Veronica Vargas

Homework 4

Due date: 10/22/2024

October 10, 2024

1 Introduction

This exercise sheet contains a series of problems designed to test and enhance your understanding of the topics covered in the course. Please ensure that you attempt all problems and provide detailed solutions where necessary. If you have any questions or need clarification, feel free to reach out your TA.

2 Exercises

Exercise 1: Fire VS Water

Read the **fireVswater pokemon.csv** available on your canvas page. The dataset presents the weight (Y) of a sample of water and fire pokemons (W) across the first $n_X = 7$ generations (X).

(a) Show with R if the groups (Water-Gen1, Fire-Gen1, Water-Gen2, ...) are balanced.

```
{r}
# are groups balanced?
fireVSwater_pokemon %>%
  group_by(generation, type) %>%
  summarise(
    count = n(),
    weight = mean(weight_kg, na.rm = T)) %>%
  ungroup()
```

	generation	type	count	weight
1	1	fire	8	50.87500
2	1	water	21	55.35714
3	2	fire	7	72.10000
4	2	water	12	53.08333
5	3	fire	3	41.30000
6	3	water	20	82.69500
7	4	fire	2	38.50000
8	4	water	7	60.37143
9	5	fire	10	56.78000
10	5	water	10	39.27000
11	6	fire	6	30.40000
12	6	water	7	54.78571
13	7	fire	3	80.43333
14	7	water	11	20.90909

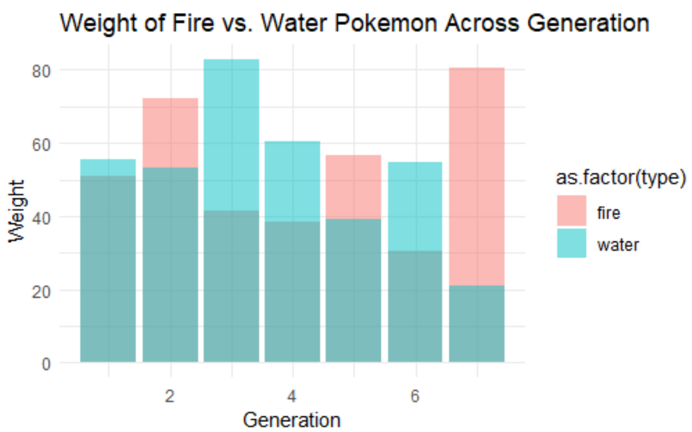
The subgroups within each generation are not balanced because they feature notably different covariate means except for those in the first generation.

(b) Compute the mean and the standard deviation for each sub-sample (Water-Gen1, Fire-Gen1,Water-Gen2, ...). Plot the evolution of the means for fire and water pokemons through generations and highlight any specific trend.

```
{r}
descriptive_pokemon <- fireVSwater_pokemon %>%
  group_by(generation, type) %>%
  summarise(
    count = n(),
    mean_weight = mean(weight_kg, na.rm = T),
    sd_weight = sd(weight_kg, na.rm = T)) %>%
  ungroup()
descriptive_pokemon

ggplot(descriptive_pokemon, aes(x = generation, y = mean_weight,
                                fill = as.factor(type))) +
  geom_bar(stat = "identity", position = "identity", alpha = 0.5,
binwidth = 0.5) +
  labs(title = "Weight of Fire vs. Water Pokemon Across Generation"
,
       x = "Generation",
       y = "Weight") +
  theme_minimal()
```

	generation	type	count	mean_weight	sd_weight
1	1	fire	8	50.87500	49.89757
2	1	water	21	55.35714	52.13879
3	2	fire	7	72.10000	86.99251
4	2	water	12	53.08333	72.89994
5	3	fire	3	41.30000	33.93626
6	3	water	20	82.69500	111.14580
7	4	fire	2	38.50000	23.33452
8	4	water	7	60.37143	122.21762
9	5	fire	10	56.78000	97.98916
10	5	water	10	39.27000	37.86406
11	6	fire	6	30.40000	27.20074
12	6	water	7	54.78571	69.13088
13	7	fire	3	80.43333	114.40919
14	7	water	11	20.90909	23.36277



Water type Pokémon weigh slightly more than fire type Pokémon on average. That said, fire type Pokémon weigh a lot more than water type Pokémon in the seventh generation.

(c) We want to study the difference of weight between fire and water types. The following estimator has been proposed: $\frac{1}{N_{fire}} \sum_{i:W_i=fire} Y_i - \frac{1}{N_{water}} \sum_{i:W_i=water} Y_i$. Describe in words the estimator and use R to apply it to the dataset.

This estimator is calculating the difference between the mean weight of fire Pokémon and the mean weight of water Pokémon. This estimator does not condition on generation.

```r
{r}
# estimator 1
fire_subsample <-
fireVSwater_pokemon[fireVSwater_pokemon$type ==
"fire",]
water_subsample <-
fireVSwater_pokemon[fireVSwater_pokemon$type ==
"water",]
estimator_1 <- mean(fire_subsample$weight_kg) -
mean(water_subsample$weight_kg)
estimator_1

[1] -1.528263
```

(d) Another estimator has been proposed: $\frac{1}{n_X}\sum_{x=1}^{n_X}(\hat{\mu}(fire, x) - \hat{\mu}(water, x$)). Describe the estimator in words and use R to apply it to the dataset.

This estimator is looking at the average difference in mean weight between fire type and water type Pokémon across all generations.

```r
{r}
# estimator 2

water_avg2 <- water_subsample %>%
  group_by(generation) %>%
  summarise(avg_weight_water = mean(weight_kg, na
.rm = TRUE)) %>%
  ungroup()

fire_avg2 <- fire_subsample %>%
  group_by(generation) %>%
  summarise(avg_weight_fire = mean(weight_kg, na
.rm = TRUE)) %>%
  ungroup()

estimator_2 <- mean(fire_avg2$avg_weight_fire -
water_avg2$avg_weight_water, na.rm = T)
estimator_2

[1] 0.5595176
```

(e) Another estimator has been proposed: $\frac{1}{n_{fire}}\sum_{i:W_i=fire}(\hat{\mu}(fire, X_i) - \hat{\mu}(water, X_i$)). Describe the estimator in words and use R to apply it to the dataset.

This estimator is calculating the difference in mean weight between fire type and water type Pokémon by multiplying it by the number fire type Pokémon in each generation and then averaging this value over the total number of fire type Pokémon across all generations. The purpose of this estimator is to calculate how much fire type Pokémon weighs on average compared to water type Pokémon.

```r
{r}
# estimator 3
water_avg3 <- water_subsample %>%
  group_by(generation) %>%
  summarise(avg_weight_water = mean(weight_kg, na.rm = TRUE),
            n_water = n()) %>%
  ungroup()

fire_avg3 <- fire_subsample %>%
  group_by(generation) %>%
  summarise(avg_weight_fire = mean(weight_kg, na.rm = TRUE),
            n_fire = n()) %>%
  ungroup()

est3_df <- full_join(water_avg3, fire_avg3, by = "generation") %>%
  mutate(est3 = est3_df$n_fire*(est3_df$avg_weight_fire -
est3_df$avg_weight_water))

estimator_3 <- (1/(nrow(fire)))*sum(est3_df$est3)
estimator_3

[1] 3.504875
```

(f) Compare the estimates obtained in (c), (d), and (e). Are they the same? If they are not tryto deduct why (*Hint: (a) and plotting the data could be useful*).

The estimators in (c), (d), and (e) are all different from each other. The estimate in (e) is much larger than the estimates in (c) and (d). This is because there are covariates in the data and none of the groups are balanced. Therefore, there is a covariate shift. Furthermore, each estimator is measuring something different. Estimator 1 measures the difference in mean weight between fire type and water type Pokémon without conditioning on generation. Estimator 2 measures the difference in mean weight between fire type and water type Pokémon by conditioning on the generation. Finally, estimator 3 measures on average how the weight of fire type Pokémon compare to the weight of water type Pokémon across all generations.

(g) Using bootstrapping create a confidence interval for the estimate resulting from estimator (e).

```
{r}
set.seed(42)

n <- 10000
diff_boot <- rep(NA, n)

for(i in 1:n){

  sample_boot <- fireVSwater_pokemon[sample(1:nrow
(fireVSwater_pokemon), nrow(fireVSwater_pokemon), replace = TRUE),]
  fire_boot <- sample_boot[sample_boot$type == "fire",]
  water_boot <- sample_boot[sample_boot$type == "water",]

  water_avg <- water_boot %>%
    group_by(generation) %>%
    summarise(avg_weight_water = mean(weight_kg, na.rm = TRUE),
n_water = n()) %>%
    ungroup()

  fire_avg <- fire_boot %>%
    group_by(generation) %>%
    summarise(avg_weight_fire = mean(weight_kg, na.rm = TRUE),
n_fire = n()) %>%
    ungroup()

  est3_df <- full_join(water_avg, fire_avg, by = "generation") %>%
    mutate(est3 = n_fire * (avg_weight_fire - avg_weight_water))

  diff_boot[i] <- (1/39) * sum(est3_df$n_fire *
(est3_df$avg_weight_fire - est3_df$avg_weight_water), na.rm = TRUE)
}
```

```
{r}
quantile(diff_boot, c(0.025, 0.975))
abs(quantile(diff_boot, 0.975) - quantile(diff_boot, 0.025))
```

```
     2.5%     97.5%
-23.23957  33.29735
    97.5%
 56.53692
```

Exercise 2: Village Leaders and Sanitation Budget

In 2004, Chattopadhyay and Duflo studied the budgetary consequences of having women, rather than men, lead village councils in West Bengal and Rajasthan, India. Each village either receives treatment (has a woman council leader) or is untreated (has a man council leader), then we observe the share of the local council budget that is allocated to providing drinking water. For the purposes of this problem, pretend we can observe the potential outcomes of the villages—how much of their budget is allocated toward water sanitation—when their council is led by a woman ($Y_i(1)$) or a man ($Y_i(0)$). These potential outcomes are shown in the table below, as well as columns for true (unobserved) individual treatment effects, τ_i, an indicator of whether or not a village receives treatment (D_i), and the Y_i we observe based on this treatment status.

Village	$Y_i(0)$	$Y_i(1)$	τ_i	D_i	Observed Y_i
Village 1	10	15	5	1	15
Village 2	15	15	0	0	15
Village 3	20	30	10	0	20
Village 4	20	15	-5	0	20

Village 5	10	20	10	0	10
Village 6	15	15	0	0	15
Village 7	15	30	15	0	15
Average	15	20	5	0.14	15.71

(a) Fill in the table with each village's individual treatment effect τ_i and add its average to the bottom of the table.

(b) Using your omniscient powers of observing all potential outcomes, what is $\frac{1}{n}\sum_{i=1}^{n} Y_i(1)$? What is $\frac{1}{n}\sum_{i=1}^{n} Y_i(0)$? What is $\frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$

The first estimator is looking at the mean budget allocation if all the villages were treated, which is equal to 20. The second estimator is looking at the mean budget allocation if none of the villages were treated, which is equal to 15. Finally, the last estimator is looking at the average treatment effect by looking at the difference between the treatment and the control groups.

(c) It appears that only Village 1 was assigned to treatment. Fill in the observed Y_is in the table based on this treatment assignment.

(d) What estimator will you use to calculate the average treatment effect estimate using observed Y_is? Calculate your estimate.

I will use the third estimator to calculate the average treatment effect because it calculates the difference in mean budget allocation between villages that were treated and villages that were untreated. Then I can simply do this for observed Yis.

Q 2d) $(1/n)\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\} = (1/n)\sum_{i=1}^{n} Y_i(1) - (1/n)\sum_{i=1}^{n} Y_i(0)$

$= (15/1) - (15+20+20+10+15+15/6)$

$\boxed{\hat{ATE} = -0.8\bar{3}}$

(e) Is your estimator from part (d) causally identified? Why or why not?

My estimator from part (d) is not causally identified because the treatment is not randomized, and it is not conditioned on any covariate variable. Therefore, this estimator is missing additional information from which to draw a more accurately representative estimate.

Exercise 3: Conditional Randomization Practice

Below is a table of data from an experiment testing a new diet dog food on chihuahua weight. The outcome of interest, Y_i, is the chihuahua's weight in pounds after the experiment. T_i indicates whether the chihuahua was treated with the new diet food (1) or not (0); this was randomized conditioned on the health of the chihuahua at the beginning of the study (Z_i). We also collected information about the chihuahua's fur type, X_i, which we coded as 1 if the chihuahua had long hair and 0 if it had short hair.

Chihuahua	Z_i	X_i	Y_i	T_i
1	Obese	0	8	1
2	Obese	1	6	1
3	Obese	1	6	0
4	Overweight	0	5	1
5	Overweight	1	6	1
6	Overweight	0	6	0

7	Overweight	1	7	0
8	Overweight	0	5	0
9	Healthy Weight	0	4	1
10	Healthy Weight	1	3	1
11	Healthy Weight	0	4	1
12	Healthy Weight	0	5	0
13	Healthy Weight	1	3	0
14	Healthy Weight	1	3	0

(a) Calculate the $\widehat{CATE}$ for each level of X_i. Calculate $\hat{\pi}(X)$ for each level of X_i.

a) $\widehat{CATE}_{X=1}$: $(1/3)(6+6+3)-(1/4)(6+7+3+3)=\boxed{0.25}$

$\hat{\pi}(X=1)=(3/7)$ $\hat{\pi}(X_1=0)=(4/7)$

$\widehat{CATE}_{X=0}$: $(1/4)(8+5+4+4)-(1/3)(6+5+5)=\boxed{-0.083}$

$\hat{\pi}(X_0=1)=(4/7)$ $\hat{\pi}(X_0=0)=(3/7)$

(b) Are these estimators $\widehat{CATE}_X$ causally identified? Why or why not? If so, calculate a standardized $\widehat{CATE}_X$.

These estimators are not causally identified because our sample was not randomly conditioned on fur type.

(c) Calculate the $\widehat{CATE}$ for each level of Z_i. Calculate $\hat{\pi}(Z)$ for each level of Z_i.

c) $\widehat{CATE}_{z=\text{Healthy Weight}}$: $(1/3)(4+3+4)-(1/3)(5+3+3)=\boxed{0}$

$\hat{\pi}(Z_{HW}=1)=(1/2)$ $\hat{\pi}(Z_{HW}=0)=(1/2)$

$\widehat{CATE}_{z=\text{overweight}}$: $(1/2)(5+6)-(1/3)(6+7+5)=\boxed{-0.5}$

$\hat{\pi}(Z_{ov}=1)=(2/5)$ $\hat{\pi}(Z_{ov}=0)=(3/5)$

$\widehat{CATE}_{z=\text{obese}}$: $(1/2)(8+6)-(1/1)(6)=\boxed{1}$

$\hat{\pi}(Z_O=1)=(2/3)$ $\hat{\pi}(Z_O=0)=(1/3)$

(d) Are these estimators $\widehat{CATE}_Z$ causally identified? Why or why not? If so, calculate a standardized $\widehat{CATE}_Z$.

These estimators are causally identified because our sample was randomly conditioned on the health of the chihuahua at the beginning of the study.

(e) Calculate the difference in sub-sample mean weight among treated and untreated chihuahuas(i.e. $\hat{\mu}(T=1)-\hat{\mu}(T=0)$). Is this estimator causally identified? Why or why not?

Q3e) $\hat{\mu}(T=1)-\hat{\mu}(T=0)$

$=(1/7)(8+6+5+6+4+3+4)-(1/7)(6+6+7+5+5+3+3)=\boxed{0.14}$

This estimator is not causally identified because treatment is not random, and it is not being conditioned for any potential covariates.

Exercise 4: Voter Turnout Experiment Analysis

Use the data "GGLsample.csv" for a sample of the data from Gerber, Green, and Larimer's voter turnout experiment. Each voter was marginally randomized into treatment or control. The relevant variables are sex, which records the individual's sex; treatment, which records whether the individual received the 'Neighbors' treatment– a postcard indicating their neighbors' voting records and that their neighbors will be informed of whether they vote in the upcoming election– or control; bintreat, a binary recoding of treatment; and voted, the outcome of interest– whether the individul voted in the election following the experiment.

(a) Estimate the $\widehat{CATE}_{male}$. Is this estimator causally identified? Why or why not?

```
{r}
cate_male <- cate_subsample %>%
  filter(sex == "male")
cate_male
```

A tibble: 1 × 6

sex <chr>	N_Treated <int>	N_Control <int>	Mean_Treated <dbl>	Mean_Control <dbl>	CATE <dbl>
male	413	2099	0.37046	0.308242	0.06221803

This estimator is not causally identified. While this estimator adjusts for the relevant covariate of sex, therefore reducing the bias in the assignment of treatment, there are still too many other variables to adjust for such as education and income. Therefore, there is too much information we do not know. As such, we cannot determine whether the treatment is truly randomized and our estimator is not causally identified.

(b) Use a bootstrap estimated sampling distribution to create a 95% confidence interval for the$CATE_{male}$ and interpret your interval in plain language.

```
{r}
quantile(diff_boot, c(0.025, 0.975))
abs(quantile(diff_boot, 0.975) - quantile(diff_boot, 0.025))

      2.5%      97.5%
0.01280856 0.11412336
     97.5%
0.1013148
```

If we repeated our *CATE* estimator many times with a sufficiently large sample size, our intervals constructed this way would include a slight increase between 0.013 and 0.114 voting outcome for males when conditioning on sex approximately 95% of the time.

(c) Estimate the $\widehat{CATE}_{female}$. Is this estimator causally identified? Why or why not?

```
{r}
cate_female <- cate_subsample %>%
  filter(sex == "female")
cate_female
```

A tibble: 1 × 6

sex <chr>	N_Treated <int>	N_Control <int>	Mean_Treated <dbl>	Mean_Control <dbl>	CATE <dbl>
female	415	2073	0.3783133	0.2855765	0.09273679

This estimator is not causally identified. While this estimator adjusts for the relevant covariate of sex, therefore reducing the bias in the assignment of treatment, there are still too many other variables to adjust for such as education and income. Therefore, there is too much information we do not know. As such, we cannot determine whether the treatment is truly randomized, and our estimator is not causally identified.

(d) Use a bootstrap estimated sampling distribution to create a 95% confidence interval for the$CATE_{female}$ and interpret your interval in plain language.

```
{r}
quantile(diff_boot, c(0.025, 0.975))
abs(quantile(diff_boot, 0.975) - quantile(diff_boot, 0.025))

       2.5%      97.5%
0.04245478 0.14427151
     97.5%
0.1018167
```

If we repeated our *CATE* estimator many times with a sufficiently large sample size, our intervals constructed this way would include a slight increase between 0.04 and 0.144 voting outcome for females when conditioning on sex approximately 95% of the time.

(e) Estimate the $\widehat{ATE}$. Is this estimator causally identified? Why or why not?

```
{r}
GGLsample <- GGLsample %>%
  mutate(binvote = recode(voted,
                          "Yes" = 1,
                          "No" = 0))

mean_treated <- mean
(GGLsample$binvote[GGLsample$bintreat ==
1])
mean_control <- mean
(GGLsample$binvote[GGLsample$bintreat ==
0])

# ATE estimate
ate <- mean_treated - mean_control

cat("ATE = ", round(ate, 2), "units")

ATE =  0.08 units
```

This estimator is not causally identified because it does not adjust for covariates. Therefore, the treatment in this estimator is not random and cannot be causally identified.

(f) Use a bootstrap estimated sampling distribution to create a 95% confidence interval for the *ATE* and interpret your interval in plain language.

```
{r}
set.seed(42)

n <- 10000
diff_boot <- rep(NA, n)

for(i in 1:n){
  sample_boot<-  GGLsample[sample(1:nrow(GGLsample), nrow(GGLsample), replace = T),]
  mean_treated <- mean(sample_boot$binvote[sample_boot$bintreat == 1], na.rm = TRUE)
  mean_control <- mean(sample_boot$binvote[sample_boot$bintreat == 0], na.rm = TRUE)

  diff_boot[i] <- mean_treated - mean_control
}
```

```
{r}
quantile(diff_boot, c(0.025, 0.975))
abs(quantile(diff_boot, 0.975) - quantile(diff_boot, 0.025))

       2.5%      97.5%
0.04257099 0.11356279
     97.5%
0.0709918
```

If we repeated our *ATE* estimator many times with a sufficiently large sample size, our intervals constructed this way would include a slight increase between 0.04 and 0.114 voting outcome approximately 95% of the time.

3 Submission Instructions

Please submit your completed exercises by **October 22** through **gradescope**. Ensure that your solutions are well-organized, clearly written, and include all necessary calculations and explanations. Questions about submission should be directed to your TA.

4 Helpful Resources

To better assist you in the completion of this exercise sheet, we suggest you to review the following material:

- **Lecture 3 -** bootstrapping estimated sampling distributions and confidence intervals
- **Lecture 6 -** estimators
- **Lecture 9 -** covariate shift and different estimator options;
- **Lecture 10 -** causal identification for marginally and conditionally randomized experiments;
- **Lab -** practicing all of the above