HW4

Vivian Bui

9/18/2021

PROBLEM 1:

a. Load the data set as a data.frame and create a single factor variable treatment that takes the value control if receiving only encouragement, safebox if receiving a safe box, and lockbox if receiving a locked box. How many individuals are in the control group? How many individuals are in each of the treatment arms? Ans:

There are 111 individuals in the control group, 117 individuals in the group received a safe box, and 195 individuals received a lock box.

```
#Load data
data.frame <- read.csv("https://raw.githubusercontent.com/dpuelz/Policy-Research-Laboratory/main/data/r

#Summary
dim(data.frame)
```

```
## [1] 423   9
```

```
summary(data.frame)
```

```
##        X           bg_female        bg_married       bg_b1_age
##  Min.   :  1.0   Min.   :0.0000   Min.   :0.0000   Min.   :17.00
##  1st Qu.:106.5   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:30.00
##  Median :212.0   Median :1.0000   Median :1.0000   Median :38.00
##  Mean   :212.0   Mean   :0.7447   Mean   :0.7541   Mean   :39.61
##  3rd Qu.:317.5   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:48.00
##  Max.   :423.0   Max.   :1.0000   Max.   :1.0000   Max.   :88.00
##
##  encouragement       safe_box        locked_box    fol2_amtinvest
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :   0.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:  40.0
##  Median :0.0000   Median :0.0000   Median :0.000   Median : 100.0
##  Mean   :0.2624   Mean   :0.2766   Mean   :0.461   Mean   : 322.2
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.: 490.0
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :5700.0
##                                                    NA's   :30
##  has_followup2
##  Min.   :0.0000
##  1st Qu.:1.0000
##  Median :1.0000
##  Mean   :0.9291
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

```
#Assign value for treatment var
data.frame$treatment <- NA
data.frame$treatment[data.frame$encouragement == 1] <- 'control'
data.frame$treatment[data.frame$safe_box == 1] <- 'safebox'
data.frame$treatment[data.frame$locked_box ==1] <- 'lockbox'

#Count
tab_treat = table(data.frame$treatment)
print(tab_treat)
```

```
##
## control lockbox safebox
##     111     195     117
```

b. Subset the data so that it contains only participants who were interviewed in 12 months during the second followup. We will use this subset for the subsequent analyses. How many participants are left in each group of this subset? Does the drop-out rate differ across the treatment conditions? What does this result suggest about the validity of this study? Ans:

In the second followup, 9 participants left the control group, 10 people left the group who received a safe box, and 11 people left the group who received a lock box.

The drop-out rate does differ across the treatment conditions: 8.11% for the control group, 8.54% for the safe box group, and 5.64% for the lock box group.

However, this result is not enough for us to evaluate and make connection to whether or not the study is valid. The influence of dropout rate in randomized control trials is reflected upon the type of missingness (i.e. whether or not the reasons of dropouts directly related to the study and susbtantially affect the possible outcomes), not the difference in rates. Suggested by Bell, Kenward, Fairclough, and Horton (link: https://www.bmj.com/content/346/bmj.e8668), 'equal dropout rates between treatment arms in a randomised controlled trial do not imply that estimates of treatment effect are unbiased'; whereas 'unequal dropout rates do not imply that estimates are biased'.

At the same time, higher chances that the people who attend the second follow up are individuals who have higher willingness to continue the study. To this end, the study lacks its ability to represent the total population, or in other words, the study has the sample bias and lack of external validity. More information might be needed to fully draw the conclusion between drop out rates and the total validity of the study.

```
followup <- subset(data.frame, data.frame$has_followup2 == 1)

#Count number of people in each group participate the follow up
table(followup$treatment)
```

```
##
## control lockbox safebox
##     102     184     107
```

```
#Number of people left in each group
left = table(data.frame$treatment) - table(followup$treatment)
print(left)
```

```
##
## control lockbox safebox
##       9      11      10
```

```
#Drop-out rate
(left[1]*100)/ tab_treat[1]
```

```
##  control
## 8.108108
```

```
(left[2]*100)/tab_treat[2]
```

```
##  lockbox
## 5.641026
```

```
(left[3]*100)/tab_treat[3]
```

```
##  safebox
## 8.547009
```

c. Does receiving a safe box increase the amount invested in health products? We focus on the outcome measured 12 months from baseline during the second follow-up. Compare the mean of amount (in Kenyan shilling) invested in health products fol2__amtinvest between each of the treatment arms and the control group. Briefly interpret the result. Ans:

Yes, the group who received safe box does see an increase in the invested amount for health products. The result shows that individuals received safe boxes invested 150.38 Kenya shilling more compared to the control group. While the group who received lock boxes invested only 49.99 Kenya shilling more compared to the control group.

```
#Get mean investment of each group during second followup
invest <- tapply(followup$fol2_amtinvest, followup$treatment, mean, na.rm=TRUE)

#Get mean investment of the control group during second followup
invest_control = mean(followup$fol2_amtinvest[followup$treatment == "control"])

#Difference
invest[-1] - invest_control
```

```
##   lockbox   safebox
##  49.99275 150.38162
```

d. Examine the balance of pre-treatment variables, gender (bg__female), age (bg__b1__age) and marital status (bg__married). Are participants in the two treatment groups different from those in the control group? What does the result of this analysis suggest in terms of the internal validity of the finding presented in the previous question? Ans:

There are differences between the participants from each group vs. control group.

Overall, both the groups received lock box and safe box treatment have more female, married, and younger participants compared to the control group. Compared to the control group, treatment groups have less than 2.7-4% in terms of female participants, -0.415%-1.63% in terms of married participants, and 2-3 ages difference in the average age of participants. These differences, though seems small, indicate that there is potential bias in our analysis that influence the internal validity - or the causal assumption we draw - from our study. For instance, there possibly a chance that group received safe box invested more not because of the difference of the treatment itself, but because of the difference in gender (more female) or marital status (which might link to the difference in financial situation) of the participants.

```
#Get pre-treatment info from each group
female = tapply(data.frame$bg_female, data.frame$treatment, sum)
age = tapply(data.frame$bg_b1_age, data.frame$treatment, mean)
married = tapply(data.frame$bg_married, data.frame$treatment, sum)

#Get pre-treatment info from the control group
female_control = sum(data.frame$bg_female[data.frame$treatment == "control"])
age_control = mean(data.frame$bg_b1_age[data.frame$treatment == "control"])
married_control = sum(data.frame$bg_married[data.frame$treatment == "control"])

#Mean difference between each group vs. control group
diff_female = tapply(data.frame$bg_female, data.frame$treatment, mean)
- mean(data.frame$bg_female[data.frame$treatment == "control"])
```

```
## [1] -0.7207207
```

```
diff_age = age - age_control
diff_married = tapply(data.frame$bg_married, data.frame$treatment, mean)
- mean(data.frame$bg_married[data.frame$treatment == "control"])
```

```
## [1] -0.7477477
```

```
#Output
print("Examine the balance of pre-treatment for each group:")
```

```
## [1] "Examine the balance of pre-treatment for each group:"
```

```
print(female)
```

```
## control lockbox safebox
##      80     146      89
```

```
print(age)
```

```
##  control  lockbox  safebox
## 41.62162 39.43590 37.98291
```

```
print(married)
```

```
## control lockbox safebox
##      83     149      87
```

```
print("Difference between each group vs. control group:")
```

```
## [1] "Difference between each group vs. control group:"
```

```
print(diff_female) #Proportion diff
```

```
##   control   lockbox   safebox 
## 0.7207207 0.7487179 0.7606838
```

```
print(diff_age) #Mean age diff
```

```
##   control   lockbox   safebox 
##  0.000000 -2.185724 -3.638716
```

```
print(diff_married) #Proportion diff
```

```
##   control   lockbox   safebox 
## 0.7477477 0.7641026 0.7435897
```

e. Does receiving a safe box or a locked box have different effects on the investment of married versus unmarried women? Compare the mean investment in health products among married women across three groups. Then compare the mean investment in health products among unmarried women across three groups. Briefly interpret the result. How does this analysis address the internal validity issue discussed in Question 4? Ans:

Either married or unmarried, female participants in the group with safe box treatment spent more on health products compared to other groups. This result helps confirm the internal validity of our study: the difference in investment of different samples we observe in (c) is caused by the difference of our treatments, not by the difference in pre-treatment characteristics we discussed in (d). The causal assumption we proposed, thus, is validated and satisfied.

```
#Subset 3 groups
safebox <- subset(data.frame, data.frame$treatment == "safebox")
lockbox <- subset(data.frame, data.frame$treatment == "lockbox")
control <- subset(data.frame, data.frame$treatment == "control")

#Mean investment made by married female in each group
safebox_married_invest = mean(safebox$fol2_amtinvest[safebox$bg_married ==1
                              & safebox$bg_female ==1], na.rm=TRUE)
lockbox_married_invest = mean(lockbox$fol2_amtinvest[lockbox$bg_married ==1
                              & lockbox$bg_female ==1], na.rm=TRUE)
control_married_invest = mean(control$fol2_amtinvest[control$bg_married ==1
                              & control$bg_female ==1], na.rm=TRUE)

print("Average investment of health products made by married women across groups:")
```

```
## [1] "Average investment of health products made by married women across groups:"
```

```r
paste("Safebox:", safebox_married_invest)
```

```
## [1] "Safebox: 557.135593220339"
```

```r
paste("Lockbox:", lockbox_married_invest)
```

```
## [1] "Lockbox: 332.432989690722"
```

```r
paste("Control:", control_married_invest)
```

```
## [1] "Control: 239.66"
```

```r
#Mean investment made by unmarried female in each group
safebox_unmarried_invest = mean(safebox$fol2_amtinvest[safebox$bg_married ==0
                              & safebox$bg_female ==1], na.rm=TRUE)
lockbox_unmarried_invest = mean(lockbox$fol2_amtinvest[lockbox$bg_married ==0
                              & lockbox$bg_female ==1], na.rm=TRUE)
control_unmarried_invest = mean(control$fol2_amtinvest[control$bg_married ==0
                              & control$bg_female ==1], na.rm=TRUE)

print("Average investment of health products made by unmarried women across groups:")
```

```
## [1] "Average investment of health products made by unmarried women across groups:"
```

```r
paste("Safebox:", safebox_unmarried_invest)
```

```
## [1] "Safebox: 264.038461538462"
```

```r
paste("Lockbox:", lockbox_unmarried_invest)
```

```
## [1] "Lockbox: 220.473684210526"
```

```r
paste("Control:", control_unmarried_invest)
```

```
## [1] "Control: 218.541666666667"
```

PROBLEM 2:

a. Estimate the following probabilities: P(event), P(any MedDiet), P(event,any MedDiet), and P (event | any MedDiet). Ans:

See below.

```r
predimed_df <- read.csv("https://raw.githubusercontent.com/dpuelz/Policy-Research-Laboratory/main/data/p
summary(predimed_df)
```

```
##     group               sex                 age           smoke
##  Length:6324        Length:6324        Min.   :49.00   Length:6324
##  Class :character   Class :character   1st Qu.:62.00   Class :character
##  Mode  :character   Mode  :character   Median :67.00   Mode  :character
##                                        Mean   :67.01
##                                        3rd Qu.:72.00
##                                        Max.   :87.00
##       bmi            waist            wth              htn
##  Min.   :19.64   Min.   : 50.0   Min.   :0.3012   Length:6324
##  1st Qu.:27.23   1st Qu.: 93.0   1st Qu.:0.5839   Class :character
##  Median :29.76   Median :100.0   Median :0.6258   Mode  :character
##  Mean   :29.97   Mean   :100.4   Mean   :0.6283
##  3rd Qu.:32.46   3rd Qu.:107.0   3rd Qu.:0.6687
##  Max.   :51.94   Max.   :177.0   Max.   :1.0000
##      diab            hyperchol           famhist             hormo
##  Length:6324        Length:6324        Length:6324        Length:6324
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       p14            toevent           event
##  Min.   : 0.000   Min.   :0.01643   Length:6324
##  1st Qu.: 7.000   1st Qu.:2.85832   Class :character
##  Median : 9.000   Median :4.78850   Mode  :character
##  Mean   : 8.678   Mean   :4.35517
##  3rd Qu.:10.000   3rd Qu.:5.79056
##  Max.   :14.000   Max.   :6.99795
```

```r
#P(event)
##M1: Calculation
p.event = sum(predimed_df$event == "Yes")/nrow(predimed_df)
print(p.event)
```

```
## [1] 0.0398482
```

```r
##M2: Using prop table
prop.table(table(predimed_df$event))
```

```
##
##        No       Yes
## 0.9601518 0.0398482
```

```r
#P(any MedDiet)

##M1: Calculation
##Find how many groups in group column
predimed_df$group <- as.factor(predimed_df$group)
levels(predimed_df$group)
```

```
## [1] "Control"        "MedDiet + Nuts" "MedDiet + VOO"
```

```
##P(any MedDiet) = 1 - P(no MedDiet) = 1 - P(Control)
p.anyMedDiet = 1 - sum(predimed_df$group == "Control")/nrow(predimed_df)
print(p.anyMedDiet)
```

```
## [1] 0.6771031
```

```
##M2: Using prop table
tab_anyMedDiet = prop.table(table(predimed_df$group))
p.anyMedDiet_2 = sum(tab_anyMedDiet[2]+tab_anyMedDiet[3])
print(p.anyMedDiet_2)
```

```
## [1] 0.6771031
```

```
#P(event, any MedDiet)

##M1: Calculation
anyMedDiet_event <- subset(predimed_df, predimed_df$event == "Yes"
                           & predimed_df$group != "Control")
p.join.event_anyMedDiet = nrow(anyMedDiet_event)/nrow(predimed_df)
print(p.join.event_anyMedDiet)
```

```
## [1] 0.0245098
```

```
##M2: Using 2-way prop table
tab_event_anyMed = prop.table(table(event = predimed_df$event,
                                    group = predimed_df$group))
p.join.event_anyMedDiet_2 = sum(tab_event_anyMed[2,2:3])
print(p.join.event_anyMedDiet_2)
```

```
## [1] 0.0245098
```

```
#P(event | anyMedDiet) = P(event, any MedDiet)/P(any MedDiet)
p.con.event_anyMedDiet = p.join.event_anyMedDiet/p.anyMedDiet
print(p.con.event_anyMedDiet)
```

```
## [1] 0.03619804
```

b. Estimate P (event | Control). Using this result and the answer from the previous question, assess whether the Mediterranean diet has an effect on the chance of a cardiac event? Ans:

The probability of having a cardiac event is lower with people who have the Mediterranean diet (3.62% versus 4.75%).

```
#P(event, Control)
Control_event <- subset(predimed_df, predimed_df$group == "Control"
                        & predimed_df$event == "Yes")
p.join.event_control = nrow(Control_event)/nrow(predimed_df)

#P(Control)
```

```
p.Control = sum(predimed_df$group == "Control")/nrow(predimed_df)

#P(event | Control) = P(event, Control)/P(Control)
p.cond.event_Control = p.join.event_control/p.Control
print(p.cond.event_Control)
```

```
## [1] 0.04750245
```

c. What additional information not given would be useful to characterize uncertainty in the effect estimate above. (Hint: Remember that this is a randomized control trial). With that information, describe the steps for characterizing this uncertainty. Ans:

The issue: the sample size of the control group is much smaller than the total sample size of all MedDiet groups. Sample size is an issue since what we compared was between Pr(event | Control) vs. Pr(event | any MedDiet). This huge different in sample size could possibly lead to Sampson's paradox - which mislead our interpretation for the fact if we only based on the calculated data.

It would also be helpful if we know whether the different types of Med Diet are mutually exclusive and/or independent or not. By knowing this, we could determine if any Med Diet means only either of the Med Diet types, or if the participants can have the mixture of all types. If participants could be on a mixture of different types of Med Diet, then we also need to know the effect of having multiple Med Diet types vs. having only one type of Med Diet in order to know exactly the effect of Med Diet treatment vs. control group.

To solve: 1/ Compare sample size in each group; 2/ If sample size is relatively equal, proceed; if not, randomly selected till each have a relatively equal size; 3/ Find probability of the event given the condition of the control and each treatment group (i.e. control, MedDiet + VOO, MedDiet + Nuts); 4/ Compare results

d. What are the effects on cardiac event likelihood of the Mediterranean diet on Female and Male subpopulations relative to the control diet? Ans:

The result shows that, on overall, for both women and men, having a Mediterranean diet led to a lower risk of having a cardiac event compared to whom with a control diet.

For instance, women with the control diet has a 0.62% of having a cardiac arrest, whereas those who has a Med Diet will only be 0.46% - 0.52% at risk. A similar pattern observed in the studied group of men: 0.92% for control diet, 0.64% - 0.82% for Med Diet.

```
#Using 3-way proportion table
prop.table(table(group = predimed_df$group,
                 event = predimed_df$event, gender = predimed_df$sex))
```

```
## , , gender = Female
##
##                  event
## group                      No         Yes
##   Control        0.188330171 0.006166983
##   MedDiet + Nuts 0.174414927 0.004585705
##   MedDiet + VOO  0.197659709 0.005218216
##
## , , gender = Male
##
##                  event
## group                      No         Yes
##   Control        0.119228336 0.009171410
##   MedDiet + Nuts 0.146584440 0.006483238
##   MedDiet + VOO  0.133934219 0.008222644
```