The question of causation

Notes

Causation vs Association

A subtle question arises after we have completed a statistical analysis. Typically the conclusion of the analysis might be something along the lines of "x and y appear to be correlated". But the real question we wanted to ask was probably: "Does x cause y to happen?"

- Correlation / Association refers to the specific dataset under consideration, and whether in that dataset there appears to be a relation between the values of x and y.
- *Causation* refers to a more fundamental, functional, relation between the two variables, that suggests that the one variable causes the other to occur.

The typical catchphrase is *"Correlation does not imply Causation"*.

There are four main reasons why a variable x might appear to be correlated to another variable y.

1. x might in fact be causing y to happen. This is called **Direct Causation**.
2. It might be y that is causing x to happen.
3. Both x and y might be caused by a third variable z (for which we might not have collected data, for example). Since they are both correlated with z, they will appear to also be correlated to each other. This is called **Common Response**.
4. There is a third variable z which, due to how the data was assembled, is correlated to x and cannot be "distinguished" as easily. In that eventuality, if z was causing y then it would appear as if x and y are correlated. In other words, we have no way to separate the effect of z from the effect on x. This is called **Confounding**.

For example, suppose we want to see if being greek has an effect on your GPA. Suppose we collected the data by visiting a sorority house, and also interviewing the football team. In this event it is likely that we will end up with a lot more greek women than men, and no non-greek women. In this case the variables "being greek" and "gender" are *confounded*. So if we see an effect on GPA, we don't really know if it is caused by gender or by being greek. This is an example of Confounding.

As an example of Common Response, suppose we compare for each state number of cigarettes bought in the state, and number of cancer deaths in the state. So these are my variables x and y, with one value for each state. It will then likely see a strong correlation between the two. The reason is simply that some states have a much higher population than others, and consequently they have more cigarette sales

and more cancer deaths. In this case this third variable, z="population in state", essentially "causes" the other two to happen, so they appear to be high at the same time and low at the same time (they just follow what z does, but we only see x and y in the graph). This is a common response case, and it is also the reason that when we want to look at these variables we always compute "per-capita" values, where we divide by the state's population.

We usually call these underlying third variables that provide an explanation for the relation between the other two "**Lurking Variables**". They are there, they explain what is going on, but for one reason or another we do not have any measurement of them, so they are hidden to us.

Each of the following is an example of two variables that are associated. For each, find an explanation for the association that is something other than "x causes y". Some are confounding cases, some are common response. Identify what the third variable z is.

1. In two groups of patients in the same amount of pain, Tylenol was given to the one group and nothing to the other. The group that had received the Tylenol reported a reduction in pain. Does that mean that taking Tylenol causes a reduction in pain? (The two variables are "taking tylenol yes/no" and "pain reduced yes/no")
2. Studying medical records, it was found that a certain anesthetic A was associated with a higher death rate, compared to other competing anesthetics. Does that mean this anesthetic is the cause of the higher death rate?
3. During WWII, it was noticed that bombers were **more** accurate and did more successful runs on days where the sky was **less clear**. Does that mean that unclear sky improves the bomber's accuracy?
4. People who use artificial sweeteners instead of sugar tend to have more weight than people who use sugar. Does this mean that artificial sweeteners cause weight gain?
5. The number of shark attacks in beaches seems to be strongly associated with the amount of ice cream sales. Explain the association.
6. Statistical studies showed that there was a strong positive association between alcohol consumption and heart disease. Does this mean alcohol causes heart disease?
7. There is a strong correlation between the per capita number of cigarettes sold in a state, and the per capita number of deaths from lung cancer in that state. Does this mean cigarette smoking causes lung cancer? (this is a different kind of problem, but think about it)

We leave this section with this xkcd comic[1]:

[1] http://xkcd.com/552/

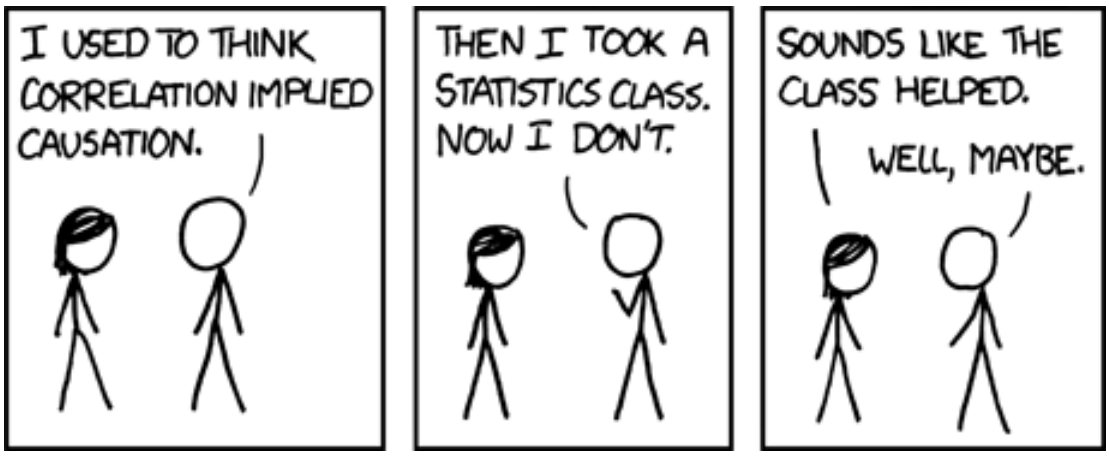

Figure 1: Correlation and Causation

Making a case for Causation

Given all these problems, how does one build a case for a causal relation between two variables? These days probably everyone would agree that smoking causes cancer, but 30 years ago this was very much not the case. So given the strong tobacco lobby pushing back, how would one build a case?

- Best option is doing a **controlled experiment**, where we can manually adjust the x values and notice the change in the y values. Of course on something like smoking this is not possible, and we are restricted to observational studies, often spanning many decades.
- A **consistent relation** among **many diverse studies** is a next strong tool. Multiple statistical groups across the globe were conducting these studies on diverse populations, thus eliminating most genetic alternative interpretations.
- The relation is **plausible**, which one might be able to confirm by experiments on rats, or at least have some theoretical explanation.
- **Higher dosage associated with stronger effects**: Those smoking more exhibited a higher risk of having cancer.
- There might be a **time effect**. Cancer takes 2-3 decades to exhibit. Male subjects started exhibiting high incidence of cancer about 20-30 years after smoking became prevalent among men. The same thing happend with women, but 20-30 years after smoking became prevalent among women.