Problem Set 1

Note on Code Solutions

Your instructor is not a software engineer. If you solve these problems in a more efficient way, that's fantastic and highly encouraged. The code solutions provide *a* way to answer the questions, not *the* way.

Causal Inference Conceptual Questions

Question 1

a. Explain the notation "$Y_i(0)$" and "$Y_i(1)$."

The potential outcome for unit i under control and the potential outcome for unit i under treatment.

b. Explain and contrast the notation "$Y_i(0)|D_i = 1$" with the notation "$Y_i(0)|d_i = 1$."

The first equation is the untreated potential outcome for unit i that hypothetically receives treatment. The second equation is the untreated potential outcome for unit i if unit i actually receives treatment. Capital letter means hypothetical. Lowercase means realized draw.

c. Explain and contrast the meaning of "$E[Y_i(0)]$" with the meaning "$E[Y_i(0)|D_i = 1]$."

The first equation is the expectation of the potential outcome under control for all units. The second is the expectation of potential outcomes under control for a randomly selected unit who received treatment in a hypothetical allocation.

d. Explain why $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] = 0$ when D_i is randomly assigned.

Under randomization, the untreated potential outcomes for a unit who hypothetically receives treatment is the same as the untreated potential outcome for a unit that hypothetically does not. This follows from the fact that $E[Y_i(0)|D_i = 1] = E[Y_i(0)]$ and $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$ under random sampling from the entire set of potential outcomes.

Question 2

Assume that we conduct a randomized experiment. Define the ATT as $E[\tau_i|D_i = 1]$. Below is the proof of the following statement, "When units are randomly assignment to treatment and control, the ATT is equal to the ATE in expectation. Equivalently, $E[\tau_i|D_i = 1] = E[\tau_i]$." Explain substantively what each line in the following proof does and why. For notation purposes, the first m observations are the treatment group, and the remaining N-m observations are the control group.

1. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = E[\frac{\sum_i^m Y_i}{m}] - E[\frac{\sum_{m+1}^N Y_i}{N-m}]$
2. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = \frac{E[Y_1] + ... + E[Y_m]}{m} - \frac{E[Y_{m+1}] + ... + E[Y_N]}{N-m}$
3. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = E[Y_i(1)|D_i = 1] - E[U_i(0)|D_i = 0]$
4. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = E[Y_i(1)] - E[Y_i(0)]$
5. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = ATE$

Line 1 uses the linear property of expectations.

Line 2 uses the linear property of expectations to show that the sum of expectations is equal to the expectation of a sum.

Line 3 substitutes in equivalent equalities.

Line 4 substitutes in the expected value of the potential outcomes under treatment and control. We can do this because randomization makes the additional terms in the decomposition cancel out.

Line 5 is a restatement of the defintion of the ATE, which we show is equal

Question 3

a. What is a standard error? What is the difference between a standard error and a standard deviation?

The standard error is a measure of statistical uncertainty surrounding a parameter estimate. It is also the standard deviation of the sampling distribution of that parameter. The standard deviation is a measure of dispersion of any distribution. Usually we refer to the standard deviation as the dispersion around an observed variable.

b. How is randomization inference used to test the sharp null hypothesis of no effect for any unit?

The sharp null hypothesis says that treated and untreated potential outcomes are identical. To form the sampling distribution, we 1. simulate a random assignment and calculate the test statistic of interest, usually a difference in means and 2. repeat this for either all possible assignments, or a large random number of assignments if the former is unfeasible. The p-value of the test statistic that is observed in the actual experiment is calculated by finding its location in the sampling distribution under the sharp null hypothesis. We reject the sharp null based on our specified significance level, usually 0.05.

c. Define a 95% confidence interval.

A 95% confidence interval is the interval around a true parameter of interest that across hypothetical infinite replications of the experiment would bracket the true parameter with 95% probability.

d. How does complete random assignment differ from block random assignment and cluster random assignment?

Complete random assignment means that each unit is assigned separately to each treatment arm such that some number m of the N total units end up in each arm. Block random assignment means that complete random assignment occurs within each block. Clustered assignment means that groups of units are assigned jointly to a treatment arm. If one unit in the cluster is assigned to treatment, all others in the cluster are also assigned to treatment. The same holds if one unit is assigned to control.

e. Designs that assign the same number of units to treatment and controlled are known as "balanced designs." Explain at least one desirable statistical property of these designs.

There are several. 1) Under some conditions balanced designs lead to less sampling variability. 2) Estimated confidence intervals are likely to be conservative, so if we reject no effect we can have greater confidence in that assessment. 3) Regression is less likely to biased under a balanced design.

Question 4

Assume that the treatment effect is constant τ. Show that the $V[Y_i(0)] = V[Y_i(1)]$ and show that $\rho[Y_i(0), Y_i(1)] = 1$.

By the property of variances $V[Y_i(1)] = V[Y_i(0)+\tau] = V[Y_i(0)]$ *because for any constant c,* $\forall c \in \mathbb{R}, V[X+c] = V[X]$

For the second part

$$\rho[Y_i(0), Y_i(1)] = \frac{Cov[Y_i(0), Y_i(1)]}{\sigma[Y_i(0)]\sigma[Y_i(1)]}$$
$$\rho[Y_i(0), Y_i(1)] = \frac{Cov[Y_i(0), Y_i(0)+\tau]}{\sigma[Y_i(0)]\sigma[Y_i(1)]}$$
$$\rho[Y_i(0), Y_i(1)] = \frac{V[Y_i(0)]}{V[Y_i(0)]}$$
$$\rho[Y_i(0), Y_i(1)] = 1$$

**Here we substitute in definitions when appropriate. This is useful to know because it tells us something about how the potential outcomes are related.*

Question 5

Consider the abstract from Andrews, Delton, and Kline (2021) below and answer the first three questions we ask about every research design as if you were the authors:

"Disaster responses are political. But can citizens make useful disaster decisions? Potential obstacles are that such decisions are complex, involve public goods, and often affect other people. Theories of political decision-making disagree on whether these problems can be overcome. We used experimental economic games that simulate disaster to test whether people are willing and able to prevent disasters for others. Groups of players face a complex task in which options that might help vary in their riskiness. Importantly, although all options are reasonable, which option is most useful depends on the experimental condition. We find that players will pay to help, can identify which option is most useful across experimental conditions, and will pay to learn how best to help. Thus, players were able to make useful and costly decisions to prevent others from experiencing disaster. This suggests that, in at least some situations, citizens may be able to make good disaster decisions."

Question 6

Based on experience, your instructor has noticed that students who study at least 3 hours a day are more likely to get good scores on exams. I, therefore, recommend in my syllabus that students study for my class for at least 3 hours a day.

a. What is the implied treatment (D_i) of the underlying experiment?

The implied treatment is assigning students to study for at least 3 hours a day.

b. State the potential outcome under control and the potential outcome under treatment for the underlying experiment in words.

($Y_i(0)$) *is the potential outcome of exam scores for unit i when studying for less than three hours a day.* $Y_i(1)$ *is the potential outcome of exam scores for unit i when studying for at least three hours a day.*

c. Since your instructor noticed this pattern, we can consider this a design in which Oski assigned a particular realization of d_i to each student. Why might you be skeptical of the causal effect I'm claiming?

There is good reason to believe that students self select into study habits, and as a result have distinctive potential outcomes. In this case $E[Y_i(0)|D_i = 0] \neq E[Y_i(0)|D_i = 1]$ *and our comparisons of these two groups will be biased. It is almost certainly the case that having all students study three hours a day would not lead to a uniform increase in scores.*

R Conceptual Questions

Question 1

Sliver is running a special on its pizza of the day. Customers can get two 12" pizzas for the price of one 18". Is this a cost-effective deal for the consumer in terms of the amount of pizza for the price? Use R to justify your answer.

No, buy the single bigger pizza.

```
circleArea <- function(diameter){
  # A function to calculate the area of a circle
  # diameter = a numeric value
  return(pi*(diameter/2)^2)
}

large_pizza <- 18
small_pizza <- 12
print(2*circleArea(small_pizza) > circleArea(large_pizza))
```

```
## [1] FALSE
```

Question 2

Using R, determine how many of the following five numbers are divisible by 2 or 7: 2, 9, 14, 21, 69

Consider using the `ifelse` function. The easiest way to solve this problem is to use the modulo operator. In R, the modulo operator is `%%`.

Three of them, 2, 14, and 69

```
# Note that on purpose, for this solution you can do it by
# hand. 2, 14, and 69 are divided by either 2 or 7
isDivisble <- function(x){
  return(ifelse(x %% 2 == 0 | x %% 7 == 0, TRUE, FALSE))
}
vec <- c(2,9, 14, 21, 69)
map(vec, .f = isDivisble)%>%
  reduce(.f = sum)
```

```
## [1] 3
```

Question 3

(Taken from Project Euler)

If we list all the natural numbers below ten that are multiples of 3 or 5, we get 3,5,6, and 9. The sum of these multiples is 23.

Find the sum of all the multiples of 3 or 5 below 1000.

```
# A loop example
multiples <- function(vec){
  # A function to find all multiples of 3 or 5 in a vector
  # vec = a vector of natural numbers
  modulos = vector(length = length(vec))
  for(i in 1:length(vec)){
    modulos[i]  = dplyr::if_else(i %% 3 == 0 | i %% 5 == 0, i,0L)
  }
  return(sum(modulos))
}
```

```r
multiples(1:999)
```

```
## [1] 233168
```

```r
## Data Frame solution
## An alternative solution with a data frame
tibble(x = 1:999)%>%
  filter(x %% 3 == 0 | x %% 5 == 0)%>%
  summarise(result = sum(x))
```

```
## # A tibble: 1 x 1
##   result
##    <int>
## 1 233168
```

Applied Questions

Question 1

Oski needs some help working at the Botanical Garden. He asks six first-year students to donate either 30 minutes or 60 minutes to improve the space. Due to time constraints, he wants to decide on a procedure to allocate slots such that an equal number of students is in the 30 minute and 60-minute groups. Oski considers two ways to do this.

Method 1: Flip a fair coin for each student. If heads, assign them to the 30-minute group. If tails, assign them to the 60-minute group.

Method 2: Put three slips of paper with 30 minutes written on them and three slips of paper with 60 minutes written on them in a box. For each student, draw a slip of paper without replacement and assign the student to the number written.

a. Using R (setting the randomization seed to 2356) show why the first method has a problem that the second method does not.

The first method does not guarantee a balanced design, which in small samples will lead to lots of practical statistical problems. In this case, the imbalanced design means only one student is in the 30-minute group.

```
set.seed(2356)

allocation1 <- NULL
for(i in 1:6){
  allocation1[i] <- rbinom(1,1,prob = c(0.5, 0.5))
}
allocation1
```

```
## [1] 0 1 0 0 0 0
```

```
allocation2 <- sample(c(0,0,0,1,1,1), 6, replace = F)
allocation2
```

```
## [1] 1 0 1 0 1 0
```

b. Using R, change the number of subjects to 600 instead of six, adjusting the number of slips of paper in the box appropriately. Does anything change with Method 1? Which method do you prefer as N gets large?

As N increases, the first method leads to a much more balanced design though still not identical. The answer to preference requires a justification. My view would be the second design is better because we guarantee balance.

```
set.seed(2356)

allocation3 <- NULL
for(i in 1:600){
  allocation3[i] <- rbinom(1,1,prob = c(0.5, 0.5))
}
sum(allocation3)
```

```
## [1] 293
```

```
allocation4 <- sample(c(rep(0,300), rep(1,300)), 600, replace = F)
sum(allocation4)
```

```
## [1] 300
```

Question 2

A study program advertises that all students who participate will improve their grades more than they otherwise would without the program. Ten students participate in the program. Compared to a normalized baseline of 0, the scores for the treatment group are $\{1, 0, 0, 4, 3\}$ and the scores for the control group are $\{-5, 4, 7, -7, -4\}$.

State the sharp null hypothesis for this design. Using R (set your randomization start seed to 1234567), show the results of a randomization test of the sharp null hypothesis of no effect of the study program with a balanced design. Do you believe the program's claims of success?

No. A test of the Sharp Null hypothesis reveals that almost 21% of possible assignments would see a result as large even when we know there is no treatment effect. Our conventional level for rejection is 0.05, which is not met here.

I present two different example solutions here. One shows the process for an experiment of any size. The second shows how to get the exact distribution, which will only work when the number of possible treatment assignments is relatively small. For a balanced design with just 40 units, there are 137,846,528,820 possible treatment assignments. As a consequence, we lean heavily on the ideas of the first method when doing randomization inference in actual research.

```
set.seed(1234567)

## true outcomes
df <- tibble(
  id = 1:10,
  y0 = c(-5, 4, 7, -7, -4,1,0,0,4,3),
  y1 = y0,
  D = c(rep(0,5), rep(1,5))

)

######## User Created Functions ##########
get_treatment_assignment <- function(N){
  random_treat <- sample(
    x = c(rep(1, N/2),
        rep(0, N/2)),
    size = N,
    replace = F)
}

get_ate <- function(df, y1, y0, d){

  ## Get groups
  y1 <- df[[y1]][d == 1]
  y0 <- df[[y0]][d == 0]

  ## Conditional Expected Values
  E_Y1 <- mean(y1, na.rm = T)
  E_Y0 <- mean(y0, na.rm = T)

  # Return the difference in means
  return(E_Y1 - E_Y0)
}

sim_dm <- function(df, to, co){
  d <- get_treatment_assignment(nrow(df))
  get_ate(df, to, co, d = d)
```

```
}

######### Randomization Inference Test ###########
## Get the observed ATE
obs_ate <- mean(df$y1[df$D==1]) - mean(df$y0[df$D==0])
dm <- NULL
num_perms <- 10000
for(i in 1:num_perms){
  dm[i] <- sim_dm(df, "y1", "y0")
}

# Get p-value by comparing the number of simulations
# where the dm was at least as large as our observed value
# This is a one-sided test in this solution
sum(dm >= obs_ate)/num_perms
```

```
## [1] 0.2023
```

```
################## Exact Solution #################
## Alternatively, since this is a small experiment
## We can do it explicitly
## There are 252 unique treatment assignments


### Get all possible assignments
all_possible_assign <- matrix(nrow = 10000, ncol = 10)

for(i in 1:10000){
  all_possible_assign[i,] <- sample(c(rep(0,5), rep(1,5)), 10, replace = F)
}

### Since we know that 10000 > 252, we are going to have
### duplicates. Cut down to just the unique rows
unique_treats <- unique(all_possible_assign)

# Confirm that we have 252 assignments
dim(unique_treats)[1] == 252
```

```
## [1] TRUE
```

```
#### Randomization Inference ######
dm <- NULL

## Randomize over all possible assignments
for(i in 1:nrow(unique_treats)){
  dm[i] <- get_ate(df, "y1", "y0", d=unique_treats[i,])
}

## p-value
sum(dm >= obs_ate)/nrow(unique_treats) # 0.2063492
```

```
## [1] 0.2063492
```