Problem Set #1: **ANSWERS**

Experiment Design: Alex & Daniel

September 11, 2018

Potential Outcomes Notation

- Explain the notation $Y_i(1)$.
  - **The potential outcome for subject i when the treatment is present. If subject i were treated, this is what we would observe. Each Y for unit i is a function whose value depends upon whether that unit is in treatment or control.**
- Explain the notation $E[Y_i(1)|d_i = 0]$.
  - **The average outcome we would have observed for the subjects who actually did not receive treatment had they received treatment.**
- Explain the difference between the notation $E[Y_i(1)]$ and the notation $E[Y_i(1)|d_i = 1]$. (Extra credit)
  - **The former is the mean of the treatment potential outcomes for all subjects, while the latter is the mean of the treatment potential outcomes only for subjects actually put in the treatment group in a particular experiment. The latter is what we would observe in a real experiment by calculating the mean of the treatment group outcome.**
- Explain the difference between the notation $E[Y_i(1)|d_i = 1]$ and the notation $E[Y_i|(1)|D_i = 1]$.
  - **d refers to the mean outcome for subjects in the treatment group for an actual realized randomization, such as the mean of Villages 3 and 7 from FE 2.7. The latter, with D, refers to the mean across all possible allocations of treatment assignment where two villages are in the treatment group. In this sense the letter could be written as $E[E[Y_i(1)|D_i = 1]]$, because we are examining the "expected treatment group mean," averaging across two villages within patterns of treatment assignment and then across these averages to get the "average average."**

FE 2.2

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

- $E[Y_i(0)] = \frac{\sum Y_i(0)}{7} = \frac{10+15+20+20+10+15+15}{7} = \frac{105}{7} = 15$
- $E[Y_i(1)] = \frac{\sum Y_i(1)}{7} = \frac{15+15+30+15+20+15+20}{7} = \frac{140}{7} = 20$
- So, $ATE = E[Y_i(1)] - E[Y_i(0)] = 5$

FE 2.3

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

	$Y1 = 15$	$Y1 = 20$	$Y1 = 30$	Marginal $Y_i(0)$
$Y0 = 10$	n: 1 %: 1/7	n: 1 %: 1/7	n: 0 %: 0/7	2/7
$Y0 = 15$	n: 2 %: 2/7	n:0 %:	n: 1 %: 1/7	3/7
$Y1 = 20$	n: 1 %: 1/7	n:0 %:	n: 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;
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 $E[Y_i(1)|Y_i(0) > 15]$.

```
t <- matrix(data = c(1,2,1,1,0,0,0,1,1), nrow = 3, ncol = 3)
rownames(t) <- c(10, 15, 20)
colnames(t) <- c(15, 20, 30)
tp <- t/sum(t)
tp
```

```
##           15        20        30
## 10 0.1428571 0.1428571 0.0000000
## 15 0.2857143 0.0000000 0.1428571
## 20 0.1428571 0.0000000 0.1428571
```

```
tp.mar <- rbind(tp, margin.table(tp, 2))
tp.mar <- cbind(tp, margin.table(tp, 1))
tp.mar
```

```
##           15        20        30
## 10 0.1428571 0.1428571 0.0000000 0.2857143
## 15 0.2857143 0.0000000 0.1428571 0.4285714
## 20 0.1428571 0.0000000 0.1428571 0.2857143
```

$$E[X|Y] = \sum_{\forall x} x * P(X = x|Y = y) \tag{1}$$

$$= \sum_{\forall x} x * \frac{P(X = x, Y = y)}{P(Y = y)} \tag{2}$$

$$E[Y_i(0)|Y_i(1) > 15] = \sum_{\forall Y_i(0)} Y_i(0) * \frac{P(Y_i(0) = Y_i(0), Y_i(1) > 15)}{P(Y_i(1) > 15)} \tag{3}$$

$$= \frac{10 * 1/7}{3/7} + \frac{15 * 1/7}{3/7} + \frac{20 * 1/7}{3/7} \tag{4}$$

$$= \mathbf{15} \tag{5}$$

$$E[Y_i(1)|Y_i(0) > 15] = \frac{15 * 1/7}{2/7} + \frac{20 * 0/7}{2/7} + \frac{30 * 1/7}{2/7} \tag{6}$$

$$= \mathbf{22.5} \tag{7}$$

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.)

```
d <- data.table(d)
d[ , tau := (y1 - y0)]
```

b. In a single paragraph, tell a story that could explain this distribution of treatment effects. *For most children, there is very little (or no) effect of being inside or outside. Howver, some are sensitive, and their visual acuity changes as a consequence of whether they play in the sun or not.*
c. What might cause some children to have different treatment effects than others? *Some children have genes that lead them to have visual acuity that is not affected by the sunlight. Other children, as children of Smeagol, have eyes that perform worse in the bright sunlight, while others, as children of Lando have eyes that perform worse in the dark.*
d. For this population, what is the true average treatment effect (ATE) of playing outside.

```
answer.d <- d[, mean(tau)]
answer.d
```

```
## [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.)

```r
answer.POe <- mean(d[ , y1[seq(1,9,2)] - y0[seq(2,10,2)]])
# or
answer.POe.0 <- d[ , mean(y1[seq(1,9,2)]) - mean(y0[seq(2,10,2)])]
answer.POe
```

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

```r
answer.POe.0
```

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

f. How different is the estimate from the truth? Intuitively, why is there a difference? *Because we've "observed" both the treatment and control potential outcomes for each child, we know what the TRUE τ is, 0.04. How different is our estimate from the truth? Well, 0.02, which doesn't seem too bad (we built it that way...). Why is it different? It is different from the truth because we the particular randomization of treatment and control led to a slightly different calculation.*

g. 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)?

```r
answer.g <- 2^10 - 2

## But there's a longer way of going about it too.
arrangements <- choose(n = 10, c(1,2,3,4,5,6,7,8,9))
sum(arrangements)
```

```
## [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 outside 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.

```r
answer.h <- d[ , mean(y1[1:5]) - mean(y0[6:10])]
answer.h
```

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

i. Compare your answer in (g) to the true ATE. Intuitively, what causes the difference? *Intuitively what has caused the difference is that there is no treatment effect (τ) among the children who chose to stay inside. So then, all the movement happens in the treatment group!

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. Discuss the strengths and weaknesses of each approach

- **Coin Flip.** *Pros:* clear what is happening; everyone understands; in expectation, balanced assignment. *Cons:* no guarantee of balance in small samples, could lead to very uneven assignments
- **Playing Cards.** *Pros:* guarantee balanced assignment. *Cons:* Persons who is assigning knows who will receive what → single blind.
- **Envalopes** *Pros:* guarantee balanced assignment and that the experimenter does not know who will receive what. *Cons:* as sample size grow, this like cards will become quite unwieldy.

b. If there are 600 people, it is going to be mighty difficult to carry around all those cards! Plus, as the size of the sample group increases, then we're going to be less concerned with slight imbalance in the treatment assignments. After all, if would be reasonably likely in the case where `n=6` to end up with a treatment group that has 4 more people in it than the control group. And with a sample of this size, this may lead to substantial bias in our estimate! However, with $n_{treat} \sim 300$ and $n_{control} \sim 300$, we're a *lot* less worried about a difference of 4 in the treatment and control distributions.
c. The expected value is the same under both of these asignment regimes! For

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? What is good or bad about it? *This is clearly a n observational study. Subjects are not randomly assigned to the treatment, which in this case is taking the preparatory class. Instead, they self-select into the treatment for unknown reasons. The fact that the students were sampled randomly from the large population is immaterial; the key issue is whether students in the sample were randomly allocated to the treatment or control group. Note that this research method is prone to bias. If students with higher potential outcomes tend to take the prep class, this research design will tend to produce upwardly biased estimates of the ATE; if students with low potential outcomes tend to take the class in order to improve what they expect to be a sub-par score, this research design will tend to produce downwardly biased estimates of the ATE.*

FE, exercise 2.8

Peisakhin and Pinto report the results of an experiment in India designed to test the effectiveness of a policy called RTIA, which allows citizens to inquire about the status of a pending request from government officials. In their study, the researchers hired confederates, slum dwellers who sought to obtain ration cards (which permit the purchase of food at low cost). Applicants for such cards must fill out a form and have their residence and income verified. The table of results and conditions is below, and there is more in *FE*.

	Bribe	RTIA	NGO	Control
Number Confederates (C)	24	23	18	21
C w/ residence verif	24	23	18	20
M days to verif	17	37	37	37
C w/ ration card 365+	24	20	3	5

a. Interpret the apparent effect of the treatments on the proportion of applicants who have their residence verified and the speed with which verification occurred. *Each of the treatments had a slight effect on the first outcome, the probability of residence verification. In the control group, this rate was 20/21 or approximately 95%. In the three treatment groups, the rate is 100%, implying an average treatment effect of approximately 100 – 95 = 5 percentage points. In terms of the median number of days until residence verification, the RTIA and NGO treatments were the same as the control group, implying an estimated ATE of 37 – 37 = 0. However, the Bribe group received their verification in only 17 days, which is 37 – 17 = 20 days faster than the control group.*

b. Interpret the apparent effect of the treatments on the proportion of applications who actually received a ration card. *In the control group, the rate was 5/21 or 24%. The NGO group fared slightly worse 3/18 = 17%. When a right to information request was filed, this rate jumped to 20/23 = 87%, which approaches the 24/24 = 100% success rate among those who paid a bribe.*

c. What do these results seem to suggest about the effectiveness of the RTIA as a way of helping slum dwellers obtain ration cards? *Although the RTIA treatment does not appear to speed the process of residency verification, it does seem to increase the probability of receiving a card by 20/23 – 5/21 = 63 percentage points over the control group, which seems like a large effect, especially for a treatment that may be implemented inexpensively by applicants.*

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 in random.

a. Critically evaluate this assumption. *This assumption may not be plausible in this application. Although lottery winners are chosen at random from the pool of players in a given lottery, this study does not compare (randomly assigned) winners and losers from a pool of lottery players. Instead, winners are compared to non-winners, where the latter group may include non-players. Winning is therefore not randomly assigned. If frequent players are more likely to win than non-players and the two groups have different potential outcomes, the comparison of the two groups may be prone to bias. For example, perhaps those that play the lottery are more conservative in the first place relative to those who do not; in this case, we may observe lottery winners also being more conservative than Americans who have not won 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, tho those who report winning little or nothing? *The assumption is not rooted in a randomization procedure because frequent players are still more likely to be winners than infrequent players. Unfortunately, without detailed information about how many tickets were purchased for each lottery, we don't know the exact probability that each subject would win. If frequent and infrequent players have different potential outcomes, the comparison is prone to bias.*

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.

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. *In plain(ish) language, the first assumption being made is "On average, the number of violent encounters those who do not read have when they actually do not read is the same as the number of violent encounters who do read would have had if they had not read." The second is "On average, the number of violent encounters those who do not read would have had if they did read is the same as the number of violent encounters those who do read have."*

In this case, those who self-select into the treatment may have distinctive potential outcomes – bookish inmates may be less prone to violence. In that case, $E[Y_i(0)|D_i = 0] > E[Y_i(0)|D_i = 1]$, meaning those who do read would have committed less violence even if they had not read. Thus, a comparison of readers and nonreaders will not tend to produce unbiased estimates of the ATE.