Problem Set 1

Anup Jha

8/27/2019

Potential Outcomes Notation

1. Explain the notation $Y_i(1)$.
   Ans: The notation $Y_i(1)$ represents the potential outcome of the ith subject if it was treated

2. Explain the notation $Y_1(1)$.
   Ans: The notation $Y_1(1)$ represents the potential outcome of the subject number 1 if it were treated

3. Explain the notation $E[Y_i(1)|d_i = 0]$.
   Ans: The notation $E[Y_i(1)|d_i = 0]$ represents the expected treated potential outcome of a subject randomly selected from the group which was not treated.

4. Explain the difference between the notation $E[Y_i(1)]$ and $E[Y_i(1)|d_i = 1]$.
   Ans: The notation $E[Y_i(1)]$ represents the expected treated potential outcome of a subject selected at random either from treatment or control group while the notation $E[Y_i(1)|d_i = 1]$ represents the expected treated potyential outcome of a subject selected at random from the treatment group

Potential Outcomes and Treatment Effects

1. Use the values in the table below to illustrate that $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)]$.
   Ans: We can represent $E[Y_i(1)] - E[Y_i(0)]$ as mean(y_1) - mean(y_0) = 2 and $E[Y_i(1) - Y_i(0)]$ can be represented as mean(tau) = 2 . Please see the r code below which does this calculation

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.
   Ans: It is not possible to collect all the necessary values to construct this table in real life as this table includes both the outcomes of with or without treatment for every subject. In the observations out of an experiment a subject can either be in treatment group or control group so either we can observe treatment outcome or control outcome of a subject and not both.

```
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

```
table[,mean(y_1)-mean(y_0)]
```

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

```
table[,mean(tau)]
```

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

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.
   Ans: We can calculate the individual treatment effect for each of the ten children by taking the y_1 - y_0 for each observation. We call this as 'tau' .

```
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.

Ans: The potential outcome schedule shows that for 8 out of the 10 children the treatment effect is 0 so there is no difference in potential outcome of treated vs untreated for these children. The child number 2 has positive treatment effect where the potential visual acquity increases from 0.1 to 0.7 when untreated vs treated respectively. The child number 5 had reverse treatment effect where the potential visual acquity decreased from 1.5 to 0.6 when untreated vs treated respectively. The reason for different treatment effect for these two children can be thought of as child number 2 benefited from the exercise when playing outside while the child number 5 might have sun allergy which might cause the visual acuity to go down if played in sun regularly. We also see that the children from number 6 to 10 have higher expected value of $Y_i(1) and Y_i(0)$ than of children numbered from 1 to 5 . So there is great chance of selection bias if random process is not utilized to create treatment and control groups.

```
(mean_y1_children_6_10 = d[child %in% 6:10,mean(y_1)])
```

```
## [1] 1.3
```

```
(mean_y1_children_1_5 = d[child %in% 1:5,mean(y_1)])
```

```
## [1] 0.76
```

```
(mean_y0_children_6_10 = d[child %in% 6:10,mean(y_0)])
```

```
## [1] 1.3
```

```
(mean_y0_children_1_5 = d[child %in% 1:5,mean(y_0)])
```

```
## [1] 0.82
```

3. For this population, what is the true average treatment effect (ATE) of playing outside.
   Ans The average treatment affect of playing outside for this population is -0.03

```
ATE_Population <- d[,mean(tau)]
ATE_Population
```

```
## [1] -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.)
   Ans: Since we have treatment group consisting of odd numbered children and control group consisting of even numbered children to calculate the ATE we can take the $mean(Y_i(1)) ;for;odd;numbered;children - mean(Y_i(0)) ;for;even;numbered;children $. We get the value of -.06 .

```
ATE_Experiment <- d[child%%2==1,mean(y_1)]-d[child%%2==0,mean(y_0)]
ATE_Experiment
```

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

5. How different is the estimate from the truth? Intuitively, why is there a difference?
   Ans: The estimate we get is -.06 while the truth is .03. Even though its double the actual the substantive effect is not much. As we see from the data that expected value of y_0 for the population is 1.06 and this effect of -.06 is much smaller. Hence the treatment effect is not substantively significant. There is a difference between true ATE and experimental ATE because the child number 2 goes into control where its y_0 is much less than its y_1 (0.6 less) while child number 5 goes into treatment group and its y_1 is much less than y_0 (0.9 less) . Their differences cancel out the effect and so the ATE from the experiment comes out close to true ATE of the population. We also see that generally the outcomes of the children number from 6 to 10 are higher than children numbered from 1 to 5 but since the odd and even are assigned to treatment group and control we have these children divided about equally to the treatment and control and hence the ATE is not much different.

```
(y0mean <- d[,mean(y_0)])
```

```
## [1] 1.06
```

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)?
   Ans: We can think that every child can be either put in control or treatment group . So for every child there are two possibilities. So total number of possibilities in which the children can be grouped into control and treatment is
   $$2 \times 2 \times 2 ... 10\ times = 2^{10} = 1024$$
   . Out of these possibilities there is one where all the children are assigned into treatment and one where all the children are assigned to control. So total number of possibilities to divide the group is $2^{10} - 2 = 1022$. In general if there were N subjects then total number of ways to divide into treatment and control would be $2^N - 2$. In our case we get the total as **1022**

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.
   Ans: To compute the difference in means for this observational data we need to take the mean of y_1 for children numbered between 1 and 5 inclusive and mean of y_0 for children numbered between 6 and 10 inclusive. R code below does that and we see that ATE for this observational study is -0.54

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

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

8. Compare your answer in (g) to the true ATE. Intuitively, what causes the difference?
   Ans: This difference is caused because of the fact that children numbered between 6 and 10 are put together in control group. While children from 1 to 5 are in treatment group. The children between 6 and 10 have higher acquity in general without treatment and so their is a selection bias with negative effect on the treatment if we take children 1 to 5 in treatment group and children 6 to 10 in treatment group. Since 6 to 10 are in control group and 1 to 5 are in treatment group we can calulate the selection bias mainly $E[Y_i(0)|d_i = 1] - E[Y_i(0)|d_i = 0]$ we get -0.48 which gets reduced from the treatment effect on treated group which is $E[Y_i(1)|d_i = 1] - E[Y_i(0)|d_i = 1]$ which is -.06 so we get the effect as -0.06 + -0.48 = -0.54 . The selection bias seems to exaggerate the treatment effect in negative side and hence might give wrong causal inference.

```
(selection.bias <- d[child <=5,mean(y_0)] - d[child > 5,mean(y_0)] )
```

```
## [1] -0.48
```

```
(ATT <- d[child <=5,mean(y_1)] - d[child <= 5,mean(y_0)] )
```

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

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!
   Ans: This is an observational study. The reason being the researcher didn't do any intervention. They didn't select a few children and administered or not administered the early childhood education program

to some. Since there is no intervention its not an experimental study but an observational study as the researchers merely observe and conclude from the data already available for them.

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!
   Ans: This is an experiment as there was an intervention but since the treatment group assignment is not random the results of this experiment might have selection bias into it and might not provide true causal conclusions. According to the field experiments book this might not be an experiment as randomization is not used to create treatment and control group.

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?
   Ans: If the government assigned students to treatment using coin toss then the experiment would become a randomized control trial and the results of this study might be more acceptable for causal question as the selection bias would not be there. But without knowing the full details of how the subjects were treated in the experiment and did the subjects communicate with each other or their were any third party interaction with the treated or untreated group we cannot conclude on excludability and Non-interefernce which can still make the causal conclusions murky (biased estimate) if the assumptions of excludability and non-interference is not correct. The coin toss also might not make equal treatment and control group in terms of number of subjects.

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.
   Ans: The statement $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ means that the expected potential untreated (listening to death metal atleast once a week) outcome of a randomly selected subject from treatment and control group is the same. Which means that on an average the potential outcome without treatment of the subjects in treatment group and control group is the same. So it means that we don't have a selection bias. So we don't have dubius treatment effect just because of groups which make into the treatment and control group. In the above question this would mean that on average the students in treatment group and control group would have same scores if they listen to death metal atleast once in a week. This mostly happens when the subjects are assigned to treatment and control group by random process.

2. Do you expect the above condition to hold in this case? Explain why or why not.
   Ans: No we don't expect this condition to hold in this case as this is an observational study and the researcher didn't really assign students at random to control and treatment group. So the control group which consists of students who listen to death metal at least once a week would comprise of generally students who are rebellious and have much more going on with their life than just studies so their potential outcome of scores even without listening to death metal would be lower than the students who are not listening to death metal atleast once a week. So the potential grade outcome of these groups without treatment would be different.So effectively we will have selection bias and the treatment effect would be combination of the causal treatment effect plus selection bias.

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.

Ans: In my view this would not give an unbiased estimate. The reason being that the treatment group consists of random people from the group which applied to MIDS but the control group is random people who don't have MIDS degree. So there might be selection bias as people admitted into MIDS are not really random assignment from the population but only random from the group which is interested in data science program. So the expected potential wage of this group might already be higher than control group even without treatment and hence creating a selection bias and hence not providing the unbiased estimate of the treatment effect. The experiment to compare the people who got through MIDS and who applied but were not admitted might be a better experiment. There are other resaons also such as excludability and non-inference which will make the estimates biased even in the better experiment. The group which was admitted to MIDS might not only have MIDS as a reason why their outcome changed. They might add more resources to study data science once they are admitted to MIDS and MIDS program might not be the sole factor which affects the outcome. The control group might still be in communication with the treatment group and they might also add resources other than MIDS to better their outcome and also might get help from the treatment group in their pursuit making the non-interference not true.