Problem Set 1: Solutions

Petra Ivanovic - pi2018 - Section 005

Due Oct 6th, 2023

This homework must be turned in on Brightspace by Oct 6th 2023. It must be your own work, and your own work only – you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. **You may consult with others, but when you write up, you must do so alone.**

Your homework submission must be written and submitted using Rmarkdown. No handwritten solutions will be accepted. You should submit:

1. A compiled PDF file named yourNetID_solutions.pdf containing your solutions to the problems.

2. A .Rmd file containing the code and text used to produce your compiled pdf named your-NetID_solutions.Rmd.

Note that math can be typeset in Rmarkdown in the same way as Latex. Please make sure your answers are clearly structured in the Rmarkdown file:

1. Label each question part

2. Do not include written answers as code comments. Write out answers and explanations separately.

3. The code used to obtain the answer for each question part should accompany the written answer. Comment your code!

Question 1. Definitions and Examples (20 points)

Answer the following questions. Be as specific and detailed as possible. Give examples.

1. **What is the fundamental problem of causal inference? (5 points)**

   - The fundamental problem of causal inference arises from the fact that we are not able to observe both potential outcomes in the same state of the world. More precisely, it is impossible to observe the value of $Y_i(1)$ and $Y_i(0)$ for the same unit. For example, when conducting an experiment we are never truly able to know what the outcome would be if the individual receiving the treatment did not receive it or if the individual in the control group did receive the treatment.

   - As a practical example we can consider the effect of a cancer drug on patient survival. If we conduct an experiment and randomly assign some patients to receive the drug and others to receive a placebo, we can only observe one outcome for each patient - they either get cured or they do not. Further, there is no way to know what would have happened if they were in the other group since we can never assign the same patient to both groups at the same time. This inability to know what would have happened if they were assigned to the other group is the fundamental problem of causal inference.

2. **Why are experiments important? (5 points)**

   - Experiments are crucial in scientific fields because they allow for randomization, replicability, and maybe most important - for control for confounding variables. Confounding variables are variables that are correlated with both our treatment and outcome and thus can significantly impact our results. Going back to our previous example, we can consider age as an confounding variable, for example regardless of which group patient is put in it is safe to assume that younger patients are more likely to be able to fight the disease compared to older patients. When doing an experiment rather than an observational study we can randomly assign patients to the treatment and control groups, and with that we can ensure that the two groups are similar on all confounding variables. This allows us to conclude that any difference in outcomes between the two groups can be attributed to the treatment. Further, "blind" experiments also allow us to conduct most bias-less studies since we can ensure that all patients are treated the same and with same expectations. Finally, it is also worth to mention replicability as a crucial benefit of en experiment as in science we always want to be able to verify our results by conducting new trials using same conditions which can only be perfectly set in experiments.

3. **What does ignorability mean? (5 points)**

   - Ignorability or "unconfoundedness" is the rule that treatment assignment must be independent of potential outcomes given the observed covarriates. More precisely, we have to aim to make it so that the only difference between our treatment and control groups is the assignment/treatment itself and no other significant differences should be present.

   - In our example this would mean people of all ages (for example), would be in both groups; we would not assign all the younger people in treatment groups since we can assume they will handle the drug better and older participants in control group but the assignment would be random and allow for all people to have same probability of receiving the drug. In other words, ignorability means that the probability of receiving the drug is the same for all patients, regardless of their potential outcome, the only difference found between treatment and control groups is the drug itself.

4. **What is SUTVA? (5 points)**

   - SUTVA stands for "Stable Unit Treatment Value Assumption" and is a set of assumptions related to Consistency when conducting experiments. First aspect of SUTVA is that there is no interference or spillover which means that manipulating one unit's treatment does not affect any other

unit's potential outcomes. This means that, in our example, effect of the drug on one patient in the treatment group does not influence the effect/outcome for any other patient in the group. Next aspect of SUTVA/consistency is having a single version of treatment for all members of the treatment group which means that all members of the group should for example receive the same drug - in no case should there be multiple different treatments within the same treatment group. While SUTVA assumptions can seem (and often are) unrealistic in practice, they are crucial for many causal inference methods.

Question 2. Bed Nets and Malaria (20 points)

Article: Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment by Jessica Cohen and Pascaline Dupas

Some economists have argued that "cost-sharing" makes it more likely that a product will be used (versus giving it away for free). Cohen and Dupas partnered with 20 Kenyan prenatal clinics to distribute subsidized anti-malarial bed nets. For each clinic, they varied the extent of the subsidy: either full (free bed-nets, $D_i = 1$) or partial (90% cheaper bed-nets, $D_i = 0$). They measure (among other things) whether women who received bed nets used them (Y_i).

1. What is $\mathbb{E}[Y_i|D_i = 0]$? (5 points)

   - $\mathbb{E}[Y_i|D_i = 0]$ represents the expected purchase / usage of bed nets for women (Y_i) who received a 90% (partial) subsidy for the bed net ($D_i = 0$) - partial subsidy. The mathematical representation of the formula, $\mathbb{E}[Y_i|D_i = 0]$ is the expected value of the outcome variable, bed net usage (Y_i) for those who received the partial subsidy for the bed nets. To be able to calculate this we would need the information about the women who received the partial subsidy in the study and how many of them purchased / used the bed nets.

2. What is $\mathbb{E}[Y_i(1)]$? (5 points)

   - $\mathbb{E}[Y_i(1)]$ represents the expected purchase / usage of bed nets for women if everyone (hypothetically) received the treatment of the full subsidy. Further, it is an estimate of the expected usage of bed nets if there was no cost-sharing and everyone received the free bed nets. Knowing this theoretical value would allow us to compare our outcome of the study and see if cost-sharing - partial subsidy treatment was impactful.

3. What is $\mathbb{E}[Y_i(1)|D_i = 0]$? (5 points)

   - $\mathbb{E}[Y_i(1)|D_i = 0]$ represents the expected purchase / usage of bed nets for women who received partial subsidies but we are considering what would have happened if they have instead received full subsidies. This represents the fundamental problem of causal inference (mentioned in question 1) as we are never able to know what would have happened if we have a different treatment to the same group. Mathematically speaking we are trying to calculate the expected value of the outcome variable (Y_i) for those who received partial subsidies, assuming they had received full subsidies instead. We cannot estimate this from our data since we can never know what would have happened under different treatment (again, the fundamental problem of causal inference.)

4. Cohen and Dupas randomized treatment at the level of the clinic, but the outcomes of interest are at the individual level. Is there a violation of consistency/SUTVA? Why or why not? Argue your case. (5 points)

   - Assumption: treatment being randomized at the level of the clinic means that entire clinic got the same treatment - e.g. all women in Clinic A got the full subsidy (free bed nets) and all women in Clinic B got partial subsidy.

   - I believe that there is no violation of consistency/SUTVA in this experiment. To be able to claim this we need to look if there is a possibility of treatment not being same for all the members of the same group or if one units' outcome could affect others. I believe in the case of a non blind trial such as this one - everyone knew what treatment they were receiving - they knew whether their bed nets were free or not, it was a good decision to randomize on the level of the clinic as randomizing on the individual level could have caused spillover. For instance, is one person knew that another got a free bed net (full subsidy) they might be less inclined to purchase one with partial subsidy since they don't consider it fair that they have to pay. Since the randomization was at the clinic level, assuming clinics are not too close in distance and people from different clinics

are not as likely to communicate, spillover should not be an issue since everyone is receiving the same conditions. Another issue that could violate the consistency/SUTVA of the trial is if women had another source of free nets apart from the trial. However, this is not likely as they are not affordable which is shown by the fact there is a need for a subsidy in the first place. Last concern that I would like to bring up is if some of the clinics were in areas with higher risk - for example if one area has more malaria cases women will be more likely to purchase bed nets regardless of the cost and this could affect the trial. However, the randomization of the experiment should ensure this does not happen.

- Due to all of the considerations I could think of being handled correctly, there should be no spillover and thus I think there is no violation of consistency or SUTVA.

Question 3. Application (Coding) (30 points)

The STAR (Student-Teacher Achievement Ratio) Project is a four year *longitudinal study* examining the effect of class size in early grade levels on educational performance and personal development.

This exercise is in part based on[1]:

Mosteller, Frederick. 1997. "The Tennessee Study of Class Size in the Early School Grades." *Bulletin of the American Academy of Arts and Sciences* 50(7): 14-25.

A longitudinal study is one in which the same participants are followed over time. This particular study lasted from 1985 to 1989 involved 11,601 students. During the four years of the study, students were randomly assigned to small classes, regular-sized classes, or regular-sized classes with an aid. In all, the experiment cost around $12 million. Even though the program stopped in 1989 after the first kindergarten class in the program finished third grade, collection of various measurements (e.g., performance on tests in eighth grade, overall high school GPA) continued through the end of participants' high school attendance.

We will analyze just a portion of this data to investigate whether the small class sizes improved performance or not. The data file name is `STAR.csv`, which is a CSV data file. The names and descriptions of variables in this data set are:

Name	Description
race	Student's race (White = 1, Black = 2, Asian = 3, Hispanic = 4, Native American = 5, Others = 6)
classtype	Type of kindergarten class (small = 1, regular = 2, regular with aid = 3)
g4math	Total scaled score for math portion of fourth grade standardized test
g4reading	Total scaled score for reading portion of fourth grade standardized test
yearssmall	Number of years in small classes
hsgrad	High school graduation (did graduate = 1, did not graduate = 0)

Note that there are a fair amount of missing values in this data set. For example, missing values arise because some students left a STAR school before third grade or did not enter a STAR school until first grade.

1. Create a new factor variable called `kinder` in the data frame. This variable should recode `classtype` by changing integer values to their corresponding informative labels (e.g., change 1 to `small` etc.). Similarly, recode the `race` variable into a factor variable with four levels (`white`, `black`, `hispanic`, `others`) by combining Asians and Native Americans as the `others` category. For the `race` variable, overwrite the original variable in the data frame rather than creating a new one. Recall that `na.rm = TRUE` can be added to functions in order to remove missing data. (5 points)

```
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.3      v readr     2.1.4
## v forcats   1.0.0      v stringr   1.5.0
## v ggplot2   3.4.1      v tibble    3.2.1
## v lubridate 1.9.3      v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

---
[1]I have provided you with a sample of their larger dataset. Empirical conclusion drawn from this sample may differ from their article.

```r
STAR <- read.csv("STAR2.csv")

# Recode classtype variable into informative labels
STAR = STAR %>% mutate(kinder = case_when(classtype == '1' ~ "small",
classtype == '2' ~ "regular",classtype == '3' ~ "regular with aid"))

STAR = STAR %>% mutate(race = case_when(race == '1' ~ "white",
race == '2' ~ "black",race == '4' ~ "hispanic", race == '3'~ "others",
race == '5'~ "others",race == '6'  ~ "others"))

head(STAR, 15)
```

```
##      race classtype yearssmall hsgrad g4math g4reading           kinder
## 1  white         2          0     NA     NA        NA          regular
## 2  black         3          0      1     NA        NA regular with aid
## 3  white         1          4      1    707       719            small
## 4  white         1          2      1     NA        NA            small
## 5  black         2          0      0     NA        NA          regular
## 6  black         2          0     NA    674       571          regular
## 7  white         3          0     NA     NA        NA regular with aid
## 8  black         1          3      1     NA        NA            small
## 9  white         2          0     NA     NA        NA          regular
## 10 black         1          4      1     NA        NA            small
## 11 white         2          0      0     NA        NA          regular
## 12 black         1          4      1     NA        NA            small
## 13 white         2          0      1    696       688          regular
## 14 white         3          0     NA     NA        NA regular with aid
## 15 white         2          0     NA     NA        NA          regular
```

2. How does performance on fourth grade reading and math tests for those students assigned to a small class in kindergarten compare with those assigned to a regular-sized class? Do students in the smaller classes perform better? Use means to make this comparison while removing missing values. Give a brief substantive interpretation of the results. To understand the size of the estimated effects, compare them with the standard deviation of the test scores. (10 points)

```r
# Calculate the means and standard deviations of test scores for
# small and regular-sized classes
small_class_mean_math <- mean(STAR$g4math[STAR$classtype == 1], na.rm = TRUE)
small_class_mean_reading <- mean(STAR$g4reading[STAR$classtype == 1], na.rm = TRUE)
small_class_sd_math <- sd(STAR$g4math[STAR$classtype == 1], na.rm = TRUE)
small_class_sd_reading <- sd(STAR$g4reading[STAR$classtype == 1], na.rm = TRUE)


regular_class_mean_math <- mean(STAR$g4math[STAR$classtype == 2], na.rm = TRUE)
regular_class_mean_reading <- mean(STAR$g4reading[STAR$classtype == 2], na.rm = TRUE)
regular_class_sd_math <- sd(STAR$g4math[STAR$classtype == 2], na.rm = TRUE)
regular_class_sd_reading <- sd(STAR$g4reading[STAR$classtype == 2], na.rm = TRUE)

# Print the means and standard deviations
cat("Small Class Mean Math Score:", small_class_mean_math, "\n")
```

```
## Small Class Mean Math Score: 709.0091
```

```r
cat("Small Class Mean Reading Score:", small_class_mean_reading, "\n\n")
```

Small Class Mean Reading Score: 724.5316

```r
cat("Small Class Standard Deviation of Math Score:", small_class_sd_math, "\n")
```

Small Class Standard Deviation of Math Score: 45.3054

```r
cat("Small Class Standard Deviation of Reading Score:", small_class_sd_reading, "\n\n")
```

Small Class Standard Deviation of Reading Score: 51.48347

```r
cat("Regular Class Mean Math Score:", regular_class_mean_math, "\n")
```

Regular Class Mean Math Score: 709.6314

```r
cat("Regular Class Mean Reading Score:", regular_class_mean_reading, "\n\n")
```

Regular Class Mean Reading Score: 720.4467

```r
cat("Regular Class Standard Deviation of Math Score:", regular_class_sd_math, "\n")
```

Regular Class Standard Deviation of Math Score: 40.81335

```r
cat("Regular Class Standard Deviation of Reading Score:", regular_class_sd_reading, "\n\n")
```

Regular Class Standard Deviation of Reading Score: 54.30644

From these results we can see that the mean scores in both math and reading are very similar in small and regular sized classes. Further, looking at the mean math score, regular class size *technically* performs a bit better with a score of 709.63 versus the score of the small class which is 709.01. However, looking at the mean reading score, small class size performs slightly better than regular size class with 724.53 versus 720.45. It is worthy to note that this differences are so so small, especially for the math scores. Looking at the standard deviations of math and reading scores we can again notice they are quite similar for both small and regular size classes. Just like it performed better with mean math score, regular size class has lower standard deviation at 40.8 (versus 45.3 in small class size) which could indicate that students in regular size class performed more similar (less variance) on the tests than in small class size. Looking at the standard deviation for the reading score, just like in mean scores, small class has 'better' - lower standard deviation at 51.48 (versus 54.31 for regular size class) which could again indicate that grades in that class for reading were more similar to each other and closer to the mean (less variance).

3. Instead of comparing just average scores of reading and math tests between those students assigned to small classes and those assigned to regular-sized classes, look at the entire range of possible scores. To do so, compare a high score, defined as the 66th percentile, and a low score (the 33rd percentile) for small classes with the corresponding score for regular classes. These are examples of *quantile treatment effects*. Does this analysis add anything to the analysis based on mean in the previous question? (Hint: You will use the quantile() function in r.) (5 points)

```r
# Calculate the 66th and 33rd percentiles for math and reading scores for small classes
small_class_quantiles_math <- quantile(STAR$g4math[STAR$classtype == 1],
                                        c(0.33, 0.66), na.rm = TRUE)
small_class_quantiles_reading <- quantile(STAR$g4reading[STAR$classtype == 1],
                                          c(0.33, 0.66), na.rm = TRUE)

# Calculate the 66th and 33rd percentiles for math and reading scores for regular-sized classes
regular_class_quantiles_math <- quantile(STAR$g4math[STAR$classtype == 2],
                                         c(0.33, 0.66), na.rm = TRUE)
regular_class_quantiles_reading <- quantile(STAR$g4reading[STAR$classtype == 2],
                                            c(0.33, 0.66), na.rm = TRUE)

# Print the quantiles
cat("Quantiles for Small Class Math Scores (33rd and 66th percentiles):",
    small_class_quantiles_math, "\n")
```

Quantiles for Small Class Math Scores (33rd and 66th percentiles): 695 727

```r
cat("Quantiles for Small Class Reading Scores (33rd and 66th percentiles):",
    small_class_quantiles_reading, "\n")
```

Quantiles for Small Class Reading Scores (33rd and 66th percentiles): 708 741

```r
cat("Quantiles for Regular Class Math Scores (33rd and 66th percentiles):",
    regular_class_quantiles_math, "\n")
```

Quantiles for Regular Class Math Scores (33rd and 66th percentiles): 696 724.64

```r
cat("Quantiles for Regular Class Reading Scores (33rd and 66th percentiles):",
    regular_class_quantiles_reading, "\n")
```

Quantiles for Regular Class Reading Scores (33rd and 66th percentiles): 705 740

Looking at the Quantiles, we can notice that for both math and reading small class size has higher scores in 66th percentile, but again they are barely noticeably larger. This could indicate that students with better scores were in smaller class size but due to such a small difference I would argue, that just like means and standard deviation, this is not enough to conclude the actual impact of the class size (or treatment). We can also notice that the score for math for 33rd percentile is slightly (by 1 point) larger in regular class. As for reading in 33rd percentile, small class size has 3 points higher score. Just like with 66th percentile this difference is too small to lead to any conclusions on which class is actually performing better. It is also worth to mention that it is likely that both/either class includes students that would maybe perform great (or badly) regardless of the class size.

4. We examine whether the STAR program reduced the achievement gaps across different racial groups. Begin by comparing the average reading and math test scores between white and minority students (i.e., Blacks and Hispanics) among those students who were assigned to regular classes with no aid. Conduct the same comparison among those students who were assigned to small classes. Give a brief substantive interpretation of the results of your analysis. (5 points)

```r
# Filter the data for students assigned to regular classes with no aid
regular_no_aid <- STAR %>%
  filter(classtype == 2)  # Assuming 2 represents regular classes with no aid

# Compare average reading and math scores between white and minority students
regular_no_aid_summary <- regular_no_aid %>%
  group_by(race) %>%
  summarize(
    Mean_Math_Score = mean(g4math, na.rm = TRUE),
    Mean_Reading_Score = mean(g4reading, na.rm = TRUE),
    Count = n()  # Add count of each race category
  ) %>%
  filter(race %in% c('white', 'black', 'hispanic'))  # Filter for white, black, and hispanic students

# Print the summary
print(regular_no_aid_summary)
```

```
## # A tibble: 2 x 4
##   race  Mean_Math_Score Mean_Reading_Score Count
##   <chr>           <dbl>              <dbl> <int>
## 1 black            699.               692.   547
## 2 white            711.               725.  1104
```

```r
# Filter the data for students assigned to regular classes with no aid
regular_no_aid_2 <- STAR %>%
  filter(classtype == 2)  # Assuming 2 represents regular classes with no aid

# Create a new column "race_group" to combine black and Hispanic as "minorities"
regular_no_aid_2 <- regular_no_aid_2 %>%
  mutate(race = ifelse(race %in% c('black', 'hispanic'), 'minorities', race))

# Compare average reading and math scores by race group
regular_no_aid_summary_2 <- regular_no_aid_2 %>%
  group_by(race) %>%
  summarize(
    Mean_Math_Score = mean(g4math, na.rm = TRUE),
    Mean_Reading_Score = mean(g4reading, na.rm = TRUE),
    Count = n()
  ) %>%
  filter(race %in% c('white', 'minorities'))  # Filter for white and minorities

# Print the summary
print(regular_no_aid_summary_2)
```

```
## # A tibble: 2 x 4
##   race       Mean_Math_Score Mean_Reading_Score Count
##   <chr>                <dbl>              <dbl> <int>
## 1 minorities            699.               692.   547
## 2 white                 711.               725.  1104
```

The above tables show mean math and reading scores for white and minority student (black and hispanic) in **regular class sizes**. The first table shows the breakdown of black and white students and second shows

white versus minority breakdown. However, we can notice that the tables are same because there are no hispanic students in these tables which indicates that there are no hispanic students in the small class size. (After some separate analysis (not shown) I did realize there are only 3 hispanic students in the entire dataset and they are all in the small class size.) Looking at this table(s) (Note: I will refer to both as one as they are the same) we can notice that black students have lower scores in both math and reading. Further, we can notice that math scores are marginally smaller at just about 12 points while reading scores are significantly smaller at 33 points. We can also notice that there is double the amount of white students versus black. It would be worth analyzing why black students have such lower reading scores compared to white students while their math scores are similar.

Note: "Other" minorities were not included as it was mentioned in discussion on Brightspace to only look at these 3 groups.

```r
# Filter the data for students assigned to small classes
small_classes <- STAR %>%
  filter(classtype == 1)  # Assuming 1 represents small classes

# Compare average reading and math scores between white and minority students
small_classes_summary <- small_classes %>%
  group_by(race) %>%
  summarize(
    Mean_Math_Score = mean(g4math, na.rm = TRUE),
    Mean_Reading_Score = mean(g4reading, na.rm = TRUE),
    Count = n()  # Add count of each race category
  ) %>%
  filter(race %in% c('white', 'black', 'hispanic'))  # Filter for white, black, and hispanic students

# Print the summary
print(small_classes_summary)
```

```
## # A tibble: 3 x 4
##   race    Mean_Math_Score Mean_Reading_Score Count
##   <chr>             <dbl>              <dbl> <int>
## 1 black              696.               699.   433
## 2 hispanic           744                760      3
## 3 white              711.               729.   986
```

```r
# Filter the data for students assigned to small classes
small_classes_2 <- STAR %>%
  filter(classtype == 1)  # Assuming 1 represents small classes

# Create a new column "race_group" to combine black and Hispanic as "minorities"
small_classes_2 <- small_classes_2 %>%
  mutate(race = ifelse(race %in% c('black', 'hispanic'), 'minorities', race))

# Compare average reading and math scores by race group
small_classes_summary <- small_classes_2 %>%
  group_by(race) %>%
  summarize(
    Mean_Math_Score = mean(g4math, na.rm = TRUE),
    Mean_Reading_Score = mean(g4reading, na.rm = TRUE),
    Count = n()
  ) %>%
  filter(race %in% c('white', 'minorities'))  # Filter for white and minorities
```

```
# Print the summary
print(small_classes_summary)
```

```
## # A tibble: 2 x 4
##   race       Mean_Math_Score Mean_Reading_Score Count
##   <chr>                <dbl>              <dbl> <int>
## 1 minorities            696.               699.   436
## 2 white                 711.               729.   986
```

The above tables show mean math and reading scores for white and minority student (black and hispanic) in **small class sizes**. The first table shows the breakdown of black, hispanic, and white students and second shows white versus minority breakdown (both black and hispanic scores combined). First thing that is worth pointing out is that hispanic students have significantly better scores than both other groups at 744 points for mean math score and 760 for mean reading score compared to white and black students whose scores are in low 700s and high 600 respectively. However, we can also notice that there are only 3 hispanic students and becasue of that the minority scores (table 2) does not have significantly higher scores than the scores black students had since there were many more black students at 433. Combining the black and hispanic students into one category, minorities, only increased the mean math score by less than 4 points and mean reading score by less than 1 point. Lets compare differences in scores between regular size class and small size class. Looking at the results of white students, the differences were minimal; smaller class size had a better mean reading score but only of 3 points while the mean math score decreased but only by 0.3 points (not even a full point). Looking at the black students the mean math score was also lower for the smaller class size but only by 2.3 points and the mean reading score was also higher but again only by about 6.3 points. The rise in joint "minority" column has about a point larger effect than the difference in performance of black students but I belive it is not a fair or valuable comparison since the reason for increase is due to performance of 3 hispanic students who are not represented at all in the regular class size.

I would not say that the small class size increased performance in either race, black or white, since the results were only marginally different and even though mean reading scores rose by few points, mean math scores decreased. Further, I would also say that the race gap was not significantly smaller in smaller class since the difference in mean math scores increased from about 12 points to about 15 points while difference in mean reading scores was lower by only 3 points (from about 33 to about 30 points.) All in all, I would say that smaller class size did neither yield significantly better performance in either race group nor decreased a race gap between the white and black students.

5. We consider the long term effects of kindergarten class size. Compare high school graduation rates across students assigned to different class types. Also, examine whether graduation rates differ by the number of years spent in small classes. Finally, as done in the previous question, investigate whether the STAR program has reduced the racial gap between white and minority students' graduation rates. Briefly discuss the results. (5 points)

```
# 1. Compare high school graduation rates across different class types
graduation_by_classtype <- STAR %>%
  group_by(kinder) %>%
  summarize(
    Graduation_Rate = mean(hsgrad, na.rm = TRUE)
  )

# Print graduation rates by class type
cat("Graduation Rates by Class Type:\n")
```

```
## Graduation Rates by Class Type:
```

```
print(graduation_by_classtype)
```

```
## # A tibble: 3 x 2
##   kinder          Graduation_Rate
##   <chr>                    <dbl>
## 1 regular                  0.823
## 2 regular with aid         0.836
## 3 small                    0.848
```

```
cat("\n")
```

```
# 2. Examine whether graduation rates differ by the number of years spent in small classes
graduation_by_yearssmall <- STAR %>%
  group_by(yearssmall) %>%
  summarize(
    Graduation_Rate = mean(hsgrad, na.rm = TRUE)
  )

# Print graduation rates by years in small classes
cat("Graduation Rates by Years in Small Classes:\n")
```

```
## Graduation Rates by Years in Small Classes:
```

```
print(graduation_by_yearssmall)
```

```
## # A tibble: 5 x 2
##   yearssmall Graduation_Rate
##        <int>           <dbl>
## 1          0           0.825
## 2          1           0.798
## 3          2           0.852
## 4          3           0.827
## 5          4           0.883
```

```
cat("\n")
```

```
# 3. Investigate whether the STAR program has reduced the racial gap in graduation rates
graduation_by_race_regular_no_aid <- regular_no_aid %>%
  group_by(race) %>%
  summarize(
    Graduation_Rate = mean(hsgrad, na.rm = TRUE)
  ) %>%
  filter(race %in% c('white', 'black', 'hispanic'))  # Filter for white, black, and hispanic students

# Print graduation rates by race
cat("Graduation Rates by Race in Regular Size Class (White, Black, Hispanic):\n")
```

```
## Graduation Rates by Race in Regular Size Class (White, Black, Hispanic):
```

```r
print(graduation_by_race_regular_no_aid)
```

```
## # A tibble: 2 x 2
##   race  Graduation_Rate
##   <chr>           <dbl>
## 1 black           0.735
## 2 white           0.856
```

```r
cat("\n")
```

```r
# 4. Investigate whether the STAR program has reduced the racial gap in graduation rates
graduation_by_race_small <- small_classes %>%
  group_by(race) %>%
  summarize(
    Graduation_Rate = mean(hsgrad, na.rm = TRUE)
  ) %>%
  filter(race %in% c('white', 'black', 'hispanic'))  # Filter for white, black, and hispanic students

# Print graduation rates by race
cat("Graduation Rates by Race in Small Size Class (White, Black, Hispanic):\n")
```

```
## Graduation Rates by Race in Small Size Class (White, Black, Hispanic):
```

```r
print(graduation_by_race_small)
```

```
## # A tibble: 3 x 2
##   race     Graduation_Rate
##   <chr>              <dbl>
## 1 black              0.771
## 2 hispanic             NaN
## 3 white              0.872
```

```r
cat("\n")
```

Looking at the first table we can notice that the students in regular classes had the lowest graduation rate while regular classed with aid had about 1.1% higher graduation rate followed by small class size students which had about 2.5% higher graduation rate than regular size class students. While 2.5% on large amount of people is quite a few people since we do not know if the the classroom size was an only contributor to this performace and since all previous statistics we have calculated yielded very low differences in performance I would say it is not significant enough to argue that the higher graduation rate is solely due to reducing class size.

In table two we can see the graduation rate by different time (years) spent in the smaller size class. While the rise in graduation rate when comparing 0 and 4 years spent in small class seems very big with almost 6% more students graduating when we look at the trend of 0, 1, 2, 3, and 4 years and graduation rate it does not seem like it is connected. If more years in small class led to larger graduation rate we would expect some kind of a positive trend - linear or non linear - but our graduation rate first decreases after year 1 then increases significantly after year 2 then decreases again after year 3 and finally increases once more for year 4. Due to this we should conduct further analysis to see if there are any other factors or confounders that could lead to this kind of varying graduation rate for our students rather than years spent in smaller class.

Table 3 shows us the break down by race for regular sized class and table 4 for small sized class. We can notice that table 4 does have an entry for hispanic students since there were 3 in that class but as we can see

graduation rate values is not specified as we do not have information of their graduation status so we can disregard it. One thing that we can notice is that for both white and black students graduation rate was higher for those in smaller class, more so for black students where increase was about 3.7% while for white was about 1.7%. This increase for black students was much more significant than that of white students (more than double the increase in percentage points) and did start to close the race gap that was about 12% for regular size class versus 10% for small class size. However, once again since no other statistics were as significant or effects as large it is worth to consider what could be alternative explanations for this increase except for decreasing the class size.

Question 4. Design Your Experiment (30 points)

Design your own experiment from start to finish. Choose an *interesting* question. Explain why observational data may give you the wrong answer. Detail the potential outcomes and a well-defined treatment. Explain the type of experiment (completely random, cluster-design, block/stratified). Will your design ensure a causal treatment effect? (Remember: Be as specific as possible and give examples.)

Context: I have been reading a lot more on my my phone recently as it removes the need to a) look for a bookstore/library that has the book I have to or want to read and b) saves the time that is needed to go and get the book and c) saves the cost of the book. I would love to know if the quality of my reading is as good as reading from a physical book - the satisfaction of finishing a book definitely is not.

Experiment setup:

1. Research Question: Does reading a physical print book result in better comprehension and information retention compared to reading the same book in an e-book format?

2. Potential outcomes:

- Y_1: Comprehension and information retention if an individual reads a book in print format.

- Y_0: Comprehension and information retention if an individual reads the same book in an e-book format.

- Causal effect of the treatment on the outcome = $Y_1 - Y_0$ = how much the outcome has changed due to the treatment in comparison to what would have happened without it.

3. Treatment:

- Format of the book -> either a physical copy of the book or an e-book copy

4. Type of the Experiment: Since our goal (as always in experiments) is to ensure the causal treatment effect we will do an entirely *Random Design*. This means that the participants will be randomly assigned to either print group - treatment or e-book group - control.

5. Steps to do the Experiment:

- Participant Selection: We will select a random, non biased group of participants of all ages, genders, races, etc. but we will ensure they all have similar baseline level of reading comprehension and retention abilities to make sure they are at the same "starting point" when it comes to doing the experiment. For example, we would not want someone who does not know how to read or someone who is a literature major because comparing their comprehensions would not be valuable regardless of the book format.
- Random Assignment: We will randomly assigned the participants using an unbias software or code/algorithm to either the control (e-book group) or treatment (physical copy group).
- Choosing/Assigning the Reading: We will select a book that has a standard text and format for both physical and e-book version and make sure both groups are reading the same version of the same book just on different mediums.
- Reading Phase: We will instruct all the participants to read the book at the same time, and in the same place with same conditions, for example a quiet room with same lighting and noise level.
- Assessment Phase: After designated time or when everyone is done reading we will give a same test, that we design to test about specific aspects of the book and its plot, to all the participants to be able to assess their comprehension and retention of the reading material.
- Data Summarization / Collection: We will collect all the tests and their results as well as any other data we want to collect on the participants such as their reading speed, genre preferences, or other prior reading habits. We will summarize this data and prepare it for further analysis

6. Data Analysis:

- We will first find all the usual descriptive statistics as we did in this homework such as means and standard deviations and common patterns that we can find in the comprehension and retention scores in both groups.
- After that we will use inferential statistics to compare the two groups, for instance we can assume that we want to keep our experiment cheap and simple so we will not have many participants so we can use a two-sample t-test to see if there is a statistically significant difference between the print book and e-book groups in terms of comprehension and retention scores.

7. Interpretation of Results:

- Finally, we will interpret the results in the terms of causal inference by making conclusions about the causal relationship between the treatment (reading format - print vs. e-book) and the outcomes (comprehension and retention). We will use our statistical analysis to make this conclusions about significance of our results but we will also take into consideration practical significance to come with best conclusions and recommendations of best reading formats (comprehension and retention wise). For instance, if the print book group demonstrates statistically and practically significantly better comprehension and retention compared to the e-book group, we can conclude that reading a physical print book has a causal effect on these outcomes.

Why to avoid observational data:

- One of the main reasons that we could not use observational data in our experiment is due to self-selection bias. In real world people would choose the format that is more convenient for them for example, a person with a severely cracked phone screen would most likely choose to read a paper book regardless of the true preference and person with a Kindle would choose the e-book without knowing if they comprehend the reading better that way. Further, even if the assignment was random, people might (and in real life will) have inherent differences affecting comprehension and retention such as pre-existing reading habits, personal preferences (CS/DS students versus literature students) or environment they are reading in.

Causal Treatment Effect:

Our experiment will ensure the following

- There is distinguishable and comparable treatment groups - those reading the physical print copy versus those reading an e-book. This will ensure that any differences we find in comprehension and retention can be attributed to the reading formats.

- The assignment to these treatment groups will be random which will allow for no self-selection bias and will ensure that observed differences in comprehension and retention are not due to any possible pre-existing preferences of the participants but only due to the format.

- Controlled environment both in the means of the reading the same book as well as surrounding in which it is being read will minimize external effects or possible confounders that could cause some to comprehend the book less well giving everyone same conditions.

- Finally, our statistical analysis will allow us to objectively and rigorously assess which group (if any) performed significantly better and thus led to better comprehension and retention due to the book format.

Ensuring all of these aspects of our experiment should (and hopefully will) establish a causal treatment effect between the reading format and comprehension/retention outcomes. The controlled experiment such as this one ensures that differences between groups are more likely due to the format and not any previously mentioned external factors.