Problem Set 3: Inference for the ATE

GOV 2003

Due at 11:59 pm (ET) on Sept 29, 2021

Instruction

Before you begin, please read the following instructions **carefully**:

- **No late submission are allowed** without a prior approval from the instructors.
- **All answers should be typed up**. We recommend the use of `Rmarkdown`. An `Rmarkdown` template for this problem set is provided. Answers to analytical solutions should also be typed up.
- **A PDF copy of your answer** including your computer code should be uploaded to `Gradescope` before the deadline. **Do not submit the markdown file itself**.
- This problem set includes a bonus question for extra credit. No deduction in the total points will be made from this question. Note that the maximum points of this problem set is 15 points. That is, if the student receives 3 points from the bonus question and 14 points from the other questions, the total points will be 15 points.

Introduction

This problem set consists of two parts. In `Part A`, we will investigate Neyman's approach to completely randomized experiments. Specifically, we will analytically drive the sampling variance of the difference-in-means estimator. This is a guided practice on how proofs work in causal inference. In `Part B`, we will investigate block randomized experiments using a (modified) data from Tennessee's Student Teacher Achievement Ratio (STAR) project by Achilles et al. (2008).

Part A: Derivation of sampling variance

Setup

In `Part A`, we will derive the sampling variance of the difference-in-means estimator under completely randomized experiments. We will first derive the finite-sample sampling variance as written in Question 4, and then use this to drive the population sampling variance in Question 5. We adopt the same setting as the one used in the lecture.

- Completly randomized experiment: n units, n_1 treated and n_0 control.
- D_i: a binary treatment of unit i
- Y_i: an outcome of unit i
- $\mathbf{O} = \{\mathbf{Y}(1), \mathbf{Y}(0)\}$: the potential outcomes
- $\widehat{\tau}_{\text{diff}} = \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i$: difference-in-means estimator

Question 1 (2 pts)

When doing proofs that involve variances and covariances, it is helpful to work with mean-zero random variables since for such a variable, B, we know that $\mathbb{V}[B] = \mathbb{E}[B^2]$. For the variance calculations, then, it is useful to work with a centered treatment variable $W_i = \frac{n}{n_1}D_i - 1$. Let's first derive some properties of this centered variable that we'll use in the proof. Show

$$\begin{aligned}
\mathbb{E}\left[W_i \mid \mathbf{O}\right] &= 0 \\
\mathbb{E}\left[W_i^2 \mid \mathbf{O}\right] &= \frac{n_0}{n_1} \\
\mathbb{E}\left[W_i W_j \mid \mathbf{O}\right] &= -\frac{n_0}{n_1(n-1)} \qquad \text{for any } i \neq j.
\end{aligned}$$

Hint: Remember that for a binary variable, $D_i^2 = D_i$. Also, you may use the fact that

$$\mathbb{E}[D_i D_j \mid \mathbf{O}] = \frac{n_1}{n} \cdot \frac{n_1 - 1}{n-1}$$

Answer 1

Observe that $\mathbb{E}\left[D_i \mid \mathbf{O}\right] = \frac{n_1}{n}$.

$$\begin{aligned}
\mathbb{E}\left[W_i \mid \mathbf{O}\right] &= \mathbb{E}\left[\frac{n}{n_1}D_i - 1 \mid \mathbf{O}\right] \\
&= \frac{n}{n_1}\mathbb{E}\left[D_i \mid \mathbf{O}\right] - 1 \\
&= \frac{n}{n_1} \cdot \frac{n_1}{n} - 1 \\
&= 0 \quad \square
\end{aligned}$$

Observe that $\mathbb{E}\left[D_i^2 \mid \mathbf{O}\right] = \mathbb{E}\left[D_i \mid \mathbf{O}\right] = \frac{n_1}{n}$ since $D_i^2 = D_i$.

$$\begin{aligned}
\mathbb{E}\left[W_i^2 \mid \mathbf{O}\right] &= \mathbb{E}\left[\frac{n^2}{n_1^2} D_i^2 - 2\frac{n}{n_1} D_i + 1 \mid \mathbf{O}\right] \\
&= \frac{n^2}{n_1^2}\mathbb{E}\left[D_i^2 \mid \mathbf{O}\right] - 2\frac{n}{n_1}\mathbb{E}\left[D_i \mid \mathbf{O}\right] + 1 \\
&= \frac{n^2}{n_1^2} \cdot \frac{n_1}{n} - \frac{2n}{n_2} \cdot \frac{n_1}{n} + 1 \\
&= \frac{n}{n_1} - 2 + 1 \\
&= \frac{n - n_1}{n_1} \\
&= \frac{n_0}{n_1} \quad \square
\end{aligned}$$

Observe that for any $i \neq j$, $\mathbb{E}[D_i D_j \mid \mathbf{O}] = \Pr(D_i = 1, D_j = 1) = \Pr(D_i = 1)\Pr(D_i = 1 \mid D_j = 1) = \frac{n_1}{n} \cdot \frac{n_1 - 1}{n - 1}$ (proof of **hint**).

$$\begin{aligned}
\mathbb{E}\left[W_i W_j \mid \mathbf{O}\right] &= \mathbb{E}\left[\left(\frac{n}{n_1} D_i - 1\right)\left(\frac{n}{n_1} D_j - 1\right) \mid \mathbf{O}\right] \\
&= \frac{n^2}{n_1^2}\mathbb{E}[D_i D_j \mid \mathbf{O}] - \frac{n}{n_1}\left(\mathbb{E}[D_i \mid \mathbf{O}] + \mathbb{E}[D_j \mid \mathbf{O}]\right) + 1 \\
&= \frac{n^2}{n_1^2} \cdot \frac{n_1}{n} \cdot \frac{n_1 - 1}{n - 1} - \frac{n}{n_1} \cdot \frac{2n_1}{n} + 1 \\
&= \frac{n(n_1 - 1)}{n_1(n - 1)} - 1 \\
&= \frac{n n_1 - n - n_1 n + n_1}{n_1(n - 1)} \\
&= -\frac{n_0}{n_1(n - 1)} \quad \square
\end{aligned}$$

Question 2 (2 pts)

Define $X_i = Y_i(1) + \frac{n_1}{n_0} Y_i(0)$. Show that we can write the estimation error of the difference in means estimator as

$$\widehat{\tau}_{\text{diff}} - \tau_{\text{fs}} = \frac{1}{n}\sum_{i=1}^{n} W_i X_i$$

Then, use this to show that we can simplify the sampling variance expression as follows,

$$\mathbb{V}\left(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right) = \frac{1}{n^2}\mathbb{E}\left\{\left(\sum_{i=1}^{n} W_i X_i\right)^2 \mid \mathbf{O}\right\}.$$

Hint: It may be helpful to rewrite the difference in means as:

$$\widehat{\tau}_{\text{diff}} = \frac{1}{n}\sum_{i=1}\left(\frac{n}{n_1} D_i - \frac{n}{n_0}(1 - D_i)\right) Y_i$$

Answer 2

$$
\begin{aligned}
\widehat{\tau}_{\text{diff}} &= \frac{1}{n_1}\sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{n}{n_1}D_i Y_i - \frac{n}{n_0}(1-D_i)Y_i\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{n}{n_1}D_i Y_i(1) - \frac{n}{n_0}(1-D_i)Y_i(0)\right\} && \text{[by consistency]} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n}{n_1}\cdot\frac{n_1}{n}(W_i+1)Y_i(1) - \frac{n}{n_0}\left\{1-\frac{n_1}{n}(W_i+1)\right\}Y_i(0)\right] && \text{[plug-in } D_i\text{]} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{(W_i+1)Y_i(1) - \left(1-\frac{n_1}{n_0}W_i\right)Y_i(0)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1)-Y_i(0)\right) + \frac{1}{n}\sum_{i=1}^{n} W_i\left(Y_i(1)+\frac{n_1}{n_0}Y_i(0)\right) \\
&= \tau_{\text{fs}} + \frac{1}{n}\sum_{i=1}^{n} W_i X_i
\end{aligned}
$$

Hence

$$
\widehat{\tau}_{\text{diff}} - \tau_{\text{fs}} = \frac{1}{n}\sum_{i=1}^{n} W_i X_i
$$

$$
\begin{aligned}
\mathbb{V}\left(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right) &= \mathbb{E}\left[\left(\widehat{\tau}_{\text{diff}} - \mathbb{E}\left[\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right]\right)^2 \mid \mathbf{O}\right] \\
&= \mathbb{E}\left[\left(\widehat{\tau}_{\text{diff}} - \tau_{\text{fs}}\right)^2 \mid \mathbf{O}\right] && \text{[}\widehat{\tau}_{\text{diff}} \text{ unbiased for } \tau_{\text{fs}}\text{]} \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i\right)^2 \mid \mathbf{O}\right] && \text{[using the result above]} \\
&= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} W_i X_i\right)^2 \mid \mathbf{O}\right] \quad \square
\end{aligned}
$$

[Bonus] Question 3 (2 pts)

Use the above results to show

$$
\mathbb{V}\left(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right) = \frac{n_0}{n(n-1)n_1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2
$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Hints:

- For any numbers $\{a_1, \ldots, a_n\}$, $\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} a_i a_j$.

- Remember that for any random variable,

$$\sum_{i=1}^{n}\left(Z_i-\overline{Z}\right)^2=\sum_{i=1}^{n}Z_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}Z_iZ_j\right)$$

Answer 3

Observe the following (proof of the second **hint**):

$$\begin{aligned}
\sum_{i=1}^{n}\left(Z_i-\overline{Z}\right)^2 &= \sum_{i=1}^{n}\left(Z_i^2-2\overline{Z}Z_i+\overline{Z}^2\right)\\
&= \sum_{i=1}^{n}Z_i^2-2\overline{Z}\sum_{i=1}^{n}Z_i+n\overline{Z}^2\\
&= \sum_{i=1}^{n}Z_i^2-2\overline{Z}n\overline{Z}+n\overline{Z}^2\\
&= \sum_{i=1}^{n}Z_i^2-n\overline{Z}^2\\
&= \sum_{i=1}^{n}Z_i^2-\frac{1}{n}\sum_{i=1}^{n}\overline{Z}^2\\
&= \sum_{i=1}^{n}Z_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}Z_iZ_j\right)\\
&= \sum_{i=1}^{n}Z_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}Z_i\right)^2
\end{aligned}$$

From the result of Question 2,

$$
\begin{aligned}
\mathbb{V}\left(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right) &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} W_i X_i\right)^2 \mid \mathbf{O}\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n} W_i^2 X_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}(W_i X_i)(W_j X_j) \mid \mathbf{O}\right] && \text{[using the first hint]} \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}\left[W_i^2 \mid \mathbf{O}\right] X_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\left[W_i W_j \mid \mathbf{O}\right] X_i X_j\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\frac{n_0}{n_1} X_i^2 - \sum_{i=1}^{n}\sum_{j\neq i}\frac{n_0}{n_1(n-1)} X_i X_j\right) && \text{[plug-in the results of Q1]} \\
&= \frac{1}{n^2}\left(\frac{n_0}{n_1}\sum_{i=1}^{n} X_i^2 - \frac{n_0}{n_1(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} X_i X_j\right) \\
&= \frac{n_0}{n(n-1)n_1}\left(\frac{n-1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} X_i X_j\right) \\
&= \frac{n_0}{n(n-1)n_1}\left[\frac{n-1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left\{\left(\sum_{i=1}^{n} X_i\right)^2 - \sum_{i=1}^{n} X_i^2\right\}\right] && \text{[using the first hint]} \\
&= \frac{n_0}{n(n-1)n_1}\left\{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right\} \\
&= \frac{n_0}{n(n-1)n_1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \square && \text{[using the second hint]}
\end{aligned}
$$

Question 4 (3 pts)

Substitute the potential outcome expressions for X_i to arrive at the desired result,

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

Hints:

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

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

  where $\overline{Y}(d) = \frac{1}{n}\sum_{i=1}^{n} Y_i(0)$.
- The last term, S_{01}, is the covariance between potential outcomes:

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

- Recall that X_i is defined in terms of potential outcomes ($Y_i(1)$ and $Y_i(0)$), and so does $\overline{X}$.

Answer 4

Observe the following result for $\overline{X}$:

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1) + \frac{n_1}{n_0}Y_i(0)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) + \frac{n_1}{n_0}\cdot\frac{1}{n}\sum_{i=1}^{n} Y_i(0) \\
&= \overline{Y}(1) + \frac{n_1}{n_0}\overline{Y}(0)
\end{aligned}$$

Now, substitute the potential outcome expressions for X_i from the result of Question 3 as below:

$$\begin{aligned}
&\mathbb{V}\left(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right) \\
&= \frac{n_0}{n(n-1)n_1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\
&= \frac{1}{n}\left\{\frac{n_0}{(n-1)n_1}\sum_{i=1}^{n}\left(Y_i(1) + \frac{n_1}{n_0}Y_i(0) - \overline{Y}(1) - \frac{n_1}{n_0}\overline{Y}(0)\right)^2\right\} \\
&= \frac{1}{n}\left[\frac{n_0}{(n-1)n_1}\sum_{i=1}^{n}\left\{Y_i(1) - \overline{Y}(1) + \frac{n_1}{n_0}\left(Y_i(0) - \overline{Y}(0)\right)\right\}^2\right] \\
&= \frac{1}{n}\left[\frac{n_0}{(n-1)n_1}\sum_{i=1}^{n}\left\{\left(Y_i(1) - \overline{Y}(1)\right)^2 + \frac{n_1^2}{n_0^2}\left(Y_i(0) - \overline{Y}(0)\right)^2 + 2\frac{n_1}{n_0}\left(Y_i(1) - \overline{Y}(1)\right)\left(Y_i(0) - \overline{Y}(0)\right)\right\}\right] \\
&= \frac{1}{n}\left\{\frac{n_0}{n_1}\cdot\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i(1) - \overline{Y}(1)\right)^2 + \frac{n_1}{n_0}\cdot\frac{1}{n-1}\left(Y_i(0) - \overline{Y}(0)\right)^2 + \frac{2}{n-1}\sum_{i=1}^{n}\left(Y_i(1) - \overline{Y}(1)\right)\left(Y_i(0) - \overline{Y}(0)\right)\right\} \\
&= \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right) \quad \square
\end{aligned}$$

Question 5 (3 pts)

Finally, derive the population sampling variance using the result $\sigma_1^2 = \mathbb{E}(S_1^2)$ and $\sigma_0^2 = \mathbb{E}(S_0^2)$.

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

Hint: By the law of total variance, $\mathbb{V}(\widehat{\tau}_{\text{diff}}) = \mathbb{E}\left[\mathbb{V}\left(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right)\right] + \mathbb{V}\left(\mathbb{E}\left[\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right]\right)$.

Answer 5

By the law of total variance and using the result from Question 4 and unbiasedness of $\widehat{\tau}_{\text{diff}}$.

$$\begin{aligned}
\mathbb{V}(\widehat{\tau}_{\text{diff}}) &= \mathbb{E}\left[\mathbb{V}\left(\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right)\right] + \mathbb{V}\left(\mathbb{E}\left[\widehat{\tau}_{\text{diff}} \mid \mathbf{O}\right]\right) \\
&= \mathbb{E}\left[\frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right)\right] + \mathbb{V}(\tau_{\text{fs}})
\end{aligned}$$

Recall that sample variance is unbiased for population variance. Hence $\sigma_1^2 = \mathbb{E}(S_1^2)$ and $\sigma_0^2 = \mathbb{E}(S_0^2)$ where $\sigma_1^2 = \mathbb{V}(Y_i(1))$ and $\sigma_0^2 = \mathbb{V}(Y_i(0))$. Using this result, the first term becomes,

$$\mathbb{E}\left[\frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right)\right] = \frac{1}{n}\left(\frac{n_0}{n_1}\sigma_1^2 + \frac{n_1}{n_0}\sigma_0^2 + 2\sigma_{01}\right)$$

where $\sigma_{01} = \mathrm{Cov}(Y_i(1), Y_i(0))$.

The second term becomes,

$$\begin{aligned}
\mathbb{V}(\tau_{\text{fs}}) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(Y_i(1) - Y_i(0)) \\
&= \frac{1}{n^2}\cdot n(\mathbb{V}(Y_i(1)) + \mathbb{V}(Y_i(0)) - \mathrm{Cov}(Y_i(1), Y_i(0))) \\
&= \frac{1}{n}(\sigma_1^2 + \sigma_0^2 - 2\sigma_{01})
\end{aligned}$$

Note that we used the fact that $Y_i(1) - Y_i(0)$ is independent from $Y_j(1) - Y_j(0)$ and thus $\mathrm{Cov}(Y_i(1) - Y_i(0), Y_j(1) - Y_j(0)) = 0$ in the second line.

Combining these two terms,

$$\begin{aligned}
\mathbb{V}(\widehat{\tau}_{\text{diff}}) &= \frac{1}{n}\left(\frac{n_0}{n_1}\sigma_1^2 + \frac{n_1}{n_0}\sigma_0^2 + 2\sigma_{01}\right) + \frac{1}{n}(\sigma_1^2 + \sigma_0^2 - 2\sigma_{01}) \\
&= \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \quad \square
\end{aligned}$$

Part B: Block randomized experiments

Setup

In `Part B`, we will investigate block randomized experiments using a modified data (`STAR.RData`) from Tennessee's Student Teacher Achievement Ratio (STAR) project by Achilles et al. (2008). Please read Ch.9.2 of Imbens and Rubin (2015) for the illustration of the project. Be aware that we are using a modified version of the data, and thus the sample size will be different from the textbook. Here, we will estimate the average treatment effect of kindergarten class size on SAT math grade (class-average) under the block randomized design. For simplicity, please make an inference on class-level (i) throughout the questions.

Let D_i denote the size of class $i \in \{1, \dots N\}$ where N is the total number of classes — $D_i = 1$ if the class size is small, and $D_i = 0$ if the class size is regular. Let Y_i denote the average SAT math grade of the students in class i. Here, we conducted a completely randomized experiment in each school (i.e, kindergarten) $j \in \{1, \dots, J\}$. Let B_i denote a block indicator ($B_i = j$ if class i is in school j), and n_j the number of classes in each school j. Note that the size of treated group and control group may different across school j's. That is, $\sum_{i:B_i=j} D_i/n_j$ may differ across different j's.

The table below describes the variables we will use.

Name	Description
stdntid	Student id
gkschid	Kindergarten School id
gktchid	Teacher id
gkclasstype	Class type, SMALL/REGULAR
g3tmathss	SAT math score

Question 6 (5 pts; 1pt for each)

(a) Write mathematical expressions of difference-in-means estimator under blocked design and sampling variance estimator of your difference-in-means estimator.

Hint: See lecture slides p.7.

(b) Compute the outcome variable (average SAT math grade for each class i) using the data. Show the outcome value for each of `gktchid` $\in \{15917103, 16118306, 16922902\}$.

(c) Calculate the point estimate of the population average treatment effect using the estimator from (a).

(d) Calculate the point estimate of the sampling variance of your estimator from (a).

(e) Now, (incorrectly) assume that the data arose from a completely randomized experiment without block randomization, again using class (i) as the units. Calculate the point estimate of the population average treatment effect and that of sampling variance under this new setting. Compare those with the estimates of (c) and (d), and briefly discuss the result.

Answer 6

(a) Let $n_{1,j}$ denote the number of small classes in school j, and $n_{0,j} = n_j - n_{1,j}$ denote that of regular classes. Within each school, we can compute the following:

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

where $\widehat{\sigma}^2_{d,j} = \frac{1}{n_{d,j}-1} \sum_{i:B_i=j} \mathbb{I}\{D_i = d\}(Y_i - \overline{Y}_{d,j})^2$ are the within-strata observed outcome variances. Here, $\overline{Y}_{0,j} = \frac{1}{n_{0,j}} \sum_{i:B_i=j} (1 - D_i) Y_i$ and $\overline{Y}_{1,j} = \frac{1}{n_{1,j}} \sum_{i:B_i=j} D_i Y_i$. Then we can aggregate these estimators using the weights $w_j = \frac{n_j}{N}$ as follows:

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

(b)

```
load("STAR.RData")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
class_data <- data %>%
  group_by(gktchid) %>%
  summarize(avg_math = mean(g3tmathss))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
class_data <- data %>%
  left_join(class_data, by = "gktchid") %>%
  select(-stdntid, -g3tmathss) %>%
  distinct()
class_data %>%
  filter(gktchid %in% c(15917103, 16118306, 16922902)) %>%
  select(avg_math)
```

```
##   avg_math
## 1 644.3636
## 2 619.9000
## 3 626.3636
```

(c)

```r
N <- nrow(class_data)
within_block_est <- class_data %>%
  mutate(D = ifelse(gkclasstype == "SMALL CLASS", 1, 0)) %>%
  group_by(gkschid) %>%
  summarize(n1j = sum(D),
            n0j = sum(1-D),
            Ybar1j = sum(D*avg_math)/n1j,
            Ybar0j = sum((1-D)*avg_math)/n0j,
            sigmahat2_1j = sum(D*((avg_math - Ybar1j)^2))/(n1j-1),
            sigmahat2_0j = sum((1-D)*((avg_math - Ybar0j)^2))/(n0j-1)) %>%
  mutate(tauhat_j = Ybar1j - Ybar0j,
         Vhat_j = sigmahat2_1j/n1j + sigmahat2_0j/n0j,
         w_j = (n1j + n0j)/N)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
tauhat_b <- sum(within_block_est$tauhat_j * within_block_est$w_j)
tauhat_b
```

```
## [1] 0.3092579
```

(d)

```r
Vhat_b <- sum(within_block_est$Vhat_j * (within_block_est$w_j)^2)
Vhat_b
```

```
## [1] 7.048841
```

(e)

```r
no_block_est <- class_data %>%
  mutate(D = ifelse(gkclasstype == "SMALL CLASS", 1, 0)) %>%
  summarise(n1 = sum(D),
         n0 = sum(1-D),
         Ybar1 = sum(D*avg_math)/n1,
         Ybar0 = sum((1-D)*avg_math)/n0,
         sigmahat2_1 = sum(D*((avg_math - Ybar1)^2))/(n1-1),
         sigmahat2_0 = sum((1-D)*((avg_math - Ybar0)^2))/(n0-1)) %>%
  mutate(tauhat_nob = Ybar1 - Ybar0,
         Vhat_nob = sigmahat2_1/n1 + sigmahat2_0/n0)
no_block_est$tauhat_nob
```

```
## [1] 2.073128
```

```r
no_block_est$Vhat_nob
```

```
## [1] 15.40513
```

Observe that block design is more efficient in general.

References

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. (2008). Tennessee's Student Teacher Achievement Ratio (STAR) project.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.