Problem Set 1

Alex, Daniel and Micah

8/27/2019

Potential Outcomes Notation

1. Explain the notation $Y_i(1)$. **This means subject i (or a general subject)'s potential outcome to treatment.**
2. Explain the notation $Y_1(1)$. **This means the subject 1's potential outcome to treatment.**
3. Explain the notation $E[Y_i(1)|d_i = 0]$. **This means the average of subjects' potential outcome to treatment for those who are in the control group.**
4. Explain the difference between the notation $E[Y_i(1)]$ and $E[Y_i(1)|d_i = 1]$. **$E[Y_i(1)]$ means the average of subjects' potential outcome to treatment across all of the subjects we have. Whereas, $E[Y_i(1)|d_i = 1]$ means the average of subjects' potential outcome to treatment for only those who are in the treatment group. The latter is only calculating the average potential outcome of a subset of the former.**

Potential Outcomes and Treatment Effects

1. Use the values in the table below to illustrate that $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - [Y_i(0)]$.
2. Is it possible to collect all necessary values and construct a table like the one below in real life? Explain why or why not. **It is not possible because in real life, we could only observe subjects actual outcome to treatment or control and never both. Therefore, y_0 and y_1 can never both realize for the same subject.**

```
kable(table)
```

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

```
results = table[, .(E_y_0 = mean(y_0), E_y_1 = mean(y_1), E_tau = mean(tau))]
results
```

```
##    E_y_0 E_y_1 E_tau
## 1:    13    15     2
```

Anser to question 1:
We know that
$E[Y_i(1)] = 15$
$E[Y_i(0)] = 13$
$E[Y_i(1) - [Y_i(0)] = E[\tau] = 2$

so $E[Y_i(1)] - E[Y_i(0)] = 15 - 13 = 2 = E[Y_i(1) - Y_i(0)]$

Visual Acuity

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.)

```
kable(d)
```

child	y_0	y_1
1	1.2	1.2
2	0.1	0.7
3	0.5	0.5
4	0.8	0.8
5	1.5	0.6
6	2.0	2.0
7	1.3	1.3
8	0.7	0.7
9	1.1	1.1
10	1.4	1.4

In this table, `y_1` means means the measured *visual acuity* if the child were to play outside at least 10 hours per week from ages 3 to 6. `y_0` means the measured *visual acuity* if the child were to play outside fewer than 10 hours per week from age 3 to age 6. Both of these potential outcomes *at the child level* would be measured at the same time, when the child is 6.

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

```
d[, tau := y_1 - y_0]
kable(d)
```

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

2. 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. **I seems that only two chidren have non-zero treatment effect with one being positive and the other being negative. Since only two of our subjects have non-zero treatment effect, one could try to explain that they are outliers in our dataset. Those individual differences could be caused by anything not related to the treatment without knowing more information about the subjects. One could say maybe subject 3 received some other treatment to correct their vision from age 3 to 6, and that subject 5 suffered some vision damage. It really could**

be anything.

3. For this population, what is the true average treatment effect (ATE) of playing outside. **It is -0.03**

4. 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? (Please describe your work.)

```
treat_outcome = d[child %% 2 == 1, mean(y_1)]
control_outcome = d[child %% 2 == 0, mean(y_0)]
est_ate = treat_outcome - control_outcome
```

Compute the average outcome of treatment by averaging the potential outcome to treatment of odd-numbered children (0.94) and compute average outcome of control by averaging the potential outcome to control of even-numbered children (1). Then subtract the outcome of control from the outcome of treatment (-0.06).

5. How different is the estimate from the truth? Intuitively, why is there a difference? **The difference is 0.03. Though random assignment allows us to have an unbiased estimate of the average treatment effect (i.e., the expectation of our estimated ATE will be equal to the true ATE), individual assingment could still vary and result in an estimate that doesn't equal to the truth for the entire population.**

6. We just considered one way (odd-even) an experiment might split the children. How many different ways (every possible ways) 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)?

```
total <- 0
for (i in 1:9) {
  total <- total + factorial(10)/(factorial(i)*factorial(10-i))
}
```

Possible ways to split $= C_1^{10} + C_2^{10} + ... + C_8^{10} + C_9^{10}$ **= 1022. The rationale is that first we calculate all the ways to have one subject in treatment and nine subjects in control, then two subjects in treatment and eight in control, so on and so forth. We do this sequentially, all the way to nine in treatment and one in control. The answer would just be the sum of all of the terms.**

7. 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.

```
treat_outcome = d[child <= 5, mean(y_1)]
control_outcome = d[child > 5, mean(y_0)]
est_ate = treat_outcome - control_outcome
```

The difference is -0.54

8. Compare your answer in (g) to the true ATE. Intuitively, what causes the difference? **We can see that the estimated ATE is a lot lower than true ATE. The difference is caused the fact that most children that have better vision happened to play less than 10 hours per week outside. Since there isn't any artificial intervention, there could be all kinds of reason why such association exists. For example, maybe children who have weak vision are encouraged to play outside more by their parents who believe playing outside is better for vision.**

Randomization and Experiments

1. Assume that researcher takes a random sample of elementary school children and compare 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 or an observational study? Explain! **This is an observational study because the children sampled were not randomly assigned to enroll in an early childhood education program**
2. 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 or an observational study? Explain! **This is still an observational study because the children sampled were still not randomly assigned to enroll in an early childhood education program**
3. Does your answer to part (2) change if we instead assume that the government assigned students to treatment and control by "coin toss" for each student? **Yes, because now children were randomly assigned to receive treatment.**

Moral Panic

Suppose that a researcher finds that high school students who listen to death metal music at least once per week are more likely to perform badly on standardized test. As a consequence, the researcher writes an opinion piece in which she recommends parents to keep their kids away from "dangerous, satanic music". Let $Y_i(0)$ be each student's test score when listening to death metal at least one time per week. Let $Y_i(1)$ be the test score when listening to death metal less than one time per week.

1. 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; but then, second, tell us the *meaning* of this statement. **The rote definition is that the average potential outcome to control for subjects in the control group is equal to the average potential outcome to control for subjects in the treatment group. The meaning of the statement is that the treatment and control groups, on average, are expected to score the same when listening to death metal at least one time per week.**
2. Do you expect the above condition to hold in this case? Explain why or why not. (describe if independent it'll hold) **The above condition probably doesn't hold if the students are not randomly assigned to listen or not listen to death metal at least one time per week. Without random assignment and in a natural setting, some confounder (guessing out of the blue, more negative worldview) might be causing both listening to death metal and lower score. Therefore, in this case, the $Y_i(0)$ is positively correlated with D_i, and we expect $E[Y_i(0)|D_i = 0] < E[Y_i(0)|D_i = 1]$ rather than $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$**

MIDS Admission

Suppose a researcher at UC Berkeley wants to test the effect of taking the MIDS program on future wages. The researcher convinces the School of Information to make admission into the MIDS program random among those who apply. The idea is that since admission is random, it is now possible to later obtain an unbiased estimate of the effect by comparing wages of those who where admitted to a random sample of people who did not take the MIDS program. Do you believe this experimental design would give you an unbiased estimate? Explain why or

why not. Assume that everybody who gets offer takes it and that prospective students do not know admission is random. **I don't believe this experimental design would give us an unbiased estimate of the effect of MIDS program on future wages if the population of interest includes general people regardless of applying to MIDS or not. Though the admission decision is random, whether to apply is still not. However, applying is necessary to be accepted; therefore, whether a person in the population of our interest is going to be in MIDS is not independent from all other aspects. For example, applying to graduate school could be an indicator of important factors that affect earnings, such as having bachelor's degree, intellectual capacity, drive, etc. Therefore, people who are in MIDS are still different from a random sample of people who did not take the MIDS program (including people did and didn't apply) in important aspects. Therefore, even with this data in hand, we will still not be able to isolate the effect of MIDS on future wages for a more general population that includes people who didn't apply.**