Introduction to Causal Inference

Lecture 6

Stratified Randomized Experiments & Pairwise Randomized Experiments

邓婉璐

WANLUDENG@TSINGHUA.EDU.CN



清华大学统计与数据科学系

Outline

- Stratified Randomized Experiments
- Pairwise Randomized Experiments
- Summary



Stratified Randomized Experiments



The Project Star

- 大班教学好还是小班教学好?

- A randomized evaluation of the effect of class size on test scores conducted in 1985 – 1986 in Tennesee called the Student/Teacher Achievement Ratio experiment.
- Very influential experiment: “one of the most important educational investigations ever carried out.”
- We use the kindergarten data.
- The unit of analysis is the teacher or class, rather than the individual student, to help justify the no-interference part of SUTVA.
- The specific outcome we analyze here is the class average score on a mathematics test.



Stratified Randomized Experiment

- An important case of a classical randomized experiment
- Group the units into J strata based on the value of covariates
- Implement a CRE within each data strata: $N(j), N_c(j), N_t(j)$ for stratum j
- Proportion of stratum j: $q(j) = N(j)/N$ ⟶ $N = N_1 + \cdots + N_j$
- Propensity score of stratum j: $\mathrm{e}(j) = N_t(j)/N(j)$
- Let $G_i \in \{1, \ldots, J\}$ denote the strata of unit i, and let $B_i(j) = I(G_i = j)$
- The assignment mechanism:

$$\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} \text{ for } \mathbf{W} \in \mathbb{W}^+$$

$$\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} | \sum_{i=1}^{N} B_i(j) \cdot W_i = N_t(j) \text{ for } j = 1, \ldots, J\}$$

- Also known as Randomized Blocks Experiments (RBEs)

The Project Star Data

- **Unit**: $N = 68$ teachers/classes in 16 schools
- **Treatment**: small class (13-17 students)
- **Control**: regular class (22-25 students)
- **Outcome**: class average score on a math test
- **Covariate for stratification**: school ID
- **Stratified randomized assignment**:

  13 schools/strata with $N_t(j) = N_c(j) = 2$

  2 schools/strata with $N_t(j) = 3, N_c(j) = 2$

  1 schools/strata with $N_t(j) = 4, N_c(j) = 2$

- **Number of possible assignments**:

$$6^{13} \times 10^{2} \times 15^{1} \approx 2 \times 10^{13}$$

$$\binom{4}{2} \quad \binom{5}{2} \quad \binom{6}{2}$$

Table 9.1. *Class Average Mathematics Scores from Project Star*

School/ Stratum	No. of Classes	Regular Classes ($W_i = 0$)	Small Classes ($W_i = 1$)
1	4	−0.197, 0.236	0.165, 0.321
2	4	0.117, 1.190	0.918, −0.202
3	5	−0.496, 0.225	0.341, 0.561, −0.059
4	4	−1.104, −0.956	−0.024, −0.450
5	4	−0.126, 0.106	−0.258, −0.083
6	4	−0.597, −0.495	1.151, 0.707
7	4	0.685, 0.270	0.077, 0.371
8	6	−0.934, −0.633	−0.870, −0.496, −0.444, 0.392
9	4	−0.891, −0.856	−0.568, −1.189
10	4	−0.473, −0.807	−0.727, −0.580
11	4	−0.383, 0.313	−0.533, 0.458
12	5	0.474, 0.140	1.001, 0.102, 0.484
13	4	0.205, 0.296	0.855, 0.509
14	4	0.742, 0.175	0.618, 0.978
15	4	−0.434, −0.293	−0.545, 0.234
16	4	0.355, −0.130	−0.240, −0.150
Average		−0.13	0.09
(S.D.)		(0.56)	(0.61)



The Case with Two Strata

- Divide the N units into 2 subsamples: females (f) and males (m)
- Implement a CRE within each data strata: $N(j), N_c(j), N_t(j)$ for the jth stratum
- Proportion of stratum j: $q(j) = N(j)/N$
- Propensity score of stratum j: $e(j) = N_t(j)/N(j)$
- Let $G_i \in \{1, \ldots, J\}$ denote the strata of unit i, and let $B_i(j) = I(G_i = j)$
- The assignment mechanism:

$$\Pr(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{S}) = \binom{N(f)}{N_t(f)}^{-1} \cdot \binom{N(m)}{N_t(m)}^{-1} \quad \text{for } \mathbf{W} \in \mathbb{W}^+$$

$$\mathbb{W}^+ = \left\{ \mathbf{W} \text{ such that } \sum_{i:G_i=f} W_i = N_t(f), \sum_{i:G_i=m} W_i = N_t(m) \right\}$$

Causal Estimands in SREs

- Average causal effect (ACE) in females:

$$\tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:G_i=f} \big(Y_i(1) - Y_i(0)\big)$$

- Average causal effect (ACE) in males:

$$\tau_{\text{fs}}(m) = \frac{1}{N(m)} \sum_{i:G_i=m} \big(Y_i(1) - Y_i(0)\big)$$

- Average causal effect (ACE) in all units:

$$\tau_{\text{fs}} = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m) = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i(1) - Y_i(0)\big)$$

Weighted average based on sex distribution in all units

Why SREs?

- SREs are ideal alternative when CREs are **not** feasible
- SREs may lead to more precise estimation of causal effect
- Studying SREs will aid interpretation & implementation in observational study

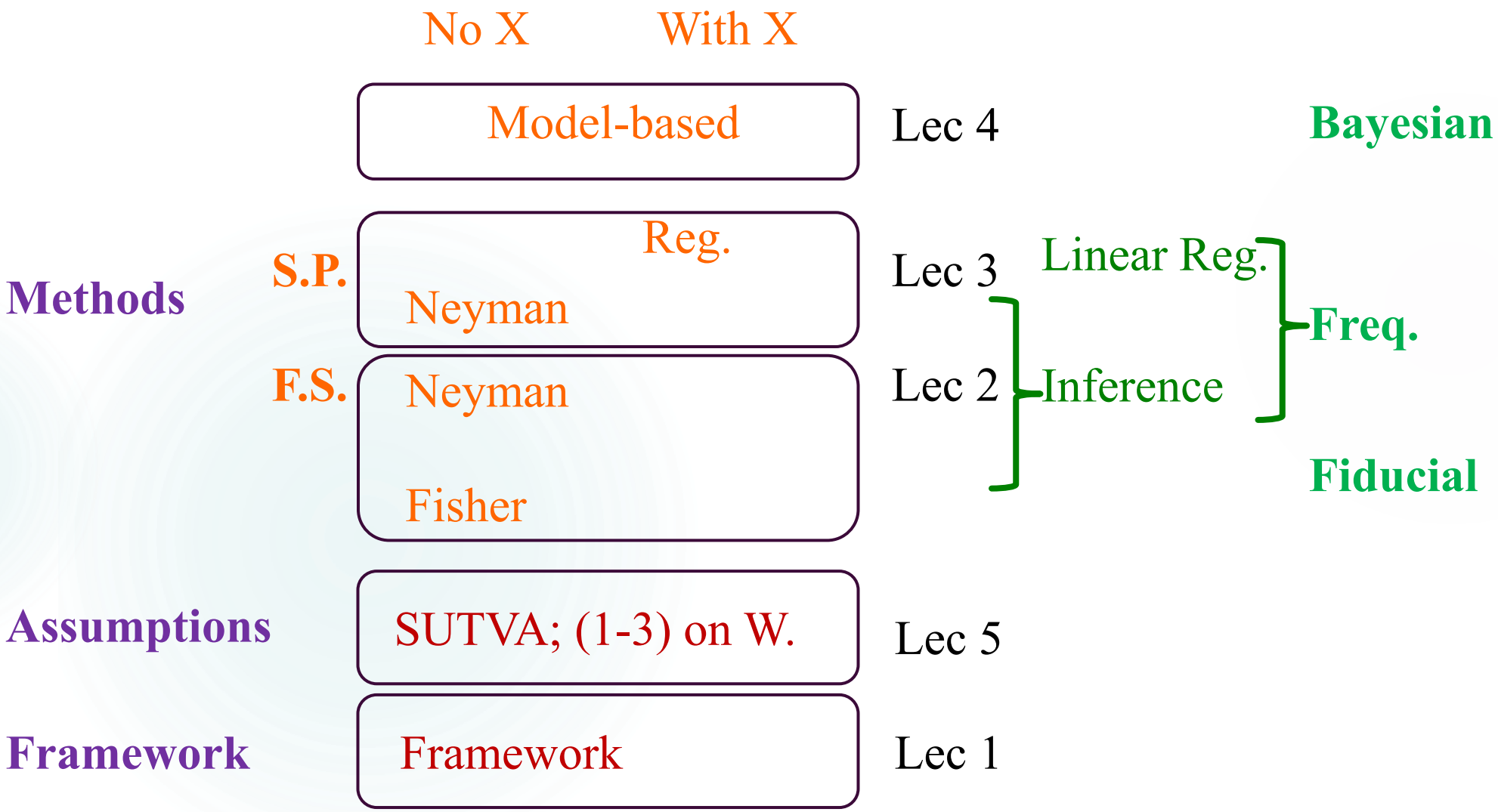
Inference Challenges of SREs

- A straightforward strategy:
  - Deal with each stratum separately
  - Summarize results from different strata to draw inclusion

- Practical challenges:
  - Sample sizes are often small in each stratum
  - Not interested in inference within each stratum: $\tau_{\mathrm{fs}}(f)$ or $\tau_{\mathrm{fs}}(m)$
  - Interested in inference across all strata: τ_{fs} or $\tau_{\mathrm{fs},t}$



Review


No X
With X
Model-based
Lec 4
Bayesian
Methods
S.P.
Reg.
Neyman
Lec 3
Linear Reg.
Freq.
F.S.
Neyman
Lec 2
Inference
Fisher
Fiducial
Assumptions
SUTVA; (1-3) on W.
Lec 5
Framework
Framework
Lec 1

Inference Approaches for SREs

- Fisher’s Exact P-Values
  - whether there exists a non-zero unit-level causal effect?
- Neyman’s Repeated Sampling Approach
  - how to estimate the average causal effect?
- Regression Methods
  - how covariates help estimate the average causal effect?
- Model-Based Inference
  - How concrete parametric models on potential outcomes help the infer the average causal effect?



Fisher's Exact P-Values

- Fisher's sharp null hypothesis:

$$H_0: \ Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$ ⇢ Same as the one in CREs

- Choice of test statistics:
  - $T^{\text{dif}}(m) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right|$ or $T^{\text{dif}}(f) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}} \right|$ ✗

Advantage: conceptually straightforward

Disadvantage: only use part of the data, may not be sensitive to the violation of H_0

Fisher's Exact P-Values

- Fisher's sharp null hypothesis:

$$H_0: \; Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$ ----→ Same as the one in CREs

- Choice of test statistics:
  - $T^{\text{dif}}(m) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right|$ or $T^{\text{dif}}(f) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}} \right|$ X
  - $$T^{\text{dif}} = \left| \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right| = \left| \frac{1}{N_{\text{t}}(f) + N_{\text{t}}(m)} \sum_{i=1}^{N} W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_c(f) + N_c(m)} \sum_{i=1}^{N} (1 - W_i) \cdot Y_i^{\text{obs}} \right|$$
    $$= \left| \frac{N_{\text{t}}(f)}{N_{\text{t}}} \cdot \overline{Y}_{\text{t}}^{\text{obs}}(f) - \frac{N_c(f)}{N_{\text{c}}} \cdot \overline{Y}_{\text{c}}^{\text{obs}}(f) + \frac{N_{\text{t}}(m)}{N_{\text{t}}} \cdot \overline{Y}_{\text{t}}^{\text{obs}}(m) - \frac{N_{\text{c}}(m)}{N_{\text{c}}} \cdot \overline{Y}_{\text{c}}^{\text{obs}}(m) \right|$$

Advantage: same as the one in CREs

Disadvantage: no longer has expectation equal to zero as it did in a CRE

Fisher's Exact P-Values

- Fisher's sharp null hypothesis:

$$H_0: \ Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$ ⇢ Same as the one in CREs

- Choice of test statistics:
  - $T^{\text{dif}}(m) = \left|\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m)\right|$ or $T^{\text{dif}}(f) = \left|\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}}\right|$ ✗
  - $T^{\text{dif}} = \left|\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right|$ ✓ / ✗
  - $T^{\text{dif},\lambda} = \left|\lambda \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f)\right) + (1-\lambda) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m)\right)\right|$ Weighted average of $T^{dif}(f)$ & $T^{dif}(m)$

Proportion of strata: $\lambda = \lambda_{\text{RSS}} \equiv N(f)/(N(f)+N(m))$ → $T^{\text{dif},\lambda_{\text{RSS}}}$

Fisher's Exact P-Values

- Fisher's sharp null hypothesis:

$$H_0: \ Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$

- Choice of test statistics:
  - $T^{\text{dif}}(m) = \left|\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m)\right|$ or $T^{\text{dif}}(f) = \left|\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}}\right|$ ✗
  - $T^{\text{dif}} = \left|\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right|$ ✓ / ✗
  - $T^{\text{dif},\lambda} = \left|\lambda \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f)\right) + (1-\lambda) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m)\right)\right|$ ✓
  - $T^{\text{rank}} = \left|\overline{R}_t - \overline{R}_c\right|$ ----→ Difference in average rank as in CREs
    Would not be attractive if there is substantial variation between strata

$$\overline{R}_t = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} R_i \qquad \overline{R}_c = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} R_i \qquad R_i = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N+1}{2}$$

Normalized rank of unit *i* in whole sample

Fisher's Exact P-Values

- Fisher's sharp null hypothesis:

$$H_0: \; Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$

- Choice of test statistics:
  - $T^{\text{dif}}(m) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right|$ or $T^{\text{dif}}(f) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}} \right|$ ✗
  - $T^{\text{dif}} = \left| \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right|$ ✓ / ✗
  - $T^{\text{dif},\lambda} = \left| \lambda \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f) \right) + (1-\lambda) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right) \right|$ ✓
  - $T^{\text{rank}} = \left| \overline{R}_t - \overline{R}_c \right|$ ✓ / ✗
  - $T^{\text{rank,stratum}} = \left| \overline{R}_t^{\text{strat}} - \overline{R}_c^{\text{strat}} \right|$ $\quad \overline{R}_t^{\text{strat}} = \frac{1}{N_{\text{t}}} \sum_{i:W_i=1} R_i^{\text{strat}} \quad \overline{R}_c^{\text{strat}} = \frac{1}{N_{\text{c}}} \sum_{i:W_i=0} R_i^{\text{strat}}$

$$R_i^{\text{strat}} = \begin{cases} \sum_{j:G_j=f} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j:G_j=f} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N(f)+1}{2}, & \text{if } G_i = f, \\ \sum_{j:G_j=m} \mathbf{1}_{Y_j^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j:G_j=m} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N(m)+1}{2}, & \text{if } G_i = m. \end{cases}$$

Normalized rank of unit i within its stratum

Fisher's Exact P-Values

- Fisher's sharp null hypothesis:

$$H_0: \ Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$

- Choice of test statistics:
  - $T^{\text{dif}}(m) = \left|\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m)\right|$ or $T^{\text{dif}}(f) = \left|\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}}\right|$ ✗
  - $T^{\text{dif}} = \left|\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right|$ ✓ / ✗
  - $T^{\text{dif},\lambda} = \left|\lambda \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f)\right) + (1-\lambda) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m)\right)\right|$ ✓
  - $T^{\text{rank}} = \left|\overline{R}_t - \overline{R}_c\right|$ ✓ / ✗
  - $T^{\text{rank,stratum}} = \left|\overline{R}_t^{\text{strat}} - \overline{R}_c^{\text{strat}}\right|$ ✓
- Get exact null distribution & P-values of the selected test statistics:

  Follow a similar procedure as in CREs: enumerate all possible values of W

Fisher's Exact P-Values for J strata

- Fisher's sharp null hypothesis:

$$H_0: \; Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$

- Choice of test statistics:
  - $T^{\text{dif}}(m) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right|$ or $T^{\text{dif}}(f) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}} \right|$ ✗
  - $T^{\text{dif}} = \left| \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right|$ ✓ / ✗
  - $T^{\text{dif},\lambda} = \left| \sum_{j=1}^{J} \lambda(j) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right) \right|$. $\lambda(j) = \frac{N(j)}{N}$ ✓
  - $T^{\text{rank}} = \left| \overline{R}_t - \overline{R}_c \right|$ ✓ / ✗
  - $T^{\text{rank,stratum}} = \left| \overline{R}_t^{\text{strat}} - \overline{R}_c^{\text{strat}} \right|$ ✓

$$R_i^{\text{strat}} = \sum_{i':G_{i'}=G_i} \mathbf{1}_{Y_{i'}^{\text{obs}} < Y_i^{\text{obs}}} + \frac{1}{2} \left(1 + \sum_{i':G_{i'}=G_i} \mathbf{1}_{Y_{i'}^{\text{obs}} = Y_i^{\text{obs}}} \right) - \frac{N(G_i)+1}{2}.$$

Example: The Project Star Data

❖ Fisher's sharp null hypothesis:

$$H_0: \; Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$

❖ Choice of test statistics:

- $T^{\text{dif}}(m) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m) \right|$ or $T^{\text{dif}}(f) = \left| \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{c}(f)^{\text{obs}} \right|$ ✗
- $T^{\text{dif}} = \left| \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \right| = 0.224 \Rightarrow p = 0.034$ ✓ / ✗
- $T^{\text{dif},\lambda} = \left| \sum_{j=1}^{J} \lambda(j) \cdot \left(\overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j) \right) \right|$. ✓

$$T^{\text{dif},\lambda_{\text{RSS}}} = 0.241 \Rightarrow p = 0.023$$

- $T^{\text{rank,stratum}} = \left| \overline{R}_t^{\text{strat}} - \overline{R}_c^{\text{strat}} \right| = 0.48 \Rightarrow p = 0.15$ ✓

Fails to reject the null hypothesis due to few outliers in the space of classes

❖ Remark: the p-values are calculated based on approximation by simulation

Neyman's Approach

❖ Causal estimand: Average causal effect (ACE) in all units

$$\tau_{\text{fs}} = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m) = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big)$$

$$\tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:G_i=f} \big(Y_i(1) - Y_i(0)\big) \qquad \tau_{\text{fs}}(m) = \frac{1}{N(m)} \sum_{i:G_i=m} \big(Y_i(1) - Y_i(0)\big)$$

$$\hat{\tau}^{\text{strat}} = \frac{N(f)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(m)$$ → Calculated & inferred in a similar way

Use only data in one stratum:

$$\hat{\tau}^{\text{dif}}(f) = \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f) = \frac{1}{N_{\text{t}}(f)} \sum_{i:G_i=f} W_i \cdot Y_i^{\text{obs}} - \frac{1}{N_{\text{c}}(f)} \sum_{i:G_i=f} (1 - W_i) \cdot Y_i^{\text{obs}}$$

Results for CREs:

$$\mathbb{E}_W\big(\hat{\tau}^{\text{dif}}(f)\big) = \tau_{\text{fs}}(f) \qquad \mathbb{V}_W\left(\hat{\tau}^{\text{dif}}(f)\right) = \frac{S_c^2(f)}{N_{\text{c}}(f)} + \frac{S_t^2(f)}{N_{\text{t}}(f)} - \frac{S_{ct}^2(f)}{N(f)}$$

Neyman's Approach

❖ Causal estimand: Average causal effect (ACE) in all units

$$\tau_{\text{fs}} = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right)$$

$$\tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:G_i=f} \left(Y_i(1) - Y_i(0)\right)$$

$$\tau_{\text{fs}}(m) = \frac{1}{N(m)} \sum_{i:G_i=m} \left(Y_i(1) - Y_i(0)\right)$$

$$\hat{\tau}^{\text{strat}} = \frac{N(f)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(m)$$

$$\mathbb{E}_W(\hat{\tau}^{\text{strat}}) = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m) = \tau_{\text{fs}}$$ ----→ Unbiased estimator

$$\mathbb{V}_W\left(\hat{\tau}^{\text{strat}}\right) = \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \mathbb{V}_W(\hat{\tau}_f) + \left(\frac{N(m)}{N_f+N(m)}\right)^2 \cdot \mathbb{V}_W(\hat{\tau}_m)$$ ← $\hat{\tau}^{\text{dif}}(f)$ & $\hat{\tau}^{\text{dif}}(m)$ are independent

$$= \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{S_c(f)^2}{N_c(f)} + \frac{S_t^2(f)}{N_t(f)} - \frac{S_{ct}^2(f)}{N(f)}\right)$$

$$+ \left(\frac{N(m)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{S_c(m)^2}{N_c(m)} + \frac{S_t^2(m)}{N_t(m)} - \frac{S_{ct}^2(m)}{N(m)}\right).$$

Estimate of the variance upper bound

$$\hat{\mathbb{V}}^{\text{neyman}} = \left(\frac{N(f)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{s_c^2(f)}{N_c(f)} + \frac{s_t^2(f)}{N_t(f)}\right)$$

$$+ \left(\frac{N(m)}{N(f)+N(m)}\right)^2 \cdot \left(\frac{s_c^2(m)}{N_c(m)} + \frac{s_t^2(m)}{N_t(m)}\right)$$

Neyman Approach for Project Star

- Causal estimand: Average causal effect (ACE) in all units

$$\tau_{\text{fs}} = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right)$$

- SRE:
  - $\hat{\tau}^{\text{dif}}(j) = \overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j)$
  - $\hat{\mathbb{V}}^{\text{neyman}}(j) = \frac{s_c(j)^2}{N_{\text{c}}(j)} + \frac{s_t(j)^2}{N_{\text{t}}(j)}$
  - $\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \hat{\tau}(j) = 0.241$
  - $\hat{\mathbb{V}}^{\text{neyman}} = \sum_{j=1}^{J} \left(\frac{N(j)}{N}\right)^2 \cdot \hat{\mathbb{V}}^{\text{neyman}}(j) = 0.092^2$
  - $\text{CI}^{0.95}(\tau_{\text{fs}}) = (0.061, 0.421)$

- CRE (incorrect, for comparison):

$$\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = 0.224$$

Standard error 0.141

$(-0.053, 0.500)$

This estimator of the sampling variance is biased due to variation in the probability of treatment between the different strata (0.5, 0.6, 0.67). There is a gain in precision from the stratification, due to variation in the average potential outcomes by stratum.



Neyman Approach: SREs vs CREs

- Estimand: $\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right) = \overline{Y}(1) - \overline{Y}(0),$
- Neyman estimator for a CRE:

$$\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} \qquad \mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{dif}}) = \frac{\sigma_c^2}{N_{\text{c}}} + \frac{\sigma_t^2}{N_{\text{t}}} \qquad p = \frac{N_t}{N}$$

- Neyman estimator for a SRE:

$$\hat{\tau}^{\text{strat}} = \frac{N(f)}{N}\cdot\hat{\tau}^{\text{dif}}(f) + \frac{N(m)}{N}\cdot\hat{\tau}^{\text{dif}}(m) = q\cdot\hat{\tau}^{\text{dif}}(f) + (1-q)\cdot\hat{\tau}^{\text{dif}}(m) \qquad q = \frac{N(f)}{N}$$

$$\hat{\tau}^{\text{dif}}(f) = \overline{Y}_{\text{t}}^{\text{obs}}(f) - \overline{Y}_{\text{c}}^{\text{obs}}(f) \qquad \hat{\tau}^{\text{dif}}(m) = \overline{Y}_{\text{t}}^{\text{obs}}(m) - \overline{Y}_{\text{c}}^{\text{obs}}(m)$$

$$\mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{strat}}) = \frac{q}{N}\cdot\left(\frac{\sigma_t^2(f)}{p} + \frac{\sigma_c^2(f)}{1-p}\right) + \frac{1-q}{N}\cdot\left(\frac{\sigma_t^2(m)}{p} + \frac{\sigma_c^2(m)}{1-p}\right) \qquad p = \frac{N_t(f)}{N(f)} = \frac{N_t(f)}{q\cdot N} = \frac{N_t(m)}{N(m)}$$

- SRE vs CRE:

$$\left(\mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{dif}}) - \mathbb{V}_{\text{sp}}(\hat{\tau}^{\text{strat}})\right) = q(1-q)\cdot\left(\frac{(\mu_c(f) - \mu_c(m))^2}{N(1-p)} + \frac{(\mu_t(f) - \mu_t(m))^2}{Np}\right) \geq 0$$

$\mathbb{E}[Y_i(0)|G_i = f]$

One is never worse off stratifying on a covariate

Regression Analysis of SREs (I)

$B_i(j) = I(G_i = j)$, $G_i \in \{1, \ldots, J\}$ denote the strata of unit i

Regression model: $Y_i^{\text{obs}} = \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \varepsilon_i$

Use $(J + 1)$ parameters

Assume that $\tau_{fs}(j) = \tau$ for all strata j

Stratum-specific intersections

Least square estimate: $(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\tau,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) \right)^2$

Population counterparts to these OLS:

$$(\tau^*, \beta^*) = \arg\min_{\tau,\beta} \mathbb{E}\left[\left(Y_i^{\text{obs}} - \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) \right)^2 \right]$$

Theorem 9.1 Suppose we conduct a stratified randomized experiment in a sample drawn at random from an infinite population. Then

(1) $\tau^* = \tau_\omega$ ⇢ $\tau_\omega = \sum_{j=1}^{J} \omega(j) \cdot \tau_{\text{sp}}(j) \Big/ \left(\sum_{j=1}^{J} \omega(j) \right)$ $\qquad \omega(j) = q(j) \cdot e(j) \cdot (1 - e(j))$

(2) $\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_\omega \right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\left(W_i - \sum_{j=1}^{J} q(j) \cdot B_i(j) \right)^2 \cdot \left(Y_i^{\text{obs}} - \tau^* \cdot W_i - \sum_{j=1}^{J} \beta_j^* \cdot B_i(j) \right)^2 \right]}{\left(\sum_{j=1}^{J} q(j) \cdot e(j) \cdot (1 - e(j)) \right)^2} \right)$

Remark: the proof follows a similar protocol as in CREs

Regression Analysis of SREs (II)

Regression model with interaction (a saturated model): Use $2J$ parameters

$$Y_i^{\text{obs}} = \tau \cdot W_i \cdot \frac{B_i(j)}{N(j)/N} + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \sum_{j=1}^{J-1} \gamma(j) \cdot W_i \cdot \left(B_i(j) - B_i(J) \cdot \frac{N(J)}{N} \right) + \varepsilon_i$$

(i) $Y_{i(j)}^{obs} = \beta_j + \tau_j \cdot W_i + \varepsilon_i$, then $\tau = N^{-1} \sum_{j=1}^{J} N(j)\tau_j$.

(ii) $Y_{i(j)}^{obs} = \beta_j + \tau \cdot W_i + \gamma_j \cdot W_i + \varepsilon_i$, with constraint $\sum_{j=1}^{J} N(j)\gamma_j = 0$.

Least square estimate: $(\hat{\tau}^{\text{ols,inter}}, \hat{\beta}^{\text{ols,inter}})$

Population counterparts to these OLS: (τ^*, β^*)

Theorem 9.2 Suppose we conduct a stratified randomized experiment in a sample drawn at random from an infinite population. Then

(1) $\tau^* = \tau_{\text{sp}}$

(2) $$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols,inter}} - \tau_{\text{sp}} \right) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^{J} q(j)^2 \cdot \left(\frac{\sigma_c^2(j)}{(1-e(j)) \cdot q(j)} + \frac{\sigma_t^2(j)}{e(j) \cdot q(j)} \right) \right)$$

Remark: the proof follows a similar protocol as in CREs

Regression Analysis for Project Star

- Causal estimand:

$$\tau_\omega = \sum_{j=1}^{J} \omega(j) \cdot \tau_{\text{sp}}(j) \Big/ \left(\sum_{j=1}^{J} \omega(j) \right) \qquad \text{or} \qquad \tau_{\text{sp}}$$

- SRE:
  - $Y_i^{\text{obs}} = \tau \cdot W_i + \sum_{j=1}^{J} \beta(j) \cdot B_i(j) + \varepsilon_i$ `lm(Y_obs ~ W + Strata)`
  - $\hat{\tau}^{\text{ols}} = 0.238 \ \ (\widehat{\text{s.e.}}\ 0.103)$
  - $Y_{i(j)}^{obs} = \beta_j + \tau_j \cdot W_i + \varepsilon_i, \quad \text{then } \tau = N^{-1} \sum_{j=1}^{J} N(j)\tau_j.$ `lm(Y_obs ~ W * Strata)`
  - $\hat{\tau}_{\text{ols,inter}} = 0.241 \ \ (\widehat{\text{s.e.}}\ 0.095).$

The two estimates for the average effect are close, with similar standard errors, consistent with limited heterogeneity in the treatment effects .

Model-Based Analysis for SRE (Not required)

The inputs

$f(\mathbf{y}^{mis}, \mathbf{Y}^{obs}, \mathbf{W}|\theta)$

$p(\theta)$

$\Big\} f(\mathbf{y}^{mis}, \mathbf{Y}^{obs}, \mathbf{W}, \theta)$

Step 2

$$f(\mathbf{Y}^{obs}, \mathbf{W}|\theta) = \int_{\mathbf{y}^{mis}} f(\mathbf{y}^{mis}, \mathbf{Y}^{obs}, \mathbf{W}|\theta)\, d\mathbf{y}^{mis}$$

$$p(\theta|\mathbf{Y}^{obs}, \mathbf{W}) = \frac{p(\theta) \cdot f(\mathbf{Y}^{obs}, \mathbf{W}|\theta)}{f(\mathbf{Y}^{obs}, \mathbf{W})}$$

Step 1

$$f(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{W}, \theta) = \frac{f(\mathbf{Y}^{mis}, \mathbf{Y}^{obs}|\mathbf{W}, \theta)}{f(\mathbf{Y}^{obs}|\mathbf{W}, \theta)} = \frac{f(\mathbf{Y}^{mis}, \mathbf{Y}^{obs}|\mathbf{W}, \theta)}{\int_{\mathbf{y}^{mis}} f(\mathbf{y}^{mis}, \mathbf{Y}^{obs}|\mathbf{W}, \theta) d\mathbf{y}^{mis}}$$

Step 3

$$f(\mathbf{Y}^{mis}, \theta|\mathbf{Y}^{obs}, \mathbf{W}) = f(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{W}, \theta) \cdot p(\theta|\mathbf{Y}^{obs}, \mathbf{W})$$

$$f(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{W}) = \int_{\theta} f(\mathbf{Y}^{mis}, \theta|\mathbf{Y}^{obs}, \mathbf{W})\, d\theta$$

$f(\mathbf{Y}^{mis}|\mathbf{Y}^{obs}, \mathbf{W})$

Step 4

The target

$f(\tau|\mathbf{Y}^{obs}, \mathbf{W})$

Model-Based Analysis for SRE (Not required)

- The 3 inputs in SRE

$$\Pr(\mathbf{W}|\mathbf{S},\mathbf{Y}(0),\mathbf{Y}(1)) = \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} \quad \text{for } \mathbf{W} \in \mathbb{W}^+$$

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Big| B_i(j), \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix}, \begin{pmatrix} \sigma_c^2(j) & 0 \\ 0 & \sigma_t^2(j) \end{pmatrix} \right)$$

$$\theta = (\mu_c(j), \mu_t(j), \sigma_c^2(j), \sigma_t^2(j), w = 0, 1, j = 1, \ldots, J).$$

Independent or hierarchical prior for θ

- Example

$$\Pr(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{S}) = \binom{N_f}{N_t(f)}^{-1} \cdot \binom{N(m)}{N_t(m)}^{-1} \quad \text{for } \mathbf{W} \in \mathbb{W}^+$$

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Big| B_i(j) = 1, \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right)$$

$$\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix} \Big| \sigma^2, \gamma_c, \gamma_t, \Sigma \sim \mathcal{N}\left(\begin{pmatrix} \gamma_c \\ \gamma_t \end{pmatrix}, \Sigma \right), \quad \begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix} \perp \begin{pmatrix} \mu_c(k) \\ \mu_t(k) \end{pmatrix} \Big| \sigma^2, \gamma_c, \gamma_t, \Sigma, j \neq k.$$

Hyper-parameters $\theta = (\gamma_c, \gamma_t, \Sigma, \sigma^2)$ hierarchical prior

Model-Based Analysis for Project Star (Not required)

- For the hierarchical prior distribution:

$$\mathbb{E}[\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{B}, \text{hierarchical}] = 0.235,$$

$$\mathbb{V}(\tau_{\text{fs}}|\mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{B}, \text{hierarchical})^2 = 0.107^2.$$

- We can also assess the evidence for variation in average potential outcomes and treatment effects by strata.

$$\begin{pmatrix}\mu_c(j)\\ \mu_t(j)\end{pmatrix}\Bigg|\sigma^2, \gamma_c, \gamma_t, \Sigma \sim \mathcal{N}\left(\begin{pmatrix}\gamma_c\\ \gamma_t\end{pmatrix}, \Sigma\right), \quad \begin{pmatrix}\mu_c(j)\\ \mu_t(j)\end{pmatrix} \perp\!\!\!\perp \begin{pmatrix}\mu_c(k)\\ \mu_t(k)\end{pmatrix}\Bigg|\sigma^2, \gamma_c, \gamma_t, \Sigma, j \neq k.$$

Hyper-parameters $\theta = (\gamma_c, \gamma_t, \Sigma, \sigma^2)$

  - E.g. we can check the posterior of Σ or its elements or new estimand based on it (e.g. the variance of the treatment effect).

- This shows the richness of model-based analysis.



Summary for SRE

- In general the stratification should be considered in design if treatment rates and potential outcomes vary systematically by stratum.
- All 4 approaches can be adapted in a fairly straightforward manner to take account of the stratification.
  - ✓ Higher power -- choosing statistics (FEP)
  - ✓ Higher precision (Neyman)
  - ✓ Appropriate model and correct interpretation (Regression)
  - ✓ A hierarchical model can be useful to take account of similarities in potential outcome distributions across strata (Model-based analysis)
- Stratification can increase precision of estimation when the strata are good predictors of the potential outcomes.



Pairwise Randomized Experiments



Pairwise Randomized Experiment

- A special case of a Stratified Randomized Experiment with $J = N/2$ strata
- One treated and one control unit (i.e., a pair) in each strata:

$$N(j) = 2, N_t(j) = N_c(j) = 1$$

- Assignment probability:

$$p(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{N/2} \binom{N(j)}{N_t(j)}^{-1} = \prod_{j=1}^{N/2} \frac{1}{2} = 2^{-N/2}, \quad \text{for } \mathbf{W} \in \mathbb{W}^+$$

$$\mathbb{W}^+ = \left\{ \mathbf{W} \;\middle|\; \sum_{i:G_i=j} W_i = 1 \text{ for } j = 1, \ldots, N/2 \right\}$$

- Intuition behind PREs:
  - Within-pair variation in potential outcomes < between-pair variation
  - If the population is relatively large, it may be possible to get very close matches of units in each pair



Children's Television Workshop Data

Table 10.1. ***Data from Youngstown Children's Television Workshop Experiment***

Pair G_i	Treatment W_i	Pre-Test Score X_i	Post-Test Score Y_i^{obs}	Normalized Rank Post-Test Score R_i
1	0	12.9	54.6	−7.5
1	1	12.0	60.6	2.5
2	0	15.1	56.5	−4.5
2	1	12.3	55.5	5.5
3	0	16.8	75.2	0.5
3	1	17.2	84.8	4.5
4	0	15.8	75.6	1.5
4	1	18.9	101.9	7.5
5	0	13.9	55.3	−6.5
5	1	15.3	70.6	−1.5
6	0	14.5	59.3	−3.5
6	1	16.6	78.4	2.5
7	0	17.0	87.0	5.5
7	1	16.0	84.2	3.5
8	0	15.8	73.7	−0.5
8	1	20.1	108.6	7.5

- Around 1970
- Evaluate the effect of an educational children's television program on reading ability as measured through test scores.



Special Notations of PREs

- Two units within each pair: unit A and B
- Potential outcomes: $(Y_{j,A}(0), Y_{j,A}(1))$ and $(Y_{j,B}(0), Y_{j,B}(1))$
- Treatment indicators: $W_{j,A}$ and $W_{j,B}$
- Observed data

$$Y_{j,A}^{\text{obs}} = \begin{cases} Y_{j,A}(0) & \text{if } W_{j,A} = 0, \\ Y_{j,A}(1) & \text{if } W_{j,A} = 1, \end{cases} \quad \text{and} \quad Y_{j,B}^{\text{obs}} = \begin{cases} Y_{j,B}(0) & \text{if } W_{j,A} = 1, \\ Y_{j,B}(1) & \text{if } W_{ji,A} = 0. \end{cases}$$

$$Y_{j,c}^{\text{obs}} = \begin{cases} Y_{j,A}^{\text{obs}} & \text{if } W_{i,A} = 0, \\ Y_{j,B}^{\text{obs}} & \text{if } W_{i,A} = 1, \end{cases} \quad \text{and} \quad Y_{j,t}^{\text{obs}} = \begin{cases} Y_{j,B}^{\text{obs}} & \text{if } W_{i,A} = 0, \\ Y_{j,A}^{\text{obs}} & \text{if } W_{i,A} = 1. \end{cases}$$

- Unit-level assignment probability: $\Pr(W_{j,A} = 1 | \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = 1/2$

treated subpopulation ≈ control subpopulation

- Average treatment effect within pair j

$$\tau_{\text{pair}}(j) = \frac{1}{2} \sum_{i:G_i=j} \big(Y_i(1) - Y_i(0)\big) = \frac{1}{2}\big(\big(Y_{j,A}(1) - Y_{j,A}(0)\big) + \big(Y_{j,B}(1) - Y_{j,B}(0)\big)\big)$$

$$\tau_{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i(1) - Y_i(0)\big) = \frac{2}{N} \sum_{j=1}^{N/2} \tau_{\text{pair}}(j)$$

Children's Television Workshop Data

Table 10.1. *Data from Youngstown Children's Television Workshop Experiment*

Pair G_i	Treatment W_i	Pre-Test Score X_i	Post-Test Score Y_i^{obs}	Normalized Rank Post-Test Score
1	0			
1	1			
2	0			
2	1			
3	0			
3	1			
4	0			
4	1			
5	0			
5	1			
6	0			
6	1			
7	0			
7	1			
8	0			
8	1			

Table 10.2. *Potential Outcomes and Covariates from Children's Television Workshop Experiment, from Table 10.1*

Pair	Unit A					Unit B				
	$Y_{i,A}(0)$	$Y_{i,A}(1)$	$W_{i,A}$	$Y_{i,A}^{obs}$	$X_{i,A}$	$Y_{i,B}(0)$	$Y_{i,B}(1)$	$W_{i,B}$	$Y_{i,B}^{obs}$	$X_{i,B}$
1	54.6	?	0	54.6	12.9	?	60.6	1	60.6	12.0
2	56.5	?	0	56.5	15.1	?	55.5	1	55.5	13.9
3	75.2	?	0	75.2	16.8	?	84.8	1	84.8	17.2
4	76.6	?	0	75.6	15.8	?	101.9	1	101.9	18.9
5	55.3	?	0	55.3	13.9	?	70.6	1	70.6	15.3
6	59.3	?	0	59.3	14.5	?	78.4	1	78.4	16.6
7	87.0	?	0	87.0	17.0	?	84.2	1	84.2	16.0
8	73.7	?	0	73.7	15.8	?	108.6	1	108.6	20.1



Causal Estimands in PREs

- Average causal effect (ACE) in strata j:

$$\tau_{\text{pair}}(j) = \frac{1}{2}\sum_{i:G_i=j}\big(Y_i(1) - Y_i(0)\big) = \frac{1}{2}\big(\big(Y_{j,A}(1) - Y_{j,A}(0)\big) + \big(Y_{j,B}(1) - Y_{j,B}(0)\big)\big)$$

- Average causal effect (ACE) in all units:

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\big(Y_i(1) - Y_i(0)\big) = \frac{2}{N}\sum_{j=1}^{N/2}\tau_{\text{pair}}(j)$$

Inference Approaches for PREs

- Fisher's Exact P-Values (same as SREs)
  - whether there exists a non-zero unit-level causal effect?
- Neyman's Repeated Sampling Approach
  - how to estimate the average causal effect?
- Regression Methods
  - how covariates help estimate the average causal effect?
- Model-Based Inference (same as SREs)
  - How concrete parametric models on potential outcomes help the infer the average causal effect?

One treated and one control unit in each strata lead to specified challenges for these two methods

Fisher's Exact P-Values

❖ Fisher's sharp null hypothesis:

$$H_0: \ Y_i(0) = Y_i(1) \text{ for } i = 1, 2, \ldots, N$$ ⇢ Same as the one in CREs & SREs

❖ Choice of test statistics:

	SREs	PREs
➢ $T^{\text{dif}}(j) = \lvert \bar{Y}_t(j) - \bar{Y}_c(j) \rvert$	✗	✗
➢ $T^{\text{dif}} = \left\lvert \bar{Y}_{\text{t}}^{\text{obs}} - \bar{Y}_{\text{c}}^{\text{obs}} \right\rvert$	✓ / ✗	✓ / ✗
➢ $T^{\text{dif},\lambda} = \left\lvert \lambda \cdot \left(\bar{Y}_{\text{t}}^{\text{obs}}(f) - \bar{Y}_{\text{c}}^{\text{obs}}(f) \right) + (1-\lambda) \cdot \left(\bar{Y}_{\text{t}}^{\text{obs}}(m) - \bar{Y}_{\text{c}}^{\text{obs}}(m) \right) \right\rvert$	✓	✓
➢ $T^{\text{rank}} = \left\lvert \bar{R}_t - \bar{R}_c \right\rvert$	✓ / ✗	✓ / ✗
➢ $T^{\text{rank,stratum}} = \left\lvert \bar{R}_t^{\text{strat}} - \bar{R}_c^{\text{strat}} \right\rvert$ ➡ $T^{\text{rank,pair}} = \left\lvert \frac{2}{N} \sum_{j=1}^{N/2} \left(\mathbf{1}_{Y_{j,1}^{\text{obs}} > Y_{j,0}^{\text{obs}}} - \mathbf{1}_{Y_{j,1}^{\text{obs}} < Y_{j,0}^{\text{obs}}} \right) \right\rvert$	✓	✓

❖ Get exact null distribution & P-values of the selected test statistics:

Follow a similar procedure as in CREs & SREs: enumerate all possible values of W

Neyman's Approach

❖ Causal estimand: Average causal effect (ACE) in all units

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right) = \frac{2}{N}\sum_{j=1}^{N/2}\tau_{\text{pair}}(j)$$

$$\tau_{\text{pair}}(j) = \frac{1}{2}\sum_{i:G_i=j}\left(Y_i(1) - Y_i(0)\right) = \frac{1}{2}\left(\left(Y_{j,A}(1) - Y_{j,A}(0)\right) + \left(Y_{j,B}(1) - Y_{j,B}(0)\right)\right)$$

$$\hat{\tau}^{\text{dif}} = \frac{1}{N/2}\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j) = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$$

Use only data in one stratum:

$$\hat{\tau}^{\text{pair}}(j) = Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}} = \sum_{i:G_i=j}(2\cdot W_i - 1)\cdot Y_i^{\text{obs}}$$

Results for PREs: $\mathbb{E}_W\left(\hat{\tau}^{\text{pair}}(j)\right) = \tau_{fs}(j)$

$$\mathbb{V}_W(\hat{\tau}^{\text{pair}}(j)) = \frac{S_c(j)^2}{N_{\text{c}}(j)} + \frac{S_t^2(j)}{N_{\text{t}}(j)} - \frac{S_{ct}(j)^2}{N(j)} = S_c(j)^2 + S_t^2(j) - \frac{S_{ct}(j)^2}{2}$$

Neyman's Approach

❖ Causal estimand: Average causal effect (ACE) in all units

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right) = \frac{2}{N}\sum_{j=1}^{N/2}\tau_{\text{pair}}(j)$$

$$\tau_{\text{pair}}(j) = \frac{1}{2}\sum_{i:G_i=j}\left(Y_i(1) - Y_i(0)\right) = \frac{1}{2}\left(\left(Y_{j,A}(1) - Y_{j,A}(0)\right) + \left(Y_{j,B}(1) - Y_{j,B}(0)\right)\right)$$

$$\hat{\tau}^{\text{dif}} = \frac{1}{N/2}\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j) = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$$

$\mathbb{E}_W\left(\hat{\tau}^{\text{dif}}\right) = \tau_{\text{fs}}$ ----→ Unbiased estimator

Cannot be estimated all the time

$$\mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right) = \frac{1}{(N/2)^2}\sum_{j=1}^{N/2}\left(S_c^2(j) + S_t^2(j) - \frac{S_{ct}^2(j)}{2}\right)$$

$\{\hat{\tau}^{\text{pair}}(j)\}_j$ are independent

Cannot be estimated in **PREs** as each pair has only one treated and control unit

Variance Estimation via Extra Assumptions

- A_1: constant treatment effect within pairs: $Y_{jA}(1) - Y_{jA}(0) = Y_{jB}(1) - Y_{jB}(0) = \tau_{\text{pair}}(j)$

$$S^2(j) = S_c^2(j) = S_t^2(j) \quad \Rightarrow \quad \mathbb{V}_W(\hat{\tau}^{\text{pair}}(j)) = 2 \cdot S^2(j)$$

- A_2: treatment effect is constant across pairs: $\tau_{pair}(j) = \tau_{\text{fs}}$ for all j, and the within-pair variances are constant $S^2(j) = S^2$ for all j

$$\mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right) = \frac{1}{(N/2)^2} \sum_{j=1}^{N/2} \left(S_c^2(j) + S_t^2(j) - \frac{S_{ct}^2(j)}{2} \right) = \frac{4}{N} \cdot S^2$$

$$\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right) = \frac{4}{N \cdot (N-2)} \cdot \sum_{j=1}^{N/2} \left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}} \right)^2$$

Summary of Neyman's Approach

Theorem 10.1 Suppose we have J pairs of units, and randomly assign one unit from each pair to the active treatment and the other unit to the control treatment. Then

(1) $\mathbb{E}_W\left(\hat{\tau}^{\text{dif}}\right) = \tau_{fs}$ ⇢ Unbiasness

(2) $$\mathbb{V}_W\left(\hat{\tau}^{\text{dif}}\right) = \frac{1}{N^2}\sum_{j=1}^{N/2}\left(Y_{j,A}(0) + Y_{j,A}(1) - \left(Y_{j,B}(0) + Y_{j,B}(1)\right)\right)^2,$$ ⇢ Variance

(3) $$\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right) = \frac{4}{N\cdot(N-2)}\cdot\sum_{j=1}^{N/2}\left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}}\right)^2$$ ⇢ Variance estimation

$$\mathbb{E}\left[\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)\right] = \mathbb{V}_W(\hat{\tau}^{\text{dif}}) + \frac{4}{N\cdot(N-2)}\cdot\sum_{j=1}^{N/2}\left(\tau_{\text{pair}}(j) - \tau\right)^2$$

⇢ = 0 if the treatment effect is constant across and within pairs

If there is heterogeneity in the treatment effects, then this sampling variance estimator is upwardly biased, and the corresponding confidence intervals will be conservative but still valid.

Neyman Estimator: PREs vs CREs

Assume the same assignment vector is achieved in a PRE and a CRE, then

- We will get same estimator for causal effect
- But, PRE usually results in a better precision

PRE

$$\hat{\tau}^{\text{dif}} = \frac{1}{8} \cdot \sum_{j=1}^{8} \hat{\tau}^{\text{pair}}(j) = 13.4$$

$$\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right) = \frac{1}{8 \cdot (8-1)} \cdot \sum_{j=1}^{8} \left(\hat{\tau}^{\text{pair}}(j) - \hat{\tau}^{\text{dif}}\right)^2 = 4.6^2$$

$$\text{CI}^{0.95}(\tau_{\text{fs}}) = \left(\hat{\tau} - 1.96 \times \sqrt{\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)}, \hat{\tau} + 1.96 \times \sqrt{\hat{\mathbb{V}}^{\text{pair}}\left(\hat{\tau}^{\text{dif}}\right)}\right) = (4.3, 22.5)$$

CRE

$$\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = 13.4$$

$$s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} \left(Y_i^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right)^2 = 18.5^2, \quad \text{and} \quad s_t^2 = 12.2^2$$

$$\hat{\mathbb{V}}^{\text{neyman}} = \frac{s_c^2}{8} + \frac{s_t^2}{8} = 7.8^2$$

$(-1.8, 28.7)$

Solution for Regression Method?

- Idea in Neyman's approach?
- Recall:

$$\tau_{\text{sp}} = \mathbb{E}_{\text{sp}}[\tau_{\text{pair}}(j)]$$

$$\hat{\tau}^{\text{dif}} = \frac{2}{N}\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j)$$

- $\hat{\tau}^{dif}$ can also be interpreted as a regression estimator, where the regression function is specified simply as a constant:

$$\hat{\tau}^{\text{pair}}(j) = \tau_{\text{sp}} + \varepsilon_j$$



Regression Analysis of PREs

basic outcome in the analysis

$$\hat{\tau}^{\text{ols}} = \frac{2}{N}\sum_{j=1}^{N/2}\hat{\tau}^{\text{pair}}(j) = \hat{\tau}^{\text{dif}}$$

M1: $\hat{\tau}^{\text{pair}}(j) = \tau + \varepsilon_j$ ⟶ No additional covariates

$$\Delta_{X,j} = \left(W_{j,A}\cdot\left(X_{j,A} - X_{j,B}\right) + (1 - W_{j,A})\cdot\left(X_{j,B} - X_{j,A}\right)\right)$$

M2: $\hat{\tau}^{\text{pair}}(j) = \tau + \beta\cdot\Delta_{X,j} + \varepsilon_j$ ⟶ With within-pair differences only

$$\overline{X}_j = \left(X_{j,A} + X_{j,B}\right)/2$$

M3: $\hat{\tau}^{\text{pair}}(j) = \tau + \gamma\cdot\overline{X}_j + \varepsilon_j$ ⟶ With pair averages only

$(\overline{X}_j - \overline{X})$

M4: $\hat{\tau}^{\text{pair}}(j) = \tau + \beta\cdot\Delta_{X,j} + \gamma\cdot(\overline{X}_j - \overline{X}) + \varepsilon_j$ ⟶ With within-pair differences & pair averages

Regression Analysis of PREs

$$\Delta_{X,j} = \left(W_{j,A} \cdot \left(X_{j,A} - X_{j,B}\right) + (1 - W_{j,A}) \cdot \left(X_{j,B} - X_{j,A}\right)\right)$$

Regression model M4: $\hat{\tau}^{\text{pair}}(j) = \tau + \beta \cdot \Delta_{X,j} + \gamma \cdot (\overline{X}_j - \overline{X}) + \varepsilon_j$

basic outcome in the analysis

$$\overline{X}_j = \left(X_{j,A} + X_{j,B}\right)/2$$

Least square estimate: $(\hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}}) = \arg\min_{\tau,\beta,\gamma} \sum_{i=1}^{N} \left(\hat{\tau}^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \overline{X})\right)^2$

Population counterparts to these OLS:

$$\left(\tau^*, \beta^*, \gamma^*\right) = \arg\min_{\tau,\beta,\gamma} \mathbb{E}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau - \beta \cdot \Delta_{X,j} - \gamma \cdot (\overline{X}_j - \mu_X)\right)^2\right]$$

Theorem 10.2 Suppose we conduct a pairwise randomized experiment in a sample of pairs drawn at random from the super-population. Then

(1) $\tau^* = \tau_{\text{sp}}$

(2) $\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_{\text{sp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_{\text{sp}}\left[\left(\hat{\tau}^{\text{pair}}(j) - \tau^* - \beta^* \cdot \Delta_{X,j} - \gamma^* \cdot (\overline{X}_j - \mu_X)\right)^2\right]\right)$

Remark: the proof follows a similar protocol as in CREs & SREs



Regression Analysis for CTW Data

M1: $\hat{\tau}^{\text{pair}}(j) = \tau + \varepsilon_j$ $\qquad \hat{\tau}^{\text{ols}} = 13.4 \quad (\widehat{\text{s.e.}}\ 4.3)$

M2: $\hat{\tau}^{\text{pair}}(j) = \tau + \beta \cdot \Delta_{X,j} + \varepsilon_j$ $\qquad \hat{\tau}_{\text{pair}}(j) = \underset{(1.5)}{9.0} + \underset{(0.6)}{5.4} \times \Delta_{X,j},$

M3: $\hat{\tau}^{\text{pair}}(j) = \tau + \gamma \cdot \overline{X}_j + \varepsilon_j$ $\qquad \hat{\tau}^{\text{pair}}(j) = \underset{(3.5)}{13.4} + \underset{(1.7)}{3.9} \times \overline{X}_j.$

M4: $\hat{\tau}^{\text{pair}}(j) = \tau + \beta \cdot \Delta_{X,j} + \gamma \cdot (\overline{X}_j - \overline{X}) + \varepsilon_j$

$$\hat{\tau}^{\text{pair}}(j) = \underset{(1.5)}{8.5} + \underset{(0.8)}{5.9} \times \Delta_{X,j} \quad \underset{(0.7)}{-1.0}\,(\overline{X}_j - \overline{X})$$

Model-Based Analysis for PRE

The inputs

$$\Pr(\mathbf{W}|\mathbf{S},\mathbf{Y}(0),\mathbf{Y}(1)) = \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} \quad \text{for } \mathbf{W} \in \mathbb{W}^+$$

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Big| B_i(j), \theta \sim \mathcal{N}\left(\begin{pmatrix} \mu_c(j) \\ \mu_t(j) \end{pmatrix}, \begin{pmatrix} \sigma_c^2(j) & 0 \\ 0 & \sigma_t^2(j) \end{pmatrix} \right)$$

$$\theta = (\mu_c(j), \mu_t(j), \sigma_c^2(j), \sigma_t^2(j), w = 0, 1, j = 1, \ldots, J).$$

Independent or hierarchical prior for θ

Step 1

$$f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W},\theta) = \frac{f(\mathbf{Y}^{\text{mis}},\mathbf{Y}^{\text{obs}}|\mathbf{W},\theta)}{f(\mathbf{Y}^{\text{obs}}|\mathbf{W},\theta)} = \frac{f(\mathbf{Y}^{\text{mis}},\mathbf{Y}^{\text{obs}}|\mathbf{W},\theta)}{\int_{\mathbf{y}^{\text{mis}}} f(\mathbf{y}^{\text{mis}},\mathbf{Y}^{\text{obs}}|\mathbf{W},\theta)\mathrm{d}\mathbf{y}^{\text{mis}}}$$

Step 2

$$f(\mathbf{Y}^{\text{obs}},\mathbf{W}|\theta) = \int_{\mathbf{y}^{\text{mis}}} f(\mathbf{y}^{\text{mis}},\mathbf{Y}^{\text{obs}},\mathbf{W}|\theta)\,\mathrm{d}\mathbf{y}^{\text{mis}}$$

$$p(\theta|\mathbf{Y}^{\text{obs}},\mathbf{W}) = \frac{p(\theta)\cdot f(\mathbf{Y}^{\text{obs}},\mathbf{W}|\theta)}{f(\mathbf{Y}^{\text{obs}},\mathbf{W})}$$

Step 3

$$f(\mathbf{Y}^{\text{mis}},\theta|\mathbf{Y}^{\text{obs}},\mathbf{W}) = f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W},\theta)\cdot p(\theta|\mathbf{Y}^{\text{obs}},\mathbf{W}).$$

$$f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W}) = \int_{\theta} f(\mathbf{Y}^{\text{mis}},\theta|\mathbf{Y}^{\text{obs}},\mathbf{W})\,\mathrm{d}\theta$$

$$f(\mathbf{Y}^{\text{mis}}|\mathbf{Y}^{\text{obs}},\mathbf{W})$$

The target $f(\tau|\mathbf{Y}^{\text{obs}},\mathbf{W})$

Step 4

Model-Based Analysis for CTW Data

- Causal estimand: τ_{fs} or $\tau_{\text{sp}} = \gamma$
- Input:
  - $$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Big| G_i = j, X_i = x, \mu(1), \ldots, \mu(N/2), \gamma, \beta, \sigma_c^2, \sigma_t^2 \sim \mathcal{N}\left(\begin{pmatrix} \mu(j) + x \cdot \beta \\ \mu(j) + \gamma + x \cdot \beta \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_t^2 \end{pmatrix} \right)$$
  - $$\begin{pmatrix} \mu(1) \\ \vdots \\ \mu(N/2) \end{pmatrix} \Big| \sigma_\mu^2, \mu \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_\mu^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_\mu^2 \end{pmatrix} \right)$$
- Output:
  - $\mathbb{E}[\tau_{\text{fs}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}, \mathbf{G}] = 8.4$
  - $\mathbb{V}(\tau_{\text{fs}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}, \mathbf{X}, \mathbf{G}) = 1.7^2$

Quite similar to those for the regression model with the covariate equal to differences in pre-treatment variables,
partly due to the choice of sampling distribution for potential outcomes.



Model-Based Analysis for CTW Data

- Model:
  - ✓ Hyper-parameter: $\boldsymbol{\theta} = (\mu, \beta, \gamma, \ln\sigma_\mu, \ln\sigma_c, \ln\sigma_t)'$
  - ✓ Prior: $\mu, \beta, \gamma \sim \mathcal{N}(0,1)$, $\sigma_\mu^{-2}, \sigma_c^{-2}, \sigma_t^{-2} \sim \chi_1^2$ (independent) => $p(\boldsymbol{\theta})$
  - ✓ Likelihood: $Y_i^{\text{obs}} \mid \boldsymbol{G}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\theta} \sim$? => $L(\boldsymbol{\theta})$
- Random Walk Metropolis-Hastings (RWMH):
  - Propose $\boldsymbol{\theta}^* \sim \mathcal{N}\left(\boldsymbol{\theta}^{(r)}, \mathbf{D}\right)$, $\mathbf{D}$ is diagonal (adapting parameter)
  - Accept $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^*$ with prob $\min\{1, p(\boldsymbol{\theta}^*)L(\boldsymbol{\theta}^*)/p(\boldsymbol{\theta})L(\boldsymbol{\theta})\}$
- Procedure to get the distribution of τ_{fs} :

  (1) Generate posterior: $\mu, \beta, \gamma, \ln\sigma_\mu, \ln\sigma_c, \ln\sigma_t \mid \boldsymbol{Y}^{\text{obs}}, \boldsymbol{G}, \boldsymbol{X}, \boldsymbol{W}$

  (2) Generate ?

  (3) Impute $Y_i^{\text{mis}} \mid \boldsymbol{Y}^{\text{obs}}, \boldsymbol{G}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{\mu}(\cdot) \sim$?

  (4) Calculate $\tau_{\text{fs}}^{(r)} = N^{-1}\sum_i (Y_i(1) - Y_i(0))$

Optional Homework

Summary



Summary

- Stratification should be considered in design if treatment rates and potential outcomes vary systematically by stratum.
- All 4 approaches can be adapted in a fairly straightforward manner in SRE.
- Pairwise randomized experiment (PRE) is a special case of a SRE.
  - ✓ Within-pair difference plays an important role in Neyman's approach and Regression approach.
  - ✓ Hierarchical prior is preferred in model-based analysis.
- Stratification (SRE & PRE) can increase precision of estimation when the strata are good predictors of the potential outcomes.



Reference

- Reference: [T1] Chapter 9, 10

- ational Psycholog
- ord, F. M. (1967), "A Pa
  Comparisons", Psychologi