Randomized Controlled Trial

Prof. Tzu-Ting Yang
楊子霆

Institute of Economics, Academia Sinica
中央研究院經濟研究所

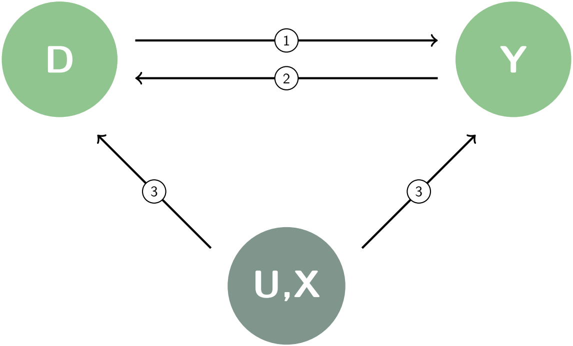
March 13, 2023

Randomized Controlled Trial: Main Idea

Randomized Controlled Trial

- The most credible identification strategy is **randomized controlled trial/experiment**
- Randomized controlled trial (RCT):
  - Each observation (e.g. individual, household, school, state, or country) is **randomly assigned** to treatment and control group

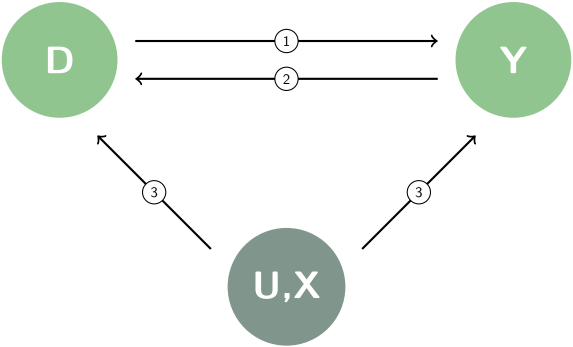
Identify Causal Effect



- There are three possible reasons why we observe treatment D is correlated with outcome Y

  1. D causes Y
  2. Y causes D
  3. Observed X or unobserved U confounding factors affect D and Y

Identify Causal Effect



- To identify causal effect of treatment, we need to make sure the observed relationship between treatment D and outcome Y is due to (1)
  - D causes Y

Randomized Controlled Trial

- RCT has two features that can help us hold “other things equal” and then eliminates selection bias
  1. Randomly assign treatment
  2. Sufficiently “large” sample size

Randomly Assign Treatment

- Randomly assign treatment (such as a coin flip) ensures that the probability of receiving treatment is unrelated to any other confounding factors (X)
- Every observation has the same probability of being assigned to the treatment group
- **Example:**
  - Suppose we can randomly assign master degree to people in Taiwan
  - This will make an individual's probability of receiving master degree (D) is unrelated to his/her family wealth or ability

Randomly Assign Treatment

- **Randomly assign treatment implies:**
  - The values of potential outcomes are **independent** of treatment assigned
  - $(\mathrm{Y}_i^{\mathbf{1}}, \mathrm{Y}_i^{\mathbf{0}})$ are independent of D_i

$$(\mathrm{Y}_i^{\mathbf{1}}, \mathrm{Y}_i^{\mathbf{0}}) \perp\!\!\!\perp D_i$$

  - Two variables are said to be **independent** means:
    - The occurrence of one variable (D) does not affect the probability of occurrence of the other $(\mathrm{Y}_i^{\mathbf{1}}, \mathrm{Y}_i^{\mathbf{0}})$

Randomly Assign Treatment

- **Intuition:**
  - Treatment assignment D is random and not based on an individual's value of potential outcome (Y_i^1, Y_i^0)

Selection Bias

$$\alpha_{\text{corr}} = \underbrace{\mathrm{E}[\mathrm{Y}_i | D_i = 1] - \mathrm{E}[\mathrm{Y}_i | D_i = 0]}_{\text{Observed Difference in Average Outcome}}$$

$$= \underbrace{\mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 | D_i = 1]}_{\text{ATT}} + \underbrace{\mathrm{E}[\mathrm{Y}_i^0 | D_i = 1] - \mathrm{E}[\mathrm{Y}_i^0 | D_i = 0]}_{\text{Selection Bias}}$$

- **Selection Bias** implies:
  - The value of potential outcomes Y_i^0 or Y_i^1 is correlated with treatment status D
    - People might select treatment according to their value of potential outcome
  - This means treatment and control groups could be different in other dimensions: **other things are NOT equal**

Randomly Assign Treatment

- Randomly assign treatment makes treatment and control group to be similar along all characteristics (including $\mathrm{Y}_i^0, \mathrm{Y}_i^1$)
  - $\mathrm{E}[\mathrm{Y}_i^{\mathbf{0}}|D_i = 1] = \mathrm{E}[\mathrm{Y}_i^{\mathbf{0}}|D_i = 0]$
  - $\mathrm{E}[\mathrm{Y}_i^{\mathbf{1}}|D_i = 1] = \mathrm{E}[\mathrm{Y}_i^{\mathbf{1}}|D_i = 0]$

- So that we can eliminate selection bias
  - $\underbrace{\mathrm{E}[\mathrm{Y}_i^0|D_i = 1] - \mathrm{E}[\mathrm{Y}_i^0|D_i = 0]}_{\text{Selection Bias}} = 0$

Sufficiently Large Sample Size

- Randomly assign treatment can ensure the **average** characteristics of two groups are similar
  - How about each group only has one individual?
- We also need **large sample size** to ensure that the group differences in individual characteristics wash out

Randomized Controlled Trial: Potential Outcome Framework

Randomized Controlled Trial

Identify ATT and ATE

- RCT Identifies ATT and ATE

$$
\begin{aligned}
&\underbrace{\mathrm{E}[\mathrm{Y}_i|D_i=1]-\mathrm{E}[\mathrm{Y}_i|D_i=0]}_{\text{Observed Difference in Average Outcome}} \\
&=\underbrace{\mathrm{E}[\mathrm{Y}_i^1-\mathrm{Y}_i^0|D_i=1]}_{\text{Causal Effect (ATT)}}+\underbrace{\mathrm{E}[\mathrm{Y}_i^0|D_i=1]-\mathrm{E}[\mathrm{Y}_i^0|D_i=0]}_{\text{Selection Bias}} \\
&=\underbrace{\mathrm{E}[\mathrm{Y}_i^1-\mathrm{Y}_i^0|D_i=1]}_{\text{Causal Effect (ATT)}}+\underbrace{0}_{\text{Selection Bias}} \\
&=\underbrace{\mathrm{E}[\mathrm{Y}_i^1-\mathrm{Y}_i^0|D_i=0]}_{\text{Causal Effect (ATU)}} \\
&=\underbrace{\mathrm{E}[\mathrm{Y}_i^1-\mathrm{Y}_i^0]}_{\text{Causal Effect (ATE)}}
\end{aligned}
$$

Randomized Controlled Trial: Estimation

Estimation and Inference in RCT

- Up until now, we've talked about identification.
- Now that we know that the ATE and ATT are identified, how will we estimate them?
- Remember: identification first, then estimation.

Estimation and Inference in RCT

Estimation

- Suppose we get a nationally representative sample: N individuals
- Randomly assign treatment (master degree)
  - N_1 individuals obtain master degree: treatment group
  - N_0 individuals do not have it ($N_0 = N - N_1$): control group
- Compare difference in monthly salary between treatment group and control group

Estimation and Inference in RCT

Estimation

- Using the analogy principle, we construct the following sample estimator:

$$\hat{\alpha}_{\mathsf{ATE}} = \bar{\mathrm{Y}}_1 - \bar{\mathrm{Y}}_0$$

- where $N_1 = \sum_i D_i$ and $N_0 = N - N_1$

$$\bar{\mathrm{Y}}_1 = \frac{1}{N_1} \sum_{D_i=1} \mathrm{Y}_i$$
$$\bar{\mathrm{Y}}_0 = \frac{1}{N_0} \sum_{D_i=0} \mathrm{Y}_i$$

Estimation and Inference in RCT

Inference

- Now we want to use **sample estimator** $\hat{\alpha}_{\text{ATE}}$ to infer whether outcomes (e.g. monthly salary) are different in treatment and control group at **population level** α_{ATE}
- Statistical inference (Hypothesis Testing) helps us answer this question

Review: Hypothesis Testing

Sample Estimator

- Suppose that the sample estimator $\hat{\alpha}_{\text{ATE}} = \bar{\mathrm{Y}}_1 - \bar{\mathrm{Y}}_0$ is $NT\$5000$
- Does it mean those who get master degree have more monthly salary than those do not hold mater degree ?
  - Note that sample estimator can be somewhat different when drawing another sample from the same population
- It's possible that this is a **chance finding**
- We need to find a sufficiently strong evidence to show this is not a chance finding

Summary of Hypothesis Testing

1. Choose a null hypothesis:
   - Since we do NOT know exact population value of α_{ATE}
     - What we can do is assume the true α_{ATE} is exactly μ
     - This is what we call the null hypothesis (H_0)
   - We usually test whether there is **no average effect** of treatment:
     - $H_0 : \alpha_{\text{ATE}} = 0$

Summary of Hypothesis Testing

2. Choose a test statistic to examine a null hypothesis:
   - We use a t-statistic to measure whether our sample estimates support/against this null hypothesis

$$t = \frac{\hat{\alpha}_{\text{ATE}} - 0}{\hat{\text{SE}}(\hat{\alpha}_{ATE})}$$

   - Estimate standard error of the sample estimator $\hat{\alpha}_{\text{ATE}}$
     - $\hat{\text{SE}}(\hat{\alpha}_{ATE}) = \hat{\sigma}_Y \sqrt{[\frac{1}{N_1} + \frac{1}{N_0}]}$
     - $\hat{\sigma}_Y$ is the estimated standard deviation of Y
   - Thus, t-statistic is the ratio of $\hat{\alpha}_{\text{ATE}}$ to its estimated standard error

Summary of Hypothesis Testing

3. Evaluate whether the sample estimator is against null hypothesis or not

   - **Goal:** Calculate **p-value**

     - **p-value:** Given null hypothesis is true, the probability of obtaining the sample estimates or more extreme ones

     - If this probability is high, it means the sample estimate might support for null hypothesis

     - If this probability is low, it means the sample estimate might be against null hypothesis

Summary of Hypothesis Testing

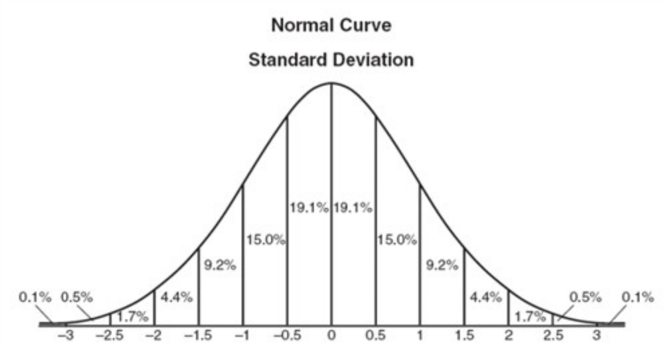
3. Evaluate whether the sample estimator is against null hypothesis or not

   - In order to calculate this probability (p-value), we need to know the distribution of the t-statistic under the null hypothesis

     - If sample size is sufficiently large, using **Central Limit Theorem (CLT)**, t-statistic will have standard normal distribution

Summary of Hypothesis Testing

Central Limit Theorem (CLT)



Summary of Hypothesis Testing

3. Evaluate whether the sample estimator is against null hypothesis or not

    - Based on standard normal distribution and sample estimator, we can get p-value
    - We reject the null hypothesis $H_0 : \alpha_{ATE} = 0$ when p-value is sufficiently low
        - We usually select an arbitrarily pre-defined threshold value θ, which is referred to as the **level of significance**
        - By convention, θ is commonly set to 0.1 or 0.05
    - If p-value is smaller than θ, we would say the sample estimate is **significantly different from the null hypothesis**

Review: Hypothesis Testing

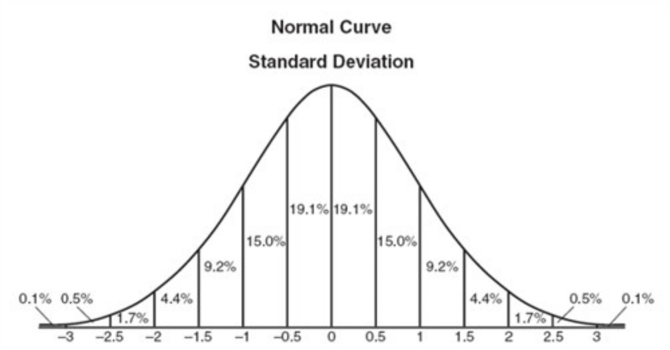
Examples

- Suppose $\hat{\mathrm{SE}}(\hat{\alpha}_{ATE})$ is 1000
- Thus, $t = \dfrac{5000}{1000} = 5$, which is even far from 2
- In this case, the p-value is less than 0.0001
- It suggests when the null hypothesis is $\mu = 0$, this result is very unlikely to happen
  - This is a strong evidence showing that the sample estimator against null hypothesis
  - The sample estimator $\hat{\alpha}_{ATE}$ is **significantly different from zero**
- Getting master degree might increase monthly salary

Review: Hypothesis Testing

Examples

- If t-statistic is 5, the p-value is less than 0.0001



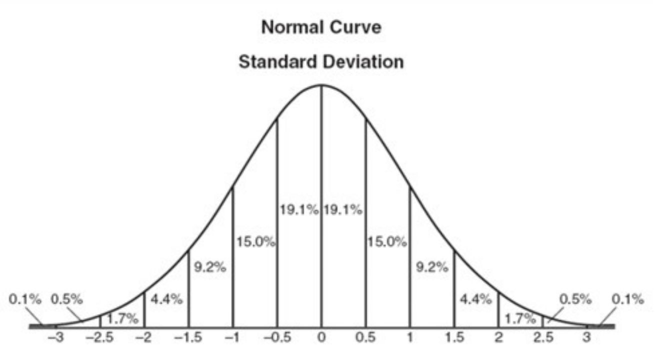
Review: Hypothesis Testing

Examples

- Instead, suppose $\hat{\mathrm{SE}}(\hat{\alpha}_{ATE})$ is 10000
- Thus, $t = \dfrac{5000}{10000} = 0.5$, which is small
- p-value is very high
- It suggests when the null hypothesis is $\mu = 0$, this result is very likely to happen
    - This is a weak evidence showing that the sample estimator against null hypothesis
    - The sample estimator $\hat{\alpha}_{ATE}$ is **insignificantly different from zero**
- Getting master degree might NOT increase monthly salary

Review: Hypothesis Testing
Examples

- If t-statistic is 0.5, the p-value is more than 0.31



STATA Command: ttest

- **ttest**: t tests (mean-comparison tests)

- One-sample t test:

```
ttest varname == # [if] [in] [, level(#)]
```

- Two-sample t test using groups:

```
ttest varname [if] [in], by(groupvar) [options]
```

STATA Command: ttest

- Please see **ttest.do**
- One-sample t test:

```
ttest age = 30 if sex==1
```

- Two-sample t test using groups:

```
ttest incwage, by(sex)
```

STATA Command: ttest

Output

```
. ttest incwage, by(sex)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
1	29,148	2421357	24899.48	4251034	2372552	2470161
2	30,852	2197090	23456.97	4120154	2151114	2243067
combined	60,000	2306039	17088.1	4185713	2272546	2339532
diff		224266.3	34178.02		157277.2	291255.3

```
    diff = mean(1) - mean(2)                                      t =   6.5617
Ho: diff = 0                                     degrees of freedom =    59998

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```