DS-UA 201: Problem Set 1

Yizhou Wan

September 29, 2020

This problem set is due at **11:59 pm on Tuesday, September 29th**. The data are on the course website.

Please upload your solutions as a .pdf file saved as `Yourlastname_Yourfirstinitial_pset1.pdf`). In addition, an electronic copy of your .Rmd file (saved as `Yourlastname_Yourfirstinitial_pset1.Rmd`) must be submitted to the course website at the same time. We should be able to run your code without error messages. Please note on your problem set if you collaborated with another student and, if so, whom. In order to receive credit, homework submissions must be substantially started and all work must be shown. Late assignments will not be accepted.

Problem 1

Researchers often want to aggregate and synthesize evidence from multiple studies of the same question to get a more precise estimate of the effect of a particular intervention. These sorts of studies are called *meta-analyses* and are very common in medical and psychological research where many experiments of varying size are often run in many different contexts.

We are going to replicate a meta-analysis using an example dataset from the *meta* suite in Stata of a series of experiments analyzing the effect of teacher expectations on student performance (*pupiliq.dta*).

From the description of the data available here (pp. 19)

This example describes a well-known study of Rosenthal and Jacobson (1968) that found the so-called Pygmalion effect, in which expectations of teachers affected outcomes of their students. A group of students was tested and then divided randomly into experimentals and controls. The division may have been random, but the teachers were told that the students identified as experimentals were likely to show dramatic intellectual growth. A few months later, a test was administered again to the entire group of students. The experimentals outperformed the controls. Subsequent researchers attempted to replicate the results, but many did not find the hypothesized effect.

Raudenbush (1984) did a meta-analysis of 19 studies and hypothesized that the Pygmalion effect might be mitigated by how long the teachers had worked with the students before being told about the nonexistent higher expectations for the randomly selected subsample of students.

We will be working with the Raudenbush data set in R. First, load the data via the function `read_dta` from the `haven` package (part of the broader `tidyverse`).

```
### Read in the tidyverse
library(tidyverse)

### Load in the haven package
library(haven)
```

```r
### Read in the pupiliq data
pupil <- read_dta("pupiliq.dta")
```

This dataset contains the results of 19 replications of the teacher expectation experiment. The relevant variables you will need are:

- `studylbl` - Author and date of the study
- `stdmdiff` - Estimated standardized average effect (difference between treated and control)
- `se` - Standard error of `stdmdiff`

Part A

Consider the case of K independent studies. Let $\widehat{\tau_i}$ denote each study i's estimator of the effect and σ_i the **known**, **constant** standard error of that estimator (assume that we know the true standard error).

One approach to a meta-analysis assumes that each $\widehat{\tau_i}$ is an unbiased estimator of a common effect parameter τ and that differences between studies are attributable to sampling error.

Consider the proposed combined estimator $\hat{\tau}$

$$\hat{\tau} = \frac{\sum_{i=1}^{K} \frac{1}{\sigma_i^2} \widehat{\tau_i}}{\sum_{i=1}^{K} \frac{1}{\sigma_i^2}}$$

Find the expectation of $\hat{\tau}$. Is it an unbiased estimator of τ?

$$E[\hat{\tau}] = E[\frac{\sum_{i=1}^{K} \frac{1}{\sigma_i^2} \widehat{\tau_i}}{\sum_{i=1}^{K} \frac{1}{\sigma_i^2}}]$$

$$E[\hat{\tau}] = \frac{1}{\frac{K}{\sigma_i^2}} E[\sum_{i=1}^{K} \frac{1}{\sigma_i^2} \widehat{\tau_i}]$$

$$E[\hat{\tau}] = \frac{1}{\frac{K}{\sigma_i^2}} \sum_{i=1}^{K} \frac{1}{\sigma_i^2} E[\widehat{\tau_i}]$$

$$E[\hat{\tau}] = \frac{\frac{K}{\sigma_i^2}}{\frac{K}{\sigma_i^2}} \widehat{\tau_i}$$

$$E[\hat{\tau}] = \widehat{\tau_i}$$

$$E[\hat{\tau}] - \widehat{\tau_i} = 0$$

Therefore, $\hat{\tau}$ is an unbiased estimator of τ.

Part B

$$var(\hat{\tau}) = var(\frac{\sum_{i=1}^{K} \frac{1}{\sigma_i^2} \widehat{\tau_i}}{\sum_{i=1}^{K} \frac{1}{\sigma_i^2}})$$

$$var(\hat{\tau}) = \frac{1}{(\frac{K}{\sigma_i^2})^2} var(\sum_{i=1}^{K} \frac{1}{\sigma_i^2} \widehat{\tau_i})$$

$$var(\hat{\tau}) = \frac{1}{\frac{K^2}{\sigma_i^4}} \sum_{i=1}^{K} \frac{1}{\sigma_i^4} \sigma^2$$

$$var(\hat{\tau}) = \frac{K\sigma^2}{K^2}$$

$$var(\hat{\tau}) = \frac{\sigma^2}{K}$$

Part C

With this estimator, $\hat{\tau}$, generate a point estimate for τ using the 19 studies in the `pupiliq.dta` dataset and construct a 95% confidence interval (assuming asymptotic normality).

```
se_mean = mean(pupil$se)
se_var = var(pupil$se) / nrow(pupil)
se_se = sqrt(se_var)
se_mean
```

```
## [1] 0.2031053
```

```
se_var
```

```
## [1] 0.0003985432
```

```
se_se
```

```
## [1] 0.01996355
```

```
ci_95_neighbors_effect <- c(se_mean - qnorm(.975)*se_se, se_mean + qnorm(.975)
*se_se)
ci_95_neighbors_effect
```

```
## [1] 0.1639774 0.2422331
```

Part D

How does your estimate from C compare to the results from the Rosenthal & Jacobson, 1968 study (study 17 in the `pupiliq.dta` dataset)?

We can see that by comapre to the stdmdiff of study 17. the data shown to the dataset lies far away from the 95% confidence interval of the estimate of τ

Part E

Suppose instead that someone suggested an alternate estimator for τ, denoted $\widehat{\tau'}$, that simply averaged all K studies.

$$E[\widehat{\tau'}] = E[\frac{1}{K}\sum_{i=1}^{K}\widehat{\tau_i}] = \frac{1}{K}E[\sum_{i=1}^{K}\widehat{\tau_i}] = \frac{1}{K}\sum_{i=1}^{K}E\,[\widehat{\tau_i}] = \frac{1}{K}K\mu = \mu$$

For variance:

$$var(\widehat{\tau'}) = var(\frac{1}{K}\sum_{i=1}^{K}\widehat{\tau_i}) = \frac{1}{K^2}var(\sum_{i=1}^{K}\widehat{\tau_i}) = \frac{1}{K^2}\sum_{i=1}^{K}\sigma^2 = \frac{K}{K^2}\sigma^2 = \frac{\sigma^2}{K}$$

Find the expectation and variance of this estimator.

the variance of $\hat{\tau} = \sum_{i=1}^{K}(\widehat{\tau'} - \widehat{\tau_i})^2 / K$

Part F

With this alternate estimator, generate a point estimate and 95% confidence interval (again assuming asymptotic normality) for τ using the 19 studies in the `pupiliq.dta` dataset.

```
se_mean = mean(pupil$se)
se_var = var(pupil$se) / nrow(pupil)
se_se = sqrt(se_var)
se_mean
```

```
## [1] 0.2031053
```

```
se_var
```

```
## [1] 0.0003985432
```

```
se_se
```

```
## [1] 0.01996355
```

```
ci_95_neighbors_effect <- c(se_mean - qnorm(.975)*se_se, se_mean + qnorm(.975)
*se_se)
ci_95_neighbors_effect
```

```
## [1] 0.1639774 0.2422331
```

Part G

How do the two estimators $\widehat{\tau'}$ and $\hat{\tau}$ compare? Which would you prefer to use and why?

I believe the second one is better because though the 95 confidence interval is similar but the second one is easier for us to find the expection and varience

Problem 2

Many studies in political science have documented an effect of *ballot order* on a candidate's vote share in an election.[1] In general, candidates that are listed first on a ballot receive a slightly higher vote share than those listed lower on the ballot. As a result, most states will randomize the order of candidates on the ballot and alter the order from ballot to ballot.

In the 2008 Democratic Primary in New Hampshire, the ballot order was the same on all ballots. Furthermore, this fixed order was decided by randomly and unformly drawing a letter of the alphabet (A-Z) and then listing all candidates alphabetically by last name starting from the randomly chosen letter (and returning back to A after Z). In the actual primary election in 2008, the letter "Z" was drawn and therefore Joe Biden was first on all ballots.

Professor Jon Krosnick of Stanford University noted in an op-ed that this process may have advantaged some candidates more than others ex-ante due to the distribution of last names on the ballot.

A total of 21 candidates were on the ballot in this election (ordered by last name below)

Name
Biden
Caligiuri
Capalbo
Clinton
Crow

[1] See, for example, Miller, Joanne M., and Jon A. Krosnick. "The impact of candidate name order on election outcomes." Public Opinion Quarterly (1998): 291-330; Ho, Daniel E., and Kosuke Imai. "Estimating causal effects of ballot order from a randomized natural experiment: The California alphabet lottery, 1978-2002." Public Opinion Quarterly 72.2 (2008): 216-240.

Dodd
Edwards
Gravel
Hewes
Hughes
Hunter
Keefe
Killeen
Koon
Kucinich
LaMagna
Laughlin
Obama
Richardson
Savior
Skok

Part A

Given the New Hampshire randomization process, what is the probability of Biden appearing as the first name on the ballot?

To make Biden the first person, we must find the possibility that the letter drawed from the pool is b or b is the first letter with a name after the drawed letter, therefore, the letters can be [T,U,V,W,X,Y,Z,A,B], where there are 9 letters. Considering that all letters have the same possibility of being drawed, so the possibility of Biden appearing as the first name on the ballot is: 9/26

Part B

What is the probability of Obama appearing as the first name on the ballot?

To make Obama the first person, we must find the possibility that the letter drawed from the pool is o or o is the first letter with a name after the drawed letter, therefore, the letters can be [M,N,O], where there are 3 letters. Considering that all letters have the same possibility of being drawed, so the possibility of Obama appearing as the first name on the ballot is: 3/26

Part C

Pollsters at the time noticed that the New Hampshire results in 2008 were significantly different from the average of polls leading up to the election. While Hillary Clinton finished 3 percentage points ahead of Barack Obama, the average of final poll estimates suggested

Obama leading Clinton by 7 percentage points.[2] In his op-ed, Krosnick suggested that Clinton may have beneffited in part from a ballot order effect. Given the New Hampshire randomization scheme, what is the probability of Hillary Clinton appearing above Barack Obama on the ballot?

To make Hillary Clinton appearing above Barack Obama, we must find the possibility that the letter drawed from the pool is C or o is the latter than C according to the order, therefore, the letters can be [P,Q,R,S,T,U,V,W,X,Y,Z,A,B,C], where there are 14 letters. Considering that all letters have the same possibility of being drawed, so the possibility is: 4/13

Problem 3

The STAR (Student-Teacher Achievement Ratio) Project was a four-year longitudinal study conducted on students in Tennessee to evaluate the impact of small class sizes on student achievement.[3] The experiment began in 1985 and followed a single group of students from kindergarten to third grade. Students were randomly assigned to one of three types of classes: small classes (13-17 students per teacher), regular classes (22-25 students per teacher), and regular classes that were also assigned a teacher aide. Regular measures of achievement and other outcomes were taken annually during the four years of the study. Additionally, follow-up studies examined outcomes at later stages (such as high-school graduation).

In this problem, we'll analyze a portion of this dataset found in the Imai *Quantitative Social Science* book. You can load it in from the `STAR.csv` file using the `read_csv` function from the `haven` package.

```
### Read in the star data
star <- read_csv("STAR.csv")
```

This dataset contains 6325 observations of students. The relevant variables you will need are:

- `race` - Student's race: (coded as 1 = white, 2 = Black, 3 = Asian, 4 = Hispanic, 5 = Native American, 6 = Other)
- `classtype` - Assigned class size treatment in kindergarten (1 = small, 2 = regular, 3 = regular with aide)
- `yearssmall` - Number of years in a small-sized class (kindergarten through 3rd grade)
- `hsgrad` - Did the student graduate from from high school (1 = did graduate, 0 = did not graduate)

---

[2] An Evaluation of the Methodology of the 2008 Pre-Election Primary Polls

[3] For more on the study, see: Mosteller, Frederick. ``The Tennessee study of class size in the early school grades.'' *The future of children* (1995): 113-127.

- `g4math` - Math score on the fourth-grade standardized test
- `g4reading` - Reading score on the fourth-grade standardized test

Part A

First, recode the `race` and `classtype` variables into factor/character variables based on the coding scheme described above. For both of these variables, you should have a new variable in your dataset that converts the numeric coding into an informative category name (e.g. "Black" instead of 2 for the `race` variable). Subset the data to only white and Black students. We'll be using this subset of the data for the remainder of the problem.

Using this dataset, create a summary table for the number of students assigned to each class type in kindergarten. Which of the three treatments has the fewest number of students assigned to it?

```
star %>% filter(race == 1 || race == 2) %>% mutate(racial = case_when(race ==
 1 ~ "White", race == 2 ~ "Black")) %>% mutate(class = case_when(classtype ==
 1 ~"small" , classtype == 2 ~"regular", classtype == 3 ~ "regular with aide
"))
```

```
## # A tibble: 6,325 x 8
##     race classtype yearssmall hsgrad g4math g4reading racial class

##    <dbl>     <dbl>      <dbl>  <dbl>  <dbl>     <dbl> <chr>  <chr>

##  1     1         3          0     NA     NA        NA White  regular with
aide
##  2     2         3          0     NA    706       661 Black  regular with
aide
##  3     1         3          0      1    711       750 White  regular with
aide
##  4     2         1          4     NA    672       659 Black  small

##  5     1         2          0     NA     NA        NA White  regular

##  6     1         3          0     NA     NA        NA White  regular with
aide
##  7     1         1          4     NA    668       657 White  small

##  8     1         3          0     NA     NA        NA White  regular with
aide
##  9     1         1          4      1    709       725 White  small

## 10     1         2          0      1    698       692 White  regular

## # ... with 6,315 more rows
```

Part B

Drop all observations with missing fourth-grade math scores **or** missing fourth-grade reading scores. Using this dataset of complete observations, estimate the average treatment effects of being assigned to a small kindergarten class versus being assigned to a regular kindergarten class (with no aide) on fourth-grade math and fourth-grade reading scores. Calculate the large-sample (Neyman) standard error and provide a 95% asymptotic confidence interval for each of your estimates. Would you reject the null of no average treatment effect on math scores at the $\alpha = 0.05$ level? Would you reject for the effect on reading scores? Interpret your findings and compare the two estimated effects. What do you conclude about the effect of small kindergarten class sizes on average test scores?

```
star %>% filter(is.na(g4math)==FALSE) %>% filter(is.na(g4reading) == FALSE)
```

```
## # A tibble: 2,344 x 6
##     race classtype yearssmall hsgrad g4math g4reading
##    <dbl>     <dbl>      <dbl>  <dbl>  <dbl>     <dbl>
##  1     2         3          0     NA    706       661
##  2     1         3          0      1    711       750
##  3     2         1          4     NA    672       659
##  4     1         1          4     NA    668       657
##  5     1         1          4      1    709       725
##  6     1         2          0      1    698       692
##  7     1         3          0     NA    733       672
##  8     1         1          4      1    740       836
##  9     1         3          0     NA    716       679
## 10     1         2          0     NA    758       836
## # ... with 2,334 more rows
```

```
smallClasses = star%>% filter(is.na(g4math)==FALSE) %>% filter(is.na(g4readin
g) == FALSE)%>% filter(classtype == 1)
regularClasses = star %>% filter(is.na(g4math)==FALSE) %>% filter(is.na(g4rea
ding) == FALSE)%>% filter(classtype ==2)


# Point Estimate
diffMath <- colMeans(smallClasses['g4math']) - colMeans(regularClasses['g4mat
h'])
diffReading <- colMeans(smallClasses['g4reading']) - colMeans(regularClasses
['g4reading'])

seMath <- sqrt(var(smallClasses['g4math'])/nrow(smallClasses['g4math'])+var(r
egularClasses['g4math'])/nrow(regularClasses['g4math']))
seReading <- sqrt(var(smallClasses['g4reading'])/nrow(smallClasses['g4reading
'])+var(regularClasses['g4reading'])/nrow(regularClasses['g4reading']))

ci_95Math <- c(diffMath - qnorm(.975)*seMath,
diffMath + qnorm(.975)*seMath)
ci_95Reading <- c(diffReading - qnorm(.975)*seReading,
```

```
diffReading + qnorm(.975)*seReading)


print(seMath)

##          g4math
## g4math 2.160204

print(seReading)

##           g4reading
## g4reading   2.65336

print(diffMath)

##      g4math
## -0.08083122

print(diffReading)

## g4reading
##  3.790515

print(ci_95Math)

## [1] -4.314753  4.153090

print(ci_95Reading)

## [1] -1.409976  8.991005
```

Through the calculation of the data, we can find that the mean difference in math score is very little while we can see slight difference on reading score. THe standard deviation of both test is pretty the same, As a result ,we can see that the 95% confidence interval range of the two test is pretty the same. As the result, I think that there is no average treatment effect on math scores but I think there do have an effect on reading score. I think that the math score has no relation to the size of class but the reading score do have the relation to.

Part C

If treatment were randomly assigned, we would expect to see balance between the different treatment conditions on pre-treatment covariates. Here, we want to investigate whether the different treatment groups have similar proportions of white and Black students. To do so, we'll perform a *balance check*. Calculate the proportion of Black students in each of the three treatment groups for kindergarten class size. Do the treatment arms all have similar proportions of Black students?

```
smallClasses = star%>% filter(classtype == 1)
regularClasses = star %>% filter(classtype ==2)
AideClass = star%>%filter(classtype == 3)
blackSmall = smallClasses%>% filter(race == 2)
whiteSmall = smallClasses%>% filter(race == 1)
```

```
blackRegular = regularClasses %>% filter(race ==2)
blackAide = AideClass%>% filter(race == 2)

smallPortion = nrow(blackSmall)/nrow(smallClasses)
regularPotion = nrow(blackRegular)/nrow(regularClasses)
aidePortion = nrow(blackAide)/nrow(AideClass)

print(smallPortion)

## [1] 0.3121053

print(regularPotion)

## [1] 0.3236098

print(aidePortion)

## [1] 0.3384133
```

Through the calculation, we can find that all three classtype have similar portion of black students: about one third of all students

Part D

In this part, we'll look instead at whether students graduate high school. Starting with the complete dataset of white and Black students, create a new dataset that removes all students with missing values of `hsgrad`. For now we'll assume that the `hsgrad` variable is missing completely at random and subset the data to only observations where that variable is non-missing.

Suppose we were interested in the effect of repeated exposure to small class sizes and not just the effect of small class sizes in kindergarten. In this new dataset, create a variable for each student that takes on a value of 1 if that student had more than 2 years of small class sizes and a value of 0 otherwise. Assuming that this new "treatment" is as-good-as-randomly assigned, estimate the average treatment effect of having 3 or 4 years of small class sizes from kindergarten to third grade on the probability that a student graduates high school. Provide a 95% asymptotic confidence interval and interpret your results. Do you reject the null of no average treatment effect at the $\alpha = .05$ level? Interpret your result.

```
graduate = star%>% filter(is.na(hsgrad)==FALSE)
smallMore = graduate %>% filter(yearssmall >2)
smallLess = graduate %>% filter(yearssmall <= 2)

moreGradMean = colMeans(smallMore['hsgrad'])
lessGradMean = colMeans(smallLess['hsgrad'])
diffGrad = moreGradMean-lessGradMean

seGrad <- sqrt(var(smallMore['hsgrad'])/nrow(smallMore['hsgrad'])+var(smallLe
ss['hsgrad'])/nrow(smallLess['hsgrad']))
```

```
ci_95Grad <- c(diffGrad - qnorm(.975)*seGrad, diffGrad + qnorm(.975)*seGrad)

diffGrad

##     hsgrad
## 0.04270638

seGrad

##            hsgrad
## hsgrad 0.01491021

ci_95Grad

## [1] 0.01348290 0.07192986
```

Through the calculation, we can find that there do exist a treatment effect on a 95% confidence interval.

Part E

Examine whether the "years of small class sizes" variable is balanced on race. Are Black students in the study any more or less likely to have more than 2 years of small class sizes from kindergarten to grade three compared to white students in the study?

```
blackMore = smallMore%>% filter(race == 2)
blackLess = smallLess%>% filter(race == 2)

morePotion = nrow(blackMore)/nrow(smallMore)
lessPotion = nrow(blackLess)/nrow(smallLess)

morePotion

## [1] 0.2506849

lessPotion

## [1] 0.2736297
```

Through the test, we can see that the portion of balck students are both close to one fourth, which means due to the data provided here, black students have the equal oppotunity as white students to have more than 2 years of small classes.

Given your findings here, is the assumption of unconfoundedness/ignorability for the "years of small class sizes" variable reasonable? Should we interpret the estimate from Part D causally?

I think the effect of "years of small class sizes" is reasonable. Because of the result of part D, we can find that students who attend small size of class are more likely to graduate than those who not. Also, base on the data from Part E, we find that the race of students isn't a factor to the class size, the ratio for small size classes are similar to the ratio of large classes