Gaurav Khanna W241 Summer 2018 Problem Set #1

Gaurav Khanna

oday

1. Potential Outcomes Notation

- Explain the notation $Y_i(1)$.

Answer: The notation denotes the outcome of the treatment(represented by 1) on subject "i" in the experiment.

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

Answer: This denotes the Expectation of $Y_i(1)$ (potential treatment outcome) when a subject is selected randomly from subjects that are not exposed to the treatment.

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

Answer: Former is Expection of $Y_i(1)$ (outcome of treatment on a subject) when a subject is chosen at random from the entire set. Latter is Expectation of $Y_i(1)$ (outcome of treatment on a subject) when a subject is chosen at random from those that were actualy 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.

Answer: Former is the Expectation of $Y_i(1)$ (outcome of treatment on a subject) when a subject is chosen at random from those that were actually treated. Latter is the Expectation of above calculated Expectation over all possible d vectors (all possible distributions assigning subjects randomly to control or treatment)

Exercise 2.7 talks about selecting 2 of 7 villages for the treatment group in a random manner (say village 3 and 7). $E[Y_i(1)|d_i = 1]$ would be the average outcome (of treatment) for these 2 villages. The selection of 2 can be done in 2C7 ways though. $E[Y_i(1)|D_i = 1]$ would be the average of the expectation for each of these ways to randomly choose 2 subjects.

Further, the groups could be all sizes from 1 to 6 from within the 7 villages. This'll give us about 1c7 + 2c7.........6c7 ways to allocate between treatment and control. $E[Y_i(1)|D_i = 1]$ would be the expectation of expectations for each of this selection vector (i.e. Expectation of $E[Y_i(1)|d_i = 1]$ over all the ways of selection a group of size $>$ 1 from the 7 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

Answer: From the table above:

$E[Y_i(1)] - E[Y_i(0)] = 39/5 - 33/5 = 6/5$

$E[Y_i(1) - Y_i(0)] = (1 + 5 + 2 + 0_2)/5 = 6/5$

Values are the same

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	1 1/7	1 1/7	0 0	2/7
15	2 2/7	0 0	1 1/7	3/7
20	1 1/7	0 0	1 1/7	2/7
Marginal $Y_i(1)$	4/7	1/7	2/7	1.0

a. Fill in the number of observations in each of the nine cells; Done

b. Indicate the percentage of all subjects that fall into each of the nine cells. **Filled the table (using fractions to be consistent with addition to 1 in the end)**

c. At the bottom of the table, indicate the proportion of subjects falling into each category of $Y_i(1)$. Done

d. At the right of the table, indicate the proportion of subjects falling into each category of $Y_i(0)$. Done

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

Answer: $E[Y_i(0)|Y_i(1) > 15] = (10(1/7) + 15(1/7) + 20(1/7))/(3/7) = 45/3 = 15$

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

Answer: $E[Y_i(1)|Y_i(0) > 15] = (15(1/7) + 30(1/7))/(2/7) = 45/2$

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.)

```
d <- data.frame(child = 1:10,
                y0 = c(1.1, 0.1, 0.5, 0.9, 1.6, 2.0, 1.2, 0.7, 1.0, 1.1),
                y1 = c(1.1, 0.6, 0.5, 0.9, 0.7, 2.0, 1.2, 0.7, 1.0, 1.1) )
knitr::kable(d)
```

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.)

```
# print(d)
# Calculating the treatment effect: y(1) - y(0)
answer.POa <- d$y1 - d$y0
print('Individual treatment effect for 10 children')
```

```
## [1] "Individual treatment effect for 10 children"
```

```
print(answer.POa)
```

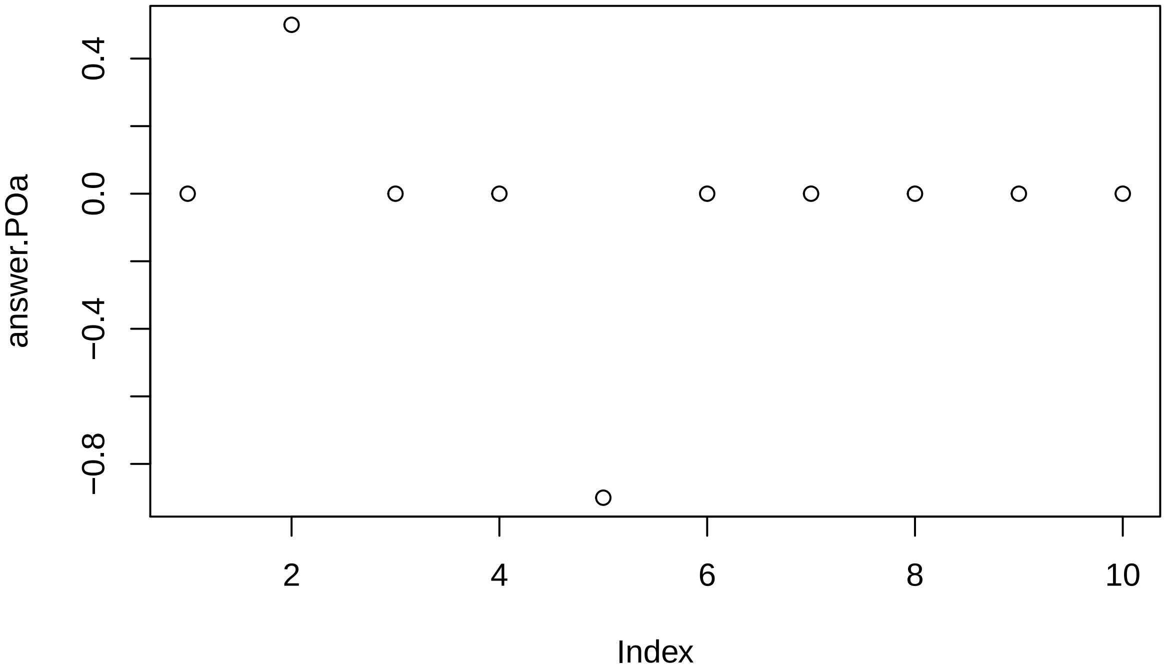
```
##  [1]  0.0  0.5  0.0  0.0 -0.9  0.0  0.0  0.0  0.0  0.0
```

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

```r
# Graph of the individual treatment effects
print('Plot of the individual treatment effects')
```

```
## [1] "Plot of the individual treatment effects"
```

```r
plot.default(answer.POa)
```



Answer: The treatment(outside play) shows no impact for 8/10 subjects. one subject shows a dramatic positive effect(.6-.1=.5) and one subject shows a somewhat less dramatic negative effect(.7-1.6 = .9). Both the impacted subjects seem to be at the edges of what appears to be the normal range (though no impact for y(0) = 2, negates this line of thinking). The effect at the edges could be errors in reporting, data capture or some other change in the context of the subjects (confounding factors...say looking at the sun for case with negative treatment effect). Entry errors are very likely as the measurements are hypothetical (at least for one of y(0) and y(1) since both cannot be measured for the same person). Overall, The individual treatment effect calculation does not lead to a good thesis on the impact of playing outside. A casual observation could be that the treatment (playing outside) has no impact on vision.

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

Answer: The differences could be due to confounding factors or a natural impact of treatment on edge cases. The edge readings could also be due to data capture errors, though its less likely as some of the measurements are hypothetical

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

Answer: The average ATE based on the hypothetical values is the sum of ITE's divided by the number (10)

```r
# Debug print(length(answer.POa))
# Get the sum of ITE's and divide by the number
answer.POd <- (Reduce("+", answer.POa))/length(answer.POa)
answer.POd
```

```
## [1] -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.)

```
# Separating out dataframe in even (control) and odd (treatment i.e. play outside)
treatment <- d[c(TRUE, FALSE), ]
treatment
```

```
##   child  y0  y1
## 1     1 1.1 1.1
## 3     3 0.5 0.5
## 5     5 1.6 0.7
## 7     7 1.2 1.2
## 9     9 1.0 1.0
```

```
control <- d[c(FALSE,TRUE), ]
control
```

```
##    child  y0  y1
## 2      2 0.1 0.6
## 4      4 0.9 0.9
## 6      6 2.0 2.0
## 8      8 0.7 0.7
## 10    10 1.1 1.1
```

```
# Each has the same number of subjects (5)

# Average for the treatment group
a_treatment <- (Reduce("+", treatment$y1))/5
# Average for the control group
a_control <- (Reduce("+", control$y0))/5
# Estimate of the ATE
answer.POe <- a_treatment - a_control
answer.POe
```

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

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

```
# Difference between the estimate and truth (both calculated above)
# Estimate is answer.POe, Truth is answer.POd
print("Difference between estimate and truth")
```

```
## [1] "Difference between estimate and truth"
```

```
answer.POe - answer.POd
```

```
## [1] -0.02
```

Answer: The estimate is 50% smaller than the truth (-.02) There's a difference as the treatment and control group membership is different (one includes the subject with positive impact, other gets the subject with negative impact)

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)?

Answer: The split can be made in 1C10 + 2C10 + 3C10. . . . . 9C10 = 1022 ways. Calculation below

```
# Adding up all the ways to choose a group of at least 1 from 10
it <- 0
n <- length(answer.POa)
for(r in 2:n-1) {
  it <- it + choose(n, r)
}
answer.POg <- it
answer.POg
```

```
## [1] 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.

Answer: The difference in means between treatment and control is -.44. Calculations below

```
# Separating out dataframe. First five (treatment, play outside for > 10 hrs per week) and rest (contro
treatment <- d[1:5,]
treatment
```

```
##   child  y0  y1
## 1     1 1.1 1.1
## 2     2 0.1 0.6
## 3     3 0.5 0.5
## 4     4 0.9 0.9
## 5     5 1.6 0.7
```

```
control <- d[6:10,]
control
```

```
##    child  y0  y1
## 6      6 2.0 2.0
## 7      7 1.2 1.2
## 8      8 0.7 0.7
## 9      9 1.0 1.0
## 10    10 1.1 1.1
```

```
# Average for the treatment group
a_treatment <- (Reduce("+", treatment$y1))/5
# Average for the control group
a_control <- (Reduce("+", control$y0))/5
# Estimate of the ATE
answer.POh <- a_treatment - a_control
answer.POh
```

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

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

Answer: The estimate is 10 times smaller (Estimate is a much larger negative number) than the true ATE. calculations are below The diffrence is due to different group membership. In this specific manner of assignment the treatment group has all the cases that are showing any impact of treatment. The control group is left with 0 ATE cases

```
# Difference between the estimate and truth (both calculated above)
# Estimate is answer.POh, Truth is answer.POd
print("Difference between estimate and truth")
```

```
## [1] "Difference between estimate and truth"
```

```
answer.POh - answer.POd
```

```
## [1] -0.4
```

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?

Answer:It is an observational study.

For the group of students sampled, there's no intervention being done, for the activity to be called an experiment. The students who are attending aftershool program may be doing it as they are behind in grades or just want to learn faster(i.e. are already getting good grades). The intention or ability of students could become a big confounding factor. It'll be an experiment if we picked from students with same grades and then randomly assigned them to after school or control.

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.

Answer:One problem with the argument is that it does not consider the number of tickets somebody is buying and how that's related to their world view. I'd think that the chances improve if we buy more tickets, so we can say that the amount of winnning depends on the investment in tickets. This argument is weakened by the fact that the chances of winning are really small in a typical lottery and one would need to buy a lot of tickets to make a dent in the outcome. The amount invested in buying lottery tickets may also be correlated to economic conditions of the subject to start with. This is also a confounding factor. The factor is reduced by random sampling of adults, but creeps back in as we're comparing winnings and they depend on buying tickets which may depend on economic factors. Again, this factor is not that big due to the small chances of the lottery. There could be other problems with the assumption. Some folks may not choose to reveal the true amount of earnings precisely due to

there views on estate taxes. The clarification notes that this is not an issue though. In crux, there are confounding factors that are not eliminated with the lottery

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?

Answer:I'd still not assume that the potential outcomes are identical. The confounding factors of money invested in buying the tickets and the economic status of subjects do not completely go away (the impact is reduced though as we're only considering subjects who are buying tickets). Here's an example, If a middle class individual, worried about estate taxes, wins a lottery, her views on taxes may not change much. A wealthy individual may be too weathly to care. Somebody poor may still be after the winnings and may not care. If the winning converts the status, they may start caring. There's too many confounding factors to clearly say if the 2 classes are comparable

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.

Answer: $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ **means that the outcome (violence) on less reading when the inmates actually read less would be the same as the violence on reading less for inmates who actually read more**

$E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1]$**? means that the outcome(violence) on reading more when the inmates actually read less would be the same as the violence on reading more for inmates who actually read more**

Together, the expressions are asking if the inmates who read less are comparable to inmates who read more? I'll be hesitant in assuming these equalities. Intuitively, It feels that the inmates who read less would have a different nature and preferences (human) than the inmates who read more. This confounding factor may lead to different results when such a population (those who naturally read less or more) are driven to do something different in the experiment. It could be that inmates who naturally read more could be very irritated and violent when asked to limit reading or may be they are so peace loving that there encounters would not change irrespective of the time spend on reading