Problem Set #1

Experiments and Causality

May 24, 2017

1. Potential Outcomes Notation

- Explain the notation $Y_i(1)$: treated potential outcome
- Explain the notation $E[Y_i(1)|d_i = 0]$: the expectation of the potential outcome of the treated group that does not receive treatment
- Explain the difference between the notation $E[Y_i(1)]$ and the notation $E[Y_i(1)|d_i = 1]$:

$E[Y_i(1)]$: the expectation of the potential outcome of the treated group $E[Y_i(1)|d_i = 1]$: the expectation of the potential outcome of the treated group that actually receives treatment

(Extra credit) - 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.

$E[Y_i(1)|d_i = 1]$: the expectation of the potential outcome of the treated group that actually receives treatment

$E[Y_i(1)|D_i = 1]$: the expectation of the potential outcome of the treated group who would be treated under some hypothetical allocation of treatments

2. FE 2.2

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

$E[Y_i(0)] = 15, E[Y_i(1)] = 20$. So $E[Y_i(0)] - E[Y_i(1)] = -5$

$E[Y_i(0) - Y_i(1)] = E[-(treatment\ effect)] = -E[treatment\ effect] = -5$

3. FE 2.3

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

$Y_i(0)$	15	20	30	Marginal $Y_i(0)$
10	1: 14.3%	1: 14.3%	0: 0%	28.6%
15	2: 28.5%	0: 0%	1: 14.3%	42.8%
20	1: 14.3%	0: 0%	1: 14.3%	28.6%
Marginal $Y_i(1)$	57.1%	14.3%	28.6%	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 $E[Y_i(0)|Y_i(1) > 15]$.

$E[Y_i(0)|Y_i(1) > 15] = \sum \frac{Y_i(0)P[Y_i(0)|Y_i(1)>15]}{P[Y_i(1)>15]}$.

With $P[Y_i(1) > 15] = 3/7$:

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

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

$P[Y_i(0) > 15] = 2/7$

$E[Y_i(1)|Y_i(0) > 15] = 15 * (1/7)/(2/7) + 30 * (1/7)/(2/7) = 45/2 = 22.5$

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.

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$treatment_effect <- d$y1 - d$y0
knitr::kable(d)
```

child	y0	y1	treatment_effect
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

child	y0	y1	treatment_effect
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

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

The distribution of the treatment effects show that 8 of 10 children have no change of visual acuity. 1 has negative effect, i.e his/her visual acuity degrades while another has a positive effect, i.e. his/her acuity improves

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

Genetics, nutrition, healthcare might cause some children to have different treatment effects than others

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

```
mean(d$treatment_effect)
```

```
## [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.)

```
# ETA = mean(y1 from the treatment group(odd-numbered children)) -
#       mean(y0 from the control group(even-numbered children))
mean(d$y1[d$child %% 2 == 1]) - mean(d$y0[d$child %% 2 == 0])
```

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

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

Comparing (e) and (d), the difference is 0.02 which is due to error in random selection.

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)?

$\binom{10}{1} + \binom{10}{2} + \binom{10}{3} + \binom{10}{4} + \binom{10}{5} + \binom{10}{6} + \binom{10}{7} + \binom{10}{8} + \binom{10}{9}$

```
binom <- function(n,k) {
  return (factorial(n) / (factorial(n-k) * factorial(k)))
}

total <- 0
for (i in 1:9) {
   total <- total + binom(10,i)
}
total
```

```
## [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.

```r
mean(d$y1[d$child <= 5]) - mean(d$y0[d$child > 5])
```

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

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

If you look at the original data, the average visual acuity for children 1-5 is much lower than the remaining group so this selection method is nonrandom and biased thus causes this difference

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) Strengths and weaknesses of each method:

- Flip coin method allows each of 6 subjects to have equal probability to be assigned to donate either 30 or 60 minutes, regardless of the previous assignment. This is simple random assignment.

Weakness: the size is too small (6) so you may still have unequal sized groups

Strength: practically very easy to execute by just flipping the coin to assign each of 6 subjects

- Both the second (card) and third (slips or paper) assignment methods are in fact the same.

Weakness: they are not practical because there are lots of prepraration when preparing the cards or envelops

Strength: They both assures that there will be only 3 assigned to 30 minutes and 3 to 60 minutes at the end.

(b) If the number of subjects were 600 instead of 6:

- Flip coin method now should have mostly equal sized groups due to large size of sampling (600)
- Both the second and third methods are become much more labor-intensive when we have to generate 600 cards or slips of paper with their envelops.

(c) If coin toss method is used, $E[D_i] = \sum d * P(D_i) = 30 * 0.5 + 60 * 0.5 = 45$. If sealed envelop method is used, $E[D_i] = 30 * (1/6 + 1/6 + 1/6) + 60*(1/2) = 45$

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 is still observational study even it's using random sampling, it's still using the SAT scores *before* the selection so it didn't intervene the students by randomly assigning them to preparatory class. And without random assignment, selection bias ($E[Y(0)|D = 0] - E[Y(0)|D = 1]$) is not 0:

$Y(0)|D = 0$ potential outcome of test score without taking the prep, if they didn't actually take the prep. This group (1) can be students who are already smart and don't need to prepare for SAT so their scores are still high

$Y(0)|D = 1$ potential outcome of test score without taking the prep, if they actually took the prep. This group can be students who are below average so even with the prep, their scores are still lower than group (1)

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:

The assumption that these 2 groups are the same to begin with is not correct:

(a) people who wins lottery can be those who want to play lottery
(b) people who don't win can be those who never play lottery because they already are wealthy and have a well-formed opinion about estate tax

So the groups are not the same (not randomly assigned) therefore we cannot understand the causal effect

Y_i(0): potential outcome of control group who is not favorable of estate tax Y_{i}(0)|D=1: potential outcome (view about estate tax) for those who don't win lottery if they actually won lottery -> counterfactual

Y_{i}(0)|D=0: potential outcome (view about estate tax) for those from those who didn't win lottery (without money) if they actually didn't win 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?

Restricting the sample to people who had played the lottery at least once during the past year eliminates the group of people who is not interested in playing the lottery and had well-formed opinion about estate tax. However, we are still left with these 2 different groups of people to select from:

- people who habitually play lottery
- people who occassional play lottery (ex: people who join their coworkers for superlotto once 1 year)

Because they are different groups to sample from, we can't 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 $E[Y_i 0|D = 1] = E[Y_i 0|D = 0]$, 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.

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.

$E[Y_i(0)|D_i = 0]$: expectation of the control potential outcome of violent encounters when reading LESS than three hours per day given the group read LESS than 3 hours.

$E[Y_i(0)|D_i = 1]$ means expectation of the control potential outcome of violent encounters when reading LESS than three hours per day given the group read MORE than 3 hours.

For people who assigned to read less than 3 hours, they may not be the same: - Some are already violent - Some can potentially already be well-read, educated and tend to be less violent so reading less doesn't make them more violent

Therefore there is difference the expectations of these 2 groups.

Now, considering these 2 groups:

$E[Y_i(1)|D_i = 0]$: expectation of the control potential outcome of violent encounters with prison staff when reading MORE than three hours per day given the group read LESS than 3 hours.

$E[Y_i(1)|D_i = 1]$: expectation of the control potential outcome of violent encounters with prison staff when reading MORE than three hours per day given the group read MORE than 3 hours

They may not be the same: - Some who were already violent because they are illiterate can become more iritated thus and more violent when forced to read more - Some who were already well-read so making them read more doesn't make them less or more violent

Therefore there is difference the expectations of these 2 groups.