6. Neyman's Repeated Sampling Approach to Completely Randomized Experiments

Chanmoo Park

July 15, 2022

Seoul National University

1. Introduction: Neyman's Approach (vs Fisher's approach)

- **Estimand** : Average Treatment Effect (ATE)

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right) = \bar{Y}(1) - \bar{Y}(0)$$

- **Null Hypothesis**

$$H_0 : \tau_{\text{fs}} = 0$$

(vs Fisher, $H_0 : Y_i(1) - Y_i(0) = 0$ for $i = 1, \ldots, N$)

- **Inference** : Large sample approximation

(vs Fisher's exact inference)

- **Assumptions**
  - Potential outcomes $Y_i(1)$, $Y_i(0)$ are fixed.
  - Assignment mechanism : Completely Randomized Experiment
  - Stability assumpation : SUTVA

1. Introduction: Neyman's Repeated Sampling Approach

- **Finite Sample Inference**
  - Size N sample is fixed ($N = N_t + N_c$)
  - The only randomness : Assignment Vector (**W**)
  - **Estimand : τ_{fs} (finite sample ATE, SATE)**

- **Super Population Inference**
  - Size N sample drawn from size N_{sp} Super Population
  - Randomness 1 : Sampling Vector (**R**)

    ; 1st Sampling (distribution generated by Simple Random Sampling)
  - Randomness 2 : Assignment Vector (**W**)

    ; 2nd Sampling (Randomization distribution)
  - **Estimand : τ_{sp} (super-population ATE, PATE)**

* Notation : Sampling Vector R
R $\in \{0, 1\}^{N_{\text{sp}}}$, where N_{sp} is usually assumed infinite but countable
$R_i = 1$ (sampled), $R_i = 0$ (not sampled)
$\sum_{i=1}^{N_{\text{sp}}} R_i = N$

2. Finite Sample Inference

- **Estimand** : $\tau_{\text{fs}} = \bar{Y}(1) - \bar{Y}(0)$
- **Estimator**

$$\hat{\tau_{\text{fs}}}^{\text{dif}} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{\text{obs}} = \bar{Y}_{\text{t}}^{\text{obs}} - \bar{Y}_{\text{c}}^{\text{obs}}$$

  - **Unbiased** (Theorem 6.1)

$$\mathbb{E}_W[\hat{\tau_{\text{fs}}}^{\text{dif}}] = \tau_{\text{fs}}$$

  - **Variance** (Theorem 6.2)

$$\mathbb{V}_W[\hat{\tau_{\text{fs}}}^{\text{dif}}] = \frac{S_c^2}{N_{\text{c}}} + \frac{S_t^2}{N_{\text{t}}} - \frac{S_{tc}^2}{N}$$

$S_c^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i(0) - \bar{Y}(0)\right)^2$; Sample variance of $Y_i(0)$'s ,

$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i(1) - \bar{Y}(1)\right)^2$; Sample variance of $Y_i(1)$'s

$S_{tc}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0))\right)^2$; Sample variance of $Y_i(1) - Y_i(0)$'s

2. Finite Sample Inference

- **Variance of $\hat{\tau}_{fs}^{dif}$: $\mathbb{V}_W[\hat{\tau}_{fs}^{dif}]$**

$$\mathbb{V}_W[\hat{\tau}_{fs}^{dif}] = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N}$$

- Neyman needed $\mathbb{V}_W[\hat{\tau}_{fs}^{dif}]$ in **test statistic** and **confidence interval**
- However, $\mathbb{V}_W[\hat{\tau}_{fs}^{dif}]$ is a function of ALL potential outcomes

$\Rightarrow$ **Never observable, Need Estimation :** $\hat{\mathbb{V}}$

- **Neyman estimator :** $\hat{\mathbb{V}}^{\text{neyman}}$

$$\hat{\mathbb{V}}^{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$$

$$s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} \left(Y_i(0) - \bar{Y}_c^{obs}\right)^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} \left(Y_i^{obs} - \bar{Y}_c^{obs}\right)^2$$

$$s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} \left(Y_i(1) - \bar{Y}_t^{obs}\right)^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} \left(Y_i^{obs} - \bar{Y}_t^{obs}\right)^2$$

- We can also check other alternative estimators in textbook.

2. Finite Sample Inference

- **Neyman estimator($\hat{\mathbb{V}}^{\text{neyman}}$) is widely used because...**

  1. **Upwardly biased** irrespective of the heterogeneity in treatment effect (under a certain condition, unbiased)
  2. **Unbiased** in infinite Super Population (later)

❶ Upward biasedness of $\hat{\mathbb{V}}^{\text{neyman}}$

- Alternative representation of $\mathbb{V}_W[\hat{\tau_{\text{fs}}}^{\text{dif}}]$

$$\mathbb{V}_W[\hat{\tau_{\text{fs}}}^{\text{dif}}] = \frac{N_{\text{t}}}{N \cdot N_{\text{c}}} \cdot S_c^2 + \frac{N_{\text{c}}}{N \cdot N_{\text{t}}} \cdot S_t^2 + \frac{2}{N} \cdot \rho_{tc} \cdot S_c \cdot S_t$$

$$\text{where, } \rho_{tc} = \frac{1}{(N-1) \cdot S_c \cdot S_t} \sum_{i=1}^{N} (Y_i(1) - \bar{Y}(1)) \cdot (Y_i(0) - \bar{Y}(0))$$

- Case 1 : Largest $\mathbb{V}_W[\hat{\tau_{\text{fs}}}^{\text{dif}}]$: $\rho_{tc} = 1$ (Perfectly positively correlated)

$$\mathbb{V}_W\left[\hat{\tau_{\text{fs}}}^{\text{dif}} \mid \rho_{tc} = 1\right] = \frac{S_c^2}{N_{\text{c}}} + \frac{S_t^2}{N_{\text{t}}} - \frac{(S_c - S_t)^2}{N}$$

- Case 2 : The most notable case (Treatment effect is additive and constant)

$$\mathbb{V}_W\left[\hat{\tau_{\text{fs}}}^{\text{dif}} \mid \rho_{tc} = 1, S_c^2 = S_t^2\right] = \frac{S_c^2}{N_{\text{c}}} + \frac{S_t^2}{N_{\text{t}}}$$

- $\hat{\mathbb{V}}^{\text{neyman}}$ **is unbiased estimator of Case 2** (Theorem 6.3)

2. Finite Sample Inference

- **Confidence Interval & Testing**
  - **Large sample approximation**

$$\text{by Central Limit Theorem, } (\hat{\tau_{\text{fs}}}^{\text{dif}} - \tau_{\text{fs}})/\sqrt{\mathbb{V}} \xrightarrow{d} N(0,1)$$

  - **Confidence Interval**

$$\text{CI}^{1-\alpha}(\tau_{\text{fs}}) = \left(\hat{\tau_{\text{fs}}}^{\text{dif}} - z_{\alpha/2}\cdot\sqrt{\hat{\mathbb{V}}}, \hat{\tau_{\text{fs}}}^{\text{dif}} + z_{\alpha/2}\cdot\sqrt{\hat{\mathbb{V}}}\right)$$

  - **Testing (**$H_0 : \tau_{\text{fs}} = 0$**, two-sided)**

$$\text{Reject } H_0 \text{ if, } \left|\frac{\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}}{\sqrt{\hat{\mathbb{V}}}}\right| > z_{\alpha/2}$$

3. Super Population Inference

- **[Recall] Super Population Inference** (Repeated Sampling)
  - Size N sample drawn from size N_{sp} Super Population
  - Randomness 1 : Sampling Vector ($\mathbf{R}$)
  - Randomness 2 : Assignment Vector ($\mathbf{W}$)
  - **Estimand : τ_{sp} (super-population ATE, PATE)**

* Notation : Sampling Vector R
$\text{R} \in \{0,1\}^{N_{\text{sp}}}$, where N_{sp} is usually assumed infinite but countable
$R_i = 1$ (sampled), $R_i = 0$ (not sampled)
$\sum_{i=1}^{N_{\text{sp}}} R_i = N$

- **Rewrite $\hat{\tau_{\text{fs}}}^{\text{dif}}$ by the Super Population representation**

$$\hat{\tau_{\text{fs}}}^{\text{dif}} = \bar{Y}_{\text{t}}^{\text{obs}} - \bar{Y}_{\text{c}}^{\text{obs}} = \frac{1}{N_{\text{t}}}\sum_{i=1}^{N} W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}}\sum_{i=1}^{N}(1-W_i)\cdot Y_i^{\text{obs}} \qquad \text{(FS)}$$

$$= \frac{1}{N_{\text{t}}}\sum_{i=1}^{N_{\text{sp}}} R_i \cdot W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}}\sum_{i=1}^{N_{\text{sp}}} R_i \cdot (1-W_i)\cdot Y_i^{\text{obs}} \qquad \text{(SP)}$$

3. Super Population Inference

- **Estimator for τ_{sp}**
  - $\hat{\tau_{\text{fs}}}^{\text{dif}}$ **is also unbiased estimator of τ_{sp}**

$$\mathbb{E}\left[\hat{\tau_{\text{fs}}}^{\text{dif}} \mid Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{E}_W\left[\hat{\tau_{\text{fs}}}^{\text{dif}} \mid R, Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right] \mid Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right]$$
$$= \mathbb{E}_{\text{sp}}\left[\tau_{\text{fs}} \mid Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right] = \tau_{\text{sp}} \quad \text{(Appendix B, p.110)}$$

- **Estimator for $\mathbb{V}[\hat{\tau_{\text{fs}}}^{\text{dif}}]$**
  - $\hat{\mathbb{V}}^{\text{neyman}}$ **is also unbiased estimator of $\mathbb{V}[\hat{\tau_{\text{fs}}}^{\text{dif}}]$**

$$\mathbb{V}\left[\hat{\tau_{\text{fs}}}^{\text{dif}} \mid Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{V}_W\left(\hat{\tau}^{\text{dif}} \mid R, Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right) \mid Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right]$$
$$+ \mathbb{V}_{\text{sp}}\left(\mathbb{E}_W\left[\hat{\tau}^{\text{dif}} \mid R, Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right] \mid Y_{\text{sp}}(1), Y_{\text{sp}}(0)\right)$$
$$= \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t} \quad \left(= \mathbb{E}\left[\hat{\mathbb{V}}^{\text{neyman}}\right]\right) \quad \text{(Appendix B, p.112)}$$

$\Rightarrow$ **Neyman showed that, $\hat{\tau_{\text{fs}}}^{\text{dif}}$ and $\hat{\mathbb{V}}^{\text{neyman}}$ is still valid under repeated sampling approach**

4. Conclusion

- Neyman's Causal Estimand was ATE (Average Treatment Effect).
- He extended arguments to the Super Population (Repeated Sampling).
- He suggested $\hat{\tau_{\text{fs}}}^{\text{dif}}$ as UE of τ_{fs} and τ_{sp}
- He suggested $\hat{\mathbb{V}}^{\text{neyman}}$ as UE of $\mathbb{V}[\hat{\tau_{\text{fs}}}^{\text{dif}}]$
- He suggested Testing & CI procedures using large sample approx.