Problem Set #1

Nishant Velagapudi

May 21st, 2018

1. Potential Outcomes Notation

- Explain the notation $Y_i(1)$. **This notation refers to the i^{th} outcome value where an intervention was made (outcome referenced by the "Y"). This can also be thought of as the potential outcome of treatment for the i^{th} example.**
- Explain the notation $E[Y_i(1)|d_i = 0]$. **This notation refers to the expected value (the E function with square brackets) of the potential outcome of the i^{th} example, conditional upon the i^{th} example is not treated.**
- Explain the difference between the notation $E[Y_i(1)]$ and the notation $E[Y_i(1)|d_i = 1]$. (Extra credit) **The first refers to the expected potential outcome of the i^{th} datapoint given the treatment. The second refers to this same potential outcome, conditional that the village was actually treated.**
- 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 former is the conditional expectation of the treated potential outcome of the i^{th} datapoint in which a treatment was actually delivered. The latter is the same conditional expectation, but the condition is that the subject would be treated under a hypothetical allocation. To summarize, D_i refers to the random variable of treatment for a subject, while d_i refers to the realization of that same random variable.**

To use exercise 2.7 as an example, the first term refers to the expected value of the treated potential outcome where the village was treated (villages 3 and 7 are the two possibilities here). The second term refers to the same expected value of the treated potential outcome where the village was hypothetically treated (we only know that this will be some two of the seven villages).

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

In this first case, $E[Y_i(1)]$ is the difference of the average of the second column (7.8) and the average of the first column (6.6), which leads to 1.2. This represents the expected value of the potential outcomes given treatment minus the expected value of the potential outcomes given non-treatment. The second case is the average of the third column - the difference of the treated and untreated potential outcomes for each of these subjects.

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	Marg. $Y_i(0)$
10	n:1 %:14.3	n:1 %:14.3	n:0 %:0	28.6%
15	n:2 %:28.9	n:0 %:0	n:1 %:14.3	42.8%
20	n:1 %:14.3	n:0 %:0	n:1 %:14.3	28.6%
Marginal $Y_i(1)$	57.1%	14.3%	28.6%	1.0

a. Fill in the number of observations in each of the nine cells; **Answers in table**
b. Indicate the percentage of all subjects that fall into each of the nine cells. **answers in table**
c. At the bottom of the table, indicate the proportion of subjects falling into each category of $Y_i(1)$. **answers at bottom of table**
d. At the right of the table, indicate the proportion of subjects falling into each category of $Y_i(0)$. **answers at right of table**
e. Use the table to calculate the conditional expectation that $E[Y_i(0)|Y_i(1) > 15]$. **Here, we take the average of all nontreated potential outcomes where the treated potential outcome was greater than 15. We can see that there are three such cases.**

$$E[Y_i(0)|Y_i(1) > 15] = \frac{1 * 10 + 1 * 15 + 1 * 20}{1 + 1 + 1}, E[Y_i(0)|Y_i(1) > 15] = 18.33$$

f. Use the table to calculate the conditional expectation that $E[Y_i(1)|Y_i(0) > 15]$. **We can see that there are only two cases where the untreated potential outcome was greater than 15: we will take the average treated potential outcome for these two examples.**

$$E[Y_i(1)|Y_i(0) > 15] = \frac{1 * 15 + 1 * 30}{1 + 1}, E[Y_i(1)|Y_i(0) > 15] = 22.5$$

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.) **To do this, we're taking the difference of $Y_i(1)$ and $Y_i(0)$ on an individual basis.**

b. In a single paragraph, tell a story that could explain this distribution of treatment effects. **In 8/10 cases the child sees no effect at all, in 1 case the child has worse eyesight as an effect of the treatment, and in the final case the child has better eyesight as an effect of the treatment. It's difficult to gauge any causal effect here: the two affected children may have had some genetic predisposition that led to different outcomes as a result of playing outside (sensitivity to low light, etc).**

c. What might cause some children to have different treatment effects than others? **Genetic factors could lead to different types of eyes. Visual acuity is linked to retinal flexibility and function - it's possible that for some subset of the population, natural light improves this function and in another subset natural light worsens this function (possibly in races/demographics more adapted to high/low natural light respectively).**

d. For this population, what is the true average treatment effect (ATE) of playing outside. **This is just the mean of the individual treatment effects.**

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.) **In this case, our treatment effects can only be measured from odd numbered children, while control effects would only be observed in even numbered children. Thus, we take the mean treated outcome within the treated group ($E[Y_i(1)|D_i = 1]$) and subtract from it the mean non-treated outcome in the control group ($E[Y_i(0)|D_i = 0]$).**

f. How different is the estimate from the truth? Intuitively, why is there a difference? **The true ATE (found in answer.POd), is slightly smaller than the estimated ATE (answer.POe) using random sampling. This is because different subsamples have to be used to calculate a treatment and control effect in practice - we can never calculate a value like answer.POd in reality.**

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 want to calculate all possible combinations that could be used in one group or another (all possible combinations in treatment or control will naturally include all possible treatments for the other group). We want combinations, not permutations, since order does not matter. We need to do this for all possible number of children (anywhere between 1 and 9, since both treatment and control have to have at least one child). The following code iterates in this range and calculates combinations and each value, summing as we go.**

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. **We assign all children above number 5 to the control group and all others to treatment (since above 5 means playing less than 10 hours per week). We then take an average vision acuity measurement in the control and treatment groups and take the difference as per the ATE formula.**

i. Compare your answer in (h) to the true ATE. Intuitively, what causes the difference? **This answer is very different than the true ATE - in fact, it's off by nearly an order of magnitude. The intuitive reason for this is obvious bias in the treatment and control group selection:**

children who can't see well may not want to be outside simply because they cannot see well.

5. Randomization and Experiments

Suppose that a reasearcher 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. An experiment requires random assignment, not only random sampling. In this case, the children or their parents were able to choose whether they wanted to attend such a program. This aspect of freedom will confound any observation of effect: we could note talented students being biased towards taking an after-school math program or being pushed by their parents to do so. Thus, we could not make a claim that after-school math programs had a causal effect on grades until we have random assignment into control and treatment groups.**

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 primary problem with this assumption is that not all people play the lottery, but the claim is accross all individuals. By interviewing a random sample of adults, a large number of people who don't play the lottery (and thus can't be represented in the winning pool) will be included in the non-winning group. It's impossible to gain any information about "people" as a whole here, because our treatment is limited to a subset of individuals.**

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? **I would also challenge this set-up: the potential outcomes of those who report winning may or may not match those who do not report winning. Winners typically spend more on tickets, which introduces a bias in the sampling. Furthermore, the size of the prize might also subset lottery players by type of ticket they buy. In this setup, since we can't assume that the samples in the winning and non-winning group are of similar demographic composition, we can't safely assume that unobserved variables are accounted for, and thus we can't imply any causal relationship between winning and opinion of estate tax.**

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.

There are four described potential outcomes in this question:

$E[Y_i(0)|D_i = 0]$ - average expected violent encounters for individuals who read less than 3 hours a day, given that they read less than three hours a day (this is observed)

$E[Y_i(0)|D_i = 1]$ - average expected violent encounters for individuals who read less than 3 hours a day, given that they read more than three hours a day (this is a hypothetical - if nonreading prisoners did read)

$E[Y_i(1)|D_i = 0]$ - average expected violent encounters for individuals who read more than 3 hours a day, given that they read less than three hours a day (this is a hypothetical - if reading prisoners did not read)

$E[Y_i(1)|D_i = 1]$ - average expected violent encounters for individuals who read more than 3 hours a day, given that they read more than three hours a day (this is observed)

The implication of the stated assumption is that reading for 3 hours a day lowers average violent encounters, instead of related outcome that prisoners who choose to read for 3 or more hours a day are less likely to be violent than prisoners who choose to read for less than three hours a day. We should hesitate to accept this assumption because it assumes that the unobserved factors with respect to committing violent encounters against prison staff are the same for prisoners who choose to read more as they are for prisoners who choose to read less. This is a faulty assumption at first glance.