Causal Inference Methods and Case Studies

STAT24630

Jingshu Wang

Lecture 7

Topic: Stratified randomized experiments

- Example: field experiment to reduce transphobia
- Stratified randomized experiment
  - Fisher's exact p-value
  - Neyman's repeated sampling approach
  - Regression analysis

Case study: durably reducing transphobia

[Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *2016*.]

- Problem background
  - Durably reducing prejudice is challenging
  - If Individuals engage in active, effortful, processing of brief messages, it can help durably change individuals' attitude
  - **Question:** Can perspective-taking – "imagining the world from another's perspective" – reduce transphobia?

- Field experiment
  - voters are assigned to a canvasser and are randomly assigned to two groups
  - Treatment group (n1 = 912): canvasser asking people to think about a time when they were judged unfairly and were guided to translate that experience to a transgender individual's experience
  - Control group (n0 = 913): canvasser asking people to recycle

Case study: durably reducing transphobia

[Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *2016*.]

- Outcome variable: feeling thermometer towards transgender people (0 − 100)
  - Measured at 5 different time points: baseline, 3 day, 3 week, 6 week, 3 month

- Durable effects at 3-month time point
  - Dropping additional observations with missing covariates n1 = 266 and n0 = 274
  - Neyman: est. = 7.99, s.e. = 2.38, 95% CI = [3.33, 12.65]
  - Simple Regression: est. = 7.99, s.e. = 2.38 (HC2)

  - Regression with covariates (age, gender, party ID, baseline thermometer rating)
    - no interaction: est. = 4.28, s.e. = 1.65 (HC2)
    - with interaction: est. = 4.35, s.e. = 1.63 (HC2)

The project STAR example

(Mosteller. 1997. Bull. Am. Acad. Arts Sci.)

- The student-Teacher Achievement Ratio Project (1985-1989)
  - More than 10,000 students involved with the cost of $12 million
  - Effects of class size in early grade levels
  - 3 arms: Small class (13-17 students), Regular-sized class (22-25 students), Regular class with aid

- Stratified randomization procedure
  - Students and teachers were randomly assigned to the one of the 3 arms
  - A school need to have enough students to allow at least one class per arm to be eligible
  - Once a school is admitted, a decision was made on the number of classes per arm
  - The unit is a teacher in a class, instead of a student to avoid violation of no interference assumption
  - Randomization of units within each school

The project STAR example

(Mosteller. 1997. Bull. Am. Acad. Arts Sci.)

- Understanding the randomization procedure
  - Two randomizations happen in the experiment
    - Randomization of teachers
    - Randomization of students
  - Our causal analysis only relies on the randomization of teachers
    - The treatment effect on a particular teacher in a particular school is comparing the test score of being randomly assigned to a type of class and the test score of being randomly assigned to another type of class
  - The randomization of students helps interpretating our results
    - Treatment effect between two arms can be explained by the classroom size difference instead of the systematic differences of students

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)

- We focus on two arms (regular classes v.s. small classes) and 16 schools that have at least two classes per arm

Stratified randomized experiment

- Basic procedure:
  1. Blocking (Stratification): create groups of similar units based on pre-treatment covariates, let $B_i \in \{1, \cdots, J\}$ be the block indicator
  2. Block (Stratified) randomization: completely randomize treatment assignment within each group
- Blocking can improve the efficiency by minimizing the variance of the potential outcomes within each strata

"Block what you can and randomize what you cannot"

Box, et al. (2005). Statistics for Experimenters. 2nd eds. Wiley

- Assignment probability

$$P(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{X}) = \begin{cases} \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} & \text{if } \sum_{i:B_i=j}^{N} w_i = N_t(j) \text{ for } j = 1, \cdots, J \\ 0 & \text{otherwise} \end{cases}$$

Fisher's exact p-value

- We still focus on the **Sharp null:** $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, N$

- Choice of test statistics:
  Denote sample means for every strata / block

$$\overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(j) = \frac{1}{N_{\mathrm{c}}(j)} \sum_{i:G_i=j} (1 - W_i) \cdot Y_i^{\mathrm{obs}}, \quad \overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(j) = \frac{1}{N_{\mathrm{t}}(j)} \sum_{i:G_i=j} W_i \cdot Y_i^{\mathrm{obs}}$$

  - Weighted combination of group mean differences across blocks

$$T^{\mathrm{dif},\lambda} = \left| \sum_{j=1}^{J} \lambda(j) \cdot \left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(j) - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(j) \right) \right|$$

    - Weights based on relative sample size $\lambda(j) = \frac{N(j)}{N}$
      sample difference is more accurate in larger strata
    - **"inverse-variance-weighting":** assume that per-strata potential outcomes sample variances $S_c^2(j) \equiv S_t^2(j) \equiv S^2$ for all j, then under stratified randomization

$$\mathbb{V}_W\left[\bar{Y}_t^{\mathrm{obs}}(j) - \bar{Y}_c^{\mathrm{obs}}(j) | \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right] = S^2 \left(\frac{1}{N_c(j)} + \frac{1}{N_t(j)} \right)$$

Fisher's exact p-value

- We still focus on the **Sharp null:** $H_0: Y_i(0) \equiv Y_i(1)$ for all $i = 1, \cdots, N$

- Choice of test statistics:
  Denote sample means for every strata / block

$$\overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(j) = \frac{1}{N_{\mathrm{c}}(j)} \sum_{i:G_i=j} (1 - W_i) \cdot Y_i^{\mathrm{obs}}, \quad \overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(j) = \frac{1}{N_{\mathrm{t}}(j)} \sum_{i:G_i=j} W_i \cdot Y_i^{\mathrm{obs}}$$

  - Weighted combination of group mean differences across blocks

$$T^{\mathrm{dif},\lambda} = \left| \sum_{j=1}^{J} \lambda(j) \cdot \left(\overline{Y}_{\mathrm{t}}^{\mathrm{obs}}(j) - \overline{Y}_{\mathrm{c}}^{\mathrm{obs}}(j) \right) \right|$$

    - Weights based on relative sample size $\lambda(j) = \frac{N(j)}{N}$
      sample difference is more accurate in larger strata
    - **"inverse-variance-weighting":** weights

$$\lambda(j) = \frac{1}{\left(\frac{1}{N_c(j)} + \frac{1}{N_t(j)} \right)} \Big/ \sum_{k=1}^{J} \frac{1}{\left(\frac{1}{N_c(k)} + \frac{1}{N_t(k)} \right)}$$

Fisher's exact p-value

- Choice of test statistics:
  - Can we simply use the two-sample mean difference statistic $T = \left|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}\right|$?
    - This is still one test statistic and we will still get valid Fisher's exact p-value if we follow the stratified randomization procedure to generate the reference distribution
    - We may not always get small value of T even wen the sharp null is true
      - Example:
        $Y_i(0) \equiv Y_i(1) = 1$ for strata 1 and $Y_i(0) \equiv Y_i(1) = 2$ for strata 2,
        $N_c(1) = N_t(1) = 5$, $N_c(2) = 15$ and $N_t(2) = 5$
        Then $\bar{Y}_t^{\text{obs}} = 1.5$ and $\bar{Y}_c^{\text{obs}} = 1.75$
      - This is the Simpson's paradox: we have different weights of the strata for the treated group and control group

  - Rank-based statistics
    - Get R_i^{strat} as the within-strata rank of each individual i (definition page 196)
    - Average difference of within-strata ranks between treatment and control

$$\left|\bar{R}_t^{\text{strat}} - \bar{R}_c^{\text{strat}}\right|$$

Neyman's repeated sampling approach

- Target: PATE or SATE $\tau = \sum_j \frac{N(j)}{N} \tau(j)$ where $\tau(j)$ is the PATE or SATE for strata j

- Analysis procedure
  1. Apply Neyman's analysis to each strata / block

$$\hat{\tau}^{\text{dif}}(j) = \overline{Y}_{\text{t}}^{\text{obs}}(j) - \overline{Y}_{\text{c}}^{\text{obs}}(j), \quad \text{and} \quad \hat{\mathbb{V}}^{\text{neyman}}(j) = \frac{s_{\text{c}}(j)^2}{N_{\text{c}}(j)} + \frac{s_{\text{t}}(j)^2}{N_{\text{t}}(j)}$$

  2. Aggregate block-specific estimates and variances

$$\hat{\tau}^{\text{strat}} = \sum_j \frac{N(j)}{N} \hat{\tau}^{\text{dif}}(j), \qquad \widehat{\mathbb{V}}(\hat{\tau}^{\text{strat}}) = \sum_j \left(\frac{N(j)}{N}\right)^2 \widehat{\mathbb{V}}^{\text{neyman}}(j)$$

- Key property:

$$\underbrace{\mathbb{V}(X)}_{\text{total variance}} = \underbrace{\mathbb{E}\{\mathbb{V}(X \mid Y)\}}_{\text{within-block variance}} + \underbrace{\mathbb{V}\{\mathbb{E}(X \mid Y)\}}_{\text{across-block variance}}$$

$$\mathbb{V}_{\text{complete}}\left(\hat{\tau}^{\text{dif}}\right) - \mathbb{V}_{\text{stratified}}(\hat{\tau}^{\text{strat}}) \geq 0$$

The project STAR results

- Fisher's exact p-value

Test statistics	P-value
Weights $\lambda(j) = \frac{N(j)}{N}$	0.034
"inverse-variance-weighting"	0.023
$\left\|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}\right\|$	0.025
Rank-based statistics	0.15

  - P-values for the first 3 are similar as most schools have 4 classes
  - Large p-value for rank-based statistics as # classes too few in most schools

- Neyman's approach
  - $\hat{\tau}^{\text{strat}} = 0.224$, $\widehat{\mathbb{V}}(\hat{\tau}^{\text{strat}}) = 0.092^2$
  - (In correct) if we analyze as if it is a completely randomized experiment
    - $\hat{\tau}^{\text{dif}} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} = 0.224$ can be a biased estimate for τ
    - $\widehat{\mathbb{V}}\left(\hat{\tau}^{\text{dif}}\right) = 0.141^2$ larger standard deviation

Linear regression

- Run separate linear regressions within each strata

- Denote $B_i(j)$ as the indicator variable of whether sample i belong to strata j
- If there are no covariates, equivalently, we can write separate linear regression models into a joint regression model

$$Y_i^{\text{obs}} = \alpha_j + \tau(j)W_i + \varepsilon_i$$

  - The underlying model for the potential outcomes

$$\mathbb{E}[Y_i(w)|\{B_i(j), j = 1, \cdots, J\}] = \alpha_j + \tau(j)w$$

  - Average causal effect for strata j is $\tau(j)$
  - The strata indicators $B_i(j)$ are treated as pre-treatment covariates
  - We need to adjust for the strata indicators as we only have conditional independence

$$\big(\boldsymbol{Y}(0), \boldsymbol{Y}(1)\big) \perp \boldsymbol{W} \mid \boldsymbol{B}(j)$$

  - The homoscedastic error assumption for the joint model is assuming that

$$\mathbb{V}[Y_i(0)|\{B_i(j), j = 1, \cdots, J\}] = \mathbb{V}[Y_i(1)|\{B_i(j), j = 1, \cdots, J\}] = \sigma^2$$