Problem Set #1

Andrew Walters

May 24, 2018

1. Potential Outcomes Notation

- Explain the notation $Y_i(1)$.

Y is a potential outcome, which is to say it is the measure of some variable in an experiment. The subscript in Y_i indicates an observable unit in an experiment, such as a person or a city. Finally the 1 in parentheses in $Y_i(1)$ indicates the treatment applied to that unit. In the case of an experiment with a single treatment group and control group, 1 would generally indicate a treatment was applied and 0 would indicate the control group.

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

$E[]$ denotes the expected value of a random variable. $d_i = 0$ denotes the treatment status of unit i. Therefore $E[Y_i(1)|d_i = 0]$ is the expectation of the potential outcome of unit i given that i is in treatment group 1. In statistics, we assume the null hypothesis that this expected value is the same as all other treatment groups.

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

While the two terms will always be equal, the inclusion of d shows that the application of the treatment effect is independent of the potential outcome.

- 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.

D_i refers to the random variable of the treatment, where as d_i refers to the actual treatment that subject i received. So $E[Y_i(1)|d_i = 1]$ means the expected value of the outcome variable for all the subjects that received the treatment while $E[Y_i(1)|D_i = 1]$ means all the subjects that COULD receive the treatment.

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

$E[Y_i(1)] - E[Y_i(0)]$

$E[Y_i(1)] - E[Y_i(0)] = \frac{1}{n}\Sigma_{i=1}^{n} Y_i(1) - \frac{1}{n}\Sigma_{i=1}^{n} Y_i(0)$

$E[Y_i(1)] - E[Y_i(0)] = \frac{1}{5}[Y_1(1) + Y_2(1) + Y_3(1) + Y_4(1) + Y_5(1)] - \frac{1}{5}[Y_1(0) + Y_2(0) + Y_3(0) + Y_4(0) + Y_5(0)]$

$E[Y_i(1)] - E[Y_i(0)] = \frac{1}{5}[6 + 8 + 12 + 5 + 8] - \frac{1}{5}[5 + 3 + 10 + 5 + 10]$

$E[Y_i(1)] - E[Y_i(0)] = \frac{1}{5}[29] - \frac{1}{5}[23] = \frac{6}{5}$

$E[Y_i(1) - Y_i(0)]$

$E[Y_i(1) - Y_i(0)] = \frac{1}{n}\Sigma_{i=1}^{n} Y_i(1) - Y_i(0)$

$E[Y_i(1) - Y_i(0)] = \frac{1}{5}[(Y_1(1) - Y_1(0)) + (Y_2(1) - Y_2(0)) + (Y_3(1) - Y_3(0)) + (Y_4(1) - Y_4(0)) + (Y_5(1) - Y_5(0))]$

$E[Y_i(1) - Y_i(0)] = \frac{1}{5}[(6 - 5) + (8 - 3) + (12 - 10) + (5 - 5) + (8 - 10)]$

$E[Y_i(1) - Y_i(0)] = \frac{1}{5}[1 + 5 + 2 + 0 - 2] = \frac{6}{5}$

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	n: 1 %: 14	n: 1 %: 14	n: 0 %: 0	%: 28
15	n: 2 %: 29	n: 0 %: 0	n: 1 %: 14	%: 43
20	n: 1 %: 14	n: 0 %: 0	n: 1 %: 14	%: 28
Marginal $Y_i(1)$	%: 57	%: 14	%: 28	1.0

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

e. $E[Y_i(0)|Y_i(1) > 15] = \frac{1}{3}[Y_3(0) + Y_5(0) + Y_7(0)] = \frac{1}{3}[20 + 10 + 15] = 15$

f. $E[Y_i(1)|Y_i(0) > 15] = \frac{1}{2}[Y_3(1) + Y_4(1)] = \frac{1}{2}[30 + 15] = \frac{45}{2}$

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.)

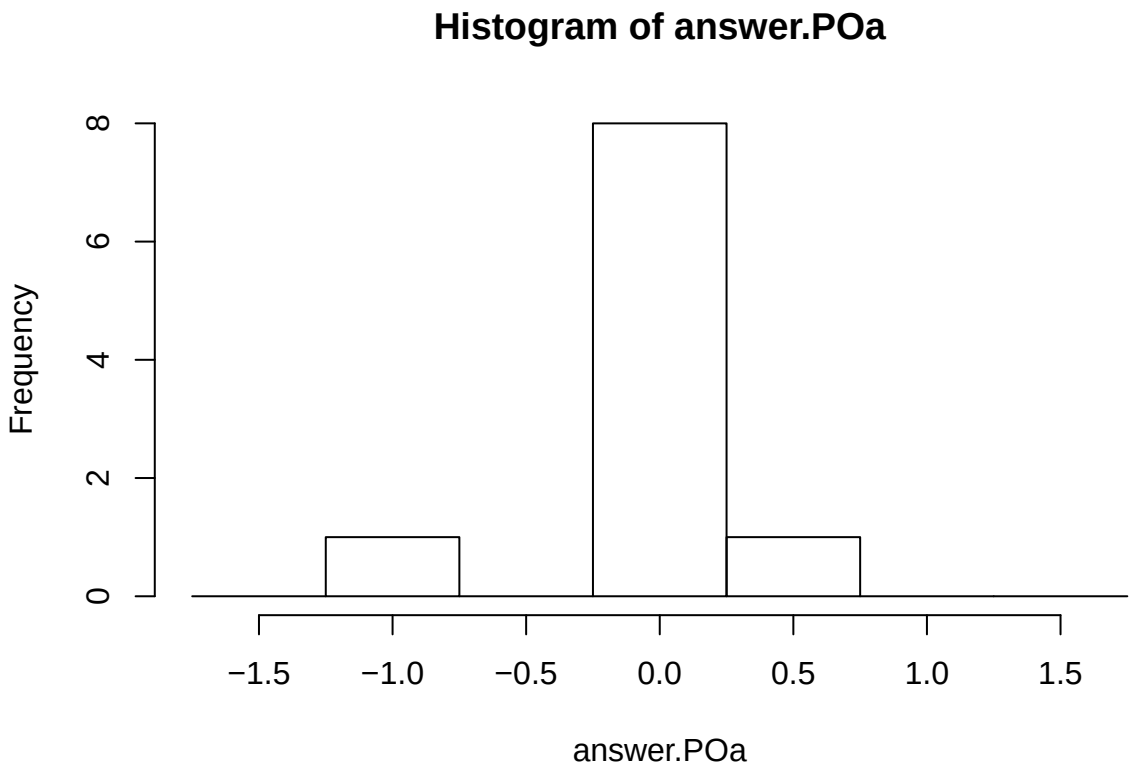
child	y0	y1
1	1.1	1.1
2	0.1	0.6
3	0.5	0.5
4	0.9	0.9
5	1.6	0.7
6	2.0	2.0
7	1.2	1.2
8	0.7	0.7
9	1.0	1.0
10	1.1	1.1

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.)

Equation for individual treatment effects: $\tau_i = Y_i(1) - Y_i(0)$.

```
answer.POa <- d$y1 - d$y0
hist(answer.POa,breaks = seq(-1.75,1.75,0.5))
```



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

In this hypothetical experiment where we get to see both states of the world, we see that the majority of children are unaffected and it is unlikely that a causal effect exists. However two children are affected by the treatment, one of whom has an increase in acuity and the other has a decrease. It is notable that the control group measurements for each child were near the extremes, and both showed acuity closer to the group mean when given the treatment. A likely consideration is that these measurements were just noise. Another possibility is that each child had a confounding factor that was affected by the treatment.

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

The two children with non-zero treatment affects could have been affected by any number of confounding factors. One example for the child who increased his or her acuity as a result of the treatment might be a result of increased blood flow. If that child had a condition that would have affected his or her eyesight had it not been for the physical benefits of additional cardio. For the child with the opposite result, they could have come into contact with a bacteria or virus that they did not encounter in the control group.

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

Equation for average treatment effect: $ATE = \frac{1}{n}\Sigma_{i=1}^{n}\tau_i$

```
answer.POd <- sum(answer.POa) / length(answer.POa)
print(c("The ATE of playing outside is: ", answer.POd))
```

```
## [1] "The ATE of playing outside is: " "-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.)

Equation for ATE in Randomly Sampled Experiment: $ATE = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$

```
treatment_outcomes <- d$y1[c(TRUE,FALSE)] #get odd children
control_outcomes <- d$y0[c(FALSE,TRUE)] #get even children
answer.POe <- sum(treatment_outcomes)/length(treatment_outcomes) - sum(control_outcomes)/length(control
print(c("The ATE of playing outside is: ", answer.POe))
```

```
## [1] "The ATE of playing outside is: " "-0.0599999999999999"
```

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

The truth, that there is (probably) no causal relationship between playing outside and eyesight, is only going to be true on average. Given a small sample size of only 10 children, a small ATE is expected noise. A t-test could tell us whether the ATE was statistically significant.

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)?

There's definitely a formal combinatorics answer for this, but I just solved it with each child like a bit which can have the state 1 or 0. And the range of values that a for a binary number with with 10 bits is $2^{10} = 1024$. Per the instructions remove the states of all-zeros and all-ones so $1024 - 2 = 1022$ possible samples.

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.

```
treatment_outcomes <- d$y1[0:5] #get odd children
control_outcomes <- d$y0[6:10] #get even children
answer.POh <- sum(treatment_outcomes)/length(treatment_outcomes) - sum(control_outcomes)/length(control
print(c("The ATE of playing outside is: ", answer.POh))
```

```
## [1] "The ATE of playing outside is: " "-0.44"
```

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

Again this a small sample size in a particularly exaggerated grouping - mostly positive examples in one sample and mostly negatives in another. Therefore the ATE would return to the true ATE of 0 as more samples were added.

5. Randomization and Experiments

Suppose that a researcher 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 because the treatment was selected by the subjects instead of by the researchers. Because of this, any comparison of the two groups outcomes would be subject to potential confounding factors. There could be many reasons that students who take after school math programs are not the same as those who don't, and only random selection followed by the treatment of after school math would ensure a valid comparison.

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.

While the winners of the lottery are random among the population that plays the lottery, not everyone plays the lottery. Adults who have not played the lottery have no chance of being selected into the treatment group, but they are still part of the control group in this "experiment".

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?

This assumption might be strong enough to conduct a natural experiment, but it is still imperfect. A person can buy multiple lottery tickets and increase their chances of winning, and it is likely that the group of lottery winners plays the lottery more often on average than those that play infrequently. This means that the groups are still not perfectly comparable in expectation.

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.

The problem being described in the two formulas is that the outcome for the group that opted into the treatment (reading) may not be the same as the outcome for the group that was assigned the treatment. The idea being that reading often is a characteristic of a well-behaved prisoner. Therefore forcing the remaining prisoners to read likely will not have the same effect size (or any effect) since the populations are fundamentally different.