# Problem Set 5

*Field Experiments*

## 1. Online advertising natural experiment.

These are simulated data (closely, although not entirely) based on a real example, adopted from Randall Lewis' dissertation at MIT.

### Problem Setup

Imagine Yahoo! sells homepage ads to advertisers that are quasi-randomly assigned by whether the user loads the Yahoo! homepage (www.yahoo.com) on an even or odd second of the day. More specifically, the setup is as follows. On any given week, Monday through Sunday, two ad campaigns are running on Yahoo!'s homepage. If a user goes to www.yahoo.com during an even second that week (e.g., Monday at 12:30:58pm), the ads for the advertiser are shown. But if the user goes to www.yahoo.com during an odd second during that week (e.g., Monday at 12:30:59), the ads for other products are shown. (If a user logs onto Yahoo! once on an even second and once on an odd second, they are shown the first of the campaigns the first time and the second of the campaigns the second time. Assignment is not persistent within users.)

This natural experiment allows us to use the users who log onto Yahoo! during odd seconds/the ad impressions from odd seconds as a randomized control group for users who log onto Yahoo! during even seconds/the ad impressions from even seconds. (We will assume throughout the problem there is no effect of viewing advertiser 2's ads, from odd seconds, on purchases for advertiser 1, the product advertised on even seconds.)

Imagine you are an advertiser who has purchased advertising from Yahoo! that is subject to this randomization on two occasions. Here is a link to (fake) data on 500,000 randomly selected users who visited Yahoo!'s homepage during each of your two advertising campaigns, one you conducted for product A in March and one you conducted for product B in August (~250,000 users for each of the two experiments). Each row in the dataset corresponds to a user exposed to one of these campaigns.

```
library(data.table)
library(stargazer)
library(dplyr)
library(sandwich)
library(lmtest)
library(ggplot2)

d1 <- data.table(fread('./data/ps5_no1.csv'))
d1
```

```
##         product_b total_ad_exposures_week1 treatment_ad_exposures_week1
##      1:         1                         4                            3
##      2:         1                         1                            1
##      3:         1                         3                            1
##      4:         0                         5                            0
##      5:         0                         1                            1
##     ---
## 499996:         1                         2                            1
## 499997:         1                         4                            0
## 499998:         0                         5                            2
## 499999:         1                         5                            1
## 500000:         0                         3                            1
```

```
##          week0 week1 week2 week3 week4 week5 week6 week7 week8 week9 week10
##      1:   5.5   6.2   0.0   0.0   0.0   0.0     0   0.0   9.7   4.1    0.0
##      2:   6.2   0.0   8.6   2.4   0.0   7.4     0   0.0   0.0   5.7    0.0
##      3:   0.0   5.3   0.0   8.1   7.8   3.3     0   0.0   9.4   0.0    0.0
##      4:   0.0   4.1   0.0   8.8   5.8   5.9     0   0.0   0.0   9.6    0.0
##      5:   7.6   3.6   4.6   5.5   7.2   7.1     0   0.0   0.0   0.0    0.0
##     ---
## 499996:   0.0   2.8   8.1   2.8   4.1   8.1     0   0.0   0.0   0.0    2.3
## 499997:   0.0   0.0   4.0   4.6   0.0   0.0     0   5.0   0.0   0.0    0.0
## 499998:   4.3   0.0   0.0   0.0   4.2   6.5     0   4.8   9.4   6.4    9.5
## 499999:   0.0   0.0   0.0   5.4   6.1   0.0    12   7.7   0.0   0.0    0.0
## 500000:   4.7   0.0   0.0   2.2   0.0   0.0     0   4.3   0.0   4.9    0.0
```

The variables in the dataset are described below:

- **product_b**: an indicator for whether the data is from your campaign for product A (in which case it is set to 0), sold beginning on March 1, or for product B, sold beginning on August 1 (in which case it is set to 1). That is, there are two experiments in this dataset, and this variable tells you which experiment the data belong to.
- **treatment_ad_exposures_week1**: number of ad exposures for the product being advertised during the campaign. (One can also think of this variable as "number of times each user visited Yahoo! homepage on an even second during the week of the campaign.")
- **total_ad_exposures_week1**: number of ad exposures on the Yahoo! homepage each user had during the ad campaign, which is the sum of exposures to the "treatment ads" for the product being advertised (delivered on even seconds) and exposures to the "control ads" for unrelated products (delivered on odd seconds). (One can also think of this variable as "total number of times each user visited the Yahoo! homepage during the week of the campaign.")
- **week0**: For the treatment product, the revenues from each user in the week prior to the launch of the advertising campaign.
- **week1**: For the treatment product, the revenues from each user in the week during the advertising campaign. The ad campaign ends on the last day of week 1.
- **week2-week10**: Revenue from each user for the treatment product sold in the weeks subsequent to the campaign. The ad campaign was not active during this time.

Simplifying assumptions you should make when answering this problem:

- The effect of treatment ad exposures on purchases is linear. That is, the first exposure has the same effect as the second exposure.
- There is no effect of being exposed to the odd-second ads on purchases for the product being advertised on the even second.
- Every Yahoo! user visits the Yahoo! home page at most six times a week.
- You can assume that treatment ad exposures do not cause changes in future ad exposures. That is, assume that getting a treatment ad at 9:00am doesn't cause you to be more (or less) likely to visit the Yahoo home pages on an even second that afternoon, or on subsequent days.

## Questions to Answer

a. Run a crosstab of total_ad_exposures_week1 and treatment_ad_exposures_week1 to sanity check that the distribution of impressions looks as it should. Does it seem reasonable? Why does it look like this? (No computation required here, just a brief verbal response.)

```
d1[, table(total_ad_exposures_week1, treatment_ad_exposures_week1)]
```

```
##                         treatment_ad_exposures_week1
## total_ad_exposures_week1     0     1     2     3     4     5     6
##                        0 61182     0     0     0     0     0     0
```

```
##                         1 36754 37215     0     0     0     0     0
##                         2 21143 42036 20965     0     0     0     0
##                         3 10683 32073 32314 10726     0     0     0
##                         4  5044 20003 30432 20223  5115     0     0
##                         5  2045 10563 20970 20793 10293  2131     0
##                         6   729  4437 10977 14771 11147  4486   750
```

**The distributions look reasonable. Each row in the lower triangle looks like it is roughly binomially distributed.**

b. Your colleague proposes the code printed below to analyze this experiment: `lm(week1 ~ treatment_ad_exposures_week1, data)` You are suspicious. Run a placebo test with the prior week's purchases as the outcome and report the results. Did the placebo test "succeed" or "fail"? Why do you say so?

```
d1[, summary(lm(week0 ~ treatment_ad_exposures_week1))]
```

```
##
## Call:
## lm(formula = week0 ~ treatment_ad_exposures_week1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.248 -2.196 -1.670  2.430  8.330
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.669685   0.006027   277.0   <2e-16 ***
## treatment_ad_exposures_week1 0.263099   0.003155    83.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.796 on 499998 degrees of freedom
## Multiple R-squared:  0.01372,    Adjusted R-squared:  0.01372
## F-statistic:  6955 on 1 and 499998 DF,  p-value: < 2.2e-16
```

**The placebo test failed. The treatment coefficient is very statistically significant, even for week 0 revenue, before it is applied. Treatment assignment is somehow being preferentially applied more to people who already buy the advertised product.**

c. The placebo test suggests that there is something wrong with our experiment or our data analysis. We suggest looking for a problem with the data analysis. Do you see something that might be spoiling the randomness of the treatment variable? How can you improve your analysis to get rid of this problem? Why does the placebo test turn out the way it does? What one thing needs to be done to analyze the data correctly? Please provide a brief explanation of why, not just what needs to be done. (*Note: This question, and verifying that you answered it correctly in part d below, may require some thinking. If we find many people can't figure it out, we will post another hint in a few days.*)

d. Implement the procedure you propose from part (c), run the placebo test for the Week 0 data again, and report the results. (This placebo test should pass; if it does not, re-evaluate your strategy before wasting time proceeding.)

e. Now estimate the causal effect of each ad exposure on purchases during the week of the campaign itself using the same technique that passed the placebo test in part (d).

f. The colleague who proposed the specification in part (b) challenges your results – they make the campaign look less successful. Write a paragraph that a layperson would understand about why your estimation strategy is superior and his/hers is biased.

g. Estimate the causal effect of each treatment ad exposure on purchases during and after the campaign, up until week 10 (so, total purchases during weeks 1 through 10).

h. Estimate the causal effect of each treatment ad exposure on purchases only after the campaign. That is, look at total purchases only during week 2 through week 10, inclusive.

i. Tell a story that could plausibly explain the result from part (h).

j. Test the hypothesis that the ads for product B are more effective, in terms of producing additional revenue in week 1 only, than are the ads for product A. (*Hint: The easiest way to do this is to throw all of the observations into one big regression and specify that regression in such a way that it tests this hypothesis.*) (*Hint 2: There are a couple defensible ways to answer this question that lead to different answers. Don't stress if you think you have an approach you can defend.*)

k. You notice that the ads for product A included celebrity endorsements. How confident would you be in concluding that celebrity endorsements increase the effectiveness of advertising at stimulating immediate purchases?

# 2. Vietnam Draft Lottery

A famous paper by Angrist exploits the randomized lottery for the Vietnam draft to estimate the effect of education on wages. (*Don't worry about reading this article, it is just provided to satisfy your curiosity; you can answer the question below without referring to it. In fact, it may be easier for you not to, since he has some complications to deal with that the simple data we're giving you do not.*)

## Problem Setup

Angrist's idea is this: During the Vietnam era, draft numbers were determined randomly by birth date – the army would literally randomly draw birthdays out of a hat, and those whose birthdays came up sooner were higher up on the list to be drafted first. For example, all young American men born on May 2 of a given year might have draft number 1 and be the first to be called up for service, followed by November 13 who would get draft number 2 and be second, etc. The higher-ranked (closer to 1) your draft number, the likelier it was you would be drafted.

We have generated a fake version of this data for your use in this project. You can find real information (here)[https://www.sss.gov/About/History-And-Records/lotter1]. While we're defining having a high draft number as falling at 80, in reality in 1970 any number lower than 195 would have been a "high" draft number, in 1971 anything lower than 125 would have been "high".

High draft rank induced many Americans to go to college, because being a college student was an excuse to avoid the draft – so those with higher-ranked draft numbers attempted to enroll in college for fear of being drafted, whereas those with lower-ranked draft numbers felt less pressure to enroll in college just to avoid the draft (some still attended college regardless, of course). Draft numbers therefore cause a natural experiment in education, as we now have two randomly assigned groups, with one group having higher mean levels of education, those with higher draft numbers, than another, those with lower draft numbers. (In the language of econometricians, we say the draft number is "an instrument for education," or that draft number is an "instrumental variable.")

Some simplifying assumptions:

- Suppose that these data are a true random sample of IRS records and that these records measure every living American's income without error.
- Assume that the true effect of education on income is linear in the number of years of education obtained.
- Assume all the data points are from Americans born in a single year and we do not need to worry about cohort effects of any kind.

### Questions to Answer

a. Suppose that you had not run an experiment. Estimate the "effect" of each year of education on income as an observational researcher might, by just running a regression of years of education on income (in R-ish, `income ~ years_education`). What does this naive regression suggest?

```
m = d2[, lm(income ~ years_education)]
(co = coeftest(m))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     -23354.64    1252.74 -18.643 < 2.2e-16 ***
## years_education   5750.48      83.34  69.000 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**The regression shows an effect of 5750.48 (83.34). The effect is statistically significant.**

b. Continue to suppose that we did not run the experiment, but that we saw the result that you noted in part (a). Tell a concrete story about why you don't believe that observational result tells you anything causal.

c. Now, let's get to using the natural experiment. We will define "having a high-ranked draft number" as having a draft number of 80 or below (1-80; numbers 81-365, for the remaining 285 days of the year, can be considered "low-ranked"). Create a variable in your dataset indicating whether each person has a high-ranked draft number or not. Using regression, estimate the effect of having a high-ranked draft number, the dummy variable you've just created, on years of education obtained. Report the estimate and a correctly computed standard error. (*Hint: Pay special attention to calculating the correct standard errors here. They should match how the draft is conducted.)

d. Using linear regression, estimate the effect of having a high-ranked draft number on income. Report the estimate and the correct standard error.

e. Divide the estimate from part (d) by the estimate in part (c) to estimate the effect of education on income. This is an instrumental-variables estimate, in which we are looking at the "clean" variation in both education and income that is due to the draft status, and computing the slope of the income-education line as "clean change in Y" divided by "clean change in X". What do the results suggest?

f. Natural experiments rely crucially on the "exclusion restriction" assumption that the instrument (here, having a high draft rank) cannot affect the outcome (here, income) in any other way except through its effect on the "endogenous variable" (here, education). Give one reason this assumption may be violated – that is, why having a high draft rank could affect individuals' income other than because it nudges them to attend school for longer.

g. Conduct a test for the presence of differential attrition by treatment condition. That is, conduct a formal test of the hypothesis that the "high-ranked draft number" treatment has no effect on whether we observe a person's income. (Note, that an earning of $0 *actually* means they didn't earn any money.)

h. Tell a concrete story about what could be leading to the result in part (g).

i. Tell a concrete story about how this differential attrition might bias our estimates.

## 3. Dinner Plates

Suppose that researchers are concerned with the health consequences of what people eat and how much they weigh. Consider an experiment designed to measure the effect of a proposal to healp people diet. Subjects are invited to a dinner and are randomly given regular-sized or slightly larger than regular sized plates. Hidden cameras record how much people eat, and the researchers find that those given larger plates eat substantially more food than those assigned small plates.

A statistical test shows that the apparent treatment effect is far greater than one would expect by chance. The authors conclude that a minor adjustment, reducing plate size, will help people lose weight.

- How convincing is the evidence regarding the effect of plate size of what people eat and how much they weight?
- What design and measurment improvements do you suggest?

## 4. Think about Treatment Effects

Throughout this course we have focused on the average treatment effect. Think back to *why* we are concerned about the average treatment effect. What is the relationship between an ATE, and some individuals' potential outcomes? Make the strongest case you can for why this is *good* measure.