Problem Set #1

Experiments and Causality

Adam Yang

September 18, 2018

1. Potential Outcomes Notation

- Explain the notation $Y_i(1)$.

$Y_i(1)$ **represents the treatment potential outcome for the** i^{th} **observation.**

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

$E[Y_i(1)|d_i = 0]$ **represents the expectation of** $Y_i(1)$ **when an observation is selected at random from the control sample set. In other words, it is the expectation of the treatment potential outcome for an observation, given that the observation is selected at random from the control sample set.**

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

$E[Y_i(1)]$ **represents the expectation of the treatment potential outcome for an observation.** $E[Y_i(1)|d_i = 1]$ **on the other hand, represents the expectation of the treatment potential outcome for an observation, given that this observation is selected at random from the to the treatment sample set. The difference between the two is that** $E[Y_i(1)]$ **involves the treatment potential outcomes of both the control and treatment sets, while** $E[Y_i(1)|d_i = 1]$ **only involves the treatment potential outcomes of the treatment set.**

- 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.

The notation $E[Y_i(1)|D_i = 1]$ **involves the idea of random assignment of subjects into the treatment and control group. In the example given in 2.7, we are told that we want to assign 2 villages to the treatment group and 5 villages to the control group. However, if we were to randomly assign 2 villages to the treatment group, there are 21 different pairs that we can come up with. Therefore,** $E[Y_i(1)|D_i = 1]$ **would be the expected value of the treatment potential outcome for an observation if it were selected from any one of those 21 potential treatment groups. The** $E[Y_i(1)|d_i = 1]$ **deals with the situation where we already know Villages 3 and 7 are in the treatment group so we don't have to worry about the other 20 potential treatment groups when calculating the expected value.**

For $E[Y_i(1)|d_i = 1]$ **assuming Villages 3 and 7 are in the treatment group:**

$\frac{30+30}{2} = 30$

For $E[Y_i(1)|D_i = 1]$**:**

$E[Y_i(1)|D_i = 1] = E[Y_i(1)] = \frac{15+15+30+15+20+15+30}{7} = 20$

2. Potential Outcomes Practice

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

	$Y_i(0)$	$Y_i(1)$	τ_i
Individual 1	5	6	1
Individual 2	3	8	5
Individual 3	10	12	2
Individual 4	5	5	0
Individual 5	10	8	-2

$E[Y_i(1)] = \frac{6+8+12+5+8}{5} = \frac{39}{5}$
$E[Y_i(0)] = \frac{5+3+10+5+10}{5} = \frac{33}{5}$
$E[Y_i(1)] - E[Y_i(0)] = \frac{39}{5} - \frac{33}{5} = \frac{6}{5}$
$E[Y_i(1) - Y_i(0)] = E[\tau_i] = \frac{1+5+2+0-2}{5} = \frac{6}{5}$
$\therefore E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)] = \frac{6}{5}$

3. Conditional Expectations

Consider the following table:

	$Y_i(0)$	$Y_i(1)$	τ_i
Individual 1	10	15	5
Individual 2	15	15	0
Individual 3	20	30	10
Individual 4	20	15	-5
Individual 5	10	20	10
Individual 6	15	15	0
Individual 7	15	30	15
Average	15	20	5

Use the values depicted in the table above to complete the table below.

$Y_i(0)$	15	20	30	Marginal $Y_i(0)$
10	n:1, 14.3%	n:1, 14.3%	n:0, 0%	2/7
15	n:2, 28.6%	n:0, 0%	n:1, 14.3%	3/7
20	n:1, 14.3%	n:0, %:0	n:1, 14.3%	2/7
Marginal $Y_i(1)$	4/7	1/7	2/7	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 $Y_i(0)$.
e. Use the table to calculate the conditional expectation that $E[Y_i(0)|Y_i(1) > 15]$.

$E[Y_i(0)|Y_i(1) > 15] = \sum y(0) \frac{Pr[Y_i(0)=y(0), Y_i(1)>15]}{Pr[Y_i(1)>15]} = \frac{10*\frac{1}{7}+15*\frac{1}{7}+20*\frac{1}{7}}{\frac{3}{7}} = 15$

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

$E[Y_i(1)|Y_i(0) > 15] = \sum y(1) \frac{Pr[Y_i(1)=y(1), Y_i(0)>15]}{Pr[Y_i(0)>15]} = \frac{15*\frac{1}{7}+20*0+30*\frac{1}{7}}{\frac{2}{7}} = 22.5$

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.)

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.)

```
# The treatment effect tau = Y_i(1) - Y_i(0)
answer.POa <- d$y1-d$y0
d$tau <- d$y1-d$y0
knitr::kable(d)
```

child	y0	y1	tau
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.

It looks like in 8/10 of the children sampled, there would be no difference between the potential outcomes of the kid's eyesight whether or not they played outside for more than 10 hours a week. Child #2 would have an eyesight of 0.1 if he were to play outside for less than 10 hours a week, and an eyesight of 0.6 if he were to play outside for more than 10 hours a week. On the other hand, child #5 would have an eyesight of 1.6 if he were to play outside for less than 10 hours a week, and an eyesight of 0.7 if he were to play outside for more than 10 hours a week. So in the case of child #2, playing outside would result in better eyesight, while for child #5, playing outside would result in worse eyesight. This makes me think that for most children, playing outside would not impact their eyesight, but for a couple of the kids, there are some unknown variables that would affect the data. For example, maybe child #2 is unlike the other children and likes to read lying down in dim light while he's indoors, but would not do the same activity if he was outside. On the other hand, maybe child #5 was born with eyes that are sensitive to sunlight and over exposure to sunlight can damage his eyesight.

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

In the previous paragraph I mentioned a couple of examples. For one, the behaviour of the child while inside compared to the behaviour of the child while outside can be different between the children. One kid might like staring at the sun when outdoors and one kid might only stare at the tv when indoors. Another reason can be the genetic predispositions of the kids. The example given above is that some kids might be born with eyes that is sensitive to sunlight so his eyesight would be worse if he spent too much time outside.

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

```
# ATE = sum all treatment effects, then divide by the number of subjects
answer.POd <- sum(d$tau)/10
paste("ATE =", answer.POd)
```

```
## [1] "ATE = -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.)

```
# The estimated ATE based on observed data is found by taking the average of the observed treatment gro
# subtract that by the average of the observed control group.

# Find average of observed treatment group
observed_treatment <- c()
for (i in c(1,3,5,7,9)){
  observed_treatment <- c(observed_treatment, d$y1[i])
}
treatment_avg <- mean(observed_treatment)

# Find average of observed control group
observed_control <- c()
for (i in c(2,4,6,8,10)){
  observed_control <- c(observed_control, d$y0[i])
}
control_avg <- mean(observed_control)

# Estimated ATE = average observed treatment - average observed control
answer.POe <- round(treatment_avg - control_avg, 3)
paste("The estimated ATE from observed data is:", answer.POe)
```

```
## [1] "The estimated ATE from observed data is: -0.06"
```

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

The estimate is 0.02 more negative than the truth. There is a difference because the sample is changed based on how you randomly assign the treatment and control groups. It is only an estimate based on data we can actually observe so we cannot expect it to get the ATE exactly right. I can imagine that by increasing the sample size, we should be able to get closer to the correct ATE.

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)?

```
# We can find this by summing up nCr where n = 10 and r is 1 to 9
sum <- 0
for (i in 1:9){
  sum <- sum + choose(10,i)
}
answer.POg <- sum
paste("The number of ways to split the children into treatment and control groups is:", answer.POg)
```

```
## [1] "The number of ways to split the children into treatment and control groups is: 1022"
```

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.

```
# Find average of observed treatment group
observed_treatment <- c()
for (i in 1:5){
  observed_treatment <- c(observed_treatment, d$y1[i])
```

```
}
treatment_avg <- mean(observed_treatment)

# Find average of observed control group
observed_control <- c()
for (i in 6:10){
  observed_control <- c(observed_control, d$y0[i])
}
control_avg <- mean(observed_control)

answer.POh <- treatment_avg - control_avg
paste("Mean treatment eyesight - mean control eyesite=", answer.POh)
```

```
## [1] "Mean treatment eyesight - mean control eyesite= -0.44"
```

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

In this case the true ATE is -0.04 and we got -0.44 which is much larger in magnitude. The difference is caused because the $Y_i(1)$ observations for kids 1 to 5 is quite a bit lower than the $Y_i(0)$ observations for kids 6 to 10. If the kids were randomly assigned into each of the groups for a carefully structured experiment, then the outcome would be more believable. However, because the kids assigned themselves into these groups, we would have to consider if there are any unobserved variables that would cause a child to both want to play outside more, and to have poorer eyesight.

5. Randomization and Experiments

Suppose that a researcher wants to investigate whether after-school math programs improve grades. The researcher randomly samples a group of students from an elementary school and then compare the grades between the group of students who are enrolled in an after-school math program to those who do not attend any such program. Is this an experiment or an observational study? Why?

This is an observational study because an experiment involves intervention where the researcher would randomly assign the students into the control or treatment groups. This is not the case in this example because the researcher is merely analyzing the data of the students who assigned themselves into the control and treatment groups by choosing whether they want to attend the after-school math programs or not. There are many factors that would cause a student to more likely be enrolled in an after-school math program and have better or worse grades which would cause bias in our resulting conclusion. For example, if poor students are forced to attend the after-school math program by the teachers, then you might wrongfully conclude that attending the after-school math program results in worse grades. On the other had, if only the most diligent and studious kids are more likely to sign up for the after-school math program, then you might overestimate the impact of the after-school math program in improving grades.

6. Lotteries

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.

Let the group that has won more than $10,000 be the treatment group.

It is true that the lottery chooses winners at random, but the population where the researcher is picking from can cause some intrinsic bias. This is mainly due to the possible inherent difference between people who play the lottery and people who don't. Furthermore, even though the lottery chooses winners at random, people self-select themselves into buckets of "likelyhood to win", for example some people buy many tickets per lottery and have a higher chance of winning, some people only buy one ticket, and others might not buy any lottery tickets.

First of all, the researcher is interviewing a random sample of adults, not a random sample of adults that play the lottery a certain amount of times a year. In that case, only people who play the lottery can possibly fall into the treatment group, while people who don't play the lottery can only fall into the control group. Therefore, if people who play the lottery and people who don't play the lottery have differing views about the estate tax, then any correlation found could be influenced by that. Furthermore, people who tend to buy more tickets for one lottery have a higher chance of winning and they may be different than people who only buy one lottery ticket. In other words, there might be some inherent difference in estate tax views between people who have higher chances of winning the lottery and people who have lower chances of winning the lottery.

Secondly, there is such a small chance of finding a random adult who has won the lottery. The researcher might end up with thousands of adults who have not won more than $10,000 and only a couple of adults who have won more than $10,000. That would mean the sample size of the treatment group is likely way to small to get a reliable correlation.

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?

No because this fix did not address my second issue with his experiment design. The sample size of the control group would be huge while the sample size of the treatment group is likely very small. You are essentially basing the opinions of the lottery winners off of a very small representation which is not ideal.

Furthermore, even though people who don't play the lottery are eliminated from the sample, you still have people who barely play the lottery next to people who play the lottery a lot. For causality to be true, you would have to assume that people who play the lottery once a year and people who play the lottery every week don't have unobserved variables that would cause different views on estate tax. Also, you must assume the same thing for people who buy many tickets per lottery and people who only buy 1 ticket per lottery. Therefore, I don't think it is 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 $E[Y_i0|D = 1] = E[Y_i0|D = 0]$, 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.

7. Inmates and Reading

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.

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.

$E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ **is the mathematical statement that essentially means: the prisoners who read more than 3 hours a day would have had the same amount of violent encounters if they actually read less than 3 hours a day when compared to those who actually do read less than 3 hours a day. I would be hesitant to assume this because maybe the prisoners who like to read are intrinsically less violent to begin win. If you force them to read less, it does not necessarily mean they will have just as many violent encounters as those who already don't like reading. Also, maybe the prisoners who are less violent get reading privaledges due to good behaviour. That's another way that the study can be biased.**

$E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1]$ **is the mathematical statement that essentially means: the prisoners who read less than 3 hours a day would have had the same amount of violent encounters if they actually read more than 3 hours a day when compared to those who actualy do read more than 3 hours a day. Similar to the above statement, the intrinsic nature of those who like to read and those who do not like to read can draw some doubt to this statement. A prisoner who likes to read might not be as violent as a prisoner who does not like to read. It's not safe to assume that if you force a prisoner to read more than 3 hours a day, they would have the same amount of violent encounters as those who already read more than 3 hours a day.**