Manzano Romulo - Problem Set 1

September 10, 2017

1 Problem Set #1

1.1 Experiments and Causality

Romulo Manzano (9Sep2017)

2 1. Potential Outcomes Notation

- Explain the notation $Y_i(1)$.

Answer Potential outcome of observation i when the treatment is applied. The value within the parenthesis indicates whether the treatment was applied (1) or not (0), whereas the subscript $_i$ refers to the ith subject treated.

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

Answer Expectation of $Y_i(1)$ when one subject is randomly selected from those subjects that didn't receive the treatment

- Explain the difference between the notation $E[Y_i(1)]$ and the notation $E[Y_i(1)|d_i = 1]$. (Extra credit)

Answer Similar to the previous definition $E[Y_i(1)|d_i = 1]$ is expectation of $Y_i(1)$ when one subject is randomly selected from those subjects that receive the treatment, whereas $E[Y_i(1)]$ refers to the expectation of $Y_i(1)$ for any randomly selected subject.

- Explain the difference between the notation $E[Y_i(1)|d_i = 1]$ and the notation $E[Y_i(1)|D_i = 1]$. Use exercise 2.7 from FE to give a concrete example of the difference.

$E[Y_i(1)|d_i = 1]$ represents the conditional expectation for a given realization of treatment assigment, wherease $E[Y_i(1)|D_i = 1]$ refers to the **expected** conditional expectation, or in other words, it refers to what the conditional expectation would be (on average) across all possible ways d_i could have been allocated

To illustrate, these is how these two expectatios would compare across realizations of the treatment assignments on excersize 2.7. $E[Y_i(0)|D_i = 0]$ vs $E[Y_i(0)|d_i = 0]$

```
In [1]: exercise27Outcomes = c(10, 12.5, 12.5, 12.5,
                               12.5, 12.5, 12.5, 15,
                               15, 15, 15, 15, 15, 15,
                               17.5, 17.5, 17.5, 17.5,
                               17.5, 17.5, 20)
        table27 <- data.frame(Realization = 1:21,
                              y_0_d_i = exercise27Outcomes,
                              y_0_D_i = mean(exercise27Outcomes))
        colnames(table27) <- c("Realization","$E[Y_{i}(0)|d_{i}=0]$ ",
                               "$E[Y_{i}(0)|D_{i}=0]$")
        table27
```

Realization	$E[Y_{i}(0) \| d_{i}=0]$	$E[Y_{i}(0) \| D_{i}=0]$
1	10.0	15
2	12.5	15
3	12.5	15
4	12.5	15
5	12.5	15
6	12.5	15
7	12.5	15
8	15.0	15
9	15.0	15
10	15.0	15
11	15.0	15
12	15.0	15
13	15.0	15
14	15.0	15
15	17.5	15
16	17.5	15
17	17.5	15
18	17.5	15
19	17.5	15
20	17.5	15
21	20.0	15

3 2. FE 2.2

Use the values depicted in Table 2.1 to illustrate that $E[Y_i(0)] - E[Y_i(1)] = E[Y_i(0) - Y_i(1)]$.

```
In [2]: table21 <- data.frame(child = 1:7,
                              y0 = c(10, 15, 20, 20,10,15,15),
                              y1 = c(15, 15, 30,15,20,15,30 ),
                              treatment = c(5,0,10,-5,10,0,15))
        table21
```

child	y0	y1	treatment
1	10	15	5
2	15	15	0
3	20	30	10
4	20	15	-5
5	10	20	10
6	15	15	0
7	15	30	15

Calculating $E[Y_i(0) - Y_i(1)]$

```
In [3]: mean(table21$y1-table21$y0)
```

5

Calculating $E[Y_i(0)] - E[Y_i(1)]$

```
In [4]: mean(table21$y1)-mean(table21$y0)
```

5

We observe equality in both calculations

4 3. FE 2.3

Use the values depicted in Table 2.1 to complete the table below.

$Y_i(0)$	15	20	30	Marginal $Y_i(0)$
10	n:%	n:%	n:%	
15	n:%	n:%	n:%	
20	n:%	n:%	n:%	
Marginal $Y_i(1)$	n:%	n:%	n:%	1.0

a. Fill in the number of observations in each of the nine cells;
b. Indicate the percentage of all subjects that fall into each of the nine cells.
c. At the bottom of the table, indicate the proportion of subjects falling into each category of $Y_i(1)$.
d. At the right of the table, indicate the proportion of subjects falling into each category of $Y_i(0)$.

Answer

$Y_i(0)$	15	20	30	Marginal $Y_i(0)$
10	1: 14.2%	1: 14.2%	0: 0%	2: 28.5%
15	2: 28.5%	0: 0%	1: 14.2%	3: 42.85.2%
20	1: 14.2%	0: 0%	1: 14.2%	2: 28.5%
Marginal $Y_i(1)$	4: 57.1%	1: 14.2%	2: 28.5%	100%

e. Use the table to calculate the conditional expectation that $E[Y_i(0)|Y_i(1) > 15]$.

Answer

We leverage the existing table21 variable and calculate the average $Y_i(0)$ for the relevant subjects

```
In [5]: table3e = table21[table21$y1 >15,]
        mean(table3e$y0)
```

15

f. Use the table to calculate the conditional expectation that $E[Y_i(1)|Y_i(0) > 15]$.

Answer

We leverage the existing table21 variable and calculate the average $Y_i(1)$ for the relevant subjects

```
In [6]: table3f = table21[table21$y0 >15,]
        mean(table3f$y1)
```

22.5

5 4. More Practice with Potential Outcomes

Suppose we are interested in the hypothesis that children playing outside leads them to have better eyesight.

Consider the following population of ten representative children whose visual acuity we can measure. (Visual acuity is the decimal version of the fraction given as output in standard eye exams. Someone with 20/20 vision has acuity 1.0, while someone with 20/40 vision has acuity 0.5. Numbers greater than 1.0 are possible for people with better than "normal" visual acuity.)

```
In [7]: table4 <- data.frame(child = 1:10,
                            y0 = c(1.1, 0.1, 0.5, 0.9, 1.6, 2.0, 1.2, 0.7, 1.0, 1.1),
                            y1 = c(1.1, 0.6, 0.5, 0.9, 0.7, 2.0, 1.2, 0.7, 1.0, 1.1) )
        table4
```

child	y0	y1
1	1.1	1.1
2	0.1	0.6
3	0.5	0.5
4	0.9	0.9
5	1.6	0.7
6	2.0	2.0
7	1.2	1.2
8	0.7	0.7
9	1.0	1.0
10	1.1	1.1

In the table, state $Y_i(1)$ means playing outside an average of at least 10 hours per week from age 3 to age 6,and state $Y_i(0)$ means playing outside an average of less than 10 hours per week from age 3 to age 6. Y_i represents visual acuity measured at age 6.

a. Compute the individual treatment effect for each of the ten children. Note that this is only possible because we are working with hypothetical potential outcomes; we could never have this much information with real-world data. (We encourage the use of computing tools on all problems, but please describe your work so that we can determine whether you are using the correct values.)

Answer

```
In [8]: table4$TreatmentEffect <- table4$y1 - table4$y0
        table4
```

child	y0	y1	TreatmentEffect
1	1.1	1.1	0.0
2	0.1	0.6	0.5
3	0.5	0.5	0.0
4	0.9	0.9	0.0
5	1.6	0.7	-0.9
6	2.0	2.0	0.0
7	1.2	1.2	0.0
8	0.7	0.7	0.0
9	1.0	1.0	0.0
10	1.1	1.1	0.0

b. In a single paragraph, tell a story that could explain this distribution of treatment effects.

Answer: One hypothesis that could this distribution of treatment effect is that there is no effect whatsoever, and that the two very pronounced treatment effects in our population are due to the various effects exposure to sunlight can have on individuals, and not the physical activity associated with playing outside. (idea further elaborated below)

c. What might cause some children to have different treatment effects than others?

Answer: One hypothesis is that the two very pronounced treatment effects in our population are due to the various effects exposure to sunlight can have on individuals with certain medical conditons such as nyctalopia and hemeralopia. These conditions relate to the ability to see in relatively low and high light respectively. Such conditions could affect vision when subjects are exposured to sun light, and depending of when measurements were taken, they could influence $Y_i(1)$

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

Answer

```
In [9]: mean(table4$TreatmentEffect)
```

-0.04

e. Suppose we are able to do an experiment in which we can control the amount of time that 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? (Again, please describe your work.)

Answer: Under such assignment, we can leverage the below formula to arrive to the ATE:
ATE = $E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1)|d_i = 1]$ - $E[Y_i(0)|d_i = 0]$
Arriving at $E[Y_i(1)|d_i = 1]$

```
In [10]: mean(table4[seq(1, nrow(table4), 2),]$y1)
```

0.9
Arriving at $E[Y_i(0)|d_i = 0]$

```
In [11]: mean(table4[seq(2, nrow(table4), 2),]$y0)
```

0.96
ATE = $E[Y_i(1)|d_i = 1]$ - $E[Y_i(0)|d_i = 0]$

```
In [12]: mean(table4[seq(1, nrow(table4), 2),]$y1) - mean(table4[seq(2, nrow(table4), 2),]$y0)
```

-0.0599999999999999

f. How different is the estimate from the truth? Intuitively, why is there a difference?

Answer Intuitively the result is different from the truth given this instatiation of the assignment doesn't necessarily reflect the expected treatment effect across all possible instatiations of the assignment.

g. 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)?

Answer We would have to calculate all possible ways to split children based on the desired number of members on the control group. That is, we would have to add all possibilities of $\frac{n!}{k!(n-k)!}$ for k between 1-9, and n = 10

```
In [13]: comb = function(n, x) {
           return(factorial(n) / (factorial(x) * factorial(n-x)))
         }
```

```
In [14]: n = 10
         totalComb= 0
         for (i in 1:9)
             {
                 totalComb = totalComb + comb(n,i)
                 print(paste('Ways to split in a control group of size '
                             ,i,'and a treatment group of size ',n-i, ': ',comb(n,i)))
             }
         (paste("There are ",totalComb,
                " possible ways to split children between control and treatment groups"))
```

```
[1] "Ways to split in a control group of size  1 and a treatment group of size  9 :  10"
[1] "Ways to split in a control group of size  2 and a treatment group of size  8 :  45"
[1] "Ways to split in a control group of size  3 and a treatment group of size  7 :  120"
[1] "Ways to split in a control group of size  4 and a treatment group of size  6 :  210"
[1] "Ways to split in a control group of size  5 and a treatment group of size  5 :  252"
[1] "Ways to split in a control group of size  6 and a treatment group of size  4 :  210"
[1] "Ways to split in a control group of size  7 and a treatment group of size  3 :  120"
[1] "Ways to split in a control group of size  8 and a treatment group of size  2 :  45"
[1] "Ways to split in a control group of size  9 and a treatment group of size  1 :  10"
```

'There are 1022 possible ways to split children between control and treatment groups'

h. 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.

Answer

```
In [15]: mean(table4[1:5,]$y1) - mean(table4[6:10,]$y0)
```

-0.44

i. Compare your answer in (h) to the true ATE. Intuitively, what causes the difference?

As observed in the original table, the true ATE is -0.04, whereas the 1-5 /6-10 assignment yields an ATE of -0.44. The difference is caused by the fact that the two observations where a non-zero treatment effect is observed are in the 1-5 group, which skews the average $Y_i(1)$ significantly (when looking at the true effect these two measurements are diluted in a higher population, as avg is sensitive to outliers). Similarly the second group (6-10) also happens to have no treatment effect whatsoever. Again, ATE is certainly affected by treatment assignment.

6 5. FE, exercise 2.5

Note that the book typically defines D to be 0 for control and 1 for treatment. However, it doesn't have to be 0/1. In particular, one can have more than two treatments, or a continuous treatment variable. Here, the authors want to define D to be the number of minutes the subject is asked to donate. (This is because "D" stands for "dosage".)

A researcher plans to ask six subjects to donate time to an adult literacy program. Each subject will be asked to donate either 30 or 60 minutes. The researcher is considering three methods for randomizing the treatment.

One method is to flip a coin before talking to each person and to ask for a 30- minute donation if the coin comes up heads or a 60- minute donation if it comes up tails.

The second method is to write "30" and "60" on three playing cards each, and then shuffle the six cards. The first subject would be assigned the number on the first card, the second subject would be assigned the number on the second card, and so on.

A third method is to write each number on three different slips of paper, seal the six slips into envelopes, and shuffle the six envelopes before talking to the first subject. The first subject would

be assigned the first envelope, the second subject would be assigned the second envelope, and so on.

a. Discuss the strengths and weaknesses of each approach.

Answer:
Coin flip: The strength of this approach is that it guarantees random assignment for each subject, furthermore, the assignment of each subject is independent of the assignment of any other participant, as there is no conditional probability on prior assignments. One weakness though is that there is no guarantees as to the sizes of the treatment groups, in this case, the chances of getting a true 50%/50% can be calculated knowing that a coin toss follows a binomial distribution

```
In [16]: dbinom(3, size=6, prob=.5)
```

0.3125

That means there is only a 30% chance that the treatment groups would be of equal (or approximately equal) size, which can result in the averages of a particular group being more or less sensitive to outliers.

Playing cards & Envelopes: The strength of these approaches is that they guarantee an exact 50%/50% split on the control groups. However, the probabilities of a subject being assigned to a given treatment group are not really equal (except for the fist subject) as the probabilties of seeing a '30' or '60' are dependent on the envelops/cards drawn for the previous subjects

b. In what ways would your answer to (a) change if the number of subjects were 600 instead of 6?

Answer The increased population size will make up for the exact 50%/50% split concern I highlighted on answer (a), this is due to the fact that not having the exact same number of subjects on each group is not as significant, given the effect of outliers on large numbers is minimized.

c. What is the expected value of D_i (the assigned number of minutes) if the coin toss method is used? What is the expected value of D_i if the sealed envelope method is used?

Answer as demonstrated below the expected value of D_i is 45 for both methods
Coin flip: we can calculate this by estimating all potential combination of treatment assignment and deriving the D_i for that treatment instance. A concise way would be to leverage the probability of observing such assignment as a weight for the average calculation of D_i

```
In [17]: n = 6
         expectation = 0
         for (i in 0:6)
             {
                 print(paste('Probability of ',i,' subject(s) in D_30 and '
                             ,n-i, ' in D_60 is ',dbinom(i, size=n, prob=.5)
                             ,' with a D_i of ',((30*i+60*(n-i))/6)))
                 expectation = expectation + (dbinom(i, size=n, prob=.5) * ((30*i+60*(n-i))/6))
             }
         print(paste('Expected value of D_i is ',expectation))
```

```
[1] "Probability of  0  subject(s) in D_30 and  6  in D_60 is  0.015625  with a D_i of  60"
[1] "Probability of  1  subject(s) in D_30 and  5  in D_60 is  0.09375  with a D_i of  55"
[1] "Probability of  2  subject(s) in D_30 and  4  in D_60 is  0.234375  with a D_i of  50"
[1] "Probability of  3  subject(s) in D_30 and  3  in D_60 is  0.3125  with a D_i of  45"
[1] "Probability of  4  subject(s) in D_30 and  2  in D_60 is  0.234375  with a D_i of  40"
[1] "Probability of  5  subject(s) in D_30 and  1  in D_60 is  0.09375  with a D_i of  35"
[1] "Probability of  6  subject(s) in D_30 and  0  in D_60 is  0.015625  with a D_i of  30"
[1] "Expected value of D_i is  45"
```

Sealed envelope method

As mentioned before,we know that the expectation is that exactly 50% of the subjects are assigned to each treatment, which means that the expected value of D_i is the simple average of the two treatments, 30 and 60, resulting in an expectated D_i of 45

7 6. FE, exercise 2.6

Many programs strive to help students prepare for college entrance exams, such as the SAT. In an effort to study the effectiveness of these preparatory programs, a researcher draws a random sample of students attending public high school in the US, and compares the SAT scores of those who took a preparatory class to those who did not. Is this an experiment or an observational study? Why?

Answer This is an observational study. We can touch on a couple of critical points to make this point. First, there is no random assignment (either naturally or designed experimentally) between those students who took the SAT score and those who didn't. In addition, there is no excludability of confounding factors, in other words, there is no guarantees that the effect of the preparatory class is isolated, and that the true underlying factor is not other unobserved characteristics of the subjects who took the preparatory class, for example, motivation (maybe the same subjects dedicated a significant amount of their time to further prepare for the exams at home).

8 8. FE, exercise 2.9

A researcher wants to know how winning large sums of money in a national lottery affect people's views about the estate tax. The research interviews a random sample of adults and compares the attitudes of those who report winning more than $10,000 in the lottery to those who claim to have won little or nothing. The researcher reasons that the lottery choose winners at random, and therefore the amount that people report having won is random.

a. Critically evaluate this assumption.
b. Suppose the researcher were to restrict the sample to people who had played the lottery at least once during the past year. Is it safe to assume that the potential outcomes of those who report winning more than $10,000 are identical, in expectation, to those who report winning little or nothing?

Clarifications

1. Please think of the outcome variable as an individual's answer to the survey question "Are you in favor of raising the estate tax rate in the United States?"

2. The hint about potential outcomes could be rewritten as follows: Do you think those who won the lottery would have had the same views about the estate tax if they had actually not won it as those who actually did not win it? (That is, is $E[Y_i 0|D = 1] = E[Y_i 0|D = 0]$, comparing what would have happened to the actual winners, the $|D = 1$ part, if they had not won, the $Y_i(0)$ part, and what actually happened to those who did not win, the $Y_i(0)|D = 0$ part.) In general, it is just another way of asking, "are those who win the lottery and those who have not won the lottery comparable?"
3. Assume lottery winnings are always observed accurately and there are no concerns about under- or over-reporting.

Answer (a): The assumption

The researchers' reasoning assumes that people who play the lottery are a representative of the larger population that may have a view on estate taxes. This assumption is clearly flawed as there is a selection bias associated with the type of individuals that spend money lottery tickets. There might be specific socio-economic or psycological factors that differentiates them from the rest of the population affected by estate taxes.

By the very nature of estate taxes, it is a subject of concern for those individuals who are concerned with passing their assets as inheritance to their relatives/individual of choosing upon their death. If we take into account that in the United States there is an exemption on the first few million dollars worth of assets before taxes are applicable, then is unrealistic to assume that the background and experiences of those who played and won more than $10k in the lottery resemble those of individuals who would otherwise be affected by, and have an opinion on estate taxes.

It is also unrealistic to assume that the motivations behind an individual playing the lottery align with disciplined living wills and estate planning, instruments familiar to those with formed opinions on estate taxes.

Answer (b): Is it safe to assume that the potential outcomes of those who report winning more than $10,000 are identical, in expectation, to those who report winning little or nothing?

Certainly not a safe assumption. First of all, the $10k cutoff seems rather arbitrary, why not define the treatment as winning more than $1k? Maybe that yields more pronounced/related results (if any).

Perhaps more important is the comparability of those who win the lottery and those who have not won the lottery. The expected outcome of someone who wins the lottery. There seem to actually be two effects in play, the fact that someone actually played the lottery and did not win it might actually affect their views on estate taxes. Maybe that population has developed a stronger view on estate taxes as a result, which would be a totally different question altogether: "Does playing and not winning the lottery affect people's views on estate taxes?" That the concept is fundamentally different than what the researchers seek to capture, and diminishes the validity of the expected value comparison across groups.

9 9. FE, exercise 2.12(a)

A researcher studying 1,000 prison inmates noticed that prisoners who spend at least 3 hours per day reading are less likely to have violent encounters with prison staff. The researcher recommends that all prisoners be required to spend at least three hours reading each day. Let d_i be 0 when prisoners read less than three hours each day and 1 when they read more than three hours each day. Let $Y_i(0)$ be each prisoner's PO of violent encounters with prison staff when reading less than three hours per day, and let $Y_i(1)$ be their PO of violent encounters when reading more than three hours per day.

a. In this study, nature has assigned a particular realization of d_i to each subject. When assessing this study, why might one be hesitant to assume that $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ and $E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1]$? In your answer, give some intuitive explanation in English for what the mathematical expressions mean.

Answer

$E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$

This formula implies that the expected PO of violent encounters when not reading at least three hours a day is on average the same across the subjects who do read at least three hours and those who don't across all potential realizations of this experiment.

However there are multiple confounding factors that are not being controlled for as part of this study. First of all, there is no random assignment to the groups, which suggests there might other factors driving the behavior observed, and that the 'natural' realization of the treatment distribution is biased by some other characteristic of the individuals.

One consideration that comes to mind is potential interference issues, where the treatment (lack thereof) of other inmates might influence the outcome of a given subject. That is, the social circle individuals belong to might exert a stronger influence on violent behavior than individual habits alone (e.g. reading)

$E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1]$

Conversely, this formula implies that the expected PO of violent encounters when reading at least three hours a day is on average the same across the subjects who do read at least three hours and those who don't across all potential realizations of this experiment.

Same arguments covered above apply for this proposition, other confounding factors might be the underlying drivers of violent behavior, and the habit of reading (or lack thereof) could be a correlated, but not causal feature.

Another way to think about this problem is selection bias, as the group that naturally engages in at least three hours of daily reading might have adopted non-violence as part of their core values, for example, people how immerse themselves on religious study might adhere to principles of non-violence and develop daily habits of rigorous study, which would explain both the time devoted to reading and the lower PO of violent encounters. This does not mean however, that if these people were to read less than three hours a day, they will engage on violent behavior.