Problem Set #1 - DATASCI W241

Greg Ceccarelli

Sept 15, 2015

1. On the notation of potential outcomes

a. Concisely, $Y_i(1)$ is the outcome if the subject i is exposed to the treatment
b. $E[Y_i(1)|d_i = 0]$ denotes the expectation of $Y_i(1)$ (i.e. the outcome of the subject i exposed to the treatment) when one subject is selected at random from the subjects that weren't treated (i.e. the counterfactual). In practice we won't actually observe $Y_i(1)$ for an untreated subject.
c. $E[Y_i(1)|d_i = 1]$ is the expectation of $Y_i(1)$ (i.e. the outcome) when one subject is selected at random from the subjects that were treated whereas $E[Y_i(1)]$ is the expected $Y_i(1)$ potential outcome for the entire set of subjects.
d. The notation $E[Y_i(1)|d_i = 1]$ is the expectation of a potential outcome for one subject selected at random (indicated by i) from those subjects given the treatement, whereas $E[Y_i(1)|D_i = 1]$ can be regarded as shorthand for $E[E[Y_i(1)|d_i = 1, d]]$ or the expectation of the expected value after summing over all possible d vectors.

2. FE, exercise 2.2

```
#set up working envrionment
setwd('/Users/ceccarelli/MIDS/DATASCI_W241/Async Material and Sample Files/Chapter 2/')
#clear variables
rm( list = ls() )
#read in tabular data
potential_outcomes.data <- read.csv("GerberGreenBook_Chapter2_Table_2_1.csv", sep=",", header = TRUE)
#create shorthand reference
p<-potential_outcomes.data
##View(potential_outcomes.data)
#Define E[Yi(1)]
eYi<-sum(p$Yi.1.)/nrow(p)
#Define E[Yi(0)]
eYo<-sum(p$Yi.0.)/nrow(p)
#Define  E[Yi(0) - Yi(1)]
ate<-sum(p$Yi.0-p$Yi.1.)/nrow(p)
#Test equivalence between E[Yi(0)] - E[Yi(1)] = E[Yi(0) - Yi(1)]
(eYo - eYi) == ate
```

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

3. FE, exercise 2.3

a.

```
#set up working envrionment
p.subset <- p[c(2:3)]
#create matrix to count observations by variable pairs
```

```
mat<-table(p.subset)
#print matrix
mat
```

```
##      Yi.1.
## Yi.0. 15 20 30
##    10  1  1  0
##    15  2  0  1
##    20  1  0  1
```

b.

```
#count number of obs in matrix
num_obs = sum(mat[,1]+mat[,2]+mat[,3])
#calc percentage of subjects in each cell (joint freq distribution)
mat.jfd <- mat / num_obs
#print new matrix
mat.jfd
```

```
##      Yi.1.
## Yi.0.        15        20        30
##    10 0.1428571 0.1428571 0.0000000
##    15 0.2857143 0.0000000 0.1428571
##    20 0.1428571 0.0000000 0.1428571
```

c.

```
#marginal distribution of Yi(1)
mat.jfd[1,]+mat.jfd[2,]+mat.jfd[3,]
```

```
##        15        20        30
## 0.5714286 0.1428571 0.2857143
```

```
#add as row in matrix
mat.jfd <- rbind(mat.jfd, mat.jfd[1,]+mat.jfd[2,]+mat.jfd[3,])
#Update Rownames
rownames(mat.jfd)[4]<-"Yi.1"
mat.jfd
```

```
##             15        20        30
## 10   0.1428571 0.1428571 0.0000000
## 15   0.2857143 0.0000000 0.1428571
## 20   0.1428571 0.0000000 0.1428571
## Yi.1 0.5714286 0.1428571 0.2857143
```

d.

```
#marginal distribution of Yi(0)
mat.jfd[,1]+mat.jfd[,2]+mat.jfd[,3]
```

```
##        10        15        20      Yi.1
## 0.2857143 0.4285714 0.2857143 1.0000000
```

```
#add as column in matrix
mat.jfd <- cbind(mat.jfd, mat.jfd[,1]+mat.jfd[,2]+mat.jfd[,3])
#update colnames
colnames(mat.jfd)[4]<-"Yi.0"
mat.jfd
```

```
##             15        20        30      Yi.0
## 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
## Yi.1 0.5714286 0.1428571 0.2857143 1.0000000
```

e.

```
#use table to calculate condition expectation that E[Yi(0)|Yi(1)>15]
marginal.e <- mat.jfd[4,3]+mat.jfd[4,2]

10*(mat.jfd[1,2]/marginal.e)+15*(mat.jfd[2,3]/marginal.e)+20*(mat.jfd[3,3]/marginal.e)
```

```
## [1] 15
```

f.

```
#use table to calculate condition expectation that E[Yi(1)|Yi(0)>15]
#
marginal.f <- mat.jfd[3,4]
15*(mat.jfd[3,1]/marginal.f)+20*(mat.jfd[3,2]/marginal.f)+30*(mat.jfd[3,3]/marginal.f)
```

```
## [1] 22.5
```

4. More practice with potential outcomes

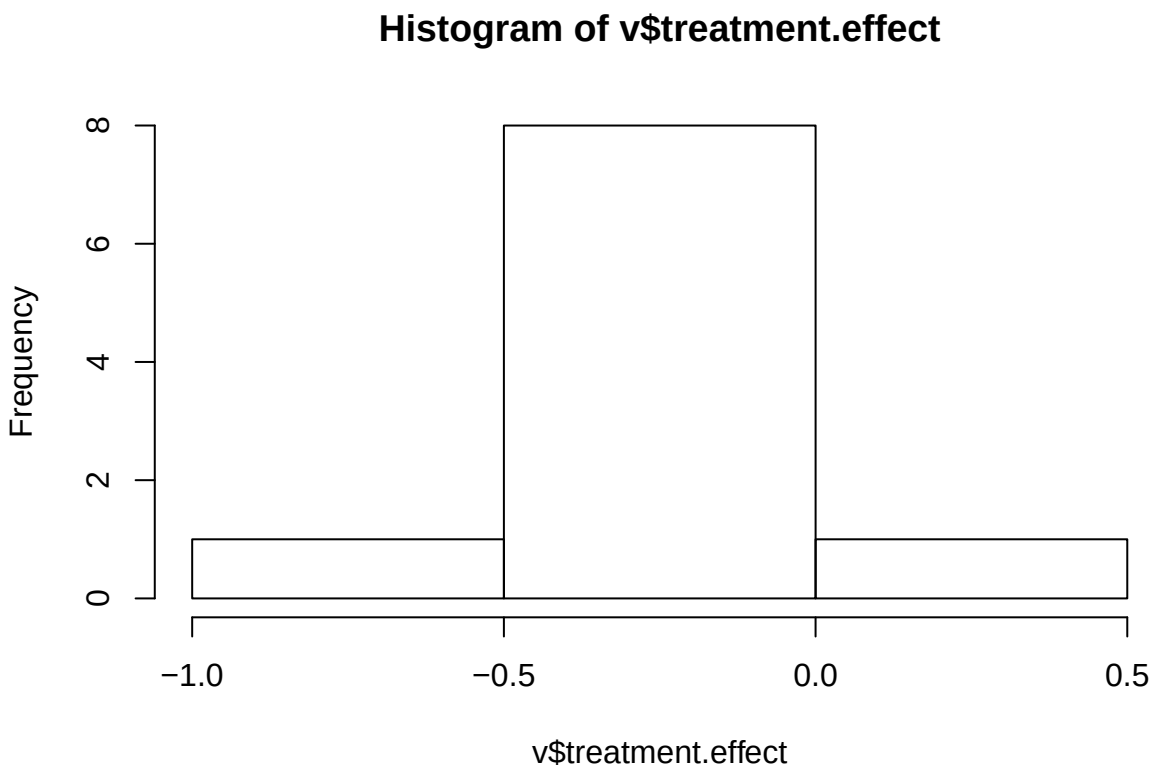
a.

```
setwd('/Users/ceccarelli/MIDS/DATASCI_W241/Async Material and Sample Files/Chapter 2/')
#read in visual acuity data
visualacuity.data <- read.csv("VisualAcuity.csv", sep=",", header = TRUE)
#create shorthand reference
v<-visualacuity.data
#compute treatment effect
v$treatment.effect <- v$Yi.1.-v$Yi.0.
v$treatment.effect
```

```
##  [1]  0.0  0.5  0.0  0.0 -0.9  0.0  0.0  0.0  0.0  0.0
```

b.

```r
#quickly review distribution of treatment effects
hist(v$treatment.effect)
```



In the case of example hypotehtical treatment effects at the child level, there are many factors that might influence different effects and explain this distribution: genetics (i.e. different rates of eye development of those children measured at age 6), differing amounts of actual outside exposure (>10 hours on average could be 10 or 40), different responses to outside exposure including accidents that could impact eyesight. Essentially, any of these factors may serve to explain why for the most part 8/10 cases the treatment effect was 0 whereas there was one positive and one negative case.

c.

```r
#test computing true average treatment effect
#(sum(v$treatment.effect) / nrow(v)) == (sum(v$Yi.1.)/nrow(v)-sum(v$Yi.0.)/nrow(v))
#print ATE
sum(v$treatment.effect) / nrow(v)
```

```
## [1] -0.04
```

d.

```r
#generate indexes
rows<-nrow(v)
even_indexes<-seq(2,rows,2)
odd_indexes<-seq(1,rows,2)
#define new dataset
v.odd <- v[1:3]
#assign odd children to treatment by setting their hypothetical control value to NA
```

```
v.odd[odd_indexes,2]<-NA
#assign even children to control by setting their hypothetical treatment value to NA
v.odd[even_indexes,3]<-NA
#Estimated average based on observed data
v.odd.ate <-  mean(v.odd$Yi.1.,na.rm = TRUE) -  mean(v.odd$Yi.0.,na.rm = TRUE)
#print ate
v.odd.ate
```

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

e. The ATE -.06 of this new experiment differs by .02 when compared to ATE -.04 of the hypothetical example. Intuitively, a "random" allocation of children to treatment effects will produce groups of children that have different average potential outcomes.

f. Two of these are not allowed: the null set and the whole set, as both result in a treatment / control with no members because we would have counted each group twice, once selecting it and once selecting all the rest. Accordingly to make different groups of 10 children (with minimum one in control and one in treatment) there are the below number of ways:

```
(1/2)*(2^10-2)
```

```
## [1] 511
```

g.

```
#generate indexes
rows<-nrow(v)
first_set<-1:5
second_set <-6:10
#define new dataset
v.subsets <- v[1:3]
#assign first set of children to treatment by setting their hypothetical control value to NA
v.subsets[first_set,2]<-NA
#assign even children to control by setting their hypothetical treatment value to NA
v.subsets[second_set,3]<-NA
#Estimated average based on observed data
v.subsets.mean.diff <-  mean(v.subsets$Yi.1.,na.rm = TRUE) -  mean(v.subsets$Yi.0.,na.rm = TRUE)
#print ate
v.subsets.mean.diff
```

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

h. Intuitively the difference is exacerbated by the fact that the populations are no longer comparable because they weren't randomly assigned a treatment and the individual potential outcomes of the children no longer equate to the average potential outcomes of the population of children.

5. FE, exercise 2.5

a. The problem is how to randomize treatment for six subjects asked to donate time to an adult literacy program. There are 3 approaches to decide amongst for randomizing the treatment: 1) coin flip 2) playing cards 3) envelopes. What are the strengths and weaknesses of each approach?

1) Strengths: Each subject has an identical probability of being assigned to the control/treatment group (30 or 60 mins respectively)

2) Weaknesses: When N is small (as it is in this case, 6), the random chance inherent to flipping a coin will become apparent. We might end up with far fewer subjects in treatment or control than anticipated

3) Strengths: Because there is no replacement of cards, you're guaranteeing that your population is receiving treatment and control groups that are identically sized

4) Weaknesses: While this method might be desirable, it would not scale well pratically and is subject to interference by the individual adminstering the cards because they have knowledge of the order

5) Strengths: Similar to method two, you're able to guarantee m of N units in treatment and control.

6) Weaknesses: Like above, this method does not scale well if the number of subjects were very large.

b. The example of simple random assignment in case a. would probably achieve the desired effect based on a larger sample size; however, none of these methods scale well. It would be more desirable to randomly permute the order of all 600 subjects and label the first m subjects as your treatment group.

c. expected value of total assigned minutes for the coin toss is:

```
((1/2)*30+(1/2)*60)+
((1/2)*30+(1/2)*60)+
((1/2)*30+(1/2)*60)+
((1/2)*30+(1/2)*60)+
((1/2)*30+(1/2)*60)+
((1/2)*30+(1/2)*60)
```

```
## [1] 270
```

expected value of the sealed envelope method (if I've understood the method correctly would be the same – you know that you'll assign 3 subjects to the control and 3 subjects to the treatment):

```
30+
30+
30+
60+
60+
60
```

```
## [1] 270
```

6. FE, exercise 2.6

This is a prime example of an observational study. The researcher, while using a random sample, did not randomly assign *the treatment* to the students in the sample. In this case, there were likely underlying factors (geographic area, parental status, socioeconomic status) about those who took a preparatory class that were not random. In order for this effort to qualify as an experiement, the researcher should select a pool of students prior to high school and then randomly assign them to attend either public or prep school which effectively controls for other factors (as mentioned above).

7. FE, exercise 2.8

a. Across treatment effects, Bribe, RTIA, NGO and Control, the proportion of applicants who receieved residence verification was collectively 98.8% - 89 applied and all but 1 received verification. The speed of verification when comparing Bribe and all other treatments (RTIA, NGO and Control) is very different. The median number of days less is 20 days when comparing Bribe with the rest.

b. By treatment, the proportion of applicants who actually received a ration card varies significantly. Bribe = 24/24 or 100% of applicants who applied received a ration card RTIA = 20/23 or 87% of applicants who applied received a ration card NGO = 3/18 or 16% of applicants received a ration card Control = 5/21 or 23.8% of applicants received a ration card

c. The results seem to suggest that the RTIA is a fairly effective measure of ensuring that ultimately those who apply or request a government service ultimately receive it.

8. FE, exercise 2.9

a. The primary issue with the researcher's assumption that the lottery chooses winners at random is that those who decide to play the lottery, unfortunately, are not. Those who decide to play the lottery, are, in general, not a random subset of the population. Those who decide to play are likely on average in a more deleterious situation than those who don't and thus aren't necessarily comparable to others,who selected at random, did not win. For this resason, the researcher could, in an ideal case, improve the ability to draw causal inferences about estate tax by first drawing a sample of U.S. adults and then randomly assigning those people "lottery winnings" which helps prevent the selection bias inherent in this example.

b. While this new design does help us with excludability, it does not allow us to completely safely assume that the poential outcomes of those who report winning are identical. We still have the issue of selection bias which impacts our ability of excludability as well as an issue of data being gathered longitudinally which may jeopardize symmetry.

9. FE, exercise 2.12(a)

a. Setting this up:

- $d_i = 0$ when prisoners read less than 3 hours each day
- $d_i = 1$ when prisoners read more than 3 hours each day
- $Y_i(0)$ = potential number of violent encounters when reading less than 3 hours a day
- $Y_i(1)$ = potential number of violent encounters when reading more than 3 hours a day

The reason why one might be hesitant to assume that $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$ and $E[Y_i(1)|D_i = 0] = E[Y_i(1)|D_i = 1]$ is that, in the first case, where we want to say the expectation of violent encounters of those who read less than 3 hours is the same as the expectation of violent encounters *had* they read more than 3 hours a day, is that since *nature* has selected them, there isn't any guarantee that those who decided to read less are equivalent to those who might decided to read more. Accordingly, E[Y_i(0)|D_i=0]$ =/= $E[Y_i(0)|D_i = 1]$ and we can't easily generalize any potential outcomes unless there is random assignment of the treatment because the expectation truly should be different on average.