Problem Set 1

Erin Choi - eyc321 - Section 2

Due Oct 7th, 2022

Definitions and Examples

1. The fundamental problem of causal inference is that in the real world, not all potential outcomes can occur; it is impossible to give any single unit both the treatment and the control. For example, if an experiment is trying to examine the effect of giving plants plant food, a single plant cannot be both given and not given plant food at the same time in one world. In reality, only one of the potential outcomes will actually occur and be observable and thus measurable - a plant can either be given plant food or not. This means it is impossible to directly compare the different results between when a unit is given the treatment or when it is assigned the control.

2. Experiments are important because by definition, they allow the researcher to control the probability distribution of treatment assignment so that it satisfies the assumptions that make the average treatment effect (ATE) identified. It is necessary to identify estimation of the ATE because, as stated in the previous answer, it is impossible to know the exact effect that the treatment has on any single unit. Having a research control treatment assignment automatically satisfies two assumptions that allow us to identify the ATE: ignorability and positivity. In other words, experiments make sure treatment assignment is independent of potential outcomes, and the probability of units being assigned treatment or control are both non-zero positive values.

3. Ignorability means that the treatment assignment process doesn't depend on potential outcomes. When the researcher determines which units are given the treatment and control, all other characteristics of the units become ignorable - thus, ignorability. It can also be called exchangeability, as the treatment and control groups that result from this process should be similar enough that the researcher can just switch, or exchange, them. For example, in an experiment that examines the effect of adding self-checkout kiosks on store reviews, letting customers choose their method of checkout could violate ignorability, as the kinds of customers who select self-checkout could systemically differ from those who choose normal checkout, causing selection bias to occur. For the ignorability assumption to be satisfied in this case, a coin should be flipped for each customer to randomly determine which checkout method they would be directed to.

4. SUTVA stands for the "Stable Unit Treatment Value Assumption." It consists of two assumptions - no spillover or interference, and use of a single version of the treatment. Spillover refers to units interacting with each other and having effects oneach other's outcomes that are not intended by the experiment itself. In an experiment that studies the effect of extra tutoring on students' exam scores, students in the treatment group sharing knowledge from their tutoring session with students in the control group would cause a spillover effect that violates SUTVA. A single version of treatment means there should not be variations of treatment given within a treatment group. In an experiment that studies the effect of COVID vaccines on COVID symptom severity, having a single, general COVID vaccine treatment group would violate SUTVA, as there are multiple brands of the vaccine that vary from each other.

Application (Coding) - STAR Project

Question 1

```
# import libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(haven)

# read in data
STAR <- read.csv("STAR.csv")

# create kinder variable, recode classtype to labels
STAR <- STAR %>% mutate(kinder = case_when(classtype == 1 ~ "small",
                                           classtype == 2 ~ "regular",
                                           classtype == 3 ~ "regular with aid"))

# recode race to have 4 levels
STAR <- STAR %>% mutate(race = case_when(race == 1 ~ "white",
                                         race == 2 ~ "black",
                                         race == 4 ~ "hispanic",
                                         race == 3 ~ "others",
                                         race == 5 ~ "others",
                                         race == 6 ~ "others"))
```

Question 2

```
# small class
# mean reading score
small_reading = mean(STAR$g4reading[STAR$kinder == "small"], na.rm = TRUE)
# mean math score
small_math = mean(STAR$g4math[STAR$kinder == "small"], na.rm = TRUE)

cat(paste("Mean Small Reading Score:", small_reading,
          "\nMean Small Math Score:", small_math))
```

```
## Mean Small Reading Score: 723.391184573003
## Mean Small Math Score: 709.185135135135
```

```r
# regular class
# mean reading score
regular_reading = mean(STAR$g4reading[STAR$kinder == "regular"], na.rm = TRUE)
# mean math score
regular_math = mean(STAR$g4math[STAR$kinder == "regular"], na.rm = TRUE)

cat(paste("Mean Regular Reading Score:", regular_reading,
          "\nMean Regular Math Score:", regular_math))
```

```
## Mean Regular Reading Score: 719.88995215311
## Mean Regular Math Score: 709.521377672209
```

```r
# estimated effect size:
# take ATE between small and regular classes for reading and math,
# divide each by standard deviation of reading and math scores respectively
sd_reading = sd(STAR$g4reading[STAR$kinder == "small" |
                                   STAR$kinder == "regular"], na.rm = TRUE)
sd_math = sd(STAR$g4math[STAR$kinder == "small" |
                             STAR$kinder == "regular"], na.rm = TRUE)

# ATE for reading and math:
# difference between small and regular mean scores for each test
ate_reading = small_reading - regular_reading
ate_math = small_math - regular_math

# divide ATE for each test by its respective SD
ees_reading = ate_reading/sd_reading
ees_math = ate_math/sd_math

## Estimated Effect Sizes:
cat(paste("Estimated Effect Sizes: \nReading:", ees_reading, "\nMath:", ees_math))
```

```
## Estimated Effect Sizes:
## Reading: 0.0667763831818957
## Math: -0.00796392696218943
```

The mean reading score of students assigned to smaller kindergarten classes is about 723.39, which is somewhat higher than that of students assigned to regular-sized classes, 719.89. Meanwhile, the mean math score of students assigned to smaller classes is very slightly lower than that of their peers assigned to regular classes: 709.19 versus 709.52.

The estimated effect size found for each kind of test reflects these comparisons in the same way. The estimated effect size for the reading test is about 0.0668, indicating that students assigned to small kindergarten classes performed somewhat better on average than students assigned to regular-sized classes. The estimated effect size for the math test is about -0.0079; it is negative and has a much smaller magnitude than the estimated effect size for reading, showing that on average, students assigned to regular-sized classes performed better on the math test, but at a smaller margin (close to 0) than that seen in the reading test.

Question 3

```
# low score (33%) cutoff
# reading test
# small
small_reading_low = quantile(STAR$g4reading[STAR$kinder=="small"],
                             probs=0.33, na.rm = TRUE)
# regular
reg_reading_low = quantile(STAR$g4reading[STAR$kinder=="regular"],
                           probs=0.33, na.rm = TRUE)

# math test
# small
small_math_low = quantile(STAR$g4math[STAR$kinder=="small"],
                          probs=0.33, na.rm = TRUE)
# regular
reg_math_low = quantile(STAR$g4math[STAR$kinder=="regular"],
                        probs=0.33, na.rm = TRUE)

# high score (66%) cutoff
# reading test
# small
small_reading_high = quantile(STAR$g4reading[STAR$kinder=="small"],
                              probs=0.66, na.rm = TRUE)
# regular
reg_reading_high = quantile(STAR$g4reading[STAR$kinder=="regular"],
                            probs=0.66, na.rm = TRUE)

# math test
# small
small_math_high = quantile(STAR$g4math[STAR$kinder=="small"],
                           probs=0.66, na.rm = TRUE)
# regular
reg_math_high = quantile(STAR$g4math[STAR$kinder=="regular"],
                         probs=0.66, na.rm = TRUE)

# Low Score Cutoffs
cat(paste("Low Score Cutoffs: \nReading, Small Class:", small_reading_low,
          "\nReading, Regular Class:", reg_reading_low,
          "\nMath, Small Class:", small_math_low,
          "\nMath, Regular Class:", reg_math_low))
```

```
## Low Score Cutoffs:
## Reading, Small Class: 705
## Reading, Regular Class: 705
## Math, Small Class: 694
## Math, Regular Class: 696
```

```
# High Score Cutoffs
cat(paste("High Score Cutoffs: \nReading, Small Class:", small_reading_high,
          "\nReading, Regular Class:", reg_reading_high,
          "\nMath, Small Class:", small_math_high,
          "\nMath, Regular Class:", reg_math_high))
```

```
## High Score Cutoffs:
## Reading, Small Class: 741
## Reading, Regular Class: 740
## Math, Small Class: 726
## Math, Regular Class: 724
```

For the reading test, the low to high score range is 705 to 741 for students assigned to small kindergarten classes and 705 to 740 for those assigned to regular-sized classes. There is a small difference of one point between the high-score cutoffs for the different class sizes, so overall, the ranges are almost identical.

For the math test, the low to high score range is 694 to 726 for students assigned to small classes and 696 to 724 for students assigned to regular classes. While the range for regular classes is slightly tighter than that for small classes, with 2 points excluded from the 33rd to 66th percentile range on either end, the ranges are also very similar overall.

Because all ranges are similar, this analysis only shows where the mean falls for student of each kindergarten class type. All mean scores, for both test and both class types, fall between the 33rd and 66th percentile, so the score distributions are not likely to be skewed.

If this quantile analysis yielded more different ranges between small and regular-sized classes, there could be more meaning added to the mean analysis. For example, the mean reading score is higher for students assigned to small classes than regular classes, but quantile analysis could have shown that the mean score for students from smaller classes falls in the low score range while the mean for those from regular classes falls in the middle range. This could indicate skewness in the small class scores and show that even though the mean is higher than that of regular classes, students assigned to smaller classes perform worse on average compared to their fellow small-class students.

Question 4

```
# compute mean reading test scores, regular classes
# white students
reg_reading_white = mean(STAR$g4reading[STAR$race == "white" & STAR$kinder == "regular"],
                         na.rm = TRUE)
# minority students (black, hispanic, not including others)
reg_reading_minority = mean(STAR$g4reading[(STAR$race == "black" | STAR$race == "hispanic") &
                                           STAR$kinder == "regular"], na.rm = TRUE)

# mean math test scores, regular classes
# white
reg_math_white = mean(STAR$g4math[STAR$race == "white" & STAR$kinder == "regular"],
                      na.rm = TRUE)
# minority
reg_math_minority = mean(STAR$g4math[(STAR$race == "black" | STAR$race == "hispanic") &
                                     STAR$kinder == "regular"], na.rm = TRUE)

# mean reading test scores, small classes
# white
small_reading_white = mean(STAR$g4reading[STAR$race == "white" & STAR$kinder == "small"],
                           na.rm = TRUE)
# minority students (black, hispanic, not including others)
small_reading_minority = mean(STAR$g4reading[(STAR$race == "black" | STAR$race == "hispanic") &
                                             STAR$kinder == "small"], na.rm = TRUE)
```

```r
# mean math test scores, small classes
# white
small_math_white = mean(STAR$g4math[STAR$race == "white" & STAR$kinder == "small"],
                        na.rm = TRUE)
# minority
small_math_minority = mean(STAR$g4math[(STAR$race == "black" | STAR$race == "hispanic") &
                                       STAR$kinder == "small"], na.rm = TRUE)

cat(paste("Regular Reading: White =", reg_reading_white, "vs. Minority =",
      reg_reading_minority,
      "\nDifference =", reg_reading_white-reg_reading_minority))
```

```
## Regular Reading: White = 725.11581920904 vs. Minority = 689.354838709677
## Difference = 35.7609804993621
```

```r
cat(paste("Regular Math: White =", reg_math_white, "vs Minority =",
      reg_math_minority,
      "\nDifference =", reg_math_white-reg_math_minority))
```

```
## Regular Math: White = 711.410364145658 vs Minority = 698.532258064516
## Difference = 12.8781060811422
```

```r
cat(paste("Small Reading: White =", small_reading_white, "vs Minority =",
      small_reading_minority,
      "\nDifference =", small_reading_white-small_reading_minority))
```

```
## Small Reading: White = 727.838815789474 vs Minority = 699.284482758621
## Difference = 28.5543330308529
```

```r
cat(paste("Small Math: White =", small_math_white, "vs Minority =",
      small_math_minority,
      "\nDifference =", small_math_white-small_math_minority))
```

```
## Small Math: White = 711.19001610306 vs Minority = 698.222222222222
## Difference = 12.9677938808375
```

For students assigned to regular classes, the racial gap was about 35.76 points in reading scores and about 12.88 points in math scores. For students assigned to small classes, the gap was about 28.55 points in reading scores and about 12.97 points in math scores.

There was a reduction of about 7 points in the reading score racial gap when students were assigned to small classes, while there was almost no difference in the math score racial gap - the gap for students assigned to small classes was actually about 0.09 greater than that for students assigned to regular classes. Thus, a small class size was effective for reducing the racial gap in reading scores, but not for math scores.

Question 5

```
# HS grad rates between class types
# small
small_grad = mean(STAR$hsgrad[STAR$kinder == "small"], na.rm = TRUE)
# regular
reg_grad = mean(STAR$hsgrad[STAR$kinder == "regular"], na.rm = TRUE)
# regular with aid
reg_aid_grad = mean(STAR$hsgrad[STAR$kinder == "regular with aid"],
                    na.rm = TRUE)

cat(paste("Graduation Rates \nSmall:", small_grad,
          "\nRegular:", reg_grad,
          "\nRegular with Aid:", reg_aid_grad))
```

```
## Graduation Rates
## Small: 0.835920177383592
## Regular: 0.825161887141536
## Regular with Aid: 0.839285714285714
```

The highest graduation rate comes from students assigned to regular-sized classes with aid, with a 83.93% graduation rate. Students assigned to small classes had a very similar graduation rate of 83.59%, and those assigned to regular classes with no aid had the lowest rate of 82.52%. There are only very small differences in graduation rates between students assigned to different class types in kindergarten.

```
# HS grad rates by number of years in small classes
# get all the possible values for number of years in small classes
all_yearssmall = sort(unique(STAR$yearssmall))
paste("Graduation Rates for")
```

```
## [1] "Graduation Rates for"
```

```
for (i in all_yearssmall) {
  i_grad = mean(STAR$hsgrad[STAR$yearssmall == i], na.rm = TRUE)
  cat(paste(i, "years in small classes:", i_grad, "\n"))
}
```

```
## 0 years in small classes: 0.82860203535083
## 1 years in small classes: 0.791044776119403
## 2 years in small classes: 0.813186813186813
## 3 years in small classes: 0.832460732984293
## 4 years in small classes: 0.877551020408163
```

The highest graduation rate, 87.76%, comes from students who spent 4 years in small classes, which is the highest number of years accounted for in this data. The lowest graduation rate, 79.10%, comes from students who spent only 1 year in a small class. Interestingly, students who spent 1 or 2 years in small classes had lower graduation rates than students who were never assigned to small classes at all. Overall, there are larger differences between graduation rates based on the number of years spent in small classes, and it appears that, while students who spent 0 years in small classes did not yield the lowest graduation rate, students who spent the most years in small classes certainly gave the highest graduation rate.

```r
# racial gap reduction of graduation rates based on class type?
# small classes
# white students
small_white_grad = mean(STAR$hsgrad[STAR$race == "white" & STAR$kinder == "small"],
                        na.rm = TRUE)
# minority students
small_minority_grad = mean(STAR$hsgrad[(STAR$race == "black" | STAR$race == "hispanic") &
                                         STAR$kinder == "small"], na.rm = TRUE)

# regular classes
# white students
reg_white_grad = mean(STAR$hsgrad[STAR$race == "white" & STAR$kinder == "regular"],
                      na.rm = TRUE)
# minority students
reg_minority_grad = mean(STAR$hsgrad[(STAR$race == "black" | STAR$race == "hispanic") &
                                       STAR$kinder == "regular"], na.rm = TRUE)

cat(paste("Graduation Rates for Regular Classes
          White:", reg_white_grad, "vs. Minority:", reg_minority_grad,
          "\nDifference:", reg_white_grad-reg_minority_grad))
```

```
## Graduation Rates for Regular Classes
##           White: 0.856962025316456 vs. Minority: 0.739583333333333
## Difference: 0.117378691983122
```

```r
cat(paste("Graduation Rates for Small Classes
          White:", small_white_grad, "vs. Minority:", small_minority_grad,
          "\nDifference:", small_white_grad-small_minority_grad))
```

```
## Graduation Rates for Small Classes
##           White: 0.867469879518072 vs. Minority: 0.74468085106383
## Difference: 0.122789028454243
```

Based on small versus regular kindergarten class sizes, there does not seem to be a reduction in the racial gap for graduation rates. The difference in graduation rates between white and minority students is greater for those assigned to small classes (12.28%) than those assigned to regular classes (11.74%). However, the graduation rates themselves are greater for both white and minority students from small kindergarten classes compared to those from regular-sized classes.

```r
# racial gap reduction of graduation rates based on years spent in small classes?
print("Graduation Rates for")
```

```
## [1] "Graduation Rates for"
```

```r
for (i in all_yearssmall) {
  i_grad_white = mean(STAR$hsgrad[STAR$yearssmall == i & STAR$race == "white"],
                      na.rm = TRUE)
  i_grad_minority = mean(STAR$hsgrad[STAR$yearssmall == i & (STAR$race == "black" |
                                                               STAR$race == "hispanic")], na.rm = TRUE)
  print(paste(i, "years in small classes: white =", i_grad_white,
              "vs. minority =", i_grad_minority))
  print(paste("Difference:", i_grad_white-i_grad_minority))
}
```

```
## [1] "0 years in small classes: white = 0.865497076023392 vs. minority = 0.726166328600406"
## [1] "Difference: 0.139330747422986"
## [1] "1 years in small classes: white = 0.811111111111111 vs. minority = 0.744186046511628"
## [1] "Difference: 0.0669250645994832"
## [1] "2 years in small classes: white = 0.84251968503937 vs. minority = 0.745454545454545"
## [1] "Difference: 0.0970651395848245"
## [1] "3 years in small classes: white = 0.878048780487805 vs. minority = 0.75"
## [1] "Difference: 0.128048780487805"
## [1] "4 years in small classes: white = 0.903073286052009 vs. minority = 0.782608695652174"
## [1] "Difference: 0.120464590399835"
```

Meanwhile, there is a difference in the racial gap for graduation rates as the number of years spent in smaller classes varies. The greatest difference in graduation rates between white and minority students (13.93%) occurs among students who spent no years in small classes. In comparison to 0 years, spending 3 or 4 years in small classes (differences of 12.8% and 12.05%, respectively) yields a small reduction of about 1-2% in the racial gap between graduation rates. The biggest reductions can be seen when comparing 0 years to 1 or 2 years spent in small classes. The racial gap in graduation rates is about 9.71% between white and minority students who spent 2 years in small classes, and the gap is only 6.69% between white and minority students who spent 1 year in small classes. While only comparing the sizes of students' kindergarten classes did not show a reduction in the graduation rate racial gap, comparing graduation rates based on number of years in small classes shows that the STAR program may have helped to reduce the racial gap somewhat, particularly for students assigned to small classes for 1 or 2 years.

Bed Nets and Malaria

1. $\mathbb{E}[Y_i|D_i = 0]$ is the mean observed outcome (whether a woman used a bed net or not) for women who were provided (or assigned) 90% cheaper bed nets.

2. $\mathbb{E}[Y_i(1)]$ is the mean potential outcome for a case in which all women were given free bed nets, regardless of whether they were actually given free or 90%-subsidized bed nets. This value is impossible to observe in reality, as not all women's observed outcomes would be the one associated with being assigned free bed nets.

3. $\mathbb{E}[Y_i(1)|D_i = 0]$ is the mean potential outcome for women if they had been provided free bed nets, given that they were actually provided 90% cheaper bed nets. This is a counterfactual: the value of the outcome for providing free bed nets would be occurring in a world in which 90% cheaper bed nets were provided for (or assigned to) any single woman instead. This is also an unobservable quantity.

4. Randomizing treatment at the clinic level, or using cluster randomization, is actually a good design choice to prevent spillover from occurring. Had Cohen and Dupas assigned treatment on an individual level, there could have been a violation of SUTVA if women at the same clinic had different bed nets and gave them to each other or swapped with one another. As long as women at one clinic cannot get bed nets from other clinics, the design does not violate the "no spillover" requirement of SUTVA. The design also satisfies the single-version-of-treatment assumption that is part of SUTVA, as there is only one version of the treatment - providing a full subsidy for bed nets for a treated clinic. The 90%-cheaper bed nets are the control, so they are not another version of the treatment. If they were, however, another version of the treatment that occurred at the same clinics that were also given free bed nets, this would be a violation of SUTVA as multiple versions of treatment would be occurring within the same cluster/level of randomization.

Let's Help a Small Business!

1. There may exist a correlation, but without an experiment, this relationship cannot be called causal. To identify this as a causal effect, randomization must occur to decide which individuals receive postcards and which do not. Then, there cannot be spillover between the people who receive postcards and who do not; people who receive postcards should not be able to affect people who don't in their decision to visit the restaurant. There should be a single version of the treatment in this experiment, which would be one version of the postcard with no variations between postcards sent to randomly chosen individuals. To satisfy ignorability, individuals should not be able to choose whether they receive a postcard or not; whether people choose to go to restaurants or not should be independent of the process of assigning the postcard-sending treatment to people. Finally, every person should have a positive chance of receiving or not receiving a postcard. There are also covariates that should be recorded, such as whether a person already knew about the restaurant before they visited during the experimental period.

2. To satisfy SUTVA, the experiment would involve sending postcards to households rather than individuals to prevent spillover from occurring between family members or housemates. Spillover could also occur via word of mouth between households and the Internet, so the experiment would be cluster-randomized; different neighborhoods/ZIP codes would be randomly selected to receive the treatment, and every household in those selected neighborhoods would receive a postcard in the mail. Again, the postcard itself is already a single version of the treatment as long as all postcards sent are identical, visually and content-wise. A coin flip would decide whether a neighborhood would be in the treatment or control group, so all households would have a 50/50 chance of receiving a postcard, satisfying positivity. By the definition of an experiment in which the researcher randomly assigns treatment and control to each unit, ignorability would also be satisfied.

   Before the postcards are sent out to neighborhoods in the treatment group, I would spend a month observing the proportion of households in each neighborhood that visit the restaurant at least once. After the postcards are sent out, I would again spend a month recording the proportion of households that went to the restaurant. During the month of the actual experiment, an individual from each household that goes to the restaurant would receive a survey after their meal that asks for their ZIP code (to assess which group they belong to), whether they've come to the restaurant before, and how they found out about the restaurant (with the postcard as one of the options). This additional information would give me data on covariates that I could control for.

   After a month, the assumptions that I mentioned were satisfied above would allow me to identify the causal effect being tested, and I would be able to estimate the ATE using a difference in means. I would calculate a difference in proportions of households that came to the restaurant before and during the experiment for each ZIP code. These differences in proportions would be averaged based on what group each neighborhood was assigned to, then I could estimate an ATE by taking the difference between the mean increase in proportion for treatment and control.

ATE

1. Estimate the ATE using the data in the table.

$$ATE_1 = E[Y|T=1] - E[Y|T=0]$$

$$E[Y|T=1] = \frac{60+75+53+69+50}{5} = 61.4$$

$$E[Y|T=0] = \frac{63+42+50+58+59}{5} = 54.4$$

$$ATE_1 = 61.4 - 54.4 = 7$$

2. Compute the dance-trainer-specific effect using the data in the table.

$$E[Y|T=1, DT=Yes] = \frac{60+75+53+69}{4} = 64.25$$

$$E[Y|T=0, DT=Yes] = \frac{63}{1} = 63$$

$$ATE_2 = E_{P(DT)}[64.25 - 63] = E_{P(DT)}[1.25] = 1.25$$

3. I believe ATE_2 estimates the ATE better. ATE_2 does not account for the dancers who are not dance trainers and does decrease the number of units for each group, particularly the control group. However, you can observe in the table of outcomes that dance trainers generally score higher than non-dance trainers, so there is less variation within each subgroup, split based on being a dance trainer or not (DT=yes versus DT=no). Even with the smaller sample size, I believe calculating ATE_2 that focuses only on dance trainers decreases the large variance that occurs in each group (treatment versus control) when calculating ATE_1 due to the difference in skill levels between those who are dance trainers and those who are not.

Design Your Experiment

"What is the effect of caffeine on short-term memory?"

This question cannot be answered based on observational data because you cannot control for people who people who regularly consume caffeine, and people may select into a certain group based on their existing behaviors.

The unit in this experiment would be an individual adult. The treatment would be one safe dose of caffeine, such as 150 mg. The outcome would be the individual's change in score on a memory task, such as an activity in which a person is shown some images to remember and is asked to recall them. This memory task would be completed once before the treatment or control is administered at the beginning of the day, then another time at the end of the day, around 8 hours later. Potential outcomes could be a big or small, negative or positive change in the memory task score.

Prior to the experiment, there would be a week during which participants would not consume caffeine. Each individual would report whether and how often they normally consume caffeine.

The experiment would use a stratified design. Since caffeine could affect people of different genders or ages differently, strata could be created based on gender and/or age ranges such as 18-27, 28-37, and so on. Strata could also be determined based on previous/regular caffeine consumption.

No spillover would occur because of the nature of the treatment and measurement method; people can't share caffeine with each other, nor can they give each other information about the memory task with others since the images they are each shown, asked to remember, and recall would vary randomly. There would be a single version of the treatment - 150 mg of caffeine, given to individuals in the same drink. (The control would be the same drink, non-caffeinated.) Thus both parts of SUTVA would be satisfied. Treatment assignment would be determined randomly within each strata by the flip of a coin, leaving no room for selection bias and satisfying ignorability. The coin flip would also give each person a 50/50 chance of being assigned to either the treatment or control group, satisfying positivity. The causal effect would be identified by satisfying all of the aforementioned necessary assumptions.

Once all data is collected, an ATE could be estimated for each strata by calculating the difference in means between treatment and control groups. If the strata were divided based on prior regular caffeine consumption versus no regular caffeine consumption, each of those strata would have a treatment and control group of about equal sizes. For each individual, the initial (pre-treatment) memory task score would be subtracted from the final (post-treatment) memory task score. Within each strata, the average difference in scores would be calculated for each group, then the control group's mean score difference would be subtracted from that of the treatment group to estimate the ATE of each block.