Notes by Vasilis Oikonomou

Chapter 2: Causality and Experiments

Main question adressed in this chapter: How do we establish causal relationships?

Key terms: observational study, treatment, outcome, association, causality, treatment group, control group, Randomized control trial (RCT), confounding factors, individual

I like to think of an exploration with data as a 3 stage process:

① Observation ⟶ ② Analysis ⟶ ③ Result

Let's break down each step and see what happens in each one.

① Observations: This is the point were you develop an idea of what you want to examine. Here there are three main things you have to establish before moving forward:

ⓐ Who is the individual I am interested in?
This is your main stakeholder. It could could be the individual, a group of people or even a collection of the US states, a country, literally anything.

ⓑ What is the treatment I want to investigate?
A treatment is the factor of interest. The treatment is the part of your observation that you believe produces an outcome on your individuals. Eg If you are thinking of

© Outcome

The outcome is the effect that you believe the treatment has on the individual. For example in the investigation of whether drinking coffee causes lung cancer, lung cancer is the outcome you think your treatment (drinking coffee) can have on the individual (in this case the people.

Punchline: Any relation that you have observed between the treatment and the outcome is called an association.

e.g. In the example of coffee drinking and lung cancer, someone observed that regular coffee drinkers tend to get lung cancer more often than people who don't drink coffee regularly. This is an association that you have established.

But establishing an association does not tell us anything about whether the treatment causes the outcome. e.g. Is coffee the reason people get lung cancer? No, but in the old days there was an association between the two.

!!! ASSOCIATION DOES NOT IMPLY CAUSATION !!!

We need to establish causality

② Analysis: This part is the most crucial of the process. Here we take the necessary steps to prove that an association is a causal relationship.

But why is association different from causation?

The answer lies in what we refer to as confounding factors. This essentially refers to other reasons which underly the relationship between treatment and outcome and for which we did not account for. For example, although we observed that people who drink coffee have a higher chance of developing lung cancer, we failed to account for the fact that (especially in older times) people who drank a lot of coffee also tended to smoke a lot (which we know is a cause for lung cancer). This was our confounding factor.

To protect ourselves from confounding factors that mislead us, we introduce ... Randomization

! Without Randomization, you cannot prove causality nomatter how obvious the association seems.

So how do we randomize our data?
We introduce Randomized Control Trials/Experiments (RCT/RCE)

Randomized Control Experiments: The process of splitting your population into what we call a treatment and a control group through a random process. without letting people know which group they are in.

Important point: My treatment and my control groups need not be of the same sizes. Take as an example the following random process for splitting people into treatment and control groups:
For each person in my population, I roll a fair die.

If I get a 1 I put this person in my treatment group. Otherwise, he goes into the control group. By the end of that process, my two groups will most probably by of different sizes but that is fine since the allocation process is random.

So what are those 2 groups?

Treatment group: They will take the treatment (say a pill)
Control group: They will not take the treatment (give the a placebo instead)

Remember: Noone should know which group they are part of.
If my outcome appears only in my treatment group but not in my control group, then I can prove causality.

Randomization helps us claim that the two groups are as similar as possible, namely that there is no reason other than the treatment for which the outcome appeared on the treatment but not on the control group.

In my mind RCTs help "even out" the effect of the confounding factors.

But can I always run an RCT?
It depends. In some cases it is impossible or even plain unethical to run an RCT. E.g. if I want to examine the effects of alcohol consumption on pregnant women, I cannot run an RCT since there is a high chance of risking the baby's health.
When researchers have to work with data that they had no hand in generating (like in the above case) then this is called an observational study

If for whatever reason you cannot randomize/generate your data and instead you have to work with data that is already there, then you perform an observational study and you can not prove causation.

(3) Results: Based on what hapenned in your analysis here you can claim whether you can prove a causal relationship (RCT) or not (observational study). Be very careful about detecting any potential confounding factors and state your findings clearly!