Causal Inference: Randomized Experiments

Ken Stiller

University of Sarajevo
October 2024
Part 2

So Far

- You learned why causal inference with observational data is hard
- You learned the differences between correlation and causation (and the role of confounders)
- You learned what a counterfactual is and became more comfortable with Potential Outcomes notation
- You were introduced to the concept of ATE
- You learned how observed mean differences across treated and control groups can contain selection bias

What is the Effect of Health Insurance on Health?

We want to compare:

- The health of someone with insurance to
- The health of **the same person** without insurance.

We begin by comparing health outcomes between those with and without insurance.
Should we do this? Are there any potential issues?

What is the Effect of Health Insurance on Health?

	Some Health Insurance	No Health Insurance	Difference
	A. Health		
Health Index	4.01	3.70	0.31 (0.03)
	B. Characteristics		
Non-White	0.16	0.17	-0.01 (0.01)
Age	43.98	41.26	2.71 (0.29)
Education	14.31	11.56	2.74 (0.10)
Employed	0.92	0.85	0.07 (0.01)
Family Income	106,467	45,656	60,810 (1,355)

The observed difference in the outcome for the treatment and control group are:

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = \underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{SelectionBias}$$

What is the Effect of Health Insurance on Health?

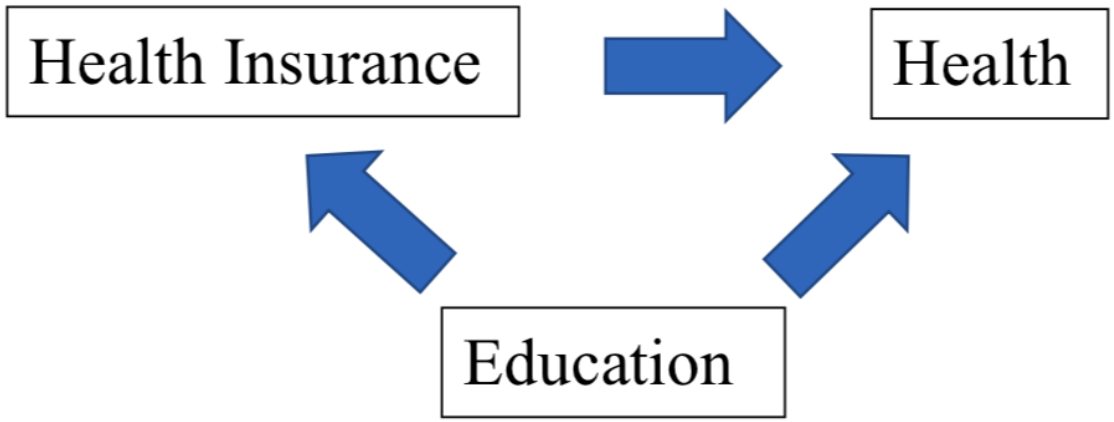
What can cause a difference in health outcomes for individuals with and without health insurance?

What is the Effect of Health Insurance on Health?

What can cause a difference in health outcomes for individuals with and without health insurance?

1. Causation: having health insurance directly leads to better health.
2. Reverse causality: the less (or more) healthy are more likely to buy insurance.
3. Confounders: e.g., the more educated tend to buy insurance more often and they know how to live healthier.
4. Any other ideas?

What is the Effect of Health Insurance on Health? Selection bias



Selection bias is when treatment is assigned in a manner that also affects the outcome. In our example confounders, e.g., education levels, may affect health. Education may also affect the choice to attain insurance. Thus, **potential outcomes differ for individuals with and without insurance.**

From observed mean differences to ATT

We would like to turn this:

$$\underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{SelectionBias}$$

Into this:

$$\underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT} + \underbrace{0}_{SelectionBias}$$

This means that:

$$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$$

This equality would make $ATT = ATE$. What assumptions do we need to claim this equality and how do we get there?

Randomization

Randomization and Selection Bias

With randomly assigned D_i there is no Selection Bias! Units i are similar on all (un)observed traits and only differ in terms of D_i.

Because both conditional expectations $E[Y_{1i}|D = 1]$ and $E[Y_{0i}|D = 0]$ come from the same underlying population, we can claim that when D_i is randomly assigned, the units are interchangeable.

Is randomization always possible? Why do we need causal inference?

Randomization

Why Randomization?

Random Assignment addresses the counterfactual problem by creating two subsamples that are identical prior to the intervention. When treatments are randomly assigned, $D_i = 1$ is a random sample of all units i and thus the potential outcomes for the treated units are identical to the potential outcomes of all units. The same for units in $D_i = 0$.

In expectation,
$E[Y_{1i}|D_i = 1] = E[Y_{1i}]$ but also $E[Y_{0i}|D_i = 0] = E[Y_{0i}]$

Now because the treatment does not affect the potential outcomes (recall the independence assumption), we are able to say that:

$$ATE = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

Randomization and Causal Estimates

If D_i was assigned at random; this would mean that:

1. $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$
2. $E[ATT] = ATE$
3. $DIGM = ATT$

With randomly assigned treatments there is No Selection Bias!

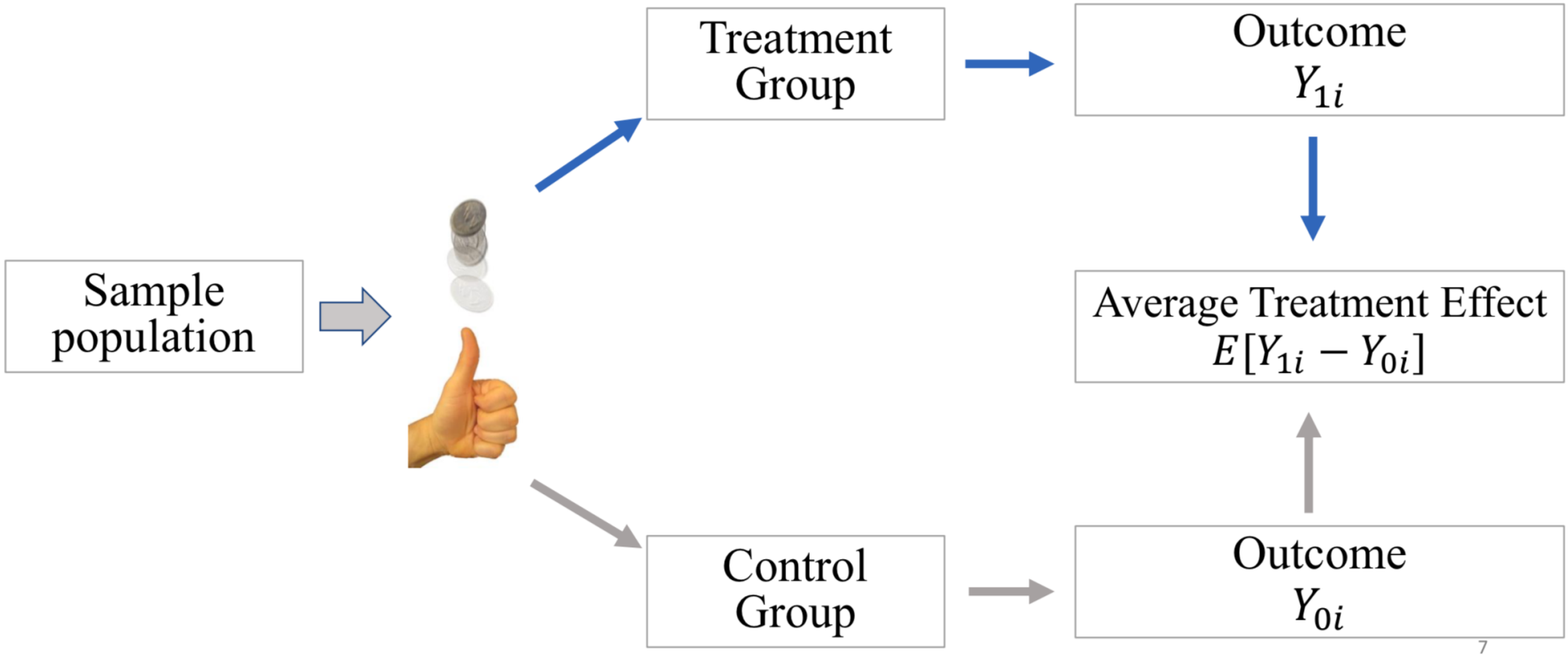
So revisiting the difference in group means (DIGM) :

$$\underbrace{E[Y_{i1} - Y_{i0}|D_i = 1]}_{ATT} + \underbrace{E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]}_{SelectionBias}$$

Is equal to:

$$\underbrace{E[Y_{i1} - Y_{i0}|D_i = 1]}_{ATT} + \underbrace{0}_{SelectionBias}$$

The Mechanics of Experiments



Randomizing the Treatment

The results we have seen showing no selection bias hinge on correctly executed **random assignment**:

- **Simple random assignment:** a procedure, such as a dice roll or coin toss, that gives each subject an identical probability of being assigned to treatment. Disadvantages when small N?
- **Complete random assignment**: m of N units are assigned to the treatment group with equal probability (one practical way to do it is permuting the order of all N subjects and label the first m subjects as the treatment group)
- **Stratified randomization or block randomization:** Suppose we observe some covariate x_j, and we know that the outcome y_i varies with x_j Then any difference in the covariate between the treatment groups will lead to a difference in the average outcome, unrelated to the actual treatment effect. This issue can be prevented by **balancing treatment assignment on these covariates**. It means partitioning the sample into groups (blocks) of different x_j, and then carrying out permutation randomization within each block.
- In practical terms, randomization is best done using statistical software like R (Randomizr)

Correct Randomization?

- Important: randomization ensures that all pre-treatment covariates (observable, unobservable) are balanced in expectation (across randomizations).
- Selection bias can't be tested (why not?), but **balance tests are often used to check for differences in observable pre-treatment covariates between the treatment and control groups**:
  - covariate-by-covariate comparison of means (e.g. via t-tests)
  - multivariate regression of treatment status (DV) on all covariates
- Balance tests useful to detect botched randomization, but can never say anything about balance in unobserved covariates, so "potentially misleading" (Imai, King, and Stuart 2008)
- What covariates would you check if you are running an experiment where you randomize free university tuition on graduating university?

Randomization and the Experimental Ideal

You might be wondering...

- Will I have to run an experiment to find causal estimates?
- Can I assign my units to treatment and control?
- Am I wasting my time learning about Causal Inference?

Causal Inference is not only about RCTs! Causal Inference is about the experimental ideal!

The aim is in most cases the same; we try to find interchangeable units (counterfactuals) that only differ in terms of their treatment status.

Always think about how to **approximate the experimental setting** (random assignment to treatment) by using observational data.

Experimental vs Observational Studies

Definition: Observational Study

An observational study is an empirical investigation of the effects of exposure to different treatment regimes, in which the investigator cannot control the assignment of treatment.

- This means that control and treatment units are not automatically exchangeable.
- Does this mean we can only work with experimental data?
- Of course not. This is why we add controls to our regression models. To adjust for the observed covariates.
- Is that good enough? What about the unobservables?
- Well, we want to make sure they are **as-if random**. That's where we will start next!
- Is balance -or exchangeability- testable? NO!