Homework 3

January 26, 2022

Instructions

- This homework is due Wednesday, February 2 at 3pm EST.
- Submit via GitHub. Remember to commit and push online so I can see it.
- Please format your homework solutions using R Markdown. You are welcome to simply add your answers below each question.
  - If the question requires a figure, make sure you have informative title, axis labels, and legend if needed.
  - Note: When I've given the framework of an answer's code, I've included the option `eval=FALSE` in the R chunk. When you start filling in your answer, you'll need to switch this to `eval=TRUE`.
- Turn in both the .rmd file and the knitted .pdf or .html file.
  - Knitting the .rmd file to a .pdf or .html file should help ensure your code runs without errors, but double check the output is what you expected.

Question 1

- Sampling variability
- Sampling distribution
- Standard deviation
- Standard error

For each of these terms, in your own words:

1. explain what it means in terms of $\hat{ATE}$.
2. explain how it relates to the other terms listed.

(Don't be afraid of repeating yourself!)

Question 2

Read the Exercise 8 prompt from GG Ch. 3. Although a natural experiment, explain in your own words how this data was generated, in effect, using a block randomized design. In other words, why should we analyze this data using what we know about estimating $\hat{ATE}$ and $\hat{SE}(\hat{ATE})$ under block randomization?

Question 3

Complete Exercise 8 from GG Ch. 3, parts a, b, c, and e.

Note: do these questions "by hand," meaning, calculate the quantities on your own in code without using pre-programmed statistical routines like `lm()`. (You can still use commands like `mean()`!)

In this code chunk, I'm just creating the dataset shown in the book for you.

```r
term_length_tx <- c(rep(0, 16),
                    rep(1, 15))
term_length_ak <- c(rep(0, 16),
                    rep(1, 18), 0)
bills_tx <- c(18,29,41,53,60,67,75,79,79,
              88,93,101,103,106,107,131,
              29,37,42,45,45,54,54,58,61,
              64,69,73,75,92,104)
bills_ak <- c(11,15,23,24,25,26,28,31,33,
              34,35,35,36,38,52,59,9,
              10,14,15,15,17,18,19,19,
              20,21,23,23,24,28,30,32,34,17)
df <- data.frame(state = c(rep("TX", 31),
                           rep("AK", 35)),
                 term_length = c(term_length_tx, term_length_ak),
                 bills = c(bills_tx, bills_ak))
```

3a

3b

3c

3e

Question 4

Now, answer Exercise 8 part d.

In addition to explaining what the question asks you to explain, calculate $\hat{ATE}$ when pooling and compare the estimated ATE to when you didn't pool in 3c.

Also, calculate $\hat{SE}(\hat{ATE})$ when pooling and compare the estimated ATE to when you didn't pool in 3e.

What are the implications if the researcher incorrectly analyzes this data?

Checking your answers

Confirm your correct results from Question 3 and your incorrect results from Question 4 using software commands available in R. I've provided the commands for you. If what you calculated in Question 3 and Question 4 matches up, then you know what's happening under the hood in these functions!

```
# with blocking
estimatr::difference_in_means(bills ~ term_length,
                              blocks = state,
                              data = df)
```

```
## Design:  Blocked
##              Estimate Std. Error   t value    Pr(>|t|)  CI Lower  CI Upper DF
## term_length -13.2168    4.74478 -2.785545 0.007079688 -22.70148 -3.732117 62
```

```
# without analyzing within-block first, then aggregating
estimatr::difference_in_means(bills ~ term_length,
                              data = df)
```

```
## Design:  Standard
##               Estimate Std. Error  t value  Pr(>|t|)  CI Lower  CI Upper
## term_length -14.51515   7.132193 -2.03516 0.0463009 -28.78439 -0.245916
##                   DF
## term_length 59.44775
```