

## Causation vs Association

---

### 1000: Introduction

---

Scientists build theories to try to explain the world around us. Part of scientific method is conjecturing theories about causal relationships and testing them against evidence. Our topic is **causal** theories. Causal theories are everywhere. For example, epidemiologists, biologists, and doctors believe that AIDS is **caused** by the HIV virus. Since it is clearly unethical to gather direct experimental evidence relevant to this theory (we can't intervene purposely to infect someone with the HIV virus), much of the evidence for this theory is the strong association between having AIDS and testing positive for the HIV virus.

Not all associations support causal theories, however. There is an association between owning a red car and getting in accidents, but few people think the red color of the car causes accidents, and in fact, a red color may be more visible than other colors, tending to reduce accidents. Instead, people who tend to be thrill-seeking risk takers; that is, people who tend to get in accidents also tend to like the color red for their cars.

To understand any science, one must know where the theory ends and the evidence begins. In this curriculum, the theories concern causal ideas and the evidence associational ones. Five of the modules in this curriculum cover causation:

- 1 Event Causation
- 2 Variable Causation
- 3 Determinism and Indeterminism
- 4 Causal Graphs
- 5 Interventions

Four of the modules in this curriculum focus primarily on association:

- 1 Relative Frequency
- 2 Conditional Relative Frequency
- 3 Independence and Association
- 4 Conditional Independence

The purpose of this module is to again emphasize the difference between **causal theories** and **associational evidence**. Unfortunately, even famous scientists sometimes confuse theory and evidence. Early in the 20th century, Karl Pearson, a famous statistician, wrote an influential book called *The Grammar of Science* in which he argued that there is no difference between causation and association, they are one and the same. Ronald Fisher, a still more famous statistician, took issue with Pearson, arguing that causation and association are related, but distinct ideas. Fisher developed the subject of experimental design in order to give scientists a practical method for determining when associations are causal and when they are not. Bertrand Russell, a brilliant philosopher writing independently at about the same time, argued that causation is a metaphysical fantasy that we superimpose on top of associations. According to Russell, only associations are real.

Causation and association are indeed quite different ideas. The whole subject of causal and statistical reasoning depends on carefully distinguishing the two ideas. Later modules in this course discuss both ideas in detail. We will also explore how causation and association are related: some lessons explore how causal structures produce associations, some explore the difficulties of moving from associations to causal hypotheses, and some discuss strategies to overcome these difficulties. This module, however, tries to separate causation and association as clearly as possible. We begin the module with two stories from life in Pearson Land, a mythical place where everyone believes that association and causation are one and the same.

---

### **2000: Stories from Pearson Land**

---

### **2100: The Government Sells Soap**

---

By the late 1960s, lung cancer had become a big public health problem in Pearson Land, killing more people than war, auto accidents, and the flu combined. The Government of Pearson Land faced a policy problem: how to reduce lung cancer at the lowest cost.

It was indisputable that smoking and lung cancer were positively associated. Lung cancer was more than 10 times as frequent among smokers than among non-smokers of the same age. Perhaps the government could reduce lung cancer by reducing smoking. Initial feasibility studies were discouraging. Getting people to stop smoking was extremely difficult even under ideal conditions, and the tobacco companies were extremely powerful and determined to resist any anti-smoking policy the government might propose.

Fortunately, Bert and Russell, two government policy makers who believed that causation is just metaphysical pixie-dust and that the real issue is always what is associated with what, realized that there was a much cheaper policy for combating lung cancer, a policy that Big Tobacco would almost certainly support. Careful data analysis showed that there was another property that had almost as high a positive association with lung cancer as did smoking: having nicotine-stained fingers!

Sensing that fame and fortune were close at hand, Bert and Russell, ran down to the Government labs to do some sleuthing about soaps that could remove nicotine stains. As luck would have it, one of the brighter scientists there, Phil Morris, had just perfected a formula that would eradicate stains after just 2 minutes of vigorous scrubbing a day. The soap produced no side effects and cost next to nothing to manufacture in huge quantities. Bert and Russell nearly sprinted to see the Surgeon General, who was looking rather depressed. "Cheer up, boss," they said, "we can drastically reduce lung cancer if we can get everyone to eliminate nicotine stains from their hands -- and we have the soap to help them do it!"

The Government quickly passed the Hand Hygiene Act, which required all citizens to wash their hands every day with Phil Morris' special soap.

---

### 2200: The Man Who Would Be Early

---

Once upon a time in Pearson Land, there was a young man named Harry who was nearly always late. Try as he might, Harry could not seem to get to any of his appointments on time. At the university, Harry took a course on research methods in social science, which, fortunately for him, had a professor who showed up late a lot too. In this class Harry learned the virtues of gathering hard data as opposed to soft opinions, and he became so enamored of the scientific method that he decided to apply it to himself. "I will study the habits of people who are early and who are late," Harry proclaimed, "and let science guide me to the land of the punctual."

Harry carefully studied dozens of people for several weeks, recording lots information about them. He analyzed his data and discovered that one property was strongly associated with being late -- hurrying! People who hurried vigorously on the way to their appointments were almost always late. People who walked at a normal pace were almost always on time! This must be right, Harry thought, I'm always hurrying and I'm always late too. Thanks to Science, I have finally found the answer -- I must stop hurrying to get to my appointments, and stroll leisurely instead!

---

### 3000: The Causal Story

---

### 3100: Introduction

---

The policies proposed in Pearson Land would not work in the real world. Harry's intervention would make him even later than usual, and Bert and Russell's hand washing campaign would do nothing to diminish the frequency of lung cancer. Why? The **associations** in Pearson Land are true enough -- hurrying to appointments really is positively associated with being late, and having fingers with nicotine stains really is positively associated with lung cancer. But these associations are not causal relationships. What happened was that Harry, Burt and Russell mistakenly believed that associations are causal relations.

---

### 3200: The Causal System for Smoking

---

Consider Burt and Russell's lung-cancer prevention program. Here is the causal story behind smoking, finger stains, and lung cancer in the real world (or "natural system"):

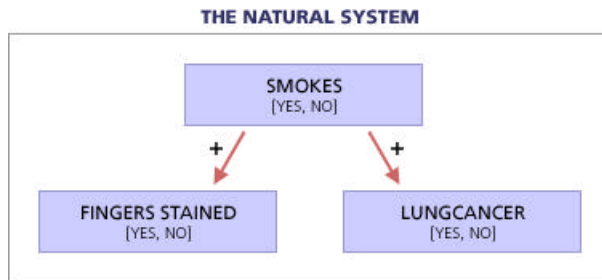


FIGURE 3200-1

In this causal graph, the red arrows mean that there is a direct causal relation from one variable to the other. We put a "+" on the arrow to mean that when the cause is assigned to have the value Yes, it makes it more likely that the effect will have the value Yes than if we had not assigned the cause to have Yes.

Causation involves action, decision, mechanisms, response to interventions, and influence. Association involves information and belief. What the causal graph gives us is some sense of what would happen **if we intervened to act** in certain ways. The causal graph above describes the "natural system" that produced the data Bert and Russell analyzed prior to suggesting their policy of hand washing. In the natural system, whether a person smokes influences whether they have stained fingers. In their post-policy world, however, **everyone** is to wash their hands with their special soap every day. We can call the world after the enactment of the Hand Hygiene Act an "experimental system." The "experiment" is produced by a particular intervention -- everyone washes the nicotine stains off their hands. What influence does smoking have on finger stains in this experimental world? What does the causal story in the "experimental system" involving finger stains, smoking and lung cancer look like?

< A link to exercises in the interactive version of this module. >

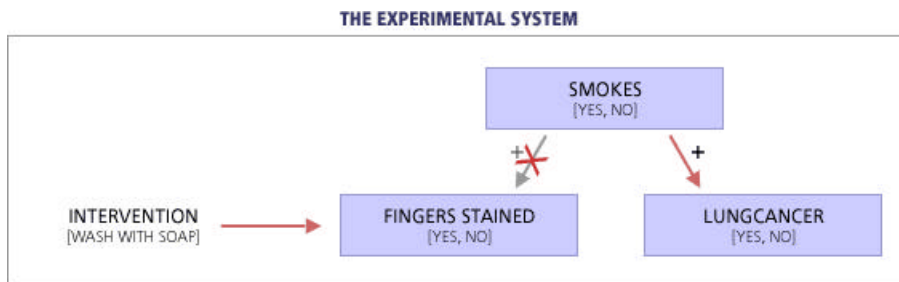


FIGURE 3200-2

In the "experimental system," smoking no longer causes finger stains, because everyone is washing their hands with special soap every day. To represent this intervention, we "X out" the arrow from **SMOKES** to **FINGERS STAINED**. **HANDS WASHED**, for sure, has no causal influence good or bad on **LUNG CANCER**. In the "experimental system," **FINGERS STAINED** and **LUNG CANCER** won't even be associated.

---

### 3300: Harry's Causal System

---

Consider the real causal story behind Harry's punctuality problem:

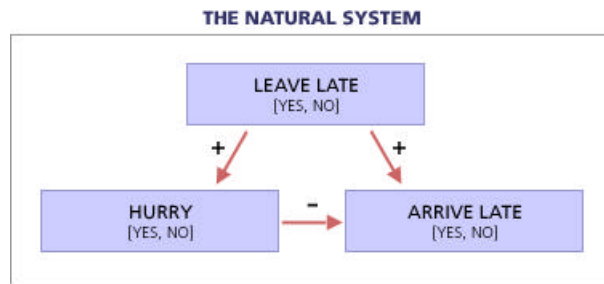


FIGURE 3300-1

Leaving late is a common cause of hurrying and arriving late, and at least partly explains why they are positively associated. The arrow from **HURRY** to **ARRIVE LATE** has a "-" on it because, if we assign someone the value Yes to the variable **Leaves Late**, then assigning **HURRY**= Yes reduces the probability of arriving late. Hurrying doesn't always prevent arriving late, of course, but if we left just a little late, sometimes it will be enough. In this story, we will assume that people who left late can rarely make it up by hurrying.

< A link to exercises in the interactive version of this module. >

Harry's policy of never hurrying can be depicted with a causal graph for the "experimental system" as follows:

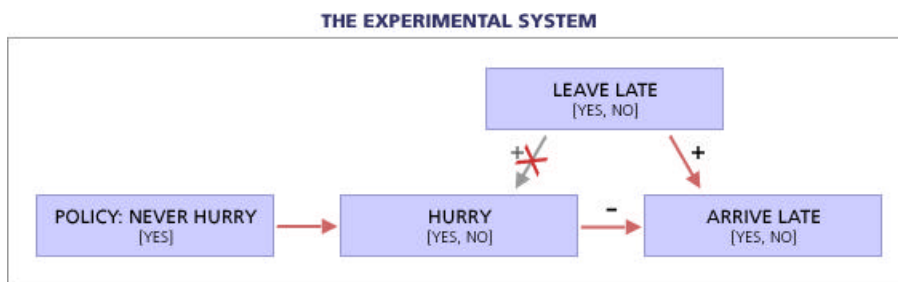


FIGURE 3300-2

In the data Harry originally collected, which was produced by the "natural system," whether people hurried was influenced causally by whether they left late. In Harry's new life-style, which we call the "experimental system," leaving late has no influence on whether Harry hurries, because Harry has decided ahead of time never to hurry! To represent the effect of Harry's intervention, we "X-out" the arrow that was in the natural system from **LEAVE LATE** to **HURRY**. What happens in Harry's experimental system? For one, he never hurries, and this will make it more likely that he will arrive late, all else being equal. Since Harry almost always leaves late, all else is not equal and now he has maximized his chances to arrive late because he is leaving late **and** not hurrying!

4000: Association

---

## 4100: Association as Information

---

In both of these stories, the correct causal theory allowed us to predict how the system would **respond** to an **intervention**, something the association alone cannot tell us. This is the fundamental difference between causation and association.

Associations cannot tell us what **will happen** if we perform a certain action or intervention, or what **would have happened** had we intervened in a certain way, but they can give us information. A property like "being male" is **associated** with another property like "having a beard" if the knowledge that someone has one property gives **information** about whether they have the other.

For example, suppose that I pick randomly out of a huge barrel containing a small chip for each American. What are the chances that I picked a male? About 50%. Now suppose I put back the first chip, stir the barrel for a few minutes, and then pick another. Before I ask you the chances of having picked a male, suppose that I first tell you that the person I picked has a beard. Now what are the chances that the person is a male, given you know they have a beard? Still about 50%? No, obviously they are a lot higher than 50%. Having a beard and being male are positively associated, because **knowing** that someone has one property raises the chances that they have the other.

So having a beard is positively associated with being male, but does having a beard cause maleness? No. Suppose, instead of learning that the person whose chip we picked was bearded naturally, we **intervene** on that person to give them a beard. No matter who they are, we put glue on their chin and attach a very realistic looking beard. Now suppose I tell you that I picked a person randomly out of the barrel, and then intervened to make sure they have a beard, waited a while, and then checked to see if they were male or female. Are the chances they are male other than 50%? No.

Having a beard and being female are associated also, but because knowing that someone has one property **lowers** the chances that they have the other, they are **negatively** associated. If knowing that someone has one property doesn't change the chances that they have another at all, then the two properties are not associated. In that case we say they are **independent**. For example, suppose I again put back the chip and spin the barrel, and then pick another chip. Suppose I tell you that the person lives east of the Mississippi. Now what are the chances that the person is a male? Still about 50%. Maybe the proportion of males living east of the Mississippi is a **tiny** bit different than the proportion of males in America overall, but being male and living east of the Mississippi are certainly nearly independent.

### Information

Learning something gives you information if it changes the state of your uncertainty. For example, if I randomly pick a card out of a regular deck and before I show it to you ask you what suit you think it is, you will be uncertain. You might tell me that there is a 1/4 chance it is Hearts, 1/4 chance it is Diamonds, 1/4 chance it is Spades, and 1/4 chance it is Clubs. Now if I tell you the card is red, I have given you information -- I have changed your state of uncertainty to one in which there is now a 1/2 chance it is Hearts, a 1/2 chance it is Diamonds, no chance it is Spades and no chance it is Clubs. If I then tell you that the card is not a Spade, I have given you no new information.

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

---

#### 4200: Harry's Associational Story

---

What Harry observed in weeks of research was a positive association between hurrying and arriving late. Among the people he observed, those who hurried had a higher chance of arriving late than those who didn't. For simplicity let's assume that overall, half of the people Harry observed were late, and half on time. Among the half that were late, the proportion of those who hurried was higher than among the half that were on time.



FIGURE 4200-1

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

As you can see, we are not asking about what would happen if someone Harry observed had been made to do something different, or what will happen if they are made to do something different the next time they set out to go to an appointment. We are asking about the informational connection between properties of people when we have not acted to alter their behavior.

---

#### 4300: Burt and Russell's Associational Story

---

By saying that there is a positive association between having nicotine stained fingers and getting lung cancer, we are making a claim about the information we gain about one property from learning about whether someone has the other.

Again, imagine we had a large barrel that contained one chip for every American. Suppose I tell you that 28% of Americans smoke.

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

---

### [5000: Cause vs. Association](#)

---

### [5100: Questions](#)

---

Causal claims involve how a system will respond to an intervention. Associational claims involve information, but are neutral with respect to how the system will react to an intervention.

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

---

### [5200: A Case Study: Media Violence and Real Violence](#)

---

The examples we've used thus far seem quite simple. It's hard to imagine how someone could think you could reduce lung cancer by having people wash their hands or that you could insure that you are always on time by never hurrying. Yet, you will find cases every in which people make causal claims based on associational evidence alone. For example, read the following excerpt from an article about violence in the movies and video games and violent acts of teenagers.

#### [Excerpt](#)

"President Clinton seems to think that the answer to the Littleton massacre can be found at the movies. Yesterday in an Oval Office ceremony, Mr. Clinton announced that he had persuaded the National Association of Theater Owners to demand that youths produce photo identification before seeing R-rated movies. And last week the President launched a \$1 million federal inquiry into entertainment industry marketing techniques...."

"Numerous correlative studies (the President cited the figure as 300, but the actual number is closer to 1,000) indicate an association between media violence and aggression."

(From "Violence Doesn't Begin in the Theater", Wall Street Journal, June 9, 1999, Jonathan Kellerman)

Clearly, President Clinton proposed an intervention that would reduce the number of teenagers who see violent movies. The reason President Clinton proposed this intervention is that he believes that reducing the number of teenagers seeing violent movies will reduce the number of teenagers who commit violent acts (like the Littleton massacre).

< A link to exercises in the interactive version of this module. >

One plausible causal explanation of an association between viewing media violence and violent acts by teenagers does not claim that viewing media violence is a cause of violent acts by teenagers. It might be the case, for example, that teenagers who have aggressive tendencies both are more likely to see violent movies and are more likely to commit violent crimes. In that case, the "natural system" would look like this:

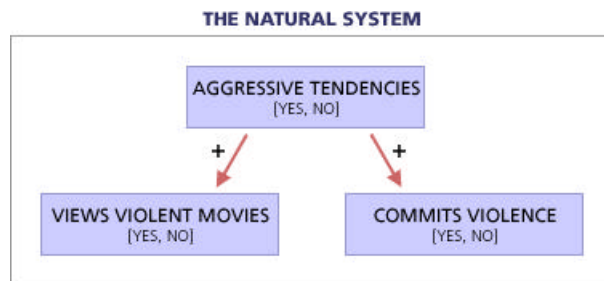


FIGURE 5200-1

Notice that if this is the natural system, then if we intervene to prevent teenagers from seeing violent movies (Clinton's proposed policy), then the "experimental system" would look like this:

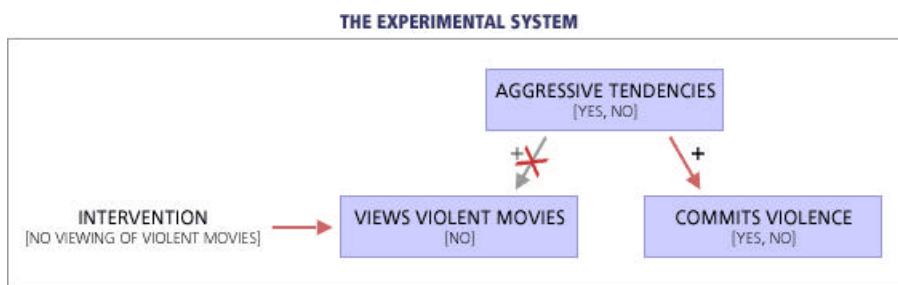


FIGURE 5200-2

In the experimental system, aggressive tendencies no longer causes teenagers to see violent movies (because they aren't allowed to). Thus, in the experimental system, viewing violent movies will not be associated with committing violent acts.

This is one of the many possible causal explanations of the observed association reported in the article. The efficacy of an intervention depends on which causal explanation is true. The associational evidence alone does not tell us whether prohibiting teenagers from viewing violent movies will actually reduce violent acts by teenagers.

---

### 5300: The Asymmetry of Causation and the Symmetry of Association

---

Association is a **symmetric**. If one property, like wealth, is associated with another, like education, then education is also associated with wealth. Put another way, if learning that an individual is wealthy changes your uncertainty about his or her level of education, then learning about an individual's educational level first is also informative about his or her wealth. This is true generally for any properties P1 and P2. If P1 **and** P2 are associated, then P1 is informative about P2, and P2 is informative about P1.

Causation is **asymmetric**. If one property, like exceptional athletic talent, is a cause of another, like wealth, that does **not** mean that wealth is a cause of exceptional athletic talent.

The simulation below illustrates these ideas. You can close and open the switches to directly change the state of the light bulb on the left and the fan on the right. Your job is to figure out the causal relations (or their absence) between the light bulb and the fan. When you have figured things out, answer the questions that follow.

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

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

In the previous exercise, your background knowledge gave you a large clue about the causal direction -- in the image below you have no such clues. Nevertheless -- you can determine the causal direction in the same way as you did in the previous exercise. Click on either object to change its state.

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

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

Because causation is asymmetric doesn't mean that symmetric cases of causation are prohibited, it only means they are not necessary. In some cases causation can go both ways. For example, losing sleep can cause anxiety, and anxiety can also cause a loss of sleep. Higher wages can cause inflation, and inflation can cause higher wages. Success causes confidence, and confidence causes success.

When a relationship is said to be **asymmetric**, it means that, from the relationship holding in one direction we cannot infer that it holds in the other. Love, unfortunately, is asymmetric. Because X loves Y doesn't mean that Y loves X. Y might love X also, but it isn't required. When a relationship is said to be **anti-symmetric**, it means that, from the relationship holding in one direction we can infer that it does **not** hold in the other. For example, the parent-child relationship is anti-symmetric. If X is a parent of Y we know that Y is not a parent of X. When a relationship is said to be **symmetric**, it means that, from the relationship holding in one direction we can infer that it also holds in the other. For example, being a sibling is a symmetric relationship.

**TABLE 5300-1: RELATIONS**

Type	Explanation
Symmetric relation R1	(A R1 B) implies (B R1 A)
Anti-symmetric relation R2	(A R2 B) implies not (B R2 A)
Asymmetric relation R3	(A R3 B) does not imply (B R3 A)

Causation is asymmetric, and association is symmetric.

---

### 6000: Summary

---

Causation involves action, mechanisms, policy, response to interventions, and influence. Association involves information. If A and B are associated, then information about A gives us information about B, and because association is symmetric, information about B gives us information about A. Association of two features cannot, by itself, establish their causal relations.

Causation can be identified through appropriate interventions or experiments. If A causes B, intervening to alter the value of A produces a change in the value of B. But, intervening on B might not produce a change in A. This is why we say that causation is asymmetric. merely being associated with B.

Knowing only that two properties are associated does **not** allow us to infer how a population will respond to an intervention involving either of these properties. This is why hand washing will not prevent lung cancer, even though the two are associated, and proceeding to appointments without hurrying will not make us punctual, even though these two are also associated.

---