HW2: Health Savings Experiments

06 September 2020

To understand why those who are poor are constrained in their ability to save for investments in preventative health, two researchers designed a field experiment in rural Kenya in which they randomly varied access to four innovative saving technologies. By observing the impact of these various technologies on asset accumulation, and by examining which types of people who benefit most from them, the researchers were able to identify key barriers to saving.

They worked with 113 ROSCAs (Rotating Savings and Credit Associations). A ROSCA is a group of individuals who come together and make regular cyclical contributions to a fund (called the "pot"), which is then given as a lump sum to one member in each cycle. In their experiment, the researchers randomly assigned 113 ROSCAs to one of the five study arms. In this exercise, we will focus on three study arms (one control and two treatment arms). The data file, `rosca.csv` is extracted from their original data, excluding individuals who have received multiple treatments for the sake of simplicity.

Individuals in all study arms were encouraged to save for health and were asked to set a health goal for themselves at the beginning of the study. In the first treatment group (*Safe Box*), respondents were given a box locked with a padlock, and the key to the padlock was provided to the participants. They were asked to record what health product they were saving for and its cost. This treatment is designed to estimate the effect of having a safe and designated s torage technology for preventative health savings.

In the second treatment group (*Locked Box*), respondents were given a locked box, but not the key to the padlock. The respondents were instructed to call the program officer once they had reached their saving goal, and the program officer would then meet the participant and open the *Locked Box* at the shop where the product is purchased. Compared to the safe box, the locked box offered stronger commitment through earmarking (the money saved could only be used for the prespecified purpose selected by the participant).

Participants were interviewed again 6 months and 12 months later. In this HW, our outcome of interest is the amount (in Kenyan shillings) spent on preventative health products after 12 months.

Descriptions of the relevant variables in the data file `rosca.csv` are:

Name	Description
`bg_female`	`1` if female, and `0` otherwise. This is a pre-treatment variable.
`bg_married`	`1` if married, and `0` otherwise. This is a pre-treatment variable.
`bg_b1_age`	Age at baseline. This is a pre-treatment variable.
`encouragement`	`1` if participant received encouragement only (control group), and `0` otherwise
`safe_box`	`1` if participant received safe box treatment, and `0` otherwise
`locked_box`	`1` if participant received lock box treatment, and `0` otherwise
`fol2_amtinvest`	Amount invested in health products at time of the second follow up
`has_followup2`	`1` if participant appears in second followup (after 12 months), and `0` otherwise

Question 0

Run the following the code chunk below to load the data set and create a new variable `treatment` that takes the value `control` if receiving only encouragement, `safebox` if receiving a safe box, and `lockbox` if receiving a locked box. We then designate that R should treat this new variable as a factor variable.

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(readr)
rosca <- read_csv("data/rosca.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   bg_female = col_double(),
##   bg_married = col_double(),
##   bg_b1_age = col_double(),
##   encouragement = col_double(),
##   safe_box = col_double(),
##   locked_box = col_double(),
##   fol2_amtinvest = col_double(),
##   has_followup2 = col_double()
## )
```

```r
rosca <- rosca %>%
  mutate(treatment = case_when(
    encouragement == 1 ~ 'control',
    safe_box == 1 ~ 'safebox',
    locked_box == 1 ~ 'lockbox'
  ))

rosca$treatment <- as.factor(rosca$treatment)
```

Question 1 (5 points)

What is the specific causal question (or questions) the researchers aimed to answer (write separate questions for each treatment)? What are the potential outcomes (there are 3)? What are the hypothesized treatment effects? For the Locked Box intervention group, what is their average missing counterfactual? How do your recommend estimating it? How would you describe the internal validity of this study? How many individuals are in the control group? How many individuals are in each of the treatment arms? Use a table to show the counts.

```
table(rosca$treatment)
```

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

Your Answer here:

What is the specific causal question (or questions) the researchers aimed to answer (write separate questions for each treatment)? 1. What is the impact of peoples health goal and savings, if they used a safebox relative to not using a safebox? 2. What is the impact of peoples health goal and savings, if they used a lockbox relative to not using the lockbox?

What are the potential outcomes (there are 3)? Control: No change in amount of money saved. Lockbox: People save a lot of money using the LockBox Safebox: People save a lot of money using a SafeBox

What are the hypothesized treatment effects? The hypothesized treatment effect would be that people in the treatment group would save money and have good health.

For the Locked Box intervention group, what is their average missing counterfactual? Factual: Treatment Group MCF: Control Group

How do your recommend estimating it? We would estimate it by looking at the control group.

How would you describe the internal validity of this study? The internal validity is high because the experiment was very randomized.

How many individuals are in the control group? How many individuals are in each of the treatment arms? Use a table to show the counts. control lockbox safebox 111 195 117

## Question 2 (6 points)

What are the drop-out rates (those for whom 12 month outcomes are missing) across the treatment and control conditions? Does the nature of either of the treatments (safebox or lockbox) suggest to you drop-out might be higher or lower in one group? Do the drop-out rates by treatment condition suggest that internal validity may be compromised? Why? Do they suggest external validity may be compromised? Why? *Hint: you can add, subtract, multiply, and divide tables in R.* Subset the data (we suggest giving the subset data a new object name) so that it contains only participants who were interviewed at 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?

```
prop.table(table(rosca$has_followup2, rosca$treatment), margin = 2)
```

```
##
##        control    lockbox    safebox
##   0 0.08108108 0.05641026 0.08547009
##   1 0.91891892 0.94358974 0.91452991
```

```
roscafu <- subset(rosca ,has_followup2 == "1")
table(roscafu$treatment)
```

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

Your Answer here:

What are the drop-out rates (those for whom 12 month outcomes are missing) across the treatment and control conditions? control lockbox safebox 1 0.91891892 0.94358974 0.91452991

Does the nature of either of the treatments (safebox or lockbox) suggest to you drop-out might be higher or lower in one group? The Lock Box has higher dropout rates.

Do the drop-out rates by treatment condition suggest that internal validity may be compromised? Why? The validity is not affected because the rates are consistent.

Do they suggest external validity may be compromised? Why? The internal validity is not compromised because the internal validity isnt.

Subset the data (we suggest giving the subset data a new object name) so that it contains only participants who were interviewed at 12 months during the second followup.
How many participants are left in each group of this subset?

control lockbox safebox 102 184 107

Question 3 (5 points)

Does receiving a safe box or lockbox increase the average amount invested in health products relative to encouragement only? We focus on the outcome measured 12 months from baseline during the second follow-up `fol2_amtinvest`. First, describe the distribution of this outcome over all study participants. Then find the average amount invested (in Kenyan shilling) in health products across the treatment and control conditions. Then calculate the differences in the the mean of amounts invested in health products between each of the treatment arms and the control group. Briefly interpret the results. What are strengths and limitations of using averages (rather than a different summary statistic) to estimate impacts given the distribution of the outcome measure?

Answer 3

```
roscafu %>%
group_by(treatment) %>%
summarize(mean(fol2_amtinvest))
```

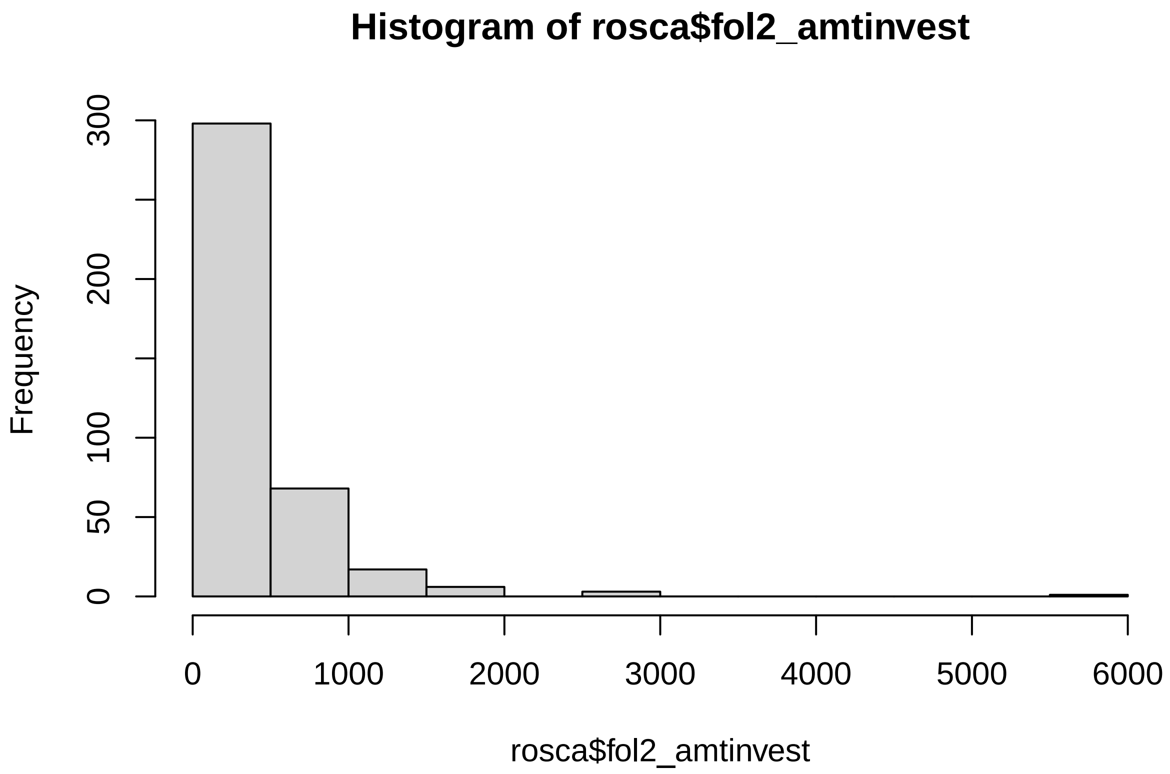
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   treatment `mean(fol2_amtinvest)`
##   <fct>                      <dbl>
## 1 control                     258.
## 2 lockbox                     308.
## 3 safebox                     408.
```

```
summary(rosca$fol2_amtinvest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     0.0    40.0   100.0   322.2   490.0  5700.0      30
```

```
hist(rosca$fol2_amtinvest)
```



Q3

Does receiving a safe box or lockbox increase the average amount invested in health products relative to encouragement only? Yes it does, the mean values are higher than the control values.

We focus on the outcome measured 12 months from baseline during the second follow-up `fol2_amtinvest`. First, describe the distribution of this outcome over all study participants. Then find the average amount invested (in Kenyan shilling) in health products across the treatment and control conditions. Then calculate the differences in the the mean of amounts invested in health products between each of the treatment arms and the control group. Briefly interpret the results.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.0    40.0   100.0   322.2   490.0  5700.0      30
```

```
The distribution of the outcome is very distributed and it is more to the left. This data is a right sk
```

What are strengths and limitations of using averages (rather than a different summary statistic) to estimate impacts given the distribution of the outcome measure?

The strength is that we can evaluate a large set of data. The limitations are that they hide disparities in our data.

Answer 3

Question 4 (8 points)

Examine the distribution of the 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 with regard to each of these three variables? What does the result of this analysis suggest about the internal validity of the findings you calculated in the previous question? Also, if you think the internal validity has been compromised, state whether you hypothesize this would lead to over- or under-estimates of the treatment effects in this case. Please provide calculations and interpretations for each of the three pre-treatment variables separately.

Answer 4

```r
#gender
tapply(rosca$bg_female, rosca$treatment, mean)
```

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

```r
#age

tapply(rosca$bg_b1_age, rosca$treatment, mean)
```

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

```r
#marital status

tapply(rosca$bg_married, rosca$treatment, mean)
```

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

Q4

Examine the distribution of the 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 with regard to each of these three variables?

The participants in the two groups are not very distributed compared to each other.

What does the result of this analysis suggest about the internal validity of the findings you calculated in the previous question?

The internal validity is high because the groups didn't have a very distributed values.

Also, if you think the internal validity has been compromised, state whether you hypothesize this would lead to over- or under-estimates of the treatment effects in this case.

I dont think it compromised because the distributions are not very far apart.

Please provide calculations and interpretations for each of the three pre-treatment variables separately.

Gender: The gender has high percentages are they all are close to each other. Age: The age has very low percentages and they are close to each other. Marital: The marital has similar trends as to gender.

Question 5 (6 points)

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 the three treatment conditions. Then compare the mean investment in health products among unmarried women the three treatment conditions. Briefly interpret the results. How does this analysis address any internal validity issues discussed in Question 4?

Answer 5

```r
roscafu %>%
  filter(bg_married == 1 & bg_female == 1) %>%
  group_by(treatment) %>%
  summarise(mean(fol2_amtinvest))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   treatment `mean(fol2_amtinvest)`
##   <fct>                      <dbl>
## 1 control                     240.
## 2 lockbox                     332.
## 3 safebox                     557.
```

```r
roscafu %>%
  filter(bg_married == 0 & bg_female == 1) %>%
  group_by(treatment) %>%
  summarise(mean(fol2_amtinvest))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   treatment `mean(fol2_amtinvest)`
##   <fct>                      <dbl>
## 1 control                     219.
## 2 lockbox                     220.
## 3 safebox                     264.
```

Q5

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 the three treatment conditions. Then compare the mean investment in health products among unmarried women the three treatment conditions. Briefly interpret the results.

The two groups do have an affect on the mean of the different groups. The means for the married people are higher then the unmarried people.

How does this analysis address any internal validity issues discussed in Question 4? The internal validity is veru high because we have randomized the people in the groups. ##5