W241 Problem Set #1

Yang Yang Qian

January 26, 2018

1. Potential Outcomes Notation

1.a.

Explain the notation $Y_i(1)$.

$Y_i(1)$ **is the potential outcome to treatment for subject** i

1.b.

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

$E[Y_i(1)|d_i = 0]$ **is the expected value of the potential outcome to treatment, when a subject** i **is selected at random from those subjects in the control group. This should not be observable in a real experiment.**

1.c.

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

$E[Y_i(1)|d_i = 1]$ **is the expected value of the potential outcome to treatment, when a subject** i **is selected at random from those subjects in the treatment group. In contrast,** $E[Y_i(1)]$ **is the expected value of the potential outcome to treatment, when a subject** i **is selected at random from the entire population. If selection is random,** $E[Y_i(1)|d_i = 1]$ **should be the same as** $E[Y_i(1)]$**.**

1.d.

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.

$E[Y_i(1)|d_i = 1]$ is the expected value of the potential outcome to treatment, when a subject i is selected at random from those subjects in the treatment group. In contrast, $E[Y_i(1)|D_i = 1]$ is the expected value of the potential outcome to treatment, when a subject i is selected at random from those subjects in the treatment group in a hypothetical allocation of treatment.

In FE 2.7, the experimenter randomly selected Villages 3 and 7 from the set of seven villages and places them into the treatment group. Therefore,

$$E[Y_i(1)|d_i = 1] = \sum_{i=1}^{7} Y_i(1) \frac{Pr[Y_i(1), d_i = 1]}{Pr[d_i = 1]}$$

$$= 15 \frac{0}{0 + 0 + \frac{2}{2}} + 20 \frac{0}{0 + 0 + \frac{2}{2}} + 30 \frac{\frac{2}{2}}{0 + 0 + \frac{2}{2}}$$

$$= 30$$

However, if we were to select randomly any two villages for treatment, then

$$E[Y_i(1)|D_i = 1] = \sum_{i=1}^{7} Y_i(1) \frac{Pr[Y_i(1), D_i = 1]}{Pr[D_i = 1]}$$

$$= 15 \frac{\frac{4}{7}}{\frac{4}{7} + \frac{1}{7} + \frac{2}{7}} + 20 \frac{\frac{1}{7}}{\frac{4}{7} + \frac{1}{7} + \frac{2}{7}} + 30 \frac{\frac{2}{7}}{\frac{4}{7} + \frac{1}{7} + \frac{2}{7}}$$

$$= 20$$

2. FE 2.2

Use the values depicted in Table 2.1 to illustrate that $E[Y_i(0)] - E[Y_i(1)] = E[Y_i(0) - Y_i(1)]$.

first,

$$\begin{aligned} E[Y_i(0)] &= \sum \frac{1}{N} Y_i(0) \\ &= \frac{1}{7}(10 + 15 + 20 + 20 + 10 + 15 + 15) \\ &= 15 \end{aligned}$$

second,

$$\begin{aligned} E[Y_i(1)] &= \sum \frac{1}{N} Y_i(1) \\ &= \frac{1}{7}(15 + 15 + 30 + 15 + 20 + 15 + 30) \\ &= 20 \end{aligned}$$

third,

$$\begin{aligned} E[Y_i(0) - Y_i(1)] &= \sum \frac{1}{N} [Y_i(0) - Y_i(1)] \\ &= \frac{1}{7}[(10 - 15) + (15 - 15) + (20 - 30) + (20 - 15) + (10 - 20) + (15 - 15) + (15 - 30)] \\ &= \frac{1}{7}(-35) \\ &= -5 \end{aligned}$$

and finally,

$$\begin{aligned} E[Y_i(0)] - E[Y_i(1)] &\overset{?}{=} E[Y_i(0) - Y_i(1)] \\ \sum \frac{1}{N} Y_i(0) - \sum \frac{1}{N} Y_i(1) &\overset{?}{=} \sum \frac{1}{N} [Y_i(0) - Y_i(1)] \\ 15 - 20 &\overset{?}{=} -5 \\ -5 &= -5 \end{aligned}$$

3. FE 2.3

Use the values depicted in Table 2.1 to complete the table below.

3.a to 3.d

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 $Y_i(0)$.

	$Y_i(1) = 15$	20	30	Marginal $Y_i(0)$
$Y_i(0) = 10$	1/7	1/7	0/7	2/7
15	2/7	0/7	1/7	3/7
20	1/7	0/7	1/7	2/7
Marginal $Y_i(1)$	4/7	1/7	2/7	7/7

	$Y_i(1) = 15$	20	30	Marginal $Y_i(0)$
$Y_i(0) = 10$	14.3 %	14.3 %	0 %	28.6 %
15	28.6 %	0 %	14.3 %	42.9 %
20	14.3 %	0 %	14.3 %	28.6 %
Marginal $Y_i(1)$	57.1 %	14.3 %	28.6 %	100 %

3.e

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

$E[Y_i(0)|Y_i(1) > 15]$ can be calculated by summing, from $i = 1$ to 7 of: Y_i0, weighted by the joint probability of $Pr[Y_i(0), Y_i(1) > 15]$, and divided by the marginal probability of $Pr[Y_i(1) > 15]$

$$
\begin{aligned}
E[Y_i(0)|Y_i(1) > 15] &= \sum_{i=1}^{7} Y_i(0) \frac{Pr[Y_i(0), Y_i(1) > 15]}{Pr[Y_i(1) > 15]} \\
&= 10 \frac{\frac{1}{7} + \frac{0}{7}}{\frac{1}{7} + \frac{0}{7} + \frac{0}{7} + \frac{1}{7} + \frac{0}{7} + \frac{1}{7}} + 15 \frac{\frac{0}{7} + \frac{1}{7}}{\frac{1}{7} + \frac{0}{7} + \frac{0}{7} + \frac{1}{7} + \frac{0}{7} + \frac{1}{7}} + 20 \frac{\frac{0}{7} + \frac{1}{7}}{\frac{1}{7} + \frac{0}{7} + \frac{0}{7} + \frac{1}{7} + \frac{0}{7} + \frac{1}{7}} \\
&= 10 \frac{\frac{1}{7}}{\frac{3}{7}} + 15 \frac{\frac{1}{7}}{\frac{3}{7}} + 20 \frac{\frac{1}{7}}{\frac{3}{7}} \\
&= 15
\end{aligned}
$$

3.f

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

$E[Y_i(1)|Y_i(0) > 15]$ can be calculated by summing, from $i = 1$ to 7 of: $Y_i 1$, weighted by the joint probability of $Pr[Y_i(1), Y_i(0) > 15]$, and divided by the marginal probability of $Pr[Y_i(0) > 15]$

$$E[Y_i(1)|Y_i(0) > 15] = \sum_{i=1}^{7} Y_i(1)\frac{Pr[Y_i(1), Y_i(0) > 15]}{Pr[Y_i(0) > 15]}$$

$$= 15\frac{\frac{1}{7}}{\frac{1}{7}+\frac{0}{7}+\frac{1}{7}} + 20\frac{\frac{0}{7}}{\frac{1}{7}+\frac{0}{7}+\frac{1}{7}} + 30\frac{\frac{1}{7}}{\frac{1}{7}+\frac{0}{7}+\frac{1}{7}}$$

$$= 15\frac{\frac{1}{7}}{\frac{2}{7}} + 20\frac{\frac{0}{7}}{\frac{2}{7}} + 30\frac{\frac{1}{7}}{\frac{2}{7}}$$

$$= 22.5$$

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.

4.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.)

$\tau_i = Y_i(1) - Y_i(0)$

child	y0	y1	t
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

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

While most children are not affected by playing outdoors, we found that increased outdoor playtime affects 2 out of 10 children's eyesight. This can be explained by the increased exposure to UV radiation causing some children's eyesight to improve, and others to worsen.

4.c

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

In some cases, the particular activities engaged by the children while playing outside can either improve their eyesight, such as peering at far-away objects, or worsen their eyesight, such as staring at the sun.

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

$ATE = E[Y_i(1) - Y_i(0)] = \frac{1}{10} \sum Y_i(1) - Y_i(0) = -0.04$

```
## [1] -0.04
```

4.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 = E[Y_i(1)|D_i=1] - E[Y_i(0)|D_i=0] = -0.06 $

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

4.f

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

The estimated ATE is within 1 standard error of the true ATE. The difference is caused by the actual values of the assigned children.

```
## [1] 0.3405877
```

## 4.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)?

Assuming we are choosing m children of 10 for treatment group, then the total number of combinations should be

$\sum_{m=1}^{9}\binom{10}{m} = \binom{10}{1}+\binom{10}{2}+\binom{10}{3}+\binom{10}{4}+\binom{10}{5}+\binom{10}{6}+\binom{10}{7}+\binom{10}{8}+\binom{10}{9} = 1022$

```
## [1] 1022
```

4.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.

$ATE = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] = -0.44$

```
## [1] -0.44
```

4.i

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

The estimated ATE from (h) is more than 1 standard error smaller than the true ATE. The difference is caused by the coincidence that the two children most affected by playing outdoors also happened to be in the treatment group.

5. FE, exercise 2.5

Note that the book typically defines D to be 0 for control and 1 for treatment. However, it doesn't have to be 0/1. In particular, one can have more than two treatments, or a continuous treatment variable. Here, the authors want to define D to be the number of minutes the subject is asked to donate. (This is because "D" stands for "dosage.")

A researcher plans to ask six subjects to donate time to an adult literacy program.

- Each subject will be asked to donate either 30 or 60 minutes.
- The researcher is considering three methods for randomizing the treatment

Method 1:

- is to flip a coin before talking to each person
    - and to ask for a 30-minute donation if the coin comes up heads
    - or a 60- minute donation if it comes up tails

Method 2:

- is to write "30" and "60" on three playing cards each,
    - and then shuffle the six cards.
- The first subject would be assigned the number on the first card,
    - the second subject would be assigned the number on the second card,
    - and so on

Method 3:

- is to write each number on three different slips of paper,
    - seal the six slips into envelopes,
    - and shuffle the six envelopes before talking to the first subject.
- The first subject would be assigned the first envelope,
    - the second subject would be assigned the second envelope,
    - and so on.

5.a

Discuss the strengths and weaknesses of each approach.

Method 1 is a form of simple random assignment. Its strength is that it is easy to perform. Its weakness is that it may create treatment or control groups that are larger than we want.

Method 2 is a form of complete random assignment. Its strength is that it will generate exactly the size treatment and control groups we want. Its weakness is it does not maintain integrity of assignment. For example, a subject may refuse to participate unless they are assigned a time he prefers. Also, the admin may be tempted to assign certain times to certain subjects for non-random reasons.

Method 3 is also a form of complete random assignment. It has the same strength as Method 2, but with additional protection for the integrity of the assignment process. Since the slips of paper are in the envelope, neither the admin nor the subject will know the assignment.

5.b

In what ways would your answer to (a) change if the number of subjects were 600 instead of 6?

If the number of subjects were increased to 600, Method 1 would be even more desireable for its ease of use. Also, as we increase the number of subjects, the less likely Method 1 will create lopsided control vs treatment groups.

Method 2 and Method 3 would be more cumbersome to impelment with physical cards. Perhaps we should consider using a statitical software package to simulate the cards and shuffling.

5.c

What is the expected value of Di (the assigned number of minutes) if the coin toss method is used?

$$E[D_i] = \sum DPr[D_i = D]$$

$$= 30\frac{1}{2} + 60\frac{1}{2}$$

$$= 45$$

What is the expected value of Di if the sealed envelope method is used?

The expected value of D_i would not change if the sealed envelope method is used.

$$E[D_i] = \sum DPr[D_i = D]$$

$$= 30\frac{3}{6} + 60\frac{3}{6}$$

$$= 45$$

6. FE, exercise 2.6

Many programs strive to help students prepare for college entrance exams, such as the SAT. In an effort to study the effectiveness of these preparatory programs, a researcher draws a random sample of students attending public high school in the US, and compares the SAT scores of those who took a preparatory class to those who did not. Is this an experiment or an observational study? Why?

This would be more like an observational study. Although the researcher used random sampling to obtain the initial batch of subjects, she did not intervene with treatment to create the variation we need in order to study causality.

7: Skip in 2017

8. FE, exercise 2.9

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.
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?

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.

8.a

Our researcher is hoping that the lottery process randomly chose winners, and so can split the randomly selected people into 2 groups, which can then be used to calculate difference in attitudes to real estate tax.However, we must be careful that winners can only be chosen from those who choose to play the lottery.Because the reseacher chose his sample from general population, there are 3 groups to consider: * played and won alot, $D_i = 1$ * played and didnt win or won little, $D_i = 0$ * didn't play, also $D_i = 0$

The expected potential outcome of $E[Y_i(0)|D_i = 1]$ **might not be the same as** $E[Y_i(0)|D_i = 0]$ **because the winners, by their desire to play, might already be different from the rest of the population.**

8.b

If the researcher limited his sample to people who had played the lottery at least once during the past year, then it is safer 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. The reason is because the population if more alike each other, since it will consist of lottery players. However, there might still be some confounding influence from frequent players vs less frequent players.

9. FE, exercise 2.12(a)

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.

a. 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.

If we assume $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ and $E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1]$, then we assume that a prisoner's PO of violent encounters is independent of whether or not he is in the more-than-three-hours-per-day group, or in the less-than-three-hours-per-day group. However, we might be hesitant to assume this independence, because there is no guarantee that nature assigned a particular realization of d_i in a random fashion. Therefore, there might be some form of systematic selection bias in the assignment. An example of such a bias might be that, prisoners who are better behaved, less likely to have violent encounters, are systematically more likely to be rewarded more read time per day.