added to drinking water in many cities). But there could be other correlates as well. We formalize some possibilities in the following table, for the correlation that higher tooth decay levels are associated with lower levels of magnesium in the town water supply. The table below lays out different issues in teasing apart causation from correlation.

CAUSAL MODEL	WHY ARE LOWER TOOTH DECAY RATES FOUND WITH HIGHER MAGNESIUM?	WOULD TOOTH DECAY RATE CHANGE IF WE ACTIVELY CHANGED MAGNESIUM LEVELS (UNDER THE MODEL)?	CAUSAL VARIABLE	DOES THE MODEL INVOKE A 'THIRD' VARIABLE?
Magnesium reduces tooth decay	magnesium reduces decay	Yes	magnesium	No
Fluoride reduces tooth decay	high magnesium is found in water with high fluoride	No	fluoride	yes
Dental hygiene reduces tooth decay	Cities with populations educated toward dental hygeine happen to be located in areas where the ground water happens to be high in magnesium	No	hygiene	Yes

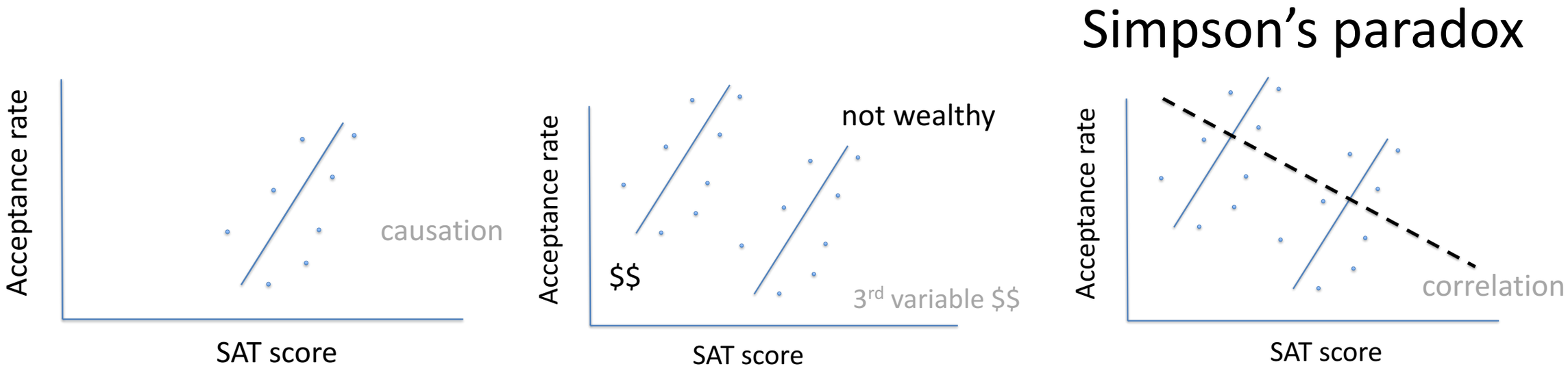
Anyone with a creative mind can generate countless causal models for a correlation. They won't all be correct, maybe none will be. To find out which are correct, a special type of data must be gathered (considered a couple chapters hence).

As an easy example, consider a hypothetical plot of the number of churches and bars in a town. Although the data are made up, a positive correlation almost certainly exists where ever bars and churches are both permitted. To argue causation from these data, we would either have to say that churches cause people to drink more (whether intentionally or unintentionally), or argue that lots of drinkers in a town causes more churches to be built (e.g., churches move in where there are sinners). Furthermore, causation would suggest either that banning bars would reduce the number of churches in the town, or that the way to cut down on the number of bars was to close down churches (depending on which way the causation went). In reality, the correlation is due to a hidden variable -- population size. That is, larger towns have more demand for churches and for bars, as well as other social institutions.



Simpson's paradox

An extreme case of the 'correlation does not imply causation' phenomenon is known as Simpson's paradox: the correlation between two variables goes in the opposite direction from their causal relationship. Consider the hypothetical case of acceptance to a university based on SAT score, as in the first panel below. Higher scores increase acceptance rate. We now add a third variable to the picture: family income. Richer families can get their kids admitted with lower SAT scores, but even here, higher scores help, as shown in the middle panel. Kids of rich families get accepted at a high rate. When we look at the association between acceptance rate and SAT score with all the data, the overall correlation now shows that SAT scores are negatively correlated with acceptance rate, shown by the dashed line.



To reiterate the point, the major difficulty with all correlations is that there are many models consistent with any correlation: the correlation between two variables may be caused by a third, fourth, or dozens of variables other than the two being compared. Thus we are left with countless alternative models in addition to the obvious ones.For example, we initially think that the correlation between cancer and residence near a power plant shows that nuclear power plants cause cancer. Then we learn that another factor, site of the power plant, may be important. It appears that the important factor is not the power plant itself, but rather some characteristic of sites chosen for power plants (one obvious possibility is that nuclear power plants are situated in low income areas that have higher cancer rates than suffered by the general population). That is, there are correlations between all sorts of other variables besides just residence and cancer.

There are many issues in society that hinge on correlations (the following table lists a couple of examples). In some cases, a correlation may identify a causal relationship, such as health defects being caused by environmental toxins. Yet because the correlational data don't reject countless alternatives models, no action is taken to correct the problem. In other cases, a correlation may be assumed to reflect the cause when it does not.

Public policy issues that involve understanding the cause of a correlation:

ISSUE	POSSIBLE CAUSATION
High cancer incidence near industrial sites, toxic waste dumps, nuclear power plants.	If the increased cancer rate is actually caused by the hazard, there would be compelling motivation for taking action. But it is often difficult to rule out the alternative explanation that those living near the hazard have different diets or for other reasons are more susceptible to cancer than the general population.
Racial differences in standardized test scores.	There are two opposing positions in this acrimonious debate: i) a person's race, per se, causes them to have low test scores, or ii) minorities often have low incomes, and it is income rather than race that determines test score. The first explanation states that a person is born with a certain intellectual ability, the second states that they acquire it.

Correlations Complicate Studying Diet and Heart Disease

The medical news over the last decade or so has been obsessed with the relationship between diet and heart disease. (Heart disease is chiefly the build-up of deposits inside blood vessels, hardening the arteries and enabling the vessels to rupture and clog.) A report that dietary fiber lowered heart attack risk led to an avalanche of pills and breakfast cereals high in fiber. More recently, a trendy topic has been iron levels in the blood. It is not clear what to make of these reports, but we can be confident that associations between diet and heart disease will continue to be the subject of studies for decades to come. However, let's consider the problems such studies pose.

Your diet consists of literally hundreds of correlated components. For example, people who eat a lot of meat also tend to eat a lot of fat, and people that eat lots of vitamin C tend to also eat much fiber. These, and numerous similar correlations, create huge problems in determining what diet you should eat to avoid heart disease. A study that found an correlation between heart disease and fat, for example, would be hard to interpret because we would not know if it was the fat, per se, or the meat that was the problem. The problem in this example is not as great as it is in other cases, because we can actually conduct experiments with human diets to explore causal relationships. But even in these experiments, it is difficult to control and randomize all relevant factors.

SECTION 7

Why Do We Bother With Correlations At All?

Given the problems with interpreting correlational data, one might reasonably ask: why do we bother with them at all if it is a causal relationship that we seek? Why not just gather data that could provide a more definite answer, or otherwise just ignore correlations? The reason is pragmatism. Correlational data are usually relatively easy and inexpensive to obtain, at least in comparison to experimental data. Also, many cause-effect relationships are so subtle that we often first learn of them through correlations detected in observational data. That is, they are often useful.