Problem Set 1

Alex, Micah, and Scott 12/01/2020

```r
library(data.table)

knitr::opts_chunk$set(echo = TRUE)
```

Potential Outcomes Notation

1. Explain the notation (Y_i(1)).

   This is the outcome for the “i”th individual if they were in the treatment group

2. Explain the notation (Y_1(1)).

   This is the outcome for the “i”th individual if they were in the control group

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

   This is the expected outcome for for the “i”th individual if they were in the treatment group, given that they are in the control group. This is not actually knowable.

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

   The former notation denotes the expected outcome of the “i”th individual if they were in the treatment group for people in both groups, whereas the latter denotes the expected outcome of the “i”th individual if they were in the treatment group given that they were actually in the treatment group.

Potential Outcomes and Treatment Effects

1. Use the values in the table below to illustrate that (E[Y_i(1)]-E[Y_i(0)] = E[Y_i(1)- [Y_i(0)]).

```r
table
```

```
##    subject y_0 y_1 tau
## 1:       1  10  12   2
## 2:       2  12  12   0
## 3:       3  15  18   3
## 4:       4  11  14   3
## 5:       5  10  15   5
## 6:       6  17  18   1
## 7:       7  16  16   0
```

```r
E_of_yi1_minus_E_of_yi0 = mean(table$y_1 - table$y_0)
E_of_yi1_minus_yi0 = mean(table$y_1) - mean(table$y_0)

"The mean of E[Y_i(1)]-E[Y_i(0)] is:"
```

```
## [1] "The mean of E[Y_i(1)]-E[Y_i(0)] is:"
```

```r
E_of_yi1_minus_E_of_yi0
```

```
## [1] 2
```

```r
"The mean of E[Y_i(1)- [Y_i(0)] is:"
```

```
## [1] "The mean of E[Y_i(1)- [Y_i(0)] is:"
```

```r
E_of_yi1_minus_yi0
```

```
## [1] 2
```

```r
"These are equal, as evidenced by the following:"
```

```
## [1] "These are equal, as evidenced by the following:"
```

```r
E_of_yi1_minus_E_of_yi0 == E_of_yi1_minus_yi0
```

```
## [1] TRUE
```

2. Is it possible to collect all necessary values and construct a table like the one below in real life? Explain why or why not.

   No. It is not possible to know both the Y_0 and Y_1 value for any given i. It is only possible to know one of these values for any given individual. One cannot have been in both the control and treatment group, and once they are in one group, we cannot know what their value would have been had they instead been in the other group.

Visual Acuity

Suppose we are interested in the hypothesis that children playing outside leads them to have better eyesight.

Consider the following population of ten 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.

```r
d <- data.table(
  child = 1:10,
  y_0 = c(1.2, 0.1, 0.5, 0.8, 1.5, 2.0, 1.3, 0.7, 1.1, 1.4),
  y_1 = c(1.2, 0.7, 0.5, 0.8, 0.6, 2.0, 1.3, 0.7, 1.1, 1.4)
)
d
```

```
##     child y_0 y_1
##  1:     1 1.2 1.2
##  2:     2 0.1 0.7
##  3:     3 0.5 0.5
##  4:     4 0.8 0.8
##  5:     5 1.5 0.6
##  6:     6 2.0 2.0
##  7:     7 1.3 1.3
##  8:     8 0.7 0.7
##  9:     9 1.1 1.1
## 10:    10 1.4 1.4
```

In this table:

- `y_1` means means the measured *visual acuity* if the child were to play outside at least 10 hours per week from ages 3 to 6’
- `y_0` means the measured *visual acuity* if the child were to play outside fewer than 10 hours per week from age 3 to age 6;
- Both of these potential outcomes *at the child level* would be measured at the same time, when the child is 6.

1. Compute the individual treatment effect for each of the ten children.

```r
ITE = d$y_1 - d$y_0
ITE
```

```
##  [1]  0.0  0.6  0.0  0.0 -0.9  0.0  0.0  0.0  0.0  0.0
```

2. Tell a “story” that could explain this distribution of treatment effects. In particular, discuss what might cause some children to have different treatment effects than others.

There are several possible explanations for the different treatment effects. One possible story is that an increase in eyesight level is due to an increase in vitamin D levels. Those who had an increase in eyesight with the treatment might have therefore lived in a warmer environment
where more skin could be exposed while playing outside. Those who saw little to no increase in eyesight with the treatment might live in a colder environment where most skin is covered by clothing. Those who saw a decrease in eyesight with the treatment might have had a habit of staring at the sun.

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

```r
true_ATE = mean(d$y_1) - mean(d$y_0)
true_ATE
```

```
## [1] -0.03
```

```
  The true average treatment effect (ATE) of playing outside is: -.03
```

4. 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? (Please describe your work.)

```r
odds = seq(1,10,2)
evens = seq(2,10,2)
new_y_0 = d$y_0[evens]
new_y_1 = d$y_0[odds]
new_ATE = mean(new_y_1) - mean(new_y_0)
new_ATE
```

```
## [1] 0.12

The estimate of the ATE under this assignment is 0.12
```

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

```r
true_ATE - new_ATE
```

```
## [1] -0.15
```

The estimate is .15 away from the truth. Intuitively, there is a difference because the estimate assumes that the average of the opposite group is what the value would have been for the given group, although this is not really the case.

6. 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
x_choose_x<- function(num_people,size_of_smaller_group) {
  x = factorial(num_people)/(factorial(num_people-size_of_smaller_group)*factorial(size_of_smaller_group
  return(x)
}

how_many_ways_order_not_important <- function() {
    groups_of_1and9 = x_choose_x(10,1)
     groups_of_2and8 = x_choose_x(10,2)
     groups_of_3and7 = x_choose_x(10,3)
     groups_of_4and6 = x_choose_x(10,4)
     groups_of_5and5 = x_choose_x(10,5)
     return(groups_of_1and9+groups_of_2and8+groups_of_3and7+groups_of_4and6+groups_of_5and5)
}
how_many_ways_order_not_important()
```

```
## [1] 637
```

There are 637 ways to split the children.

7. 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.

```r
half_and_half <- function(){
  treatment = seq(1,5)
  control = seq(6,10)
  new_y_0 = d$y_0[control]
  new_y_1 = d$y_0[treatment]
  new_ATE = mean(new_y_1) - mean(new_y_0)
  return(new_ATE)
}
half_and_half()
```

```
## [1] -0.48

If we do and observational study as described, the difference in means, or ATE, from the resulting obser
```

8. Compare your answer in (7) to the true ATE. Intuitively, what causes the difference?

```r
true_ATE-half_and_half()
```

```
## [1] 0.45

The answer in part 7 is .45 higher than the true ATE, Intuitively, this difference can be caused by a sa
representative (made more likely by the small sample size), dependent data, non-randomized separation of
causes.
```

Randomization and Experiments

1. Assume that researcher takes a random sample of elementary school children and compare the grades of those who were previously enrolled in an early childhood education program with the grades of those who were not enrolled in such a program. Is this an experiment, an observational study, or something in between? Explain!

   This is an observational study because there is no treatment being aplied to a randomized group, compared to the other randomized group that does not recieve the treatment. Instead, existing data is observed and used. There is no experimental environment or randomization to ensure no confounding factors. The study is therefore observational.

2. Assume that the researcher works together with an organization that provides early childhood education and offer free programs to certain children. However, which children that received this offer was not randomly selected by the researcher but rather chosen by the local government. (Assume that the government did not use random assignment but instead gives the offer to students who are deemed to need it the most) The research follows up a couple of years later by comparing the elementary school grades of students offered free early childhood education to those who were not. Is this an experiment, an observational study, or something in between? Explain!

   This is something in between, but closer to an observational study. A treatment is being applied here instead of just observed, but the application mimics observational data. The only difference between this study and one that is fully observational is that the researcher is offering the free education instead of someone else.

3. Does your answer to part (2) change if we instead assume that the government assigned students to treatment and control by “coin toss” for each student? Why or why not?

   Yes. If the students are randomly assigned to treatment and control by coin toss, the study would be experimental. Who randomizes the groups does not matter - it only matters that the groups are randomized.

Moral Panic

Suppose that a researcher finds that high school students who listen to death metal music at least once per week are more likely to perform badly on standardized test. :metal: As a consequence, the researcher writes an opinion piece in which she recommends parents to keep their kids away from “dangerous, satanic music”.

- Let the potential outcomes to control, (Y_i(0)), be each student’s test score when listening to death metal at least one time per week.
- Let (Y_i(1)) be the test score when listening to death metal less than one time per week.

1. Explain the statement (E[Y_i(0)|D_i=0] = E[Y_i(0)|D_i=1]) in words. First, state the rote english language translation – i.e. “The expected value of …” – but then, second, tell us the *meaning* of this statement. Rote english traslation: The expected value of the “i”th individual if they were in the control group given that they were in the control group equals the expected value of the “i”th individual if they were in the control group given that they were actually in the treatment group.

   Translation in context: The expected value of the “i”th student if they had listened to death metal at least once a week given that they actually did listen to death metal at least once a week (so in essence, the actual measurement of student i, who was in the control group, after the experiment) is equal to the the expected value of the “i”th student if they listened to death metal at least once a week given that they actually listened to death metal less than once a week.

   More concise meaning, as stated by the textbook: Students that do not recieve the treatment have the same expected untreated potential outcome that the treatment group would have if it were untreated.

   Actual meaning in contect: Students that listen to death metal at least once per week have the same expected grades as the students who listen to death metal less than once a week if they were to have listened to death metal at least once per week.

   Even more concisely: There is no underlying difference between the two groups, and the only difference is whether or not they recieved the treatment

2. Do you expect that this circumstance actually matches with the meaning that you’ve just written down? Why or why not?

   If the researcher’s finding was due to a randomly controlled experiment then yes, I expect that the circumstance actually matches with the meaning that I have written down. This is because if it were a randomly controlled experiment, the only difference between the two groups would be the treatment. If, however, the researcher’s findings were based on an observational study rather than an experimental study, then no, I do not expect the circumstance to match with the meaning that I’ve just written down becasue I would expect that the measured treatment effect might be due to some underlying difference between the students who listen to death metal at least once a week and the students who listen to death metal less than once per week.