Problem Set #1

Experiments and Causality

January 21, 2019

1. Potential Outcomes Notation

- Explain the notation $Y_i(1)$.
  - **It is the potential outcome if subject 'i' is exposed to the treatment**
- Explain the notation $E[Y_i(1)|d_i = 0]$.
  - **It is the treated potential outcome if subject 'i' is untreated**
- Explain the difference between the notation $E[Y_i(1)]$ and the notation $E[Y_i(1)|d_i = 1]$. (Extra credit)
  - $E[Y_i(1)]$ **is just the expectation that we see the potential outcome if subject 'i' is exposed to the treatment. However, this would encompass outcomes where we see the treated outcome even if the subject was not necessairily treated. That is where** $E[Y_i(1)|d_i = 1]$ **is different, by setting the conditional expectation we are only looking at potential treated outcomes from the treatment group.**
- 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 primary difference lies in the lowercase 'd' versus the uppercase 'D'. The 'd' refers to an actual datapoint inside a dataset. Whereas the 'D' is a random variable that alludes to a hypothetical treatment or lack thereof.**
  - **If we look at the example from exercise 2.7 and follow the assumption that villages 3 and 7 are treated the equation unfolds as follows:** $E[Y_i(1)|d_i = 1] = 30$ **because we already know the exact data points that are in the treatment in the case of** $d_i = 1$**. Contrastly,** $E[Y_i(1)|D_i = 1]$ **implies a hypothetical situation and since we do not know which villages will be chosen for treatment** $E[Y_i(1)|D_i = 1] = 20$ **which is the average of** $Y_i(1)$ **when** $D_i = 1$**.**

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

Answer:

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

$$\frac{6+8+12+5+8}{5} - \frac{5+3+10+5+10}{5} = \frac{1+5+2+0-2}{5}$$

$$\frac{39}{5} - \frac{33}{5} = \frac{6}{5}$$

$$\frac{6}{5} = \frac{6}{5}$$

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	n:1	n:0	0.29
	14.3	14.3	0.00	
15	n:2	n:0	n:1	0.42
	28.5	0.00	14.3	
20	n:1	n:0	n:1	0.29
	14.3	0.00	14.3	
Marginal $Y_i(1)$	0.57	0.14	0.29	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 $E[Y_i(0)|Y_i(1) > 15]$. -$E[Y_i(0)|Y_i(1) > 15] = \sum_{i=1}^{7} \frac{Pr[Y_i(0), Y_i(1) > 15]}{Pr[Y_i(1) > 15]}$

```
prob_greater_than_fifteen = 0.14+0.29
top_frac = (1/7)*10 + (1/7)*15 + (1/7)*20
bot_frac = prob_greater_than_fifteen
expected_value = top_frac/bot_frac
cat("The expected value for part e is: ",expected_value)
```

```
## The expected value for part e is:  14.95017
```

f. Use the table to calculate the conditional expectation that $E[Y_i(1)|Y_i(0) > 15]$. -$E[Y_i(1)|Y_i(0) > 15] = \sum_{i=1}^{7} \frac{Pr[Y_i(1), Y_i(0) > 15]}{Pr[Y_0(1) > 15]}$

```
prob_greater_than_fifteen = 0.29
top_frac = (1/7)*15 + (1/7)*30
bot_frac = prob_greater_than_fifteen
expected_value = top_frac/bot_frac
cat("The expected value for part f is: ",expected_value)
```

```
## The expected value for part f is:  22.16749
```

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.)

-$\tau = y1 - y0$

```
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

```
answer.P0a <- d$tau
cat("Please refer to the table above for answer.P0a")
```

```
## Please refer to the table above for answer.P0a
```

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

-A team of research scientists is conducting an observational study in order to determine whether or not playing in the sun more often leads to better visual acuity. They take a random sample of ten children of age 3 and measure their eyesight. Eyesight is measured again at age 6. The researchers conclude that overall playing outside does not improve visual acuity. Two data points do stand out. James saw his aquity improve by +0.5 and Lily saw her aquity decrease by -0.9. One researcher recalls that James had been playing in the mud the day of his eye test when he was three and had been rubbing his left eye a lot. After some digging it was discovered that most members in Lily's family have astigmatism which causes eyesight to deteriorate irrespective of sunlight. The researchers conclude that overall playing outside does not improve eyesight.

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

-There could be a plethora of reasons why some children have different treatment effects than others. Maybe eyesight develops slowly in some children versus others. It could be possible that children were not focused on the eye exam that was conducted. What if the test was in english and that was not their native language? Similar to Lily's case there could be genetics at play.

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

```
answer.POd <- mean(answer.POa)
cat("The ATE for part d is: ", answer.POd)
```

```
## The ATE for part d is:  -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.)

-ATE in this case can be calculated by the following equation: $ATE = \frac{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i(0)]}{5}$

```
control = c(0.1, 0.9, 2.0, 0.7, 1.1)
treatment = c(1.1, 0.5, 0.7, 1.2, 1.0)
answer.POe <- (sum(treatment) - sum(control))/5
cat("The ATE for part e is: ",answer.POe)
```

```
## The ATE for part e is:  -0.06
```

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

-The estimate is lower by 0.02 which is fairly close to the actual value. The difference arises from our method of random assignment. Because of the way we split, we see one delta value in the treatment and one delta value in the control so the overall ATE is still relatively low.

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)?

$TotalNumberOfWays = \binom{10}{9} + \binom{10}{8} + \binom{10}{7} + \binom{10}{6} + \binom{10}{5} + \binom{10}{4} + \binom{10}{3} + \binom{10}{2} + \binom{10}{1}$

```
answer.POg <- dim(combn(10,9))[2] + dim(combn(10,8))[2] + dim(combn(10,7))[2] + dim(combn(10,6))[2] + di
cat("The total number of different ways: ", answer.POg)
```

```
## The total number of different ways:  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.

```
control = c(2.0, 1.2, 0.7, 1.0, 1.1)
treatment = c(1.1, 0.6, 0.5, 0.9, 0.7)
answer.POh <- (sum(treatment) - sum(control))/5
cat("The ATE for part h is: ",answer.POh)
```

```
## The ATE for part h is:  -0.44
```

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

-The estimate is lower by 0.40 which is significantly lower than the actual value. The difference arises from our method of random assignment. This caused both subjects that appear to show treatment effects to be in the treatment group. Hence, we get a more significant ATE relative to the actual ATE we have for the entire sample.

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. There is no intervention from the researcher. He simply randomly samples a population and compares students who go to an afterschool program versus those that do not.**

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.

-The assumption makes the claim that the randomization that the lottery uses for choosing winners somehow validates the randomization in the study being conducted. If we assume that the randomization method the lottery system uses for choosing a winner is valid, there is no way to guarantee that this experiment inherits that randomization. The reporting from the winners comes from a subset of people, which are those that have played the lottery. The losers, however, could have either played the lottery or not have played the lottery. Now we can see a form of selection bias in play in this experiment from the way the winners are different from the losers. Hence, it is not truly random and the results cannot be trusted.

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?

-Under this assumption the population becomes more similar for winners and losers. Hence, it is safer to assume that the potential outcomes of those reporting 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.

-The mathematical expressions basically means that hypothetically, regardless of whether an inmate reads or not, their expected potential outcome for having violent encounters with the prison staff does not change. Going into more detail we can also say, that the average number of violent encounters inmates have are the same between the individuals who do not read would have had if they did read and those that do read. We would be hesitant to believe this because there are several factors unaccounted for that are at play. It may not explicitly be the act of reading that is driving the number of violent encounters, but rather core differences in the types of people that choose to partake in reading. An example could be that inmates spend more time in their cell to read and, hence, simply have less interation with the prison guards. The inmates that do read are more educated and believe violence to be barbaric. Additionally, given that this is an observational study, the researcher had no control over randomizing those that do read versus those that do read introducing a form of selection bias.