Field Experiments Problem Set 1

Ted Dunmire

Wednesday May 27, 2015

1. On the notation of potential outcomes:

   a. Explain the notation

   $$Y_i(1)$$

   This notation expresses the outcome if the ith subject is exposed to the treatment. Whatever that may be. That is, this is the measured outcome for subject i if i is given the treatment. In the village example, this would be the equivalent of asking "what is the outcome of the ith village if the village has been exposed to the treatment (having a female village head)."

   b. Explain the notation

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

   This notation expresses the expected value of Yi(1) when a the ith subject is selected at random from the sample not given the treatment. The di expression is a conditional (often read as "given that") and thus designates the sample group to draw from at random. In the village example, this would be equivalent to saying "the expected outcome of the ith village if the ith village has been exposed to the treatment given that the village is selected at random from those villages not treated." This is obviously hypothetical.

   c. Explain the difference between the notation

   $$E[Y_i(1)]$$

   and notation

   $$E[Y_i(1)|d_i = 1]$$

   The difference is about the conditions for the group. E[Yi(1)] is a random variable meaning the expected value of Yi(1) when one subject is sampled at random. This is like having a sample of all outcomes Yi(1) (all subjects are given treatment) and then selecting one at random, where the expected value is then the average. The difference in the second expression is the di=1 defines a subgroup. In this case, the subgroup is defined as subjects that were given the treatment. The difference here is that an added condition is attached to the expected value of the second expression.
   Another way to think about it would be the first expression gives the expected value if all subjects were treated (this is hypothetical, so the equivalent of if all villages had female village heads), while the second expression is the expected value if the subjects ACTUALLY were given the treatment (so only the expected value of the subjects who actually received the treatment in the real world).

2. FE, exercise 2.2: Use values in table 2.1 to illustrate

   $$E[Y_i(0)] - E[Y_i(1)] = E[Y_i(0) - Y_i(1)]$$

   First I will create the table

```
Yi_0 <- c(10, 15, 20, 20, 10, 15, 15)
Yi_1 <- c(15, 15, 30, 15, 20, 15, 30)
table_2.1 <- data.frame(Yi_0, Yi_1)
table_2.1
```

```
##   Yi_0 Yi_1
## 1   10   15
## 2   15   15
## 3   20   30
## 4   20   15
## 5   10   20
## 6   15   15
## 7   15   30
```

Ok, now for the first part. I must show the first part of the expression. The expected value is the average, so I will take the mean of both columns of observations, then I will subtract the treatment from the control group to obtain the left side of the expression.

```
E_Yi_0 <- mean(table_2.1$Yi_0)
E_Yi_1 <- mean(table_2.1$Yi_1)
result <- E_Yi_0 - E_Yi_1
result
```

```
## [1] -5
```

Ok, now I have **-5** as the result. Now hopefully the right side of the expression will be the same. So for this I have the expected value (or average) of the difference between the control and treatment for each observation. I will create a new column in my table to store this difference and take the mean.

```
table_2.1$diff <- table_2.1$Yi_0 - table_2.1$Yi_1
result2 = mean(table_2.1$diff)
result2
```

```
## [1] -5
```

Excellent. I have also obtained **-5** here for this result. Thus the two expressions yield the same result.

3. FE exercise 2.3: Use values depicted in Table 2.1 to complete the table below

   a. Number of observations in each of the nine cells

I'll use a frequency table

```
Yi_0 <- c(10, 15, 20, 20, 10, 15, 15)
Yi_1 <- c(15, 15, 30, 15, 20, 15, 30)
table_2.1 <- data.frame(Yi_0, Yi_1)
table_2.1
```

```
##   Yi_0 Yi_1
## 1   10   15
## 2   15   15
## 3   20   30
## 4   20   15
## 5   10   20
## 6   15   15
## 7   15   30
```

```
my_table <- table(table_2.1$Yi_0, table_2.1$Yi_1)
my_table
```

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

b. Indicate percentage of all subjects that fall into each of the nine cells. Aka the joint frequency distribution

```
my_prop_table <- prop.table(my_table)
my_prop_table
```

```
##
##          15     20     30
##   10 0.1429 0.1429 0.0000
##   15 0.2857 0.0000 0.1429
##   20 0.1429 0.0000 0.1429
```

c. Proportion of subjects falling into each category of Yi(1). Aka the marginal distribution of Yi(1)

This is going to be the sum of the individual columns. So this will be 0.1428571 + 0.2857143 + 0.1428571 for the first and so on. I will use colSums to make it easy.

```
marg_Yi_1 <- colSums(my_prop_table)
marg_Yi_1
```

```
##     15     20     30
## 0.5714 0.1429 0.2857
```

d. Indicate the proportion of subjects falling into each category of Yi0 (the marginal distribution of Yi(0))

Similar to above this is the sum of the rows. Here I'll just use rowsSum to do the same thing.

```
marg_Yi_0 <- rowSums(my_prop_table)
marg_Yi_0
```

```
##     10     15     20
## 0.2857 0.4286 0.2857
```

Now just to make things pretty I'll bind the rows to the joint frequency distribution

```
my_prop_table <- rbind(my_prop_table, marg_Yi_1)
my_prop_table <- cbind(my_prop_table, marg_Yi_0)
```

```
## Warning: number of rows of result is not a multiple of vector length (arg
## 2)
```

```
my_prop_table[4,4] <- 1
my_prop_table
```

```
##                15     20     30 marg_Yi_0
## 10         0.1429 0.1429 0.0000    0.2857
## 15         0.2857 0.0000 0.1429    0.4286
## 20         0.1429 0.0000 0.1429    0.2857
## marg_Yi_1 0.5714 0.1429 0.2857    1.0000
```

e. Use the table to calculate the

$$E[Y_i(0)|Y_i(1) > 15]$$

I will use the conditional expectation here. The denominator Pr[Y_i(1) > 15] = 3/7. The numerator is the sum of the joint probability of Y_i(0) and Y_i(1) > 15, which can be expressed as Pr[Y_i(0) | Y_i(1) > 15], which is (10*(1/7) + 15*(1/7) + 20 * (1/7)). So I will run the following:

```
(10*(1/7) + 15*(1/7) + 20 * (1/7))/(3/7)
```

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

To yield an expected value of **15**

f. E[Y_i(1) | Y_i(0) > 15]

I will use the same procedure as above with a conditional expectation. The denominator is 2/7 and the numerator is (25*(1/7) + 30 (1/7)). Then running the code:

```
(15*(1/7) + 30 * (1/7))/(2/7)
```

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

Which yields **22.5** as the expected value.

4. Based on provided table of visual acuity. Let's make the table first

```
Yi_0 <- c(1.1, 0.1, 0.5, 0.9, 1.6, 2.0, 1.2, 0.7, 1.0, 1.1)
Yi_1 <- c(1.1, 0.6, 0.5, 0.9, 0.7, 2.0, 1.2, 0.7, 1.0, 1.1)
visual <- data.frame(Yi_0, Yi_1)
```

a. Compute individual treatment effect

I will subtract the observed control from the observed treatment to find the treatment effect for each subject. I will store this in a new column.

```
visual$effect <- visual$Yi_1 - visual$Yi_0
visual
```

```
##    Yi_0 Yi_1 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
## 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
```

Now we can clearly see the effect of the treatment in each of the subjects.

b. Describe the distribution of treatment effects.

The distribution contains a lot of zeros. This would indicate that the treatment had no effect on these subjects. According to doctors the main influences on visual acuity are the shape and functioning of the retina and the brain. Playing outside might cause some changes (such as new neural pathways being formed or an injury) but this distribution would suggest that other factors are probably more influential than playing outside in changing visual acuity.

c. What is the true average treatment effect (ATE)

Here I will take the mean of the effect column I generated earlier and this will be the average treatment effect for this group.

```
ATE <- mean(visual$effect)
ATE
```

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

I can see the true ATE is **-0.04**

d. odd numbered children assigned to treatment and even numbered to control. What is estimate of ATE reached under this assignment?

I will create a new table and assign the children to the appropriate columns. Where they would have information from the true table that does not exist I will put an NA. Then I will take the mean of each of these columns to get an average and take the difference to find the treatment effect.

```
Yi_1_2 <- c(1.1, NA, 0.5, NA, 0.7, NA, 1.2, NA, 1.0, NA)
Yi_0_2 <- c(NA, 0.1, NA, 0.9, NA, 2.0, NA, 0.7, NA, 1.1)
est_ATE <- mean(Yi_1_2, na.rm = T) - mean(Yi_0_2, na.rm = T)
est_ATE
```

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

Now I will get an estimated ATE of **-0.060**

e. How different is the estimate from the truth and why?

The true value is -0.04 and the estimated value is -0.060, meaning a difference of -0.020 (the estimated value is 50% "higher" than the true value). Intuitively this is because there is variation from sample to sample. I did not have all the observations here, so there will be some variation just due to chance. Also in this case I did not observe the true treatment effect for children 2 and 5, who actually showed an effect from the treatment.

f. How many ways can we split the children into treatment and control groups?

This probably is asking from 10 subjects, how many combinations can be made so that there are always two groups and that there is always at least 1 person in each group. For this I will sum the possible combinations (the dim function is getting the count of each combination) for 10 choose 1 through 10 choose 9.

```
count = 0
for (i in 1:9) {
  count = count + dim(combn(10,i))[2]
}
count
```

```
## [1] 1022
```

This yields a total of **1022** different ways to configure the groups.

g. Observational study.

Again make separate columns to represent the two groups, take the means, and then find the difference between the "treatment" and "control"

```
Yi_0_3 <- c(2.0, 1.2, 0.7, 1.0, 1.1)
Yi_1_3 <- c(1.1, 0.6, 0.5, 0.9, 0.7)
est_ATE_2 <- mean(Yi_1_3) - mean(Yi_0_3)
est_ATE_2
```

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

h. Compare to true ATE and what causes the difference

For this we get a result of **-0.44** which is actually quite different than the true ATE of -0.04 (in fact it's more than 10 times larger). This is an observational study so perhaps the children who have better eyesight choose to play inside for whatever reason (maybe their better eyesight allows them to see in the dark and play video games more effectively, as an example). As this is an observational study I really have no way of knowing why these two groups of children are systematically different.

5. FE exercise 2.5

   a. The strength of the coin flip is that it is totally random (assuming a fair coin). A problem with this approach is that there is the researcher has no control over how many participants are assigned to the different groups. This could lead to a situation where 1 person is in the control vs 5 in the case (or the reverse) which may not provide good data. There is also a small but real chance that the researcher winds up with no participants in one of the groups. The shuffling approach guarantees distribution of the assignment, which is an advantage. However, it also relies on the researcher performing the shuffling to produce randomness. It would be possible that the researcher would not shuffle the cards to produce a truly "random" assignment (for example, by just cutting the deck in half). The weakness here is the researcher has the ability to introduce bias in the assignment. For the shuffle in the envelopes, much like the original shuffling example, the researcher controls the distribution and guarantees good case and control groups. Here there is the added advantage that it is more difficult for the researcher to know with certainty which envelopes contain which assignments. So one would probably be more confident in the random assignment of participants. However, there is still a chance the researcher could interfere with the randomization by not shuffling properly.

b. Now, if there are more subjects this changes things. For the coin example, there is now a much smaller chance that there will be a group (case or control) that is heavily skewed. It is certainly still possible to have unequal groups but now it could be something like 150 versus 450 which is probably more reasonable for comparison purposes. In the shuffling case, there is now a lot more work added to the part of the researcher, who now has to shuffle a lot more cards. There is also still the issue of ensuring the researcher shuffles to produce a random output. It would be possible to imagine a situation where a researcher does not fully randomize the deck due to the large number of cards. For the envelope case, there is even more work for the researcher to do. This will also require the researcher to properly shuffle the envelopes, although again the researcher will probably have less ability to impact assignment because he/she will not know with certainty which assignment is in which envelope.

c. If the coin toss method is used, assuming a fair coin each subject has an expected value of $0.5*30* + *0.5*60$ minutes or 45 minutes for each subject. There are 6 subjects so the average (expected value) would be **45 minutes**. If the envelope method is used, exactly half of the subjects would have 30 minutes and the other half would have 60 minutes. The expected value would then be $0.5*30* + *0.5*60$ (representing assignment) and the expected value for the envelope method would be **45 minutes**. Therefore, in both cases the expected value is 45 minutes.

6. FE, exercise 2.6

   This is an observational study. At first glance this may appear to be an exeperiment, after all there is a clear treatment (the attendance of a preparatory class). However, the researcher knows nothing about what might differentiate the students who took the class from those who didn't, they may be systematically different. For example, it would be easy to imagine students who took the preparatory class were more highly motivated than those who didn't, which meant they studied longer. Perhaps they came from different socieoeconomic backgrounds and the students who took the class come from families with higher incomes and advantages than those who did not take the class.

   Without knowing beforehand how the groups might differ and ensuring that the **only** feature that varies between the groups is the treatment of taking a preparatory class, this has to be considered an observational study, not an experiment.

7. FE, exercise 2.8

   a. The treatments seem to have little impact on the confederates getting their residence verified. For the bribe, RTIA, and NGO groups, 100% of their residencies were verified. In the case of the control 20/21 = 95.24% of the confederates had their residence verified. So it would seem that the treatment "guarantees" getting a residence verified but it does not mark a large difference. However, the treatment did have an effect on the median number of days to get the residence verified. In the case of the control, NGO, and RTIA groups it took 37 days to get the residence verified (this is the median). The bribe group only took 17 days, a (37-17)/37 = 54.10% drop. This certainly seems to be different from the others. In conclusion then the treatments do not seem to have a big impact on the number of confederates who got their residence verified but the bribe treatment had a large impact on the median number of days it took to get the residence verified.

   b. When examining who actually received their card, I used the following code to generate the proportion of the original group that received their ration card within 1 year

```
confed <- c(24, 23, 18, 21)
rationCard <- c(24, 20, 3, 5)
rationCard/confed
```

```
## [1] 1.0000 0.8696 0.1667 0.2381
```

Going from left to right these correspond to the Bribe, RTIA, NGO, and Control groups respectively. It should be apparent there is a big difference between the Bribe and RTIA groups, who saw 100% and 86.96% of their confederates get a card, than with the NGO and Control groups who saw 16.67% and 23.80%. It would seem Bribe and RTIA treatments have a big impact and increase the number of confederates getting their ration card compared to the Control and the NGO.

c. The RTIA results suggest essentially no improvement in the number of confederates who get their residence verified, nor does it improve the median number of days to get the residence verified. However, it does seem RTIA increases the number of confederates who received a ration card within a year, so this would seem to suggest the RTIA is a good (although not perfect or necessarily the best) way to get a ration card.

8. FE exercise 2.9

   a. This is not a good assumption. The researcher interviews a random sample of adults. The researcher makes no distinction about adults who actually PLAY the lottery. It is entirely possible that there will be subjects who report earning no lottery money because they do not play the lottery. This would mean the groups are systematically different and would not make for a good comparison. Another possible source of systematic difference is class. Wealthy people tend not to play the lottery and could have a different opinion on the estate tax than poorer people who do play. Here class or socioeconomic status could act as a confounding factor that is really the cause rather than the lottery.

   b. This is certainly better because now people actually played the lottery. However, I think one should still be critical or suspicious of this. One reason is just because someone reported they played the lottery once a year or more that probably hides a lot of variation between subjects. There are going to be people who are playing weekly versus those who play very infrequently. Once again socioeconomic status could be a confounding factor that both contributes to this lottery behavior and informs opinions on the estate tax. Also the frequency at which one plays the lottery could impact the randomness assumption. This would be because people who play the lottery more often are more likely to win, leading to a selection problem where the groups are probably different.

9. FE exercise 2.12(a)

   a. What this is addressing is the issue of random assignment. If there is random assignment, then the term

   $$E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1] = E[Y_i(0)]$$

   The same goes for Yi(1). Intuitively what this is saying is that each subject has an equal chance of being selected for treatment. If this is the case, selection for treatment in a hypothetical experiment (the Di term) is independent of the potential outcomes for the Yi term. If there is not random assignment (in this experiment "nature selects") there is what is called selection bias. In this example, this would translate to subjects who are maybe less violent choosing to read more anyway. This would cause the researcher to exaggerate the effect of the treatment.