Problem Set 1

Prakhar Maini

01/11/2020

```
library(data.table)

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

Potential Outcomes Notation

1. Explain the notation $Y_i(1)$.
2. Explain the notation $Y_1(1)$.
3. Explain the notation $E[Y_i(1)|d_i = 0]$.
4. Explain the difference between the notation $E[Y_i(1)]$ and $E[Y_i(1)|d_i = 1]$

ANSWER

1) $Y_i(1)$:= Potential Outcome for subject i if the subject is exposed to the treatment (i.e. treated)

2) $Y_1(1)$:= Potential Outcome for subject 1 if the subject is treated

3) $E[Y_i(1)|d_i = 0]$:= Expected value of *Treated potential outcome* for subject i who is *not* treated

4) Difference between $E[Y_i(1)]$ and $E[Y_i(1)|d_i = 1]$:= $E[Y_i(1)]$ refers to the expected value of potential outcome for subject i if treated where as $E[Y_i(1)|d_i = 1]$ refers to the expected value of $Y_i(1)$ when subject i is selected from the set of treated subjects

Potential Outcomes and Treatment Effects

```
table <- data.table(
  subject = 1:7,
  y_0 = c(10, 12, 15, 11, 10, 17, 16),
  y_1 = c(12, 12, 18, 14, 15, 18, 16),
  tau = c(2, 0, 3, 3, 5, 1, 0)
)
```

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

ANSWER We look at the following table

id	y_0	y_1	tau
1	10	12	2
2	12	12	0
3	15	18	3

id	y_0	y_1	tau
4	11	14	3
5	10	15	5
6	17	18	1
7	16	16	0

From this table we see that

$$E[Y_i(1)] = \frac{12+12+18+14+15+18+16}{7} = 15$$

$$E[Y_i(0)] = \frac{10+12+15+11+10+17+16}{7} = 13$$

So $E[Y_i(1)] - E[Y_i(0)] = 2$.

Now looking at the RHS we have:

$$E[Y_i(1) - Y_i(0)] = \frac{2+0+3+3+5+1+0}{7} = 2$$

This way we see that the parity holds between RHS and LHS of the equation $E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)]$. This is also true due to the linearity of expactation property.

In real life, a given subject i will either be treated or not treated. If the suject is treated, we will observe $Y_i(1)$ and if not we will observe $Y_i(0)$ but we will not be able to observe both. Hence, in real life, we will not be able to construct such a full table.

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.

```
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)
)
```

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.

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.
3. For this population, what is the true average treatment effect (ATE) of playing outside.
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.)
5. How different is the estimate from the truth? Intuitively, why is there a difference?
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)?
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.
8. Compare your answer in (7) to the true ATE. Intuitively, what causes the difference?

ANSWER Individual treatment effect for a child i will be $\tau_i = Y_i(1) - Y_i(0)$. This way we can construct the following table.

child	y_0	y_1	τ
1	1.2	1.2	0
2	0.1	0.7	0.6
3	0.5	0.5	0
4	0.8	0.8	0
5	1.5	0.6	-0.9
6	2.0	2.0	0
7	1.3	1.3	0
8	0.7	0.7	0
9	1.1	1.1	0
10	1.4	1.4	0

In order to explain this distribution of treatment effect, we can putforth following arguments:

We see no effect for majority of kids because the effect of sun-light is small on visual acuity. However, some children might be playing around reflective surfaces or playing games which leads them to stare at sun for long time which impair their visual acuity. Some kids might be coming from not so financially well-off homes and are not able to get the right nutritions. This might impairs their visual acuity with or without playing outside for 10 hours a week. Some children while playing outside might be able to know about their existing eye-condition or deterioration in eye-sight and will be able to seek help (via change in diet and/or seeking medical help) at the right time which enables them to have better visual acuity.

Looking at the data, we see that the true ATE can be calculated as following:

$$ATE_{true} = \frac{0.6 - 0.9}{10} = -0.03$$

If we assign odd-numbered children to treatment group, we will be able to only observe $Y_i(1)$ for them and only $Y_i(0)$ for rest of the kids. Keeping this in mind we can do folloing computation

$$E[ATE] = E[Y_i(1)|d_i = 1] - E[Y_i(0)|d_i = 0]$$

looking at the real data we have

$E[Y_i(1)|d_i = 1] = \frac{1.2+0.5+0.6+1.3+1.1}{5} = 0.94$ and $E[Y_i(0)|d_i = 0] = \frac{0.1+0.8+2.0+0.7+1.4}{5} = 1.0$

Hence our estimate of ATE will be -0.06

We see from above calculations that the estimated ATE is 2x in size compared to ATE_{true} and in the same direction. Intuitively, we expect some variation between the estimated value and true value because of the small size of population (i.e. 10). If the overall population size was bigger, we expect less variation to occur in calculated ATE vis-a-vis True-ATE due to sampling.

Coming to the number of possible ways to split 10 children in 2 groups, there are 2^{10} ways to do that (a binary choice for each children). However, due to the constraint that atleast 1 child should be in both test and control set, we need to remove 2 edge cases (i.e. all children in either test or control). Hence with the given constraint, we have $2^{10} - 2 = 1022$ ways of splitting 10 children into two groups.

If we only do observational study and children 1-5 choose to be in the treatment set and 6-10 in control we will have $E[Y_i(1)] = \frac{1.2+0.7+0.5+0.8+0.6}{5} = 0.76$ and $E[Y_i(0)] = \frac{2.0+1.3+0.7+1.1+1.4}{5} = 1.30$. This way the observational study will estimate the effect size as $0.76 - 1.3 = -0.54$.

If we compare the result of the observational study (-0.54) with the ATE_{true} (-0.03), we see a huge difference. This underscores the importance of randomization in an experiment when calculating the effect size. With out randomization and explicit intervention, kids with good eye-sight might have chosen to play outside more and kids with worse eye-sight might have chosen to not play out more. This resulted in a heavily biased effect size in the observational study.

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!
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!
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?

ANSWER In the first case, though researcher has taken a random sample, the resulting difference in grades can't be causally related with enrollment in the early childhood education program. There might be systemic factors (such as socio-economic status etc.) which might be confounding the results becasue the enrollment to the childhood program may or may not be randomized. In that sense, this study is an observational study and the inference might be biased.

In the second case, we know that the core problem of assigning student to enrollment was done based on a systemic factor (i.e. need) and not randomly. This way, the core problem remains the same as the first study. The same factors that define "need" might be responsible for the the difference in ATE between test and control. The second study is also an observational study only where the inference might be biased.

Finally, in the third study, the student assignment to enrollment was random (i.e. based on coin flip). This solves the core problem. Now that we have random assignment and intervention (childhood education program) in place, the third study is an experiment and we expect the inference to be unbiased.

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

ANSWER The statement $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ literally means that the **Expected value of treated potential outcome among subjects that receive treatment is same as the Expected value of treated potenntial outcome among subjects that don't receive treatment**. This statement is true only under random assignment of subjects in treatment and control groups. Essentially, this statement implies that random assignment into treatment group or control group is conveys no information about the potential value $Y_i(0)$ (i.e. potential value of outcome if the subject i is not treated). This is true for $Y_i(1)$ as well. Random assignment ensures that observed expected value $E[Y_i(0)|D_i = 0]$ is equal to unobserved expected value $E[Y_i(0)|D_i = 1]$ which is in turn equal to the expected value $E[Y_i(0)]$.

In the present circumstance, the attribution of students to test (listeing to death metal less than once / week) and control (listening to death meatal at least once / week) group is not random. There might be unobserved heterogeneity that might cause students from a cetain back-ground to listen to death metal more (e.g. divorced parents etc.). In that case, students in the control set are not representative of general student population. This will mena that the equality won't hold and we might have selection bias when calculating the treatment effect.