Data 8, Chapter 2: Causality and Experiments

Notes prepared by Vasilis Oikonomou
Digitized by Suraj Rampure

The main question addressed in this chapter is: **How do we establish casual relationships?**

Key terms

observational study
outcome
causality
control group
confounding factors
treatment
association
treatment group
Randomized control trial (RCT)
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. Observation

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

a. **Who is the individual I am interested in?**

This is your main stakeholder. It could be the individual, a group of people or even a collection of the US states, a country… literally anything.

b. **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**.

c. 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 believe your treatment (drinking coffee) can have on the individual (in this case, humans).

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

For instance, in the example of drinking coffee and lung cancer, someone observed that regular coffee drinkers tend to get lung cancer more often than people who do not 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. For example, 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 step is the most important part of the process. Here, we take the necessary steps to prove that an association is a casual relationship.

But why is association different than 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 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 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 the idea of **randomization. Without randomization, you cannot prove causality, no matter 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 control groups need not be of the same size. 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 place the person in my treatment group. Otherwise, he/she goes into the control group.

By the end of the process, my two groups will most probably be of different sizes, but that is fine since the allocation process is random.

So what are these two groups?

Treatment group: Those who will take the treatment (e.g. a pill)
Control group: Those who will not take the treatment (e.g. give them a placebo)

Remember, no one should know which group they are a part of. If the outcome appears only in the treatment group and not in the control group, then we 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 a RCT?
It depends. In some cases, it is impossible or even plain unethical to run an RCT. For example, 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 (such as in the above case) 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 already exists, you can perform an **observation study** and you **cannot prove causation**.

3. Results

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