PSC4375: Randomized Experiments

Week 1: Lecture 2

Prof. Weldzius

Villanova University

Slides Updated: 2025-01-22

Changing minds on gay marriage

- Question: can we effectively persuade people to change their minds?
- Hugely important question for political campaigns, companies, etc.
- Psychological studies show it isn't easy.
- **Contact Hypothesis**: outgroup hostility diminished when people from different groups interact with one another.
- Today we'll explore this question the context of support for gay marriage and contact with a member of the LGBT community.
  - $Y_i =$ support for gay marriage (1) or not (0)
  - $T_i =$ contact with member of the LGBT community (1) or not (0)

Causal effects and counterfactuals

- What does "T_i causes Y_i" mean? $\rightsquigarrow$ **counterfactuals**, "what if"
- Would citizen i have supported gay marriage if they had contact with a member of the LGBT community?
- Two **potential outcomes**:
    - $Y_i(1)$: would i have supported gay marriage if they **had** contact with a member of the LGBT community?
    - $Y_i(0)$: would i have supported gay marriage if they **didn't have** contact with a member of the LGBT community?
- **Causal effect** for citizen i: $Y_i(1) - Y_i(0)$
- **Fundamental problem of causal inference**: only one of the two potential outcomes is observable.

Sigma notation

- We will often refer to the **sample size** (number of units) as n.
- We often have n measurements of some variable: $(Y_1, Y_2, ..., Y_n)$
- We often want sums: how many in our sample support gay marriage?

$$Y_1 + Y_2 + ... + Y_n$$

* Notation is a bit clunky, so we often use the **Sigma notation**:

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + ... + Y_n$$

* $\sum_{i=1}^{n}$ means sum each value from Y_1 to Y_n

Averages

- The **sample average** or **sample mean** is simply the sum of all values divided by the number of values.
- Sigma notation allows us to write this in a compact way:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

* Suppose we surveyed 6 people and 3 supported gay marriage:

$$\bar{Y} = \frac{1}{6}(1+1+1+0+0+0) = 0.5$$

Quantity of interest

- We want to estimate the average causal effects over all units:

$$\text{Sample Average Treatment Effect (SATE) } = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$$

* Why can't we just calculate this quantity directly? * What we can estimate instead:

$$\text{Difference in means } = \bar{Y}_{treated} - \bar{Y}_{control}$$

- $\bar{Y}_{treated}$: observed average outcome for treated group - $\bar{Y}_{control}$: observed average outcome for control group * When will the difference-in-means be a good estimate of the SATE?

Randomized control trials (RCTs)

- **Randomize control trial**: each unit's treatment assignment is determined by chance
  - e.g., flip a coin; draw read and blue chips from a hat; etc.
- Randomization ensures **balance** between treatment and control group.
  - Treatment and control group are identical **on average**
  - Similar on both observable and unobservable characteristics.
- Control group $\approx$ what would have happened to treatment group if they had taken control
  - $\bar{Y}_{control} \approx \frac{1}{n}\sum_{i=1}^{n} Y_i(0)$
  - $\bar{Y}_{treated} - \bar{Y}_{control} \approx$ SATE

Some potential problems with RCTs

- **Placebo effects**:
    - Respondents will be affected by any intervention, even if they shouldn't have any effect
- **Hawthorne effects**:
    - Respondents act differently just knowing that they are under study.

Balance checking

- Can we determine if randomization "worked"?
- If it did, we shouldn't see large differences between treatment and control group on **pretreatment variable**
  - Pretreatment variable are those that are unaffected by treatment
- We can check in the actual data for some pretreatment variable X
  - $\bar{X}_{treated}$: average value of variable for treated group
  - $\bar{X}_{control}$: average value of variable for control group
  - Under randomization, $\bar{X}_{treated} - \bar{X}_{control} \approx 0$

Multiple treatments

- Instead of 1 treatment, we might have multiple **treatment arms**:
  - Control condition
  - Treatment A
  - Treatment B
  - Treatment C, etc.
- In this case, we will look at multiple comparisons:
  - $\bar{Y}_{treated,A} - \bar{Y}_{control}$
  - $\bar{Y}_{treated,B} - \bar{Y}_{control}$
  - $\bar{Y}_{treated,A} - \bar{Y}_{treated,B}$
- If treatment arms are randomly assigned, these differences will be good estimators for each causal contrast.