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 causal 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:

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

This is your main stakeholder. An "individual" does not mean "a person", it means unique entity of any kind! Your individuals could be sections of Data 8, dogs in Berkeley, or mountain ranges.

2. What is the treatment I want to investigate?

A treatment is the factor of interest - something that you've observed that you believe produces an **outcome** on your **individuals**. Later on, we will learn how to design our experiment so that we can safely claim that the treatment **causes** a particular outcome**.**

1. **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, the **outcome** is lung cancer, and the **treatment** is what you think might have caused it (drinking coffee). The **outcome** is experienced by the **individuals**.

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

Experimental design is crucial in our efforts to establish a causal relationship between a treatment and an outcome.

Suppose we want to test the effect of a pill that claims to improve students' performance in math. We choose to take the following approach. We will form two groups; one group will receive the pill whose effect we want to test while the other will receive a placebo (a harmless pill that has no effect whatsoever). We will call these two groups, the **treatment group** and the **control group** respectively. Then, we will ask both groups to take the same math exam and compare the performance of the two groups. If the group that took the real pill performs better on the exam, then we can claim that the pill does have an effect.

In the procedure we have left one crucial detail unspecified: How are the control and treatment groups chosen?

The reason this matters is the following. For our experiment to be successful, we need to make sure that the treatment and control groups do not differ in ways other than the treatment. If this was the case, that would be catastrophic for our experiment.

In the above example, imagine if all the members of the treatment group were math Phds while everyone in the control group were high school seniors. Then clearly, the difference in performance between the two groups should have been attributed to the individuals' education level and not the effect of the pill!

Usually such differences are not as obvious which makes them even more dangerous. In the Statistics jargon such an underlying difference between the two groups (other than the treatment) is called a confounding factor, because it might confound you (that is, mess you up) when you try to reach a conclusion.

Naturally, the question that arises is the following: How can we be reasonably confident that there are no systematic differences between our treatment and control groups?

The answer is **Randomization**! Randomization is the process through which we assign individuals to a group based on a random process. You can imagine flipping a coin for every individual that participates in your study and assigning them to the treatment group if the result is heads and sending them to the control group otherwise.

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.You can think of this process as helping “even out” the effect of the confounding factors.

Without randomization, you cannot prove causality, no matter how obvious the association seems. We call experiments that use randomization **Randomized Control Trials/Experiments (RCT/RCE)**.

Important point: The treatment and control groups need not be of the same size.

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.

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 causal relationship (RCT) or not (observational study). Be very careful about detecting any potential confounding factors and state your findings clearly!

Practice Problems[1]

1. The Public Health Service studied the relationship between exercise and heart disease, in a large sample of representative households. For men and women in each age group, those who had frequently exercised since adolescence were on average less likely to have heart disease. However, those who had did not frequently exercise since adolescence were less likely to have heart disease than those who had *recently* started exercising.
    a. Why did they study men and women and the different age groups separately?
    b. The lesson seems to be that you should ideally exercise since adolescence, but if you're an adult who doesn't exercise, you shouldn't start. Do you agree?

2. California is evaluating a new program to rehabilitate prisoners before their release; the object is to reduce the rate at which prisoners will be back in prison within two years. Prisoners can opt into the program, which involves several months of military-style "boot camp" with strict discipline. 70% of those who did not participate in the program returned to prison after release, while 50% of those who did participate returned to prison. According to a prison spokesperson, "Those who complete boot camp are less likely to return to prison than other inmates."
    a. What is the treatment in the experiment? The treatment group? The control group?
    b. Is the spokesperson's statement based on an observational study or a randomized controlled experiment?
    c. Is the spokesperson's statement correct?
    d. Does the boot camp reduce the rate at which prisoners return to prison?

3. According to a study done at Kaiser Permanente in Walnut Creek, CA, users of oral contraceptives (birth control) have a higher rate of cervical cancer than non-users, even after adjusting for age, education, and marital status. Investigators concluded that the pill causes cervical cancer.
    a. Is this an observational study or a randomized controlled experiment?
    b. Why did the investigators correct for age, education, and marital status?
    c. Women using the pill likely differed from non-users in another factor which influences cervical cancer. What could that factor be?
    d. Were the conclusions of the study justified by the data?
    e. Another study at the University of Copenhagen surveyed 1.8 Million women in Denmark, and found that women using the pill have higher rates of *Breast* Cancer then those not using the pill. There is no obvious confounding factor between birth control use and breast cancer. In this case, can we establish causation?

[1] Adapted from Freedman, Pisani, and Purves' Statistics, 4th edition.

Answers

1.
   a. To correct with confounding variables associated with gender and age.
   b. No! People may be more likely to start exercising if they find out they have or are at risk for heart disease, and thus we can't conclude from the data that beginning to exercise causes heart disease.

2.
   a. The treatment is the boot camp. The treatment group is the group of prisoners who attend the boot camp. The control is those who do not.
   b. It is an observational study, since the prisoners could choose to attend the boot camp or not. The prisoners were not randomly assigned into the treatment group or control group.
   c. Yes, the prisoners who attend the boot camp are less likely to return to prison.
   d. No, we cannot establish causality since the prisoners who chose to attend the boot camp may have already been less likely to return to prison.

3.
   a. An observational study.
   b. Accounting for these things helps eliminate some common confounding factors.
   c. Women taking the pill are more likely to be sexually active, which can lead to sharing HPV, which increases rates of cervical cancer.
   d. No, because the difference shown was not between random groups - they could differ in more than just their contraceptive use (e.g. sexual activity/HPV transmission).
   e. No, because the difference shown was not between random groups. It's important to note, though, that an inability to establish causality does not imply that there is *no* causality, just that we can't prove it. Some medical professionals believe that oral contraceptives may increase the risk of breast cancer (see the linked article).