Causal Inference

Lab 3
Difference-in-Means Estimator and Neyman Variance Estimator

Xiangyu Song

Washington University in St. Louis

January 30, 2026

DiM, SATE and PATE

- N samples randomly chosen from a super-population
- **Sample** Average Treatment Effect (SATE):

$$\text{SATE} = \frac{1}{N} \sum_{i=1}^{N} [Y_i(1) - Y_i(0)]$$

- **Population** Average Treatment Effect (PATE):

$$\text{PATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$$

- Difference-in-Means (DiM) estimator:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - D_i) Y_i$$

DiM, SATE and PATE, cont.

- $\hat{\tau}$ is an unbiased estimator of SATE in randomized experiments

$$\mathbb{E}\left[\hat{\tau} \mid \mathcal{O}\right] = \frac{1}{N_1} \sum_{i=1}^{N} \mathbb{E}\left[D_i \mid \mathcal{O}\right] Y_i(1) - \frac{1}{N_0} \sum_{i=1}^{N} \mathbb{E}\left[(1 - D_i) \mid \mathcal{O}\right] Y_i(0)$$
$$= \frac{1}{N} \sum_{i=1}^{N} \left(Y_i(1) - Y_i(0)\right)$$
$$= \text{SATE}$$

- $\hat{\tau}$ is an unbiased estimator of PATE if treatment is **randomly** assigned in the population

$$\mathbb{E}\left[\hat{\tau}\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{\tau} \mid \mathcal{O}\right]\right]$$
$$= \mathbb{E}\left[\text{SATE}\right]$$
$$= \text{PATE}$$

Sampling Variance of DiM Estimator

- Difference-in-Means (DiM) estimator:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - D_i) Y_i$$

- Sampling variance of DiM estimator at finite sample level

$$\text{Var}\left(\hat{\tau} \mid \mathcal{O}\right) = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{\tau_i}^2}{N}$$

- ▶ S_0^2 and S_1^2 are in-sample variances of $Y_i(0)$ and $Y_i(1)$
- ▶ $S_{\tau_i}^2$ is the in-sample variance of individual treatment effects
- ▶ None of above are observable
- ▶ Can also be written as

$$\text{Var}\left(\hat{\tau} \mid \mathcal{O}\right) = \frac{1}{N} \left(\frac{N_1}{N_0} S_0^2 + \frac{N_0}{N_1} S_1^2 + 2S_{01} \right)$$

Sampling Variance of DiM Estimator

Prove:

$$\text{Var}\left(\hat{\tau} \mid \mathcal{O}\right) = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{\tau_i}^2}{N}$$

Also check Imbens and Rubin (2015), Section 6 Appendix A

Sampling Variance of DiM estimator, cont.

- Difference-in-Means (DiM) estimator:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - D_i) Y_i$$

- Sampling variance of DiM estimator at population level

$$\text{Var}\,(\hat{\tau}) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1}$$

  ▶ σ_0^2 and σ_1^2 are population-level variances of $Y_i(0)$ and $Y_i(1)$
  ▶ None of above are observable
  ▶ Can be estimated using sample variances of observed outcomes in treatment and control groups

Neyman Estimator

- Neyman estimator for sampling variance

$$\widehat{\text{Var}}\left(\hat{\tau}\right) = \frac{\hat{\sigma}_0^2}{N_0} + \frac{\hat{\sigma}_1^2}{N_1}$$

  - $\hat{\sigma}_d^2$ are sample variances of observed outcomes within each group $d \in \{0, 1\}$.

$$\hat{\sigma}_d^2 = \frac{1}{N_d - 1} \sum_{i=1}^{N} \mathbb{1}_{\{D_i = d\}} \left(Y_i - \bar{Y}_1^{\text{obs}}\right)$$

  - $\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\tau}\right) \mid \mathcal{O}\right] = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$ and
    $\mathbb{E}\left[\text{Var}\left(\hat{\tau}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\tau}\right) \mid \mathcal{O}\right]\right] = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$

- Estimating sampling variance
  - At finite sample level, Neyman estimator is **conservative**

$$\text{Var}\left(\hat{\tau} \mid \mathcal{O}\right) \leq \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} = \mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\tau}\right) \mid \mathcal{O}\right]$$

  - At population level, Neyman estimator is **unbiased**

$$\text{Var}\left(\hat{\tau}\right) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} = \mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\tau}\right)\right] \quad \text{(proved in class)}$$

Neyman Estimator

Prove:

$$\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\tau}\right) \mid \mathcal{O}\right] = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$$

$$\mathbb{E}\left[\text{Var}\left(\hat{\tau}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\tau}\right) \mid \mathcal{O}\right]\right] = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$$

Recap

- Difference-in-Means (DiM) estimator
  - ▶ Unbiased for both SATE and PATE
- Sampling variance of DiM estimator unobserverable for both
  - ▶ Finite-sample: $\text{Var}(\hat{\tau} \mid \mathcal{O})$
  - ▶ Population-level: $\text{Var}(\hat{\tau})$
- Neyman estimator for sampling variance
  - ▶ Conservative for finite-sample level sampling variance
  - ▶ Unbiased for population-level sampling variance