Problem Set 2

Experiments and Causality

1. What happens when pilgrims attend the Hajj pilgrimage to Mecca?

On the one hand, participating in a common task with a diverse group of pilgrims might lead to increased mutual regard through processes identified in *Contact Theories*. On the other hand, media narritives have raised the spectre that this might be accompanied by "antipathy toward non-Muslims". Clingingsmith, Khwaja and Kremer (2009) investigates the question.

Using the data here, test the sharp null hypothesis that winning the visa lottery for the pilgrimage to Mecca had no effect on the views of Pakistani Muslims toward people from other countries. Assume that the Pakistani authorities assigned visas using complete random assignment. Use, as your primary outcome the `views` variable, and as your treatment feature `success`. If you're ambitious, write your fucntion generally so that you can also evaluate feeligns toward specific nationalities.

```
d <- read.csv("./data/Clingingsmith.2009.csv", stringsAsFactors = FALSE)
```

a. Using either `dplyr` or `data.table`, group the data by `success` and report whether views toward others are generally more positive among lottery winners or lottery non-winners.

b. But is this a meaningful difference, or could it just be randomization noise? Conduct 10,000 simulated random assignments under the sharp null hypothesis to find out. (Don't just copy the code from the async, think about how to write this yourself.)

c. How many of the simulated random assignments generate an estimated ATE that is at least as large as the actual estimate of the ATE?

As an aside, there is a really nice way to produce literate reports in RStudio. Just like we're writing our code in "code chunks" where things are evaluated, you can also include objects into the printed space by creating an inline chunk.

For example, suppose we have an object `x` which has in it the values `1, 2, and 5.`. We can create this objects in the normal way:

```
x <- c(1,2,5)
```

But we can also write what the average of this vector is using an inline call. The mean of this vector is 2.6666667.

d. What is the implied *one-tailed* p-value?

e. How many of the simulated random assignments generate an estimated ATE that is at least as large *in absolute value* as the actual estimate of the ATE?

f. What is the implied two-tailed p-value?

2. Randomization Inference Practice

McElhoe and Conner (1986) suggest using a *new* instrument called a "Visiplume" measure pollution. The EPA has a standard method for measuring pollution. Because they're good scientists, McElhoe and Conner want to validate that their instrument is measuring the same levels of pollution as the EPA instrument.

To do so, they take six readings – one with each instrument – at a single site. The recorded response is the ratio of the Visiplume reading to the EPA standard reading, and the values that are recorded are: 0.950, 0.978, 0.762, 0.733, 0.823, and 1.011.

Suppose that we want to test the question, "Do the Visiplume readings and the EPA standard readings produce similar enough estimates?"

a. How would you structure the sharp-null hypothesis – that Visiplume and the EPA reaings are the same – in this case?

b. Suppose that our summary of the data is the sum of the ratios. That is, in the test that we conducted, we obsered $0.95 + ... + 1.011 = 5.257$. Using randomization inference, test the sharp-null hypothesis that you formed in the first part of the question. Produce a histogram of the test statistic under the sharp null that compares against the 5.257 value from the test, and also produce a two-sided p-value.

3. Term Limits Aren't Good.

Naturally occurring experiments sometimes involve what is, in effect, block random assignment. For example, Rocio Titiunik , in this paper studies the effect of lotteries that determine whether state senators in TX and AR serve two-year or four-year terms in the aftermath of decennial redistricting. These lotteries are conducted within each state, and so there are effectively two distinct experiments on the effects of term length.

The "thoery" in the news (such as it is), is that legislators who serve 4 year terms have more time to slack off and not produce legislation. If this were true, then it would stand to reason that making terms shorter would increase legislative production.

One way to measure legislative production is to count the number of bills (legislative proposals) that each senator introduces during a legislative session. The table below lists the number of bills introduced by senators in both states during 2003.

```
library(foreign)

d <- read.dta("./data/Titiunik.2010.dta")
head(d)
```

```
##   term2year bills_introduced texas0_arkansas1
## 1         0               18                0
## 2         0               29                0
## 3         0               41                0
## 4         0               53                0
## 5         0               60                0
## 6         0               67                0
```

a. Using either `dplyr` or `data.table`, group the data by state and report the mean number of bills introduced in each state. Does Texas or Arkansas seem to be more productive? Then, group by two- or four-year terms (ignoring states). Do two- or four-year terms seem to be more productive? **Which of these effects is causal, and which is not?** Finally, using `dplyr` or `data.table` to group by state and term-length. How, if at all, does this change what you learn?

b. For each state, estimate the standard error of the estimated ATE.

c. Use equation (3.10) to estimate the overall ATE for both states combined.

d. Explain why, in this study, simply pooling the data for the two states and comparing the average number of bills introduced by two-year senators to the average number of bills introduced by four-year senators leads to biased estimate of the overall ATE.

e. Insert the estimated standard errors into equation (3.12) to estimate the stand error for the overall ATE.

f. Use randomization inference to test the sharp null hypothesis that the treatment effect is zero for senators in both states.

g. **IN Addition:** Plot histograms for both the treatment and control groups in each state (for 4 histograms in total).

3. Cluster Randomization

Use the data in *Field Experiments* Table 3.3 to simulate cluster randomized assignment. (*Notes: (a) Assume 3 clusters in treatment and 4 in control; and (b) When Gerber and Green say* ***simulate'', they do not mean****run simulations with R code'', but rather, in a casual sense "take a look at what happens if you do this this way." There is no randomization inference necessary to complete this problem.*)

```
## load data
d <- read.csv("./data/ggChapter3.csv", stringsAsFactors = FALSE)
```

a. Suppose the clusters are formed by grouping observations {1,2}, {3,4}, {5,6}, ... , {13,14}. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to treatment.

b. Suppose that clusters are instead formed by grouping observations {1,14}, {2,13}, {3,12}, ... , {7,8}. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to treatment.

c. Why do the two methods of forming clusters lead to different standard errors? What are the implications for the design of cluster randomized experiments?

4. Sell Phones?

You are an employee of a newspaper and are planning an experiment to demonstrate to Apple that online advertising on your website causes people to buy iPhones. Each site visitor shown the ad campaign is exposed to $0.10 worth of advertising for iPhones. (Assume all users could see ads.) There are 1,000,000 users available to be shown ads on your newspaper's website during the one week campaign.

Apple indicates that they make a profit of $100 every time an iPhone sells and that 0.5% of visitors to your newspaper's website buy an iPhone in a given week in general, in the absence of any advertising.

a. By how much does the ad campaign need to increase the probability of purchase in order to be "worth it" and a positive ROI (supposing there are no long-run effects and all the effects are measured within that week)?

b. Assume the measured effect is 0.2 percentage points. If users are split 50:50 between the treatment group (exposed to iPhone ads) and control group (exposed to unrelated advertising or nothing; something you can assume has no effect), what will be the confidence interval of your estimate on whether people purchase the phone?

- **Note:** The standard error for a two-sample proportion test is $\sqrt{p(1-p) * (\frac{1}{n_1} + \frac{1}{n_2})}$ where $p = \frac{x_1+x_2}{n_1+n_2}$, where x and n refer to the number of "successes" (here, purchases) over the number of "trials" (here, site visits). The length of each tail of a 95% confidence interval is calculated by multiplying the standard error by 1.96.

c. Is this confidence interval precise enough that you would recommend running this experiment? Why or why not?

d. Your boss at the newspaper, worried about potential loss of revenue, says he is not willing to hold back a control group any larger than 1% of users. What would be the width of the confidence interval for this experiment if only 1% of users were placed in the control group?

5. Sports Cards

Here you will find a set of data from an auction experiment by John List and David Lucking-Reiley (2000).

```
d2 <- read.csv("./data/listData.csv", stringsAsFactors = FALSE)
head(d2)
```

```
##   bid uniform_price_auction
## 1   5                     1
## 2   5                     1
## 3  20                     0
## 4   0                     1
## 5  20                     1
## 6   0                     1
```

In this experiment, the experimenters invited consumers at a sports card trading show to bid against one other bidder for a pair trading cards. We abstract from the multi-unit-auction details here, and simply state that the treatment auction format was theoretically predicted to produce lower bids than the control auction format. We provide you a relevant subset of data from the experiment.

In this question, we are asking you to produce p-values and confidence intervals in three different ways: (1) Using a `t.test`, using a regression, and using randomization inference.

a. Using a `t.test`, compute a 95% confidence interval for the difference between the treatment mean and the control mean.

b. In plain language, what does this confidence interval mean?

c. Regression on a binary treatment variable turns out to give one the same answer as the standard analytic formula you just used. Demonstrate this by regressing the bid on a binary variable equal to 0 for the control auction and 1 for the treatment auction.

d. Calculate the 95% confidence interval you get from the regression. There is a built in,

```
??confidence
```

e. On to p-values. What p-value does the regression report? Note: please use two-tailed tests for the entire problem.

f. Now compute the same p-value using randomization inference.

g. Pull the same p-value from the `t.test`.

h. Compare the two p-values in parts (e) and (f). Are they much different? Why or why not? How might your answer to this question change if the sample size were different?