Psychological Phenomenons in Various Studies

Jaimie Chin & Yijia Chen

September 29, 2020

This project aims to explore the phenomenon of various studies that exhibit biased results. The three phenomenons we will explore are:

- Pygmalion Effect
- Ballot Order Effect
- Biased Treatment Effect

Part 1: Pygmalion effect - Biased & Unbiased Estimators in Meta-Analyses

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 portion of the project aims to describe 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.

Load & Read the Data

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`).

```r
### 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 we will use in this project are:

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

Mathematical Intuition of Unbiased Estimators

Considering the case of K independent studies: Let $\hat{\tau}_i$ 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 $\hat{\tau}_i$ 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 $\hat{\tau}$

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

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

$$\begin{aligned}
\mathbb{E}[\hat{\tau}] &= \mathbb{E}\left[\frac{\sum_{i=1}^K \frac{1}{\sigma_i^2}\hat{\tau}_i}{\sum_{i=1}^K \frac{1}{\sigma_i^2}}\right] && (1)\\
&= \frac{1}{\sum_{i=1}^K \frac{1}{\sigma_i^2}}\mathbb{E}\left[\sum_{i=1}^K \frac{1}{\sigma_i^2}\hat{\tau}_i\right] && (2)\\
&= \frac{1}{\sum_{i=1}^K \frac{1}{\sigma_i^2}}\sum_{i=1}^K \frac{1}{\sigma_i^2}\mathbb{E}\left[\hat{\tau}_i\right] && (3)\\
&= \frac{1}{\sum_{i=1}^K \frac{1}{\sigma_i^2}}\sum_{i=1}^K \frac{1}{\sigma_i^2}\tau && (4)\\
&= \tau\frac{\sum_{i=1}^K \frac{1}{\sigma_i^2}}{\sum_{i=1}^K \frac{1}{\sigma_i^2}} && (5)\\
&= \tau && (6)
\end{aligned}$$

$\mathbb{E}[\hat{\tau}] - \tau = 0$ therefore $\hat{\tau}$ is an unbiased estimator of τ

Mathematical Intuition of the Variance of an Unbiased Estimator

Find the variance of $\hat{\tau}$ (under the assumption that σ_i is known for all i).

$$
\begin{aligned}
Var[\hat{\tau}] &= Var\left[\frac{\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\hat{\tau}_i}{\sum_{i=1}^{K}\frac{1}{\sigma_i^2}}\right] && (7)\\
&= \frac{1}{\left(\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\right)^2} Var\left[\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\hat{\tau}_i\right] && (8)\\
&= \frac{1}{\left(\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\right)^2}\left[\sum_{i=1}^{K}\frac{1}{(\sigma_i^2)^2}Var(\hat{\tau}_i)\right] && (9)\\
&= \frac{1}{\left(\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\right)^2}\left[\sum_{i=1}^{K}\frac{\sigma_i^2}{(\sigma_i^2)^2}\right] && (10)\\
&= \frac{1}{\left(\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\right)^2}\left[\sum_{i=1}^{K}\frac{1}{\sigma_i^2}\right] && (11)\\
&= \frac{1}{\sum_{i=1}^{K}\frac{1}{\sigma_i^2}} && (12)
\end{aligned}
$$

Creating a Point Estimator & Variance for Multiple Studies

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).

```
#point estimate calculation based off of estimator
pt_1 <- (sum((pupil$stdmdiff)/(pupil$se)^2)) / (sum(1/(pupil$se)^2))
pt_1
```

```
## [1] 0.06034335
```

```
#standard error calculated from variance
se <- sqrt((1)/(sum((1)/(pupil$se)^2)))
se
```

```
## [1] 0.03648548
```

```
#Construction of 95% confidence interval
ci_95 <- c(pt_1 - qnorm(.975)*se, pt_1 + qnorm(.975)*se)
ci_95
```

```
## [1] -0.01116687  0.13185357
```

```
#p-value
ci_95_p_value <- 2*(1 - pnorm(abs(pt_1/se)))
ci_95_p_value
```

```
## [1] 0.09814773
```

Comparing Point Estimate with Rosenthal & Jacobson, 1968 study (study 17 in the `pupiliq.dta` dataset)

Our calculated point estimate is considerably less than the standardized difference in means result from the Rosenthal & Jacobson, 1968 study. The standardized difference in means result from the Rosenthal &

Jacobson, 1968 study also falls outside of the calculated 95% confidence interval of our point estimate. In fact, the Rosenthal & Jacobson effect (.3) is about 5 times more extreme.

```r
print(pupil[17,]$stdmdiff)
```

```
## [1] 0.3
## attr(,"label")
## [1] "Standardized difference in means"
## attr(,"format.stata")
## [1] "%9.0g"
```

```r
print(c(pupil[17,]$stdmdiff - abs(qnorm(.025))*pupil[17,]$se,
  pupil[17,]$stdmdiff + abs(qnorm(.025))*pupil[17,]$se))
```

```
## [1] 0.02756501 0.57243499
```

The standard error of that single estimate is also about 3.8 times greater than the meta-analysis estimate.

```r
### Calculate the variance of tau_hat
var_tau_hat <- function(sigma_i){
  return(1/sum(1/sigma_i^2))
}

### Find the standard error
pupil[17,]$se/sqrt(var_tau_hat(pupil$se))
```

```
## [1] 3.809735
## attr(,"label")
## [1] "Standard error of stdmdiff"
## attr(,"format.stata")
## [1] "%10.0g"
```

This tells us that if we use estimator $\hat{\tau}$ to construct a meta-analysis on pupiliq.dta, the Rosenthal & Jacobson, 1968 study would not be an expected value. While Rosenthal & Jacobson would reject the null of zero effect under an α of 0.05, our meta-analysis estimate would fail to reject as the 95% confidence interval includes zero. Given the greater precision of the meta-analysis estimate, we would prefer that estimate under the assumption that all studies are unbiased estimators for the same treatment effect parameter (e.g. no hidden effect modifiers across studies).

Mathematical Intuition of Averaged Estimator

For the purposes of this project, we wanted to explore an alternative estimator for τ, denoted $\widehat{\tau'}$, that simply averaged all K studies.

$$\widehat{\tau'} = \frac{1}{K}\sum_{i=1}^{K}\hat{\tau}_i$$

We document the mathematical intuition behind finding the expectation and variance of this estimator.

$$\mathbb{E}\left[\widehat{\tau'}\right] = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\hat{\tau}_i\right]$$

$\frac{1}{K}$ is a constant so it can be factored out of the expectation (Homogeneity).

$$\mathbb{E}\left[\widehat{\tau'}\right] = \frac{1}{K}\mathbb{E}\left[\sum_{i=1}^{K}\hat{\tau}_i\right]$$

Distribute expectation further through the sum ($\sum_{i=1}^{K}$) by the first property of expectations where the sum of expectations is equal to the expectation of the sum.

$$\mathbb{E}\left[\widehat{\tau'}\right] = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}\left[\hat{\tau}_i\right]$$

The summation $\sum_{i=1}^{K}$ becomes a constant where i^K is equal to K. We can assume (as stated in the preamble) that each $\hat{\tau}_i$ is an unbiased estimator of a common effect parameter τ. So the expectation of the unbiased estimator $\hat{\tau}_i$ is equal to τ. τ becomes a constant which we can separate from the summation.

$$\mathbb{E}\left[\widehat{\tau'}\right] = \frac{K\tau}{K}$$

Simplify.

$$\mathbb{E}\left[\widehat{\tau'}\right] = \tau$$

To fine the variance of $\widehat{\tau'}$

$$Var\left[\widehat{\tau'}\right] = Var\left[\frac{1}{K}\sum_{i=1}^{K}\hat{\tau}_i\right]$$

$\frac{1}{K}$ is a constant so it can be factored out of the variance. We square $\frac{1}{K}$ because variance of a constant * random variable is equal to the constantˆ2 * the variance of the random variable.

$$Var\left[\widehat{\tau'}\right] = \frac{1}{K^2}Var\left[\sum_{i=1}^{K}\hat{\tau}_i\right]$$

If we assume that each observation is independent, the variance of the sum = the sum of variances.

$$Var\left[\widehat{\tau'}\right] = \frac{1}{K^2} * \sum_{i=1}^{K} * Var\left[\hat{\tau}_i\right]$$

The summation $\sum_{i=1}^{K}$ becomes a constant where i^K is equal to K.Variation of an individual observation is σ^2 because it comes from a distribution.

$$Var\left[\widehat{\tau'}\right] = \frac{K\sigma^2}{K^2}$$

Simplify.

$$Var\left[\widehat{\tau'}\right] = \frac{\sigma^2}{K}$$

### Generating a Point Estimate for the Alternative Estimator

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

```r
#Point estimate for alternate estimator
pt_2 <- (sum(pupil$stdmdiff)) / 19
print(pt_2)
```

```
## [1] 0.1636842
```

```
#standard error calculation
se_2 <- sqrt(((sum(pupil$se))^2 / 19))
print(se_2)
```

```
## [1] 0.8853153
```

```
#confidence interval calculation
ci_95_2 <- c(pt_2 - qnorm(.975)*se_2, pt_2 + qnorm(.975)*se_2)
print(ci_95_2)
```

```
## [1] -1.571502  1.898870
```

```
#p-value calculation
ci_95_2_p_value <- 2*(1 - pnorm(abs(pt_2/se_2)))
ci_95_2_p_value
```

```
## [1] 0.8533169
```

Comparing Estimators $\widehat{\tau'}$ and $\hat{\tau}$

In the case of both estimators, we fail to reject the null hypothesis (keep the null) since both confidence intervals contain 0 (If we assume that the null hypothesis = no difference in means = no effect).

However, the estimator $\widehat{\tau'}$ generates a greater point estimate and standard error (by a factor of about 1.4) which generates a confidence interval that spans across a larger distribution of potential estimated outcomes.

```
### Calculate the variance of tau_prime_hat
var_tau_prime_hat <- function(sigma_i){
  return(sum(sigma_i^2)/(length(sigma_i)^2))
}

sqrt(var_tau_prime_hat(pupil$se))/sqrt(var_tau_hat(pupil$se))
```

```
## [1] 1.383695
```

The p-value for $\widehat{\tau'}$ tells us that there is an 85.33% chance that our estimated results are random. On the other hand, The p-value for $\hat{\tau}$ tells us that there is an 29.52% chance that our estimated results are random. Both p-values are greater than the $\alpha = .05$, so we fail to reject the null (accept the null hypothesis).

The estimator $\hat{\tau}$ also generates a smaller point estimate and standard error which generates a confidence interval that spans across a smaller distribution of potential estimated outcomes. This tells us that the more results from the pupiliq.dta will fall within the 95% confidence interval when we use the estimator $\widehat{\tau'}$.

The estimator $\widehat{\tau'}$ would represent the pupiliq.dta better than $\hat{\tau}$, which is why I would prefer to use $\widehat{\tau'}$ (also because it was easier to code in R).

Part 2: Ballot Order Effect

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.

[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.

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

Outlining Probabilities in the Ballot Order Effect

Given the New Hampshire randomization process, the probability of Biden appearing as the first name on the ballot is 9 times out of the 26 possible randomly chosen letters of the alphabet (and returning back to A after Z). If any of the letters A, B, T, U, V, W, X, Y, or Z were chosen, Biden would have appeared first on the ballot. The New Hampshire randomization process allows Biden to appear as the first name on the ballot 34.62% (roughly 1/3) of the time.

On the other hand, the probability of Obama appearing as the first name on the ballot is 3 times out of the 26 possible randomly chosen letters of the alphabet (and returning back to A after Z). If any of the letters M, N, or O were chosen, Obama would have appeared first on the ballot. The New Hampshire randomization process allows Obama to appear as the fist name on the ballot 11.54% (roughly 1/9) of the time.

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, the probability of Hillary Clinton appearing above Barack Obama on the ballot is 14 times out of the 26 possible randomly chosen letters of the alphabet (and returning

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

back to A after Z). If any of the letters A, B, C, P, Q, R, S, T, U, V, W, X, Y, or Z were chosen, Clinton would have appeared above Obama on the ballot. The New Hampshire Randomization process allows Clinton to appear above Obama on the ballot 53.85% (more than 1/2) of the time.

Part 3: Bias Treatment in Protected and Unprotected Groups

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 and will be accessed in this project as 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 we will use in our analysis 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)
- `g4math` - Math score on the fourth-grade standardized test
- `g4reading` - Reading score on the fourth-grade standardized test

Recoding Variables of Interest for Treatment Bias Analysis

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

Using this subset, we will create a summary table for the number of students assigned to each class type in kindergarten and discover which of the three treatments has the fewest number of students assigned to it.

```
library(dplyr)
library(tidyverse)

#recode 'race' and 'classtype'
star <- star %>% mutate(Race=recode(race,
               "1" = "white",
               "2" = "Black",
               "3" = "Asian",
               "4" = "Hispanic",
               "5" = "Native American",
               "6" = "Other"),
               ClassType=recode(classtype,
                                 "1" = "small",
```

[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.

```
                                "2" = "regular",
                                "3" = "regular with aide"
                                )
               )

#Create a subset with only white and Black students
star_subset <- star %>% filter(Race == "white" | Race == "Black")

#Group and summarize the # of students assigned to each classtype
star_subset %>% group_by(ClassType) %>% summarize(Assigned.Students = n(), .groups = 'keep')
```

```
## # A tibble: 3 x 2
## # Groups:   ClassType [3]
##   ClassType         Assigned.Students
##   <chr>                         <int>
## 1 regular                        2182
## 2 regular with aide              2223
## 3 small                          1887
```

The “small” class size treatment has the fewest number of students assigned to it.

Estimating Average Treatment Effects for Math and Reading Scores of Biased Groups

For this analysis, we chose to drop all observations with missing fourth-grade math scores **or** missing fourth-grade reading scores.

Using this dataset of complete observations, we 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.

We calculate the large-sample (Neyman) standard error and provide a 95% asymptotic confidence interval for each of your estimates.

```
star_subset_score <- star_subset %>% filter(!is.na(g4math), !is.na(g4reading))
```

```
#Average treatment effect on math score
math_small_class_effect <- mean(star_subset_score$g4math[star_subset_score$ClassType == "small"])
math_regular_class_effect <- mean(star_subset_score$g4math[star_subset_score$ClassType == "regular"])
math_class_effect <- math_small_class_effect - math_regular_class_effect
print(math_class_effect)
```

```
## [1] -0.1995642
```

```
#Average treatment effect on reading score
reading_small_class_effect <- mean(star_subset_score$g4reading[star_subset_score$ClassType == "small"])
reading_regular_class_effect <- mean(star_subset_score$g4reading[star_subset_score$ClassType == "regula
reading_class_effect <- reading_small_class_effect - reading_regular_class_effect
print(reading_class_effect)
```

```
## [1] 3.728538
```

```
#Neyman large sample error on math scores
var_math_class_effect <- var(star_subset_score$g4math[star_subset_score$ClassType == "small"])/sum(star_

math_se_class_effect <- sqrt(var_math_class_effect)

print(math_se_class_effect)
```

```
## [1] 2.169283
```

```
#Neyman large sample error on reading scores
var_reading_class_effect <- var(star_subset_score$g4reading[star_subset_score$ClassType == "small"])/su

reading_se_class_effect <- sqrt(var_reading_class_effect)

print(reading_se_class_effect)
```

```
## [1] 2.659981
```

```
# 95% confidence interval for math score(asymptotic)
ci_95_math_class_effect <- c(math_class_effect - qnorm(.975)*math_se_class_effect, math_class_effect + 
print(ci_95_math_class_effect)
```

```
## [1] -4.451281  4.052152
```

```
# 95% confidence interval for reading score(asymptotic)
ci_95_reading_class_effect <- c(reading_class_effect - qnorm(.975)*reading_se_class_effect, reading_cla
print(ci_95_reading_class_effect)
```

```
## [1] -1.484928  8.942004
```

```
#p-value for math score
p_value_math_score <- 2*(1 - pnorm(abs(math_class_effect/math_se_class_effect)))
p_value_math_score
```

```
## [1] 0.9267016
```

```
#p-value for reading score
p_value_reading_score <- 2*(1 - pnorm(abs(reading_class_effect/reading_se_class_effect)))
p_value_reading_score
```

```
## [1] 0.161
```

Considering the 95% confidence interval on the average treatment effect on both math and reading score, we would fail to reject the null (accept the null hypothesis) of no average treatment effect for both math and reading scores at the $\alpha = .05$ level. If we assume that the null has the average treatment effect of 0, then the null lies within the 95% confidence interval for both math and reading scores. If the null lies within the confidence interval, then there is a chance that the null is a possible estimated outcome.

Furthermore, the p-values tell us that there is a 92.67% chance that the result small class size (compared to regular class size with no aid) on math scores is random and that there is a 16.1% chance that the result of small class size (compared to regular class size with no aid) on reading scores is random. Both p-values are greater than the $\alpha = .05$, so we can quite confidently accept the null for both math and reading scores.

If we compare the estimated treatment effects of a small class size on math and reading score, the treatment barely has a negative effect (barely 1 point) on math scores and a slight positive effect (roughly 4 points) on reading scores. If we observe the Neyman standard error, the standard error of the difference-in-means for reading scores is similar to the standard error of the difference in means for the math scores. This tells us that there is roughly a 2~3 point variation and uncertainty regarding the average treatment effect for reading scores. Both standard errors are relatively large considering the large sample size and tell us that there is uncertainty to the effects of the treatment. In conclusion, it is rather unlikely that a small kindergarten class size has an effect on average test scores when compared to a regular class size with no aid.

Balance Check of Biased Groups

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*. We do this by calculating the proportion of Black students in each of the three treatment groups for kindergarten class size and assess if treatment arms all have similar proportions of Black students.

```
star_subset_score <- star_subset_score %>% mutate (race = case_when (race == 1 ~ FALSE,
                                                race == 2 ~ TRUE))

star_subset_score %>% group_by(ClassType) %>% summarize(Total_Students = n(), Black_Students = sum(race
```

```
## # A tibble: 3 x 4
## # Groups:   ClassType [3]
##   ClassType         Total_Students Black_Students Proportion
##   <chr>                      <int>          <int>      <dbl>
## 1 regular                      829            124      0.150
## 2 regular with aide            786            113      0.144
## 3 small                        718            113      0.157
```

Each treatment group of class type has relatively the same proportion of Black students (about 14% to 16% of each treatment group were Black students).

Estimating Average Treatment Effects for High School Graduation of Biased Groups

In this part, we'll look at whether students graduate high school. Starting with the complete dataset of white and Black students, we 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.

In this section, we are interested in the effect of repeated exposure to small class sizes and not just the effect of small class sizes in kindergarten. Therefore, in this new dataset, we 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, we 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.

```
#new subset where students with missing values of 'hsgrad' are removed
star_grad_subset <- star_subset %>% filter(hsgrad != "NA")

#New column variable created to recode  # of years spent in a small classroom
star_grad_subset <- star_grad_subset %>% mutate(YearsSmall=recode(yearssmall,
               "0" = "0",
               "1" = "0",
               "2" = "0",
               "3" = "1",
               "4" = "1"))

#estimate average treatment effect of having 3 or 4 years of small class sizes from kindergarten to thi
more_small_years_effect <- mean(star_grad_subset$hsgrad[star_grad_subset$YearsSmall == "1"])
less_small_years_effect <- mean(star_grad_subset$hsgrad[star_grad_subset$YearsSmall == "0"])
small_years_effect <- more_small_years_effect - less_small_years_effect
print(small_years_effect)
```

```
## [1] 0.04270223
```

```
#standard error
var_small_years_effect <- var(star_grad_subset$hsgrad[star_grad_subset$YearsSmall == "1"]) / sum(star_g

years_small_se_effect <- sqrt(var_small_years_effect)
```

```
# 95% confidence interval on hsgrad probability
ci_95_small_years_effect <- c(small_years_effect - qnorm(.975)*years_small_se_effect, small_years_effec
print(ci_95_small_years_effect)
```

```
## [1] 0.01343459 0.07196987
```

```
#p-value calculation
p_value_small_years_effect <- 2*(1 - pnorm(abs(small_years_effect/years_small_se_effect)))
p_value_small_years_effect
```

```
## [1] 0.004241257
```

We can reject the null of no average treatment effect at the $\alpha = .05$ level since our confidence interval does not contain 0 (assuming that the null = 0). The p-value also tells us there is only a 0.42% chance that our results are random. The p-value is less than the $\alpha = .05$ level as well, so we can quite confidently reject the null. Although we can reject the null of no average treatment effect at the $\alpha = .05$ level, the estimate average effect only has a slight positive effect (0.043 times more likely) on whether or not the student graduated high school. The confidence interval, though it does not contain 0, is also relatively close to 0, which shows that the positive effect of having 3 or 4 years of small class sizes only has a slight effect on the probability of graduated high school.

Balance Check on New Calculated Variable

For the final part of our analysis, we examine whether the "years of small class sizes" variable is balanced on race. We wish to discover whether 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.

Given our findings here, we want to see if the assumption of unconfoundedness/ignorability for the "years of small class sizes" variable is reasonable in order to interpret our previous estimate causally.

```
star_grad_subset <- star_grad_subset %>% mutate(race = case_when (race == 1 ~ FALSE,
                                                                  race == 2 ~ TRUE))

star_grad_subset %>% group_by(YearsSmall) %>% summarize(Total_Students = n(), Black_Students = sum(race
```

```
## # A tibble: 2 x 4
## # Groups:   YearsSmall [2]
##   YearsSmall Total_Students Black_Students Proportion
##   <chr>               <int>          <int>      <dbl>
## 1 0                    2309            634      0.275
## 2 1                     729            183      0.251
```

According to the balance check, Black students in the study were less likely to have more than 2 years of small class sizes from kindergarten to grade three compared to white students in the study. Black students who had more than 2 years of small class sizes made up only 25.10% (roughly 1/4) of all the students who had more than 2 years of small class sizes. On the other hand, white students who had more than 2 years of small class sizes made up 74.90% (roughly 3/4) of all the students who had more than 2 years of small class sizes. Approximately, for every 3 white students who had more than 2 years of small class sizes, only 1 Black student had more than 2 years of small class sizes.

Given these findings, the assumption of unconfoundedness/ignorability for the "years of small class sizes" variable becomes more unreasonable. If we assume ignorability or unconfoundaedness, we assume that the probability of receiving treatment is independent of potential outcomes. If this were the case, then the proportion of Black students having more than 2 years of small class sizes would be similar to the proportion of white students having more than 2 years of small class sizes. This is not the case in this study and therefore, it is difficult to assume unconfoundedness/ignorability.

If we can not assume unconfoundedness/ignoarability, we should not interpret our estimate causally. There may be a confounding variable, like SES (socioeconomic status) or location, that may have an effect on both the "years of small class sizes" (treatment) and the probability of graduating high school (effect). Perhaps Black students go to school in a location there are no small class sizes to offered and the rate of students graduating high school is lower. Or, the socioeconomic status of Black students prevents them from being in a small class for more years and also causes them to drop out of high school.In the prescence of confounding variables, we can not assume that treatment is independent of the potential outcomes when both may be dependent on a variable that wasn't considered in the study. Thus, we should not interpret our previous estimate causally as the treatment of the years of small class sizes is likely not an accurate predictor of the probability of graduating high school if we can not assume unconfoundedness/ignorability.