Problem Set 1

Emanuel Mejía

12/01/2021

Contents

1 Potential Outcomes Notation

1.1 Explain the notation $Y_i(1)$.

Answer: ...

$Y_i(1)$ refers to the potential outcome (for whatever variable we're measuring) of the i^{th} subject if this subject were to be part of the treatment group in the experiment ($d_i = 1$).

1.2 Explain the notation $Y_1(1)$.

Answer: ...

$Y_1(1)$ refers specifically to the potential outcome of the first subject in the list if it were to be part of the treatment group in the experiment ($d_i = 1$).

1.3 Explain the notation $E[Y_i(1)|d_i = 0]$.

Answer: ...

$E[Y_i(1)|d_i = 0]$ denotes the expectation of the potential outcome if the subject were treated ($Y_i(1)$), but conditional on villages that were NOT treated ($d_i = 0$), meaning the hypothetical expectation of $Y_i(1)$ when one subject is chosen at random from the list of non-treated subjects.

1.4 Explain the difference between the notation $E[Y_i(1)]$ and $E[Y_i(1)|d_i = 1]$

Answer: ...

$E[Y_i(1)]$ corresponds to the expected value of the potential outcome of any subject if this subject were to receive the treatment, and its main difference with $E[Y_i(1)|d_i = 1]$ is that the latter takes into account to compute this expectation only the subjects that we already know that are on the treatment group, while the first one computes it for every subject in the experiment (doesn't matter if it happens to be in treatment or control).

2 Potential Outcomes and Treatment Effects

2.1 Illustration

Use the values in the table below to illustrate that $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)]$.

```
table
```

```
##    subject y_0 y_1 tau
## 1:       1  10  12   2
## 2:       2  12  12   0
## 3:       3  15  18   3
## 4:       4  11  14   3
## 5:       5  10  15   5
## 6:       6  17  18   1
## 7:       7  16  16   0
```

```
# Use this code chunk to show your code work
E_Y1 <- table[ , mean(y_1)]
E_Y1
```

```
## [1] 15
```

```
E_Y0 <- table[ , mean(y_0)]
E_Y0
```

```
## [1] 13
```

```
E_tau <- table[ , mean(tau)]
E_tau
```

```
## [1] 2
```

Answer: ...

On the left side of the equation we have:

$$\begin{aligned}
E[Y_i(1)] - E[Y_i(0)] &= \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0) \\
&= 15 - 13 \\
&= 2
\end{aligned}$$

While on the right side:

$$\begin{aligned}
E[Y_i(1) - Y_i(0)] &= E[\tau] \\
&= \frac{1}{N}\sum_{i=1}^{N} \tau_i \\
&= 2
\end{aligned}$$

Therefore $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)]$.

2.2 Data Possibilities

Is it possible to collect all necessary values and construct a table like the one provided in real life? Explain why or why not?

Answer: ...

It is NOT possible to collect all the data necessary to construct this kind of table in real life because each subject will either receive the treatment, so we will only be able to see $Y_i(1)$ materialized or, on the other hand, remain in the control group, and the only value we will observe is the realization of $Y_i(0)$. But we will never be able to see both of them for any given subject and therefore we won't be able to compute τ either.

3 Visual Acuity

3.1 Treatment effect

Compute the individual treatment effect for each of the ten children.

```
# Use this code chunk to show your code work
d[ , tau := y_1 - y_0]
d
```

```
##     child y_0 y_1  tau
##  1:     1 1.2 1.2  0.0
##  2:     2 0.1 0.7  0.6
##  3:     3 0.5 0.5  0.0
##  4:     4 0.8 0.8  0.0
##  5:     5 1.5 0.6 -0.9
##  6:     6 2.0 2.0  0.0
##  7:     7 1.3 1.3  0.0
##  8:     8 0.7 0.7  0.0
##  9:     9 1.1 1.1  0.0
## 10:    10 1.4 1.4  0.0
```

Answer: …

The individual treatment effects are contained in the vector above and we can notice that for every child but two of them (#2 and #5) there is no effect.

3.2 Story time

Tell a “story” that could explain this distribution of treatment effects. In particular, discuss what might cause some children to have different treatment effects than others.

```
# Use this code chunk to show your code work (if needed)
d[,tau]
```

```
##  [1]  0.0  0.6  0.0  0.0 -0.9  0.0  0.0  0.0  0.0  0.0
```

Answer: …

As previously said, there is only two children with individual effects different than zero. So the first result we might state is that for most children there doesn’t seem to be any effect of playing outside on eyesight. But there are two kids that have inverse individual effects between each other.

One plausible story is that these two kids have different health conditions leading to their specific result, meaning that child #2 has certain condition that leads to better acuity when being outside, perhaps lack of light inside the house decreases her eyesight or dust accumulated in closed spaces. While child #5 has the opposite effect, it might be that UV-light severely harms her eyes or something in the specific environment her house has this effect.

Not a doctor here, but the point is that there are other plausible causes to observe these outcomes other than the one we’re trying to measure.

3.3 True ATE

For this population, what is the true average treatment effect (ATE) of playing outside.

```
# Use this code chunk to show your code work
ATE <- d[ , mean(tau)]
ATE
```

```
## [1] -0.03
```

Answer: ...

True Average Treatment Effect is $ATE = -0.03$

3.4 Even-Odd split

Suppose we are able to do an experiment in which we can control the amount of time these children play outside for three years. We happen to randomly assign the odd-numbered children to treatment and the even-numbered children to control. What is the estimate of the ATE you would reach under this assignment? (Please describe your work.)

```
# Use this code chunk to show your code work
d[child %% 2 == 1 , treat := 1]
d[child %% 2 == 0 , treat := 0]
d
```

```
##     child y_0 y_1  tau treat
##  1:     1 1.2 1.2  0.0     1
##  2:     2 0.1 0.7  0.6     0
##  3:     3 0.5 0.5  0.0     1
##  4:     4 0.8 0.8  0.0     0
##  5:     5 1.5 0.6 -0.9     1
##  6:     6 2.0 2.0  0.0     0
##  7:     7 1.3 1.3  0.0     1
##  8:     8 0.7 0.7  0.0     0
##  9:     9 1.1 1.1  0.0     1
## 10:    10 1.4 1.4  0.0     0
```

```
E_Y0_d0 <- d[treat == 0, mean(y_0)]
E_Y0_d0
```

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

```
E_Y1_d1 <- d[treat == 1, mean(y_1)]
E_Y1_d1
```

```
## [1] 0.94
```

```
ATE_hat <- E_Y1_d1 - E_Y0_d0
ATE_hat
```

```
## [1] -0.06
```

Answer: . . .

Under this assumption we no longer "know" all the values, meaning we only know $Y_i(1)$ for subjects under treatment and $Y_i(0)$ for subjects in control group and our best ATE estimate would be computed as: $ATE = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$, so we will only use the known values to compute the expectations.

In this case we have the average of $Y_i(1)$ for treated subjects $E[Y_i(1)|D_i = 1] = 0.94$ and the average of $Y_i(0)$ for control subjects $E[Y_i(0)|D_i = 0] = 1$. So, the estimated average treatment effect comes from the difference, meaning $\widehat{ATE} = -0.06$.

3.5 Biased or Unbiased?

How different is the estimate from the truth? In your own words, why is there a difference? Does this mean that the estimator is a biased or an unbiased estimator? Does this mean that the estimate is biased or unbiased?

```r
# Use this code chunk to show your code work
ATE_diff <- ATE_hat - ATE
ATE_diff
```

```
## [1] -0.03
```

Answer: . . . In this specific case we have a difference of -0.03 which would mean an absolute difference of 100% which might seem really large. Although, as long as we're truly randomly assigning children to treatment or control, our estimator should be unbiased around the true value, and if a particular estimate (a singular computation) seems biased is just because we happened to randomly choose a sample with these results since we don't know every potential outcome but only the ones we can see.

3.6 How many splits are possible?

We just considered one way (odd-even) an experiment might split the children. How many different ways (every possible way) are there to split the children into a treatment versus a control group (assuming at least one person is always in the treatment group and at least one person is always in the control group)?

```r
# Use this code chunk to show your code work
N <- 10

com_num <- NA
for (i in 1:9){
  com_num[i] <-factorial(N)/(factorial(i)*factorial(N-i))
}

com_num
```

```
## [1]  10  45 120 210 252 210 120  45  10
```

Answer: . . .

The vector above contains the quantity of possible combinations for each splitting method, from only 1 kid in treatment (9 in control) up to 9 kids in treatment (1 in control). So the total number of possible ways to split between groups is 1022

3.7 Thinking about your assignment strategy

Given there are as many ways to assign as you answered in the last sub-question, can you provide a rationale for why you might prefer one assignment strategy over another?

For concreteness, suppose that either (a) you can have a treatment assignment where one and only one of the kids is randomly assigned to treatment; or (b) you can have a treatment assignment where exactly five of the kids are randomly assigned to treatment.

As a small hint, you might note that $\left\{\left[\sum_{i=1}^{n} Y_i(1)|d_i = 1\right] - \left[\sum_{j=1}^{n} Y_j(0)|d_j = 0\right]\right\} \equiv ATE$ is an estimator and there are some properties of estimators that we care about.

To make the question tractable, suppose that if you were to increase the size of the population procedure (a) would keep a single kid in treatment, while procedure (b) would keep 50% of the sample in treatment and 50% of the sample in control.

Answer: … One of the estimators' properties is consistency which basically means that the bigger the sample the closer the estimate will tend to be to the population's true parameter. Although, in this specific example we have an issue to do that.

In order to compute the estimated ATE we actually need to estimate two potential outcomes first ($Y_i(1)$ and $Y_i(0)$), and we can't increase both at the same time, because increasing the sample for one of them will consequently decrease the other.

Given this issue, it seems like choosing 5-5 children for treatment and control groups respectively will give a well balanced estimator for both of them, and therefore a better estimator for the ATE.

3.8 Compute the MSE of these two designs

Because you have the entire population of kids, their entire scheduled of potential outcomes, and two proposed sampling procedures: conduct a simulation study. First, calculate all of the possible treatment effects that you might observe under each design. Then, compute the mean-squared error of each design. Which design – the one where you have a single kid in treatment, or the one where you have five kids in treatment – produces a lower MSE? (Hint the `combn` function might help you with your subsetting.)

```r
# Use this code chunk to show your code work
library(combinat)

# Creating a table with rows for each 1-9 combination
len_1_9 <- dim(combn(10,1))[2]
tbl_effects_1_9 <- data.table(
  'com' = 1:len_1_9
)

# Computing the expected outcome for each combination
for (i in 1:len_1_9){
  selected <- combn(10,1)[,i]
  d[ , treat := 0]
  d[child %in% selected, treat := 1]
  tbl_effects_1_9[i, treated := selected]
  tbl_effects_1_9[i, E_Y0_d0 := d[treat == 0, mean(y_0)]]
  tbl_effects_1_9[i, E_Y1_d1 := d[treat == 1, mean(y_1)]]
}
# Computing estimated ATE and errors
```

```
tbl_effects_1_9[ , ATE_est := E_Y1_d1 - E_Y0_d0]
tbl_effects_1_9[ , ATE_error := ATE_est - ATE]
tbl_effects_1_9[ , SQ_error := ATE_error**2]
tbl_effects_1_9
```

```
##     com treated   E_Y0_d0 E_Y1_d1     ATE_est   ATE_error    SQ_error
##  1:   1       1 1.0444444     1.2  0.15555556  0.18555556 0.034430864
##  2:   2       2 1.1666667     0.7 -0.46666667 -0.43666667 0.190677778
##  3:   3       3 1.1222222     0.5 -0.62222222 -0.59222222 0.350727160
##  4:   4       4 1.0888889     0.8 -0.28888889 -0.25888889 0.067023457
##  5:   5       5 1.0111111     0.6 -0.41111111 -0.38111111 0.145245679
##  6:   6       6 0.9555556     2.0  1.04444444  1.07444444 1.154430864
##  7:   7       7 1.0333333     1.3  0.26666667  0.29666667 0.088011111
##  8:   8       8 1.1000000     0.7 -0.40000000 -0.37000000 0.136900000
##  9:   9       9 1.0555556     1.1  0.04444444  0.07444444 0.005541975
## 10:  10      10 1.0222222     1.4  0.37777778  0.40777778 0.166282716
```

```
# Computing MSE
MSE_1_9 <- tbl_effects_1_9[ , mean(SQ_error)]
MSE_1_9
```

```
## [1] 0.2339272
```

```
# Creating a table with rows for each 5-5 combination
len_5_5 <- dim(combn(10,5))[2]
tbl_effects_5_5 <- data.table(
  'com' = 1:len_5_5
)

# Computing the expected outcome for each combination
for (i in 1:len_5_5){
  selected <- combn(10,5)[,i]
  d[ , treat := 0]
  d[child %in% selected, treat := 1]
  tbl_effects_5_5[i, treated := paste(selected, collapse = " ")]
  tbl_effects_5_5[i, E_Y0_d0 := d[treat == 0, mean(y_0)]]
  tbl_effects_5_5[i, E_Y1_d1 := d[treat == 1, mean(y_1)]]
}
# Computing estimated ATE and errors
tbl_effects_5_5[ , ATE_est := E_Y1_d1 - E_Y0_d0]
tbl_effects_5_5[ , ATE_error := ATE_est - ATE]
tbl_effects_5_5[ , SQ_error := ATE_error**2]
tbl_effects_5_5[1:15, ]
```

```
##     com    treated E_Y0_d0 E_Y1_d1 ATE_est ATE_error SQ_error
##  1:   1  1 2 3 4 5    1.30    0.76   -0.54     -0.51   0.2601
##  2:   2  1 2 3 4 6    1.20    1.04   -0.16     -0.13   0.0169
##  3:   3  1 2 3 4 7    1.34    0.90   -0.44     -0.41   0.1681
##  4:   4  1 2 3 4 8    1.46    0.78   -0.68     -0.65   0.4225
##  5:   5  1 2 3 4 9    1.38    0.86   -0.52     -0.49   0.2401
##  6:   6 1 2 3 4 10    1.32    0.92   -0.40     -0.37   0.1369
##  7:   7  1 2 3 5 6    1.06    1.00   -0.06     -0.03   0.0009
##  8:   8  1 2 3 5 7    1.20    0.86   -0.34      -0.31   0.0961
##  9:   9  1 2 3 5 8    1.32    0.74   -0.58      -0.55   0.3025
## 10:  10  1 2 3 5 9    1.24    0.82   -0.42      -0.39   0.1521
## 11:  11 1 2 3 5 10    1.18    0.88   -0.30      -0.27   0.0729
## 12:  12  1 2 3 6 7    1.10    1.14    0.04       0.07   0.0049
## 13:  13  1 2 3 6 8    1.22    1.02   -0.20      -0.17   0.0289
## 14:  14  1 2 3 6 9    1.14    1.10   -0.04      -0.01   0.0001
## 15:  15 1 2 3 6 10    1.08    1.16    0.08       0.11   0.0121
```

```
# Computing MSE
MSE_5_5 <- tbl_effects_5_5[ , mean(SQ_error)]
MSE_5_5
```

```
## [1] 0.08987778
```

Answer: …

For the 1-9 design (1 child in treatment, 9 in control) we have a $MSE_{1-9} = 0.2339272$, while for the 5-5 design we have a $MSE_{5-5} = 0.0898778$. Meaning that the latter design will perform better on average.

3.9 Observational study

Suppose that we decide it is too hard to control the behavior of the children, so we do an observational study instead. Children 1-5 choose to play an average of more than 10 hours per week from age 3 to age 6, while Children 6-10 play less than 10 hours per week. Compute the difference in means from the resulting observational data.

```
# Use this code chunk to show your code work
d[child < 6, treat := 1]
d[child >= 6 , treat := 0]
d
```

```
##     child y_0 y_1  tau treat
##  1:     1 1.2 1.2  0.0     1
##  2:     2 0.1 0.7  0.6     1
##  3:     3 0.5 0.5  0.0     1
##  4:     4 0.8 0.8  0.0     1
##  5:     5 1.5 0.6 -0.9     1
##  6:     6 2.0 2.0  0.0     0
##  7:     7 1.3 1.3  0.0     0
##  8:     8 0.7 0.7  0.0     0
##  9:     9 1.1 1.1  0.0     0
## 10:    10 1.4 1.4  0.0     0
```

```
E_Y0_d0_O <- d[treat == 0, mean(y_0)]
E_Y1_d1_O <- d[treat == 1, mean(y_1)]

ATE_hat_O <- E_Y1_d1_O - E_Y0_d0_O
ATE_hat_O
```

```
## [1] -0.54
```

Answer: …

In this observational case we have an estimated $ATE = -0.54$

3.10 Observational ATE

Compare your answer in `Observational study` to the true ATE. In your own words what causes the difference? Does this mean that the estimator is a biased or an unbiased estimator? Does this mean that the estimate is biased or unbiased?

```r
# Use this code chunk to show your code work (if needed)
ATE_diff_O <- ATE_hat_O - ATE
ATE_diff_O
```

```
## [1] -0.51
```

Answer: …

In this observational case we have a difference of -0.51 which would mean an absolute difference of 1700% which is much larger than in the first example (3.5).

In this specific scenario, we can't ensure that this estimator is unbiased because we are not randomly assigning children to each group (we're just taking existing data), in fact we don't even know why each child pertains to each group and even if we repeat the study we won't systematically get any closer to the true population's value, so if a specific estimate happens to be close to the true population's value it would be just matter of luck because a biased estimator will produce biased estimates.

The only way this estimator and its estimates could be unbiased is if the original data owner actually ran the experiment and randomly assigned people to each group.

4 Randomization and Experiments

The following questions can be a little bit challenging. This is because the argument that you are being asked to make is based on the rote application of a definition. To begin with, it is useful for you to define what you mean when you write about either *an experiment* or *an observational study*. Then, with these definitions on hand, use the definitions to answer the following questions.

4.1 Define your terms

- **An experiment is:** A study where, in order to answer the question of interest, there is a systematic intervention by randomly applying a form of treatment (or remaining without any treatment) before measuring anything.
- **An experiment provides the following statistical guarantees:** When properly designed and by ensuring random selection of the sample to be treated and the sample to be controlled this kind of study ensures that the Average Treatment Effect estimator will be unbiased, consistent and normally asymptotic, while as efficient as possible (with the available resources). This ensures that any causal question can be assessed through this method with statistical guarantee.
- **An observational study is:** A study where, in order to answer the question of interest, available observed data is gathered and recorded without any systematic intervention on any parameter of the subjects.
- **An observational study provides the following statistical guarantees:** By ensuring random selection of the sample under this observation, estimators may be developed to generalize sample parameters to population parameters that satisfy the desired estimator-properties as well. This allows to answer descriptive or even correlation questions about the population parameters (but NOT causal questions) with statistical guarantee.

4.2 Does a random, iid sample produce an unbiased treatment effect estimate?

Assume that a researcher takes a random sample of elementary school children and compares the grades of those who were previously enrolled in an early childhood education program with the grades of those who were not enrolled in such a program. Is this an experiment, an observational study, or something in between?

Answer: … By looking at the previous definitions this can be cataloged as an *Observational Study* since there is no intervention, only data collection of what's already there. And while a proper design of this study might assess descriptive questions about student grades or number of kids enrolled in early childhood education programs, and even the correlation between these two, it cannot evaluate causation between these two in neither way (nor grades explaining enrolling, nor the other way around).

4.3 What if an official agency produces the idd sample?

Assume that the researcher works together with an organization that provides early childhood education and offer free programs to certain children. However, which children that received this offer was not randomly selected by the researcher but rather chosen by the local government. (Assume that the government did not use random assignment but instead gives the offer to students who are deemed to need it the most) The research follows up a couple of years later by comparing the elementary school grades of students offered free early childhood education to those who were not. Is this an experiment, an observational study, or something in between? Explain!

Answer: … This particular study cannot be classified as an actual experiment, because even when there is indeed an intervention, there is not a totally random allocation of the subjects to either treatment or control groups. Ït might be cataloged as *something in between*, perhaps at most as a *quasi-experiment*.

4.4 What if someone else randomly assigns

If the government assigned students to treatment and control by “coin toss”, rather than simply sampling the population, would you say that the study is experimental or observational? Why? What, if any guarantees does this process provide?

Answer: … In my opinion this could be listed as *experiment* because there is indeed an intervention and the subject allocation is also random. Although it might enter in the *natural occurring experiment* category because the assigning is made by a third party (government) and not directly by the researcher.

5 Moral Panic

5.1 Explain the statements

Explain the statement $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ in words. First, state the rote English language translation. Second, tell us the *meaning* of this statement. A full points solution will use the term "potential outcomes" twice.

Answer: . . . This states that the expected non-treated potential outcome among villages that don't receive treatment is equal to the expected non-treated potential outcome among villages that end up receiving treatment.

What this actually means is that whether a subject ends up receiving treatment or not is independent of the expected potential outcome of not receiving it. Meaning that (when the allocation is properly designed) whether a subject is placed in treatment or control has nothing to do with their expected potential outcome of not receiving the treatment.

5.2 Can you believe it

Do you expect that this circumstance actually matches with the meaning that you've just written down? Why or why not?

Answer: . . . Not at all. There doesn't seem to be independence of subject allocation with the expected control potential outcome.

Data "found" don't ensure this independence is present, because we don't really know how subjects were allocated to either treatment or control. So, this equality might not be satisfied and the study might be invalid because group assignment might be provoking different outcomes just by this classification and not because of the actual effect of the treatment.

Perhaps we're understanding this the other way around, sad teens that are doing poorly on their tests need this music to relax and find peace, or perhaps we're in a spurious correlation situation where both of them are related or even caused by some other confounder(s).

So, it seems this study needs a true experiment before determining that this correlation implies a causal statement. Long live the :metal:!