Causal Frameworks II: Potential Outcomes

Lecture 2 - Introduction to Causal Inference

Kevin Li

Another Causal Framework

In the last class, we discussed how exogeneity allows us to identify causal effects.

- Thus, our goal in causal inference is to achieve exogeneity by controlling/accounting for all confounders.

In today's class, we will introduce another causal framework - potential outcomes. We will also link the two together.

Potential Outcomes

Imagine two hypothetical worlds that are identical, except one thing - whether individual i gets the treatment or not.

$$\begin{cases} \text{World 0} & i \text{ gets the treatment (treated) } D_i = 0 \\ \text{World 1} & i \text{ does not get the treatment (untreated) } D_i = 1 \end{cases}$$

Now, let us find the outcome value Y in both worlds:

$$\begin{cases} \text{World 0} & D_i = 0, \text{ Outcome value: } Y_i(0) \\ \text{World 1} & D_i = 1, \text{ Outcome value: } Y_i(1) \end{cases}$$

These are potential outcomes Y in both hypothetical worlds.

Individual Causal Effect

We know that:

- $Y_i(0)$ is the outcome Y in the hypothetical world i does not get treatment.
- $Y_i(1)$ is the outcome Y in the hypothetical world i gets treatment.

These two hypothetical worlds are identical, except the treatment. Thus, any difference between the outcomes in these two worlds must be the **causal effect** for individual i.

$$\tau_i = Y_i(1) - Y_i(0)$$

We denote causal effects with the greek letter τ (tau).

Counterfactuals

In real life, we do not have hypothetical parallel world (shocking).

An individual i either gets, or does not get treatment. That means that we don't actually observe in real life one of the hypothetical worlds.

- If i gets treated $D_i = 1$ in real life, we will never observe their hypothetical outcome $Y_i(0)$ when they are untreated.
- If i does not get treatment $D_i = 0$ in real life, we will never observe their hypothetical outcome $Y_i(1)$ when they are treated.

The potential outcome $Y_i(1)$ or $Y_i(0)$ that we do not observe is called a **counterfactual**.

Fundamental Problem of Causal Inference

Issue: The causal effect requires we know/measure both potential outcomes:

$$\tau_i = Y_i(1) - Y_i(0)$$

The **fundamental problem of causal inference** is that we never observe both of these potential outcomes, meaning we cannot calculate the true treatment effect τ_i.

Solution: Our goal in causal inference is thus to estimate/approximate the missing counterfactual, so we can estimate the causal effect.

Group Estimands

Estimating individual counterfactuals $Y_i(0)$ or $Y_i(1)$ can be both difficult, and inaccurate.

Instead, we often estimate group counterfactuals:

1. The average $Y_i(0)$ counterfactuals of all treated units.
2. The average $Y_i(1)$ counterfactuals of all untreated units.

This means we are not actually estimating τ_i (the causal effect for individual i), but the average τ of groups.

Also - practicality aside, it just makes sense to look at group-level causal effects since they are more generalisable. Its not as useful to know the effect on Ben or Ava alone, when we can estimate the average effect on everyone.

Causal Estimands

There are three causal estimands (effects we want to estimate) we are typically interested in:

1. Average Treatment Effect τ_{ATE}. This is the average treatment effect $\mathbb{E}[\tau_i]$ for all individuals i.
2. Average Treatment Effect on the Treated τ_{ATT}. This is the average treatment effect $\mathbb{E}[\tau_i|D_i = 1]$ for only individuals i who actually receive the treatment in real life.
3. Local/Conditional Average Treatment Effect $\tau_{\text{LATE}}/\tau_{\text{CATE}}$. This is the average treatment effect for only individuals i who meet a set of criteria. For example, the treatment effect for only females in our study.

We will introduce strategies to estimate these values throughout the next few classes.

Relationship with Exogeneity

How does this definition of causal effects relate to exogeneity (which we covered last class)?

We know that $Y_i(1)$ and $Y_i(0)$ come from hypothetical worlds that are identical to each other **except** the treatment D_i.

That also implies that between these two hypothetical worlds, the confounder values are equal (there are no differences between confounders).

Thus, the potential outcomes definition of causality implies exogeneity as well, and the reverse is also true: randomisation allows us to find missing counterfactuals and causal effects.

Stable Unit Treatment Value Assumption

The potential outcomes framework depends on the Stable Unit Treatment Value Assumption (SUTVA).

- Individual A's potential outcomes are not affected by if individual B receives or does not receive treatment.

Common violations:

- **Spillover**: if we are studying the effect of a new curriculum on student performance, if Student A's friend gets the new curriculum, that could affect Student A's performance.
- **Saturation**: If we are studying the effects of a vaccine, if enough of individual A's surrounding community gets the vaccine, individual A's outcomes will change because of herd immunity.