

Problems with Causal Discovery

1000: Introduction

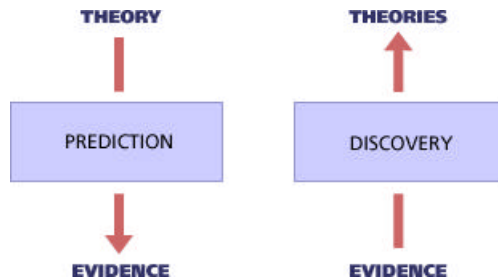


FIGURE 1000-1

Scientists typically face two problems which are just different sides of the same coin. One problem is prediction: beginning with a theory about the world, they must derive testable predictions from this theory. For example, an economist might have developed a theory about the relationship between Income Tax Rates and Economic Growth. To test this theory with observable evidence, he or she needs to derive a prediction from the theory and then collect data relevant to this prediction. For example, the economist might predict that, when Federal Income Tax Rates are cut, then economic growth (measured by quarterly growth in US GNP) will accelerate.

Another problem is discovery: beginning with evidence, scientists must try to discover what set of theories explain (or predict) this evidence. For example, suppose a different economist, who began with no theory of how Tax Rates and GNP relate, came across 20 years (80 quarters) of data on Income Tax Rates and growth in GNP. Her discovery problem is to articulate all of the theories that explain or predict these data.

In qualitative causal science, we deal with qualitative causal theories (causal graphs), and qualitative statistical evidence (associations). In prediction mode, we start with a causal theory and make predictions about associational evidence. In discovery mode, we start with associational data and try to articulate all the causal theories that explain or predict the observed associations.

In the three modules on "Causation to Association," we examined how causal graphs make predictions about unconditional and conditional associations among the variables in the graph. For example, we found that, when two variables are causally connected in a causal graph, then they are predicted to be associated.

Different graphs, however, and this is a crucial point in the subject of causal and statistical reasoning, can make the **same predictions**. For example, both $X \rightarrow Y$ and $X \leftarrow Y$ predict that X and Y are associated.

This module focuses on the problem of causal discovery. That is, when we don't know the causal relationships among the variables, how do we use the **data** we have about associations among those variables to try to discover the causal story? We know that association and causation are different. Causal information is more useful, since it enables us to predict the outcome of an intervention. But associations are the crudest way of describing the data that we get from the world. We can observe whether there is an association between taking a drug and recovering from a disease. We observe whether there is an association between whether a house contains lead paint and the results of aptitude tests given to the children who live there. In the drug case, the challenge is to decide whether the drug is a **cause** of the cure or not, and in the lead paint case, the challenge is to decide whether lead paint **causes** mental impairment.

There are many different reasons why causal discovery is hard. For example, the associational data that we have isn't always sufficient to determine just one unique graph. Another potential problem is that we might have measurement error in our data. In this module, we will sketch a few of the most common problems for causal discovery, but we won't provide solutions. That will have to wait for later modules.

2000: Underdetermination

Before discussing the problems specific to causal discovery, we will digress a little on the general topic of "underdetermination of theories by evidence." In some sense, all problems of discovery can be conceived of as problems of underdetermination.

We say that theory is "underdetermined" by data if the more than one theory can explain the data. For example, suppose you live in a house with three roommates, Bill, Richard, and Joe. It's Friday, and you come home from work thirsty and ready to celebrate the beginning of long weekend. You open the fridge to get the beer you bought yesterday, and find it is gone.

You have three theories that explain this data. Theory 1: Bill drank it. Theory 2: Richard drank it. Theory 3: Joe drank it. Before getting more information, your data underdetermines your theories. Any one of them (plus others not in this set) could explain the missing beer. You might be able to narrow the theories down by gathering more evidence. You might go to the recycling bin, and find the empty beer bottle on the top of the heap. OK, you say, it can't be Richard, he never recycles. So now you are down to Theories 1 and 3, but you are still uncertain which is true.

In literally all of science, it takes more evidence than we can ever practically gather to guarantee that any one particular theory is the right one. Most of the time, many different theories will explain the evidence that we in fact possess. In philosophy this situation is known as the **underdetermination of theory by evidence**.

Below is a simulation designed to illustrate the problem of underdetermination of theory by evidence. Suppose you are gambling with dice. The outcome you will be asked to discover is very simple: what happened on the last roll of a pair of dice? Normally, you have direct access to this information: you just look and see what faces landed "up." Suppose you can't see the dice, however. Your "evidence" in this exercise is incomplete and indirect. Your job is to use the evidence you do have to eliminate as many possibilities as you can, eventually eliminating all but the true outcome. When working through this simulation, try to think about not only the situation when you stop collecting data, but also what would happen if we forced you to stop after three, one, or even no pieces of data. What would you be able to conclude about the values of the dice?

[< A link to a Java applet in the interactive version of this module. >](#)

This simulation demonstrates a harsh reality about human knowledge: our evidence for a belief seldom justifies only that one belief. When you had limited evidence about the dice, you couldn't pick out which sides actually came up. You could rule out **some** possibilities, but even after ruling those out, there were usually several remaining possibilities that were consistent with the evidence.

This is the typical situation faced by scientists trying to find discover causal structures in the world. Although evidence we collect can rule out many possibilities, it is rare that it rules out all but one. In the next few sections, we will face exactly this situation in reasoning from associational evidence to causal theories. Causal theories (and specifically causal graphs) are almost always underdetermined by the associational data we possess.

[< A link to exercises in the interactive version of this module. >](#)

[3000: Causal Underdetermination](#)

[3100: Associations Underdetermine Causal Theories](#)

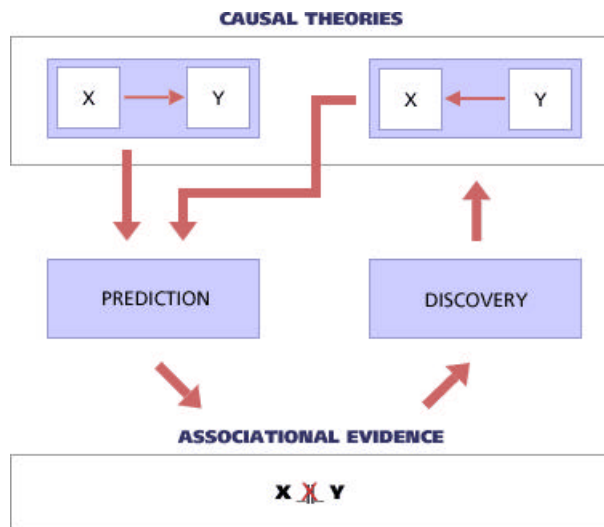


FIGURE 3100-1

The essential problem for causal discovery from associations, however, is the problem of underdetermination: many causal graphs can explain the same pattern of associations. In the figure above, both causal graphs predict the same association between X and Y . Starting with either causal theory, we derive the same unique prediction about the patterns we should observe in the data. Starting with the association, however, we can only discover the **set** of causal theories in the upper box.

Put generally: **associational data undetermines causal theories.**

For example, suppose we observe that there is a positive association between education and income at age 50. People with more education tend to make more money. What causal theory explains this association? Well, for one, more education might cause more income (Graph 1 in the figure below). It also might be, however, that one's parents are a common cause of education and income (Graph 2). Parents who act to make sure their children get a lot of education might also be better able to help get them jobs that pay well.

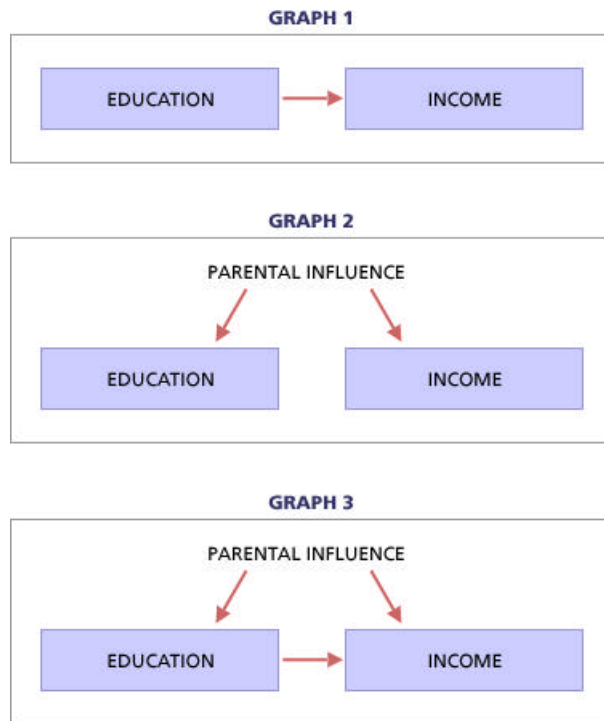


FIGURE 3100-2

It also might be that the observed association is produced by a combination of both causal connections (Graph 3). With only this one association as data, we cannot distinguish among these three causal theories (and others that we haven't presented).

3200: An Abstract Example

In general, we can use the theory of d-separation to predict which variables will be associated (and which will be conditionally associated) in data generated by a causal process that we can represent with a causal graph. For example:

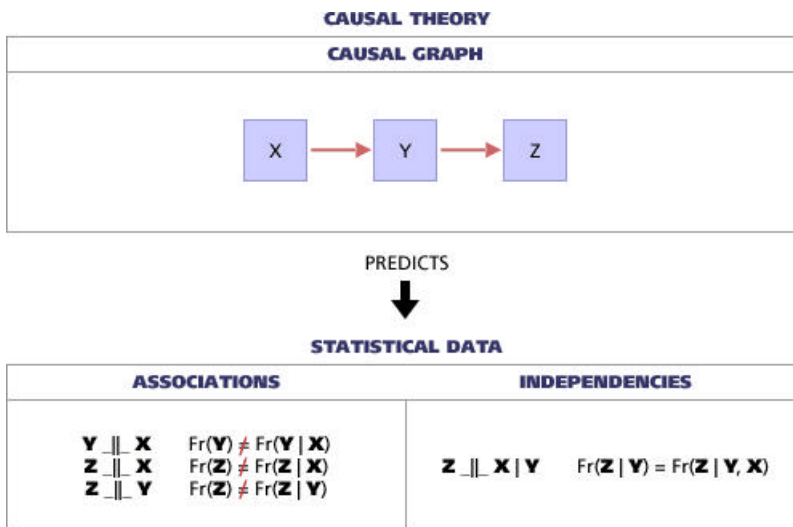


FIGURE 3200-1

To belabor the point, the problem is that more than just this graph explains the same set of independencies and associations :

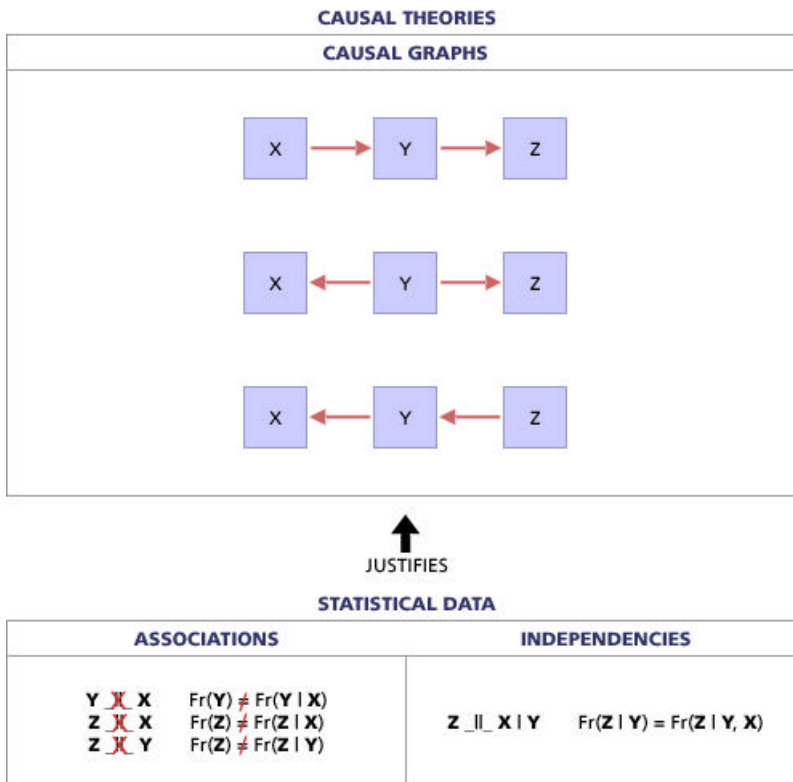


FIGURE 3200-2

All of the causal graphs in the top of the box predict exactly the associations and independencies in the bottom.

Are these the only causal graphs among just **X**, **Y**, and **Z** that predict these associations and independencies?

[< A link to exercises in the interactive version of this module. >](#)

3300: Reducing Underdetermination

What can we do to reduce the underdetermination? Some of the possible solutions are covered in other modules ("Confounding" and "Experiments"). But one thing that we can do right now is to consider adding information beyond just the associations. For example, if **A** and **B** are associated, then even if we know that there are no (unmeasured) common causes of them, we still don't have enough information to distinguish between these three causal theories:

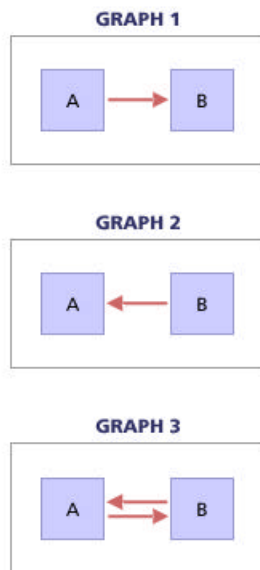


FIGURE 3300-1

However, suppose that we also know that **A** always precedes **B** in time. In that case, because later events cannot cause earlier events, the two theories in which **B** is a cause of **A** are ruled out.

[< A link to exercises in the interactive version of this module. >](#)

The number of possible graphs (and therefore the degree of the underdetermination) goes down if we are willing to make more auxiliary assumptions. We might have information about:

- + Time order of the variables
- + Which variables could not plausibly cause other variables
- + Whether there are unmeasured common causes
- + Whether there are any cycles (direct or indirect) in the graph

Extra knowledge or assumptions can only help to reduce the number of alternative graphs. With enough knowledge or enough assumptions, we can of course reduce the underdetermination completely.. For example, if **A** and **B** are independent, and we assume that associations from multiple causal connections don't balance out, then the only possible graph among these two variables is:



FIGURE 3300-2

Understanding that associational data underdetermines causal theories is important, but it is by no means the end of the story; rather it is just the beginning. What we really want is a grip on exactly which causal theories explain the associations in the data we have and which do not.

[4000: Problems with Causal Discovery](#)

[4100: Underdetermination and Unmeasured Common Causes](#)

Let us take a closer look at the problem of underdetermination in causal discovery. Consider the simplest case: we have two variables, **X** and **Y**, and we know that they are associated. Furthermore, we can even suppose that **X** occurs before **Y** in time (i.e., we're taking advantage of the strategy from the previous section of adding some non-associational information). Even with this knowledge, we still face the problem of underdetermination, because all three causal graphs below are possible.

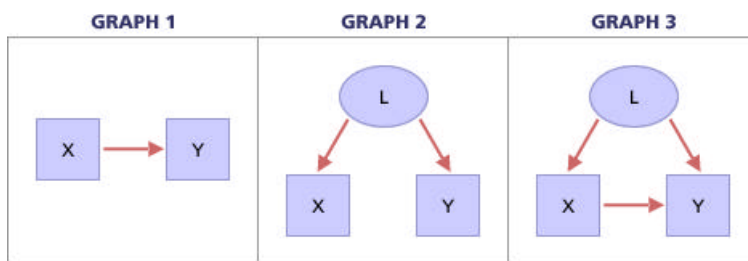


FIGURE 4100-1

In fact, this way of portraying things makes the situation look much better than it actually is. In alternatives 2 and 3, L is shown as an unmeasured common cause of X and Y. But "L" is just a placeholder that has no specific content. There might be more than one common cause. There might be two, or three, or an astronomical number. In this notation, the unmeasured common cause "L" stands for: "there is **at least one** unmeasured common cause." So really alternatives 2 and 3 are shorthand for an incredibly long list of theories.

In general, the causal relationship between two variables will **always** be underdetermined if we **only** know that the two variables are associated (and possibly that one of them occurs before the other). Remember that two variables are associated only if they are causally connected. Therefore, we know that at least one of the following is true:

- 1 X causes Y;
- 2 Y causes X; or
- 3 there is a common cause of X and Y.

But knowing which variable came first will only rule out either possibility #1 or possibility #2. One of those two will always remain, and no amount of temporal information will ever rule out #3 (since there could always be a common cause that came before both of the variables). The possibility of a common cause means that associational information will **always** underdetermine the causal theory. The only way to avoid underdetermination of the causal relationship between two variables is to make the **theoretical** assumption that we have found all of the common causes.

[4200: Data Problems](#)

[4210: Measurement Error](#)

Although the problem of underdetermination is general, there are specific varieties of it. For example, there can be problems with the data. So far, we have implicitly assumed that we can at least learn which variables are associated and which are independent. But actual scientists aren't just handed associations and independencies, they must go out and collect data from the world. This section and the next one explore some of the problems that can creep in during the data-collection step.

Measurement Error

There are actually two different ways in which measurement error can affect our data. To capture the first route, we need to draw a (rough) distinction between conceptual variables and measurement variables. Roughly speaking, measurement variables are variables that we can actually measure, and conceptual variables are the (typically broader) concepts/variables that we think are the actual causal factors.

Consider an example. Suppose we want to know what causes a person to commit a crime. Suppose we think one potential causal factor is an individual's **RELATIONSHIP WITH FATHER**, with values good, average, or poor. We might think, for example, that people with a poor relationship with their fathers have lower self-esteem as a result, and so are more likely to commit crimes. **RELATIONSHIP WITH FATHER**, however, is a conceptual variable, not a measurement one, since we cannot directly measure the value of **RELATIONSHIP WITH FATHER**. Instead, we can obtain the value of a measurement variable, such as **NUMBER OF FIGHTS WITH FATHER IN PREVIOUS MONTH**. We can (almost) directly measure this second variable, and we think that it is directly related to the conceptual variable in which we are interested.

The problem is that the associations we get among the measurement variables may not accurately reflect the associations among the conceptual variables. For example, suppose that **NUMBER OF FIGHTS WITH FATHER** is associated with **NUMBER OF CRIMES COMMITTED**. This association might lead us to believe that **RELATIONSHIP WITH FATHER** is associated with **CRIME**. However, the association between the measurement variables is entirely consistent with the conceptual variables being independent (if, for example, committing crimes causes one to have fights with one's father). Therefore, the use of measurement variables can lead to misleading data.

Consider another example. Suppose we believe that taking a pledge of virginity delays the time at which teenagers become sexually active. The two conceptual variables in this examples are: **TOOK VIRGINITY PLEDGE** and **AGE AT ONSET OF SEXUAL ACTIVITY**. How might we measure these conceptual variables, however. In both cases, we might simply interview people and ask them to self-report 1) whether they took a virginity pledge, and 2) what age they were when they began sexual activity. Even if we conducted the interview confidentially, could we be sure that the age reported for number 2 would be the right age, or might subjects who had pledged virginity have a motivation to exaggerate how old they were when they became sexually active? In this case, the difference between the conceptual variable and the measurement variable might be quite important.

?

Even if we can directly measure our conceptual variables, we still face a different kind of measurement error: namely, that we might not accurately record the values of the variables for all of the individuals. Suppose, for example, that we have a population in which blond hair and smoking are associated. In particular, we can suppose that:

- + $\text{Fr}(\text{HAIR COLOR} = \text{Blond}) = .5$;
- + $\text{Fr}(\text{SMOKING} = \text{Yes}) = .5$; and
- + $\text{Fr}(\text{SMOKING} = \text{Yes} \mid \text{HAIR COLOR} = \text{Blond}) = .6$

Since $\text{Fr}(\text{SMOKING} = \text{Yes})$ is different from $\text{Fr}(\text{SMOKING} = \text{Yes} \mid \text{HAIR COLOR} = \text{Blond})$, the two properties are associated. But if there are 500 people in the population and we **misrecord** 25 of the blond-haired smokers as non-smokers, then the two frequencies will be equal. In other words, if we do not correctly record the value for each individual, then our data will be that smoking and blond hair are independent, even though they are actually associated.

The example we've given here is perhaps a bit deceiving. After all, it's quite simple to find out whether someone smokes or has blond hair -- we just look! So, why would we be worried about measurement error? The problem is that we usually are considering variables that are not so easy to measure. For example, we might want to know whether one's yearly income causes crime. But how do we measure the variable **YEARLY INCOME**? One possibility would be to ask people how much money they make. But people can lie, or they might make a mistake in calculating their income, or they might just be sloppy. Or what about measuring a variable such as **NUMBER OF NIGHTS OF BINGE DRINKING IN THE PAST MONTH**? There is very little reason to think that every individual will truthfully report their value for this variable, and so we are faced with measurement error, which can lead to inaccurate data.

[< A link to exercises in the interactive version of this module. >](#)

4220: Selection Bias

There are two kinds of selection bias that can lead to problems with our data: treatment selection bias, and sample selection bias. Although both of these kinds of problems make causal discovery hard, they do so for different reasons, and so we will consider them separately.

Treatment Selection Bias

In most experiments, one variable is designated as the **treatment** and another as the **response**. For example, in studies to assess the impact of a computer tutor for algebra on educational achievement, the treatment is whether or not a student used the computer tutor, and the response might be the amount the student learned about algebra, for example. Treatment bias arises when the assignment of treatment depends upon some other cause of the response **besides** the treatment itself.

A common form of treatment selection bias arises when subjects in an experiment get to choose their own treatment. For example, suppose we offer each student in a class of 100 algebra students a choice as to whether they use a computer tutor or not. Suppose that the students who have already become comfortable with computers choose to use it, and the others choose not to. If being comfortable with computers is associated with being good at learning algebra, then the experiment will have a treatment selection bias. Any association between tutor use and algebra learning might not be from the causal influence of the tutor, but rather from the fact that tutor users began the study better at algebra learning!

For another example, we might want to know whether a particular drug stops hair loss in men. Now suppose that hair loss is actually caused by stress, and the drug is ineffective. That is, suppose that the causal graph is:



FIGURE 4220-1

As the experimenters, we control who gets the drug and who doesn't. If we only give the drug to people who are not stressed, then the drug will be associated with hair loss, even though there is not, in the natural setting, a causal connection between **DRUG** and **HAIR LOSS**. Why? Because **DRUG** is now an effect of **STRESS** (since **STRESS** = No causes **DRUG** = Yes). We have an association that is induced by the particular way we chose who got the drug. If we had selected people randomly to receive the drug, then taking the drug and hair loss would have been independent.

In general, treatment selection bias will be a problem when the criteria we use to determine which individuals receive some treatment are associated with some other cause of the response **besides** the treatment itself. For a more detailed discussion of treatment selection bias, you can also look at the module on "Experiments," and particularly the sections on randomized assignment of treatment.

< [A link to exercises in the interactive version of this module.](#) >

Sample Selection Bias

Roughly, sample selection bias occurs when having a certain value of one of the variables under study makes it more likely to be excluded from the study itself. For example, suppose that we want to know whether a particular drug will cause lower blood pressure.

Furthermore, let us assume that the actual causal graph is:

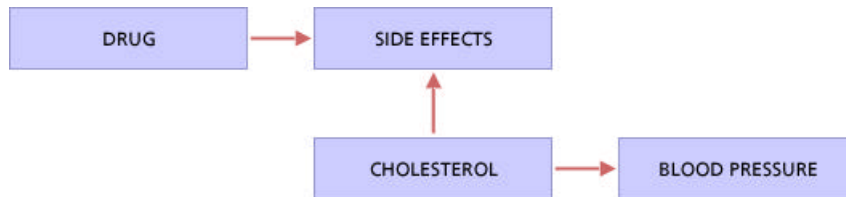


FIGURE 4220-2

If the side effects are sufficiently bad, then people who suffer from them won't continue being part of the study. That is, all of the individuals in our study/sample will have the value **SIDE EFFECTS = No**. Since every individual in our study has **SIDE EFFECTS = No**, the associations and independencies in our sample will all actually be **conditional on SIDE EFFECTS = No**. In this case, **SIDE EFFECTS** is a collider in the actual graph. As we learned in an earlier module, conditioning on a collider can induce an association along that undirected path. So, **DRUG** and **BLOOD PRESSURE** will be associated in our sample, even though there is no causal connection between them! If we use the association we would observe between **DRUG** and **BLOOD PRESSURE** to conclude that the drug helps lower **BLOOD PRESSURE**, we would be mistaken because of Sample Selection Bias.

5000: Interactive Exploration

In this section you will use the Causality Lab to explore hands-on how causal theories are underdetermined by observed patterns of associations and conditional associations. The problem you will explore involves the variables **FISH POPULATION** and **ACIDITY OF LAKE WATER** throughout the United States. For this exploration, we will assume that none of the data problems from the previous sections exist; that is, there is no measurement error, or selection bias.

In the Causality Lab exercise below, you will first need to determine whether there is an association between **FISH POPULATION** (with possible values Normal and Low) and **ACIDITY OF LAKE WATER** (with possible values Normal and High). You will then be asked to construct several causal theories among these two variables, all of which successfully predict whether there is or is not an association between these two variables.

In order to do these exercises, you will need to go over the Causality Lab Manual.

[< A link to exercises in the interactive version of this module. >](#)

In fact, **FISH POPULATION** and **ACIDITY OF LAKE WATER** are associated, i.e., they are not "independent." If high acidity caused a low fish population in a lake, then we would expect to see this dependency. However, we would **also** expect to see this same dependency if a low fish population was responsible for a high acidity.

This latter possibility is indeed plausible, even though it seems far fetched. If a large number of fish die-off in a lake because of industrial pollution, and the decaying fish produce excess nitrogen content, then the vegetation in the lake will rise. An increase in the vegetation in a lake does cause an increase in the lake's acidity; such lakes are known as "brown-water lakes". So, it is plausible that **FISH POPULATION** is a direct cause of **ACIDITY OF LAKE WATER**.

You can use the Causality Lab to learn the independencies between variables in the entire population (as opposed to just in a sample). This is something you can seldom do in a real population. For practical reasons, we have to take samples. Because learning the independencies in an entire population can't be done by measuring, the software lets you consult an "Oracle" with special knowledge of the entire population.

[< A link to exercises in the interactive version of this module. >](#)

So, for a causal hypothesis to be consistent with the evidence you now have about the population, that hypothesis must imply that **FISH POPULATION** and **ACIDITY OF LAKE WATER** are associated.

The Causality Lab has tools that let you see what alternative causal hypotheses imply with respect to independences in the population. You can use this tool to see if a causal hypothesis you conjecture (a causal graph), makes the right prediction about which independencies actually hold in the population.

[< A link to exercises in the interactive version of this module. >](#)

So, all three causal graphs predict that we will see the same independencies in the population of lakes. They each predict that the two variables will be dependent. But, those associations are the only evidence you have to work with! So, the evidence won't let you choose one over another!

Think about the problem that we were originally faced with: we wanted to learn if changing acidity in lakes would affect the fish population. Empirically, we learned that there is an association between acidity of lakes and the population of fish in the lakes. The important question is: What is the causal relation? What we have learned here tells us that the association alone cannot choose between the three possible causal hypotheses. The associational evidence underdetermines the theories.

6000: Summary

Representing causal claims involves clarifying what we mean by them. For example, when we draw a causal graph and include an edge from variable X to variable Y , we need to be precise about what the presence of the edge means, and what we are claiming about the world when we draw such an edge. Discovering causal claims involves moving from associational statements, for example that smokers are ten times likelier to get lung cancer than non-smokers, to causal claims that such statements support, for example that smoking causes lung cancer.

Discovery is hard for several reasons. For one, evidence is nearly always limited. That is, we can rarely acquire all of the data that we want. Furthermore, even when we can acquire all of the associational data that we want, the possibility of unmeasured common causes means that the causal connections underlying every observed association are always underdetermined. There are also often problems with our data, such as measurement error or sample selection bias. These problems compound the difficulties that we would almost certainly face purely because of theoretical underdetermination.

Nevertheless, causal discovery is not always impossible, as we have sources of data that can help us. The first is the kind of data we have been discussing: statistical data, such as tables that record the values of a variable for each individual in a population. The second is knowledge about the process by which the data were gathered. Did we consult census data? Do an experiment, etc.? This knowledge can help us account for data problems. The third source of information is background knowledge conceived of more generically. We know that a person's income cannot cause his age, for example. Although such knowledge rarely **uniquely** determines the causal knowledge we are seeking, it almost always reduces the possibilities we need to consider.
