Module 3: Inference for the Average Treatment Effect

Fall 2021

Matthew Blackwell

Gov 2003 (Harvard)

Where are we? Where are we going?

- Fisher: use sharp null to fill in science table + permutation tests.
    - No way to estimate or infer about "average" effects, just the sharp null.
- Neyman: use the difference in means as an **estimator** of the ATE.
    - No assumptions to fill in the potential outcomes.
    - No exact derivation of the randomization distribution.
    - $\rightsquigarrow$ asymptotic approximations.
- What's common: the focus on **randomization** as generating variation in estimators.

Social pressure effect

- Gerber, Green, and Larimer (APSR, 2008)

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______

Social pressure results

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Typical reporting of the Neighbors vs Control effect:

$$\text{estimate} = \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i \approx 8.1$$

$$\text{standard error} = \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_0^2}{n_0}} \approx 0.27$$

$$95\%\ \text{CI} = [\text{est} - 1.96 \cdot SE,\ \text{est} + 1.96 \cdot SE] \approx [7.57, 8.63]$$

- Can this analysis be justified by randomization?

1/ Completely randomized experiments

Estimand of interest

- Common estimand in experiments: **sample average treatment effect**

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

- Neyman/our goals:
    - We want to find an estimator that is **unbiased** for the SATE.
    - But also derive the **sampling variance** of the estimator.
- Properties of the estimators across repeated samples from:
    - the randomization distribution.
    - the randomization distribution + sampling from the population.

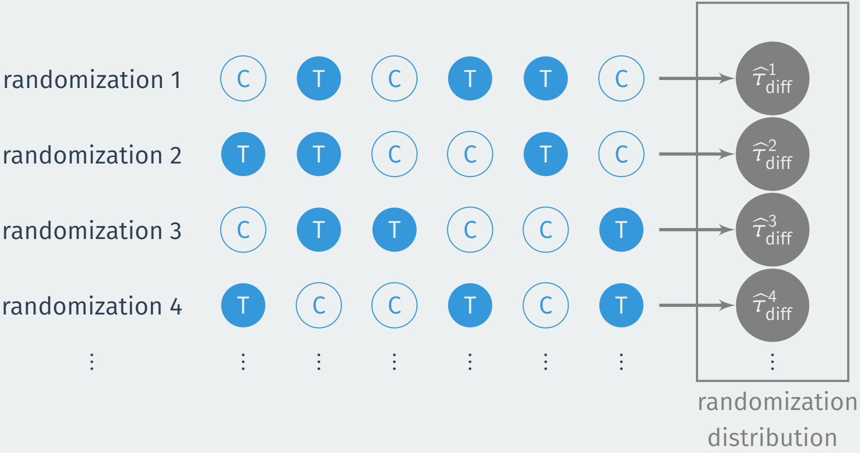
Finite sample results

- Setting: **completely randomized experiment**
  - n units, n_1 treated and n_0 control.
- Natural estimator for the SATE, **difference-in-means**:

$$\widehat{\tau}_{\text{diff}} = \underbrace{\frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i}_{\text{mean among treated}} - \underbrace{\frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i) Y_i}_{\text{mean among control}}$$

- Conditional on the sample, $\widehat{\tau}_{\text{diff}}$ only varies because of D_i

Repeated samples/randomizations



- **Randomization distribution** = sampling distribution of this estimator.

Finite-sample properties

- How does $\widehat{\tau}_{\text{diff}}$ across randomizations?
- Key properties of the randomization distribution we'd like to know:
  - **Unbiasedness**: is mean of the randomization distribution equal to the true SATE?
  - **Sampling variance**: variance of the randomization distribution?
- Use these properties to construct confidence intervals, conduct tests.

Unbiasedness

- In a completely randomized experiment, $\widehat{\tau}_{\text{diff}}$ is unbiased for τ_{fs}
- Let $\mathbf{O} = \{\mathbf{Y}(1), \mathbf{Y}(0)\}$ be the the potential outcomes.

$$
\begin{aligned}
\mathbb{E}_D[\widehat{\tau}_{\text{diff}} \mid \mathbf{O}] &= \frac{1}{n_1}\sum_{i=1}^{n} \mathbb{E}_D[D_i Y_i \mid \mathbf{O}] - \frac{1}{n_0}\sum_{i=1}^{n} \mathbb{E}_D[(1-D_i)Y_i \mid \mathbf{O}] \\
&= \frac{1}{n_1}\sum_{i=1}^{n} \mathbb{E}_D[D_i Y_i(1) \mid \mathbf{O}] - \frac{1}{n_0}\sum_{i=1}^{n} \mathbb{E}_D[(1-D_i)Y_i(0) \mid \mathbf{O}] \\
&= \frac{1}{n_1}\sum_{i=1}^{n} \mathbb{E}_D[D_i \mid \mathbf{O}]Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n} \mathbb{E}_D[(1-D_i) \mid \mathbf{O}]Y_i(0) \\
&= \frac{1}{n_1}\sum_{i=1}^{n} \left(\frac{n_1}{n}\right) Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n} \left(\frac{n_0}{n}\right) Y_i(0) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0) = \tau_{\text{fs}}
\end{aligned}
$$

- Note: number treated/control doesn't matter for unbiasedness!

Finite-sample sampling variance

- Sampling variance of the difference-in-means estimator is:

$$\mathbb{V}_D(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{\tau_i}^2}{n},$$

- S_0^2 and S_1^2 are the in-sample variances of $Y_i(0)$ and $Y_i(1)$, respectively.

$$S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(0) - \overline{Y}(0))^2 \qquad S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(1) - \overline{Y}(1))^2$$

- Here, $\overline{Y}(d) = (1/n)\sum_{i=1}^{n} Y_i(0)$.
- Last term is the in-sample variation of the individual treatment effects:

$$S_{\tau_i}^2 = \frac{1}{n-1}\sum_{i=1}\left(Y_i(1) - Y_i(0) - \tau_{\text{fs}}\right)^2$$

- None of these are directly observable!

Finite-sample sampling variance

$$\mathbb{V}_D(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{\tau_i}^2}{n}$$

- If the treatment effects are constant across units, then $S_{\tau_i}^2 = 0$.
    - $\rightsquigarrow$ in-sample variance is largest when treatment effects are constant.
- Intuition looking at two-unit samples:

	$i=1$	$i=2$	Avg.
$Y_i(0)$	10	-10	0
$Y_i(1)$	10	-10	0
τ_i	0	0	0

	$i=1$	$i=2$	Avg.
$Y_i(0)$	-10	10	0
$Y_i(1)$	10	-10	0
τ_i	20	-20	0

- Both have $\tau_{\text{fs}} = 0$, first has constant effects.
- In first setup, $\widehat{\tau}_{\text{diff}} = 20$ or $\widehat{\tau}_{\text{diff}} = -20$ depending on the randomization.
- In second setup, $\widehat{\tau}_{\text{diff}} = 0$ in either randomization.

Estimating the sampling variance

- We can use sample variances within levels of D_i to estimate S_0^2 and S_1^2:

$$\widehat{\sigma}_d^2 = \frac{1}{n_d - 1} \sum_{i=1}^{n} \mathbb{1}\{D_i = d\} \left(Y_i - \overline{Y}_d\right)^2$$

- Here, $\overline{Y}_0 = (1/n_0) \sum_{i=1}^{n} (1 - D_i) Y_i$ and $\overline{Y}_1 = (1/n_1) \sum_{i=1}^{n} D_i Y_i$.
- But what about $S_{\tau_i}^2$?

$$S_{\tau_i}^2 = \frac{1}{n-1} \sum_{i=1} \Big(\underbrace{Y_i(1) - Y_i(0)}_{???} - \tau_{\text{fs}} \Big)^2$$

- What to do?

Bounding the sampling variance

- First approach: find the worst possible (largest) variance.
- We can rewrite the variance as:

$$\mathbb{V}(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}) = \frac{1}{n}\left(\frac{n_1}{n_0}S_0^2 + \frac{n_0}{n_1}S_1^2 + 2S_{01}\right)$$

- Last term is the **covariance** between potential outcomes:

$$S_{01} = \frac{1}{n-1}\sum_{i=1}^{n}\left\{Y_i(1) - \overline{Y}(1)\right\}\left\{Y_i(0) - \overline{Y}(0)\right\}$$

- We can use the **Cauchy-Schwarz** inequality: $S_{01} \leq S_0 S_1$

$$\mathbb{V}(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}) \leq \frac{1}{n}\left(\frac{n_1}{n_0}S_0^2 + \frac{n_0}{n_1}S_1^2 + 2S_0S_1\right) = \frac{n_0 n_1}{n}\left(\frac{S_0}{n_0} + \frac{S_1}{n_1}\right)^2$$

- Upper bound that is only a function of identified parameters.

Conservative variance estimation

- Usual variance estimator is the Neyman (or robust) estimator:

$$\widehat{\mathbb{V}} = \frac{\widehat{\sigma}_0^2}{n_0} + \frac{\widehat{\sigma}_1^2}{n_1}, \qquad \mathbb{E}\left[\widehat{\mathbb{V}} \mid \mathbf{O}\right] = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$$

- Notice that $\widehat{\mathbb{V}}$ is biased for $\mathbb{V}(\widehat{\tau}_{\text{diff}} \mid \mathbf{O})$, but that bias is always positive.
- Leads to **conservative inferences**:
    - Standard errors, $\sqrt{\widehat{\mathbb{V}}}$ will be at least as big as they should be.
    - Confidence intervals using $\sqrt{\widehat{\mathbb{V}}}$ will be at least wide as they should be.
    - Type I error rates will still be correct, power will be lower.
    - Both will be exactly right if treatment effects are constant.

Inference in the Neyman approach

- If n is large, CLT will imply $\widehat{\tau}_{\text{diff}}$ will be approximately normal.
- Formulate confidence intervals in the usual way:

$$\text{CI}^{95}(\tau_{\text{fs}}) = \left(\widehat{\tau}_{\text{diff}} - 1.96 \cdot \widehat{\mathbb{V}}^{1/2},\ \widehat{\tau}_{\text{diff}} + 1.96 \cdot \widehat{\mathbb{V}}^{1/2}\right)$$

- Testing very similar to standard normal-approximation tests:

$$H_0 : \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0) = 0 \qquad T = \frac{\widehat{\tau}_{\text{diff}}}{\sqrt{\widehat{\mathbb{V}}}} \overset{a}{\sim} N(0,1)$$

- Contains more situations than the sharp null, but...
  - Fisher tests might not be well-powered against $\tau_{\text{fs}} = 0$ alternatives.
- Can improve approximations using t-distribution.
  - Works since $\widehat{\mathbb{V}}$ will be approximately χ^2_{n-1} in large samples.

Population estimands

- What if we want to make inference to a (super)population?
  - n units are a **simple random sample** from the population.
  - $Y_i(1), Y_i(0)$ are now random variables (induced by sampling)
- New goal: inference for the PATE, $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$.
  - Average of the SATEs across repeated samples: $\text{PATE} = \mathbb{E}[\text{SATE}]$.
- Difference-in-means is **unbiased** across repeated samples:

$$\mathbb{E}[\widehat{\tau}_{\text{diff}}] = \underbrace{\mathbb{E}\{\mathbb{E}_D[\widehat{\tau}_{\text{diff}} \mid \mathbf{O}]\}}_{\text{iterated expectations}} = \underbrace{\mathbb{E}[\tau_{\text{fs}}]}_{\text{SATE unbiasedness}} = \tau$$

Population sampling variance

- What about the sampling variance of $\widehat{\tau}_{\text{diff}}$ when estimating the PATE?
    - Variation comes from random sample **and** random assignment.
- It turns out that the sampling variance of the estimator is simply:

$$\mathbb{V}(\widehat{\tau}_{\text{diff}}) = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} = \frac{\mathbb{V}[Y_i(0)]}{n_0} + \frac{\mathbb{V}[Y_i(1)]}{n_1}$$

    - Here, σ_0^2 and σ_1^2 are the population-level variances of $Y_i(1)$ and $Y_i(0)$.
- The variance of τ_i term drops out $\rightsquigarrow$ higher variance for PATE than SATE.

Estimating pop. sampling variance

$$\mathbb{V}(\widehat{\tau}_{\text{diff}}) = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1},$$

- Notice that the Neyman estimator $\widehat{\mathbb{V}}$ is now unbiased for $\mathbb{V}(\widehat{\tau}_{\text{diff}})$:

$$\widehat{\mathbb{V}} = \frac{\widehat{\sigma}_0^2}{n_0} + \frac{\widehat{\sigma}_1^2}{n_1}$$

- Two interpretations of $\widehat{\mathbb{V}}$:
    1. Unbiased estimator for sampling variance of the traditional estimator of the PATE
    2. Conservative estimator for the sampling variance of the traditional estimator of the SATE

2/ Block randomized experiments

Block randomized experiments

- Basic idea: run completely randomized experiments within strata defined by covariates.
- Main motivation: **more efficient** than standard design (ie, lower SEs)
- George Box: “Block what you can and randomize what you cannot.”
- We will compare variance of blocked designs to complete randomization.
  - Some confusion in the literature: can blocking hurt?
  - Care needed: comparison depends on sample assumptions (Pashley & Miratrix, 2021, JEBS)

Simple two block example

- GOTV mailer experiment:
  - We have n households with registered voters.
  - Complete randomization: choose n_1 households to get mailers.
  - Outcome, Y_i: turnout in election.
- What if we have data from the voter file: **previous turnout**.
  - Create blocks: $V_i = 1$ if voted in last election, $V_i = 0$ otherwise.
  - n_{v} is the number of previous voters,
  - $n_{\text{nv}} = n - n_{\text{nv}}$ is the number of previous nonvoters.
- SATEs within blocks defined by V_i:

$$\tau_{\text{v, fs}} = \frac{1}{n_{\text{v}}} \sum_{i:V_i=1} \{Y_i(1) - Y_i(0)\} \qquad \tau_{\text{nv, fs}} = \frac{1}{n_{\text{nv}}} \sum_{i:V_i=0} \{Y_i(1) - Y_i(0)\}$$

- Iterated expectations gives us:

$$\tau_{\text{fs}} = \underbrace{\left(\frac{n_{\text{v}}}{n_{\text{v}} + n_{\text{nv}}}\right)}_{\text{fraction voters}} \tau_{\text{v, fs}} + \underbrace{\left(\frac{n_{\text{nv}}}{n_{\text{v}} + n_{\text{nv}}}\right)}_{\text{fraction nonvoters}} \tau_{\text{nv, fs}}$$

Block randomized design

- **Block/stratified randomized experiment**:
    - Completely randomized experiment in each block.
    - Choose $n_{1,\text{v}}$ voters to be treated, $n_{0,\text{v}} = n_{\text{v}} - n_{1,\text{v}}$ control.
    - Choose $n_{1,\text{nv}}$ nonvoters to be treated, $n_{0,\text{nv}} = n_{\text{nv}} - n_{1,\text{nv}}$ control.
- Probability of treatment in each group called the **propensity score**:
    - Prob. of treatment for voters: $\mathbb{P}(D_i = 1 \mid V_i = 1) = p_{\text{v}} = n_{1,\text{v}}/n_{\text{v}}$
    - Prob. of treatment for nonvoters: $\mathbb{P}(D_i = 1 \mid V_i = 0) = p_{\text{nv}} = n_{1,\text{nv}}/n_{\text{nv}}$
- Blocking ensures balance across blocks:
    - When $p_{\text{v}} = p_{\text{nv}}$, distribution of treatment is exactly the same in each block.
    - With complete randomization, treatment might be very imbalanced across V_i.
    - No possibility of “chance” imbalances skewing the estimates.

Estimators in blocked designs

- Within-strata difference in means:

$$\widehat{\tau}_{\mathrm{v}} = \overline{Y}_{1,\mathrm{v}} - \overline{Y}_{0,\mathrm{v}} = \frac{1}{n_{1,\mathrm{v}}} \sum_{i:V_i=1} D_i Y_i - \frac{1}{n_{0,\mathrm{v}}} \sum_{i:V_i=1} (1 - D_i) Y_i$$

$$\widehat{\tau}_{\mathrm{nv}} = \overline{Y}_{1,\mathrm{nv}} - \overline{Y}_{0,\mathrm{nv}} = \frac{1}{n_{1,\mathrm{nv}}} \sum_{i:V_i=0} D_i Y_i - \frac{1}{n_{0,\mathrm{nv}}} \sum_{i:V_i=0} (1 - D_i) Y_i$$

- Unbiased for the within-strata SATEs: $\mathbb{E}[\widehat{\tau}_{\mathrm{v}} \mid \mathbf{O}] = \tau_{\mathrm{v}}$
- $\rightsquigarrow$ unbiased estimator for the overall SATE:

$$\widehat{\tau}_{\mathrm{b}} = \left(\frac{n_{\mathrm{v}}}{n}\right) \widehat{\tau}_{\mathrm{v}} + \left(\frac{n_{\mathrm{nv}}}{n}\right) \widehat{\tau}_{\mathrm{nv}}$$

  - Equivalent to the regular difference in means if $p_{\mathrm{v}} = p_{\mathrm{nv}} = 1/2$.
  - Otherwise, standard $\widehat{\tau}_{\mathrm{diff}}$ under block design will be **biased**.

Sampling variance of blocking estimator

- Each block is a completely randomized experiment so we have:

$$\mathbb{V}(\widehat{\tau}_{\text{v}} \mid \mathbf{O}) = \frac{S^2_{1,\text{v}}}{n_{1,\text{v}}} + \frac{S^2_{0,\text{v}}}{n_{0,\text{v}}} - \frac{S^2_{\tau_i,\text{v}}}{n_{\text{v}}}$$

  - $S^2_{d,\text{v}}$ are the within-block sample variances of the potential outcomes

- Finite sample variance of the blocked estimator:

$$\mathbb{V}(\widehat{\tau}_b \mid \mathbf{O}) = \left(\frac{n_{\text{v}}}{n}\right)^2 \mathbb{V}(\widehat{\tau}_v \mid \mathbf{O}) + \left(\frac{n_{\text{nv}}}{n}\right)^2 \mathbb{V}(\widehat{\tau}_{\text{nv}} \mid \mathbf{O})$$

- Use the conservative variance estimators from each strata:

$$\widehat{\mathbb{V}}_b = \left(\frac{n_{\text{v}}}{n}\right)^2 \left(\frac{\widehat{\sigma}^2_{1,\text{v}}}{n_{1,\text{v}}} + \frac{\widehat{\sigma}^2_{0,\text{v}}}{n_{0,\text{v}}}\right) + \left(\frac{n_{\text{nv}}}{n}\right)^2 \left(\frac{\widehat{\sigma}^2_{1,\text{nv}}}{n_{1,\text{nv}}} + \frac{\widehat{\sigma}^2_{0,\text{nv}}}{n_{0,\text{nv}}}\right)$$

  - $\widehat{\sigma}^2_{d,\text{v}}$ are the within-strata **observed outcome variances**

General blocking notation

- Blocks, $j \in \{1, \dots, J\}$.
  - Block indicator $B_i = j$ if i is in block j.
  - Sizes: $n_j > 2$ and proportions $w_j = n_j / n$.
  - Number treated in each block: $n_{1,j}$ and $n_{0,j} = n_j - n_{1,j}$
- Within-block estimators:

$$\widehat{\tau}_j = \frac{1}{n_{1,j}} \sum_{i:B_i=j} D_i Y_i - \frac{1}{n_{0,j}} \sum_{i:B_i=j} (1 - D_i) Y_i, \qquad \hat{\mathbb{V}}(\widehat{\tau}_j) = \frac{\widehat{\sigma}^2_{1,j}}{n_{1,j}} + \frac{\widehat{\sigma}^2_{0,j}}{n_{0,j}}$$

- Aggregate blocking estimators:

$$\widehat{\tau}_b = \sum_{j=1}^{J} w_j \widehat{\tau}_j, \qquad \hat{\mathbb{V}}(\widehat{\tau}_b) = \sum_{j=1}^{J} w_j^2 \hat{\mathbb{V}}(\widehat{\tau}_j)$$

Efficiency of blocking

- Efficiency of block versus CR depends on the sampling scheme.
    - Usually blocking will be more efficient/lower variance, but not always.
- Finite sample difference in sampling variances:

$$\mathbb{V}(\widehat{\tau}_{CR} \mid \mathbf{O}) - \mathbb{V}(\widehat{\tau}_b \mid \mathbf{O}) = \frac{1}{n-1}\left[B - W\right]$$

- Measures of between- and within-block variation:

$$B = \sum_{j=1}^{J}\left(\frac{n_j}{n}\right)\left\{\overline{Y}_j(1) + \overline{Y}_j(0) - (\overline{Y}(1) + \overline{Y}(0))\right\}^2$$

$$W = \sum_{j=1}^{J}\frac{n_j}{n}\frac{n-n_j}{n}\mathbb{V}(\widehat{\tau}_k \mid \mathbf{O})$$

- Difference can be positive or negative (blocking can hurt or help)
    - **Blocking is better when outcomes vary a lot across blocks, not much within blocks** (blocks are predictive of outcome, so usually the case)
    - Blocking always more efficient for PATE under stratified sampling

How to block

- Discrete covariates $\rightsquigarrow$ blocks by unique combinations.
- Alternative: create blocks by creating homogeneous groups in $\mathbf{X}$.
    - Choose distance metric such as Mahalanobis distance:

$$M(\mathbf{X}_i, \mathbf{X}_k) = \sqrt{(\mathbf{X}_i - \mathbf{X}_k)\widehat{\mathbb{V}}(\mathbf{X})^{-1}(\mathbf{X}_i - \mathbf{X}_k)}$$

    - Difficult/impossible to find optimal blocks in general, but “greedy” algorithms exist.
    - Possible to get optimal blocks with **pair matching** ($J = n/2$).

Matched pair design

- Keep blocking for efficiency until each block is size 2.
- **Matched pair design**:
    - Create $J = n/2$ pairs of similar units with outcomes (Y_{1j}, Y_{2j})
    - Random assignment:
        - $W_j = 1$ if first unit is treated
        - $W_j = -1$ if second unit is treated
- Unbiased difference in means estimator:

$$\widehat{\tau}_p = \frac{1}{J} \sum_{j=1}^{J} W_j (Y_{1j} - Y_{2j})$$

- Within-pair variance estimator not feasible (why?)
- Across-pair variance estimator (conservative for SATE):

$$\widehat{\mathbb{V}}(\widehat{\tau}_p) = \frac{1}{J(J-1)} \sum_{j=1}^{J} \{W_j (Y_{1j} - Y_{2j} - \widehat{\tau}_p)\}^2$$