**Business Experimentation and Causal Methods**

**Assignment 1**

**Prof. Fradkin**

**Due Date: Noon ET, Monday March 22**

In this assignment, we will begin using R and we will also do a problem about potential outcomes.

Starting instructions:

1. Before you begin this assignment, please watch the ‘Introduction to R’ lecture.
2. If you want an R refresher, Please go to rstudio.cloud, ‘Assignment 1’, and open up practice\_code.R. Practice\_code.R contains a record of the commands used in the ‘Introduction to R’ lecture. Take some time to run these commands and make sure they work for you and that you understand what you do.

NOTE: If you’d like to use Rstudio on your own computer rather than rstudio.cloud, please download the files from Google Drive or from rstudio.cloud to your own computer. Ask on Slack if you need help.

Submission instructions:

Please write your answers in the file assignment1\_notebook.Rmd (details below). Once you’re done press the ‘knit’ button in rstudio.cloud or Rstudio. This will create a pdf file. Please download the file and submit it on Gradescope.

[**Problem 1: Using R to analyze the class list and assign treatment. (70 points)**](#_bxhsily681c2) **1**

[**Problem 2: Wayfair Question and Answer Page (20 points)**](#_yd951qx7zohv) **2**

## Problem 1: Using R to analyze the class list and assign treatment. (70 points)

Our dataset will be the list of students in the class. Our goal for this assignment is the following. The script assignment2\_notebook.Rmd will guide you through each of these steps.

|  |  |
| --- | --- |
| ***Variables Definitions for class\_data.csv*** | |
| *name* | *Student’s name (last name, first name)* |
| *user\_id* | *Student’s user id* |
| *email* | *Student’s email* |
| *treatment\_group* | *Assigned treatment group for each student.* |
| *section* | *Student’s assigned section.* |
|  |  |

1. Read the data. (10 points)
2. Find the record with your own name. (10 points)
3. Find the treatment group associated with your name.
4. Create a new column called ‘section\_code’ that takes the value of 1 if the student is in section ‘A1’ and 0 otherwise. The purpose is to use this column to calculate what share of students is in each section. (10 points)
5. Calculate the share of students that has each treatment group. (10 points)
6. Plot a histogram of the treatment group using the ggplot function. (10 points)
7. Assign each student to a treatment or control group using simple (bernoulli) randomization. Calculate the share of students in the control that are in section A1. (20 points)
8. BONUS: Refine the plot in 5) to use a different theme, label the axes, and change the color based on the section (A1 or B1).
   1. HINT: use either the theme\_bw() or theme\_classic() theme.
   2. Use the commands xlab() and ylab() to label the axes.
   3. Use the fill property of aes to change the colors.
   4. Here is a useful link: <https://rstudio-pubs-static.s3.amazonaws.com/179803_5612c0bc54d54696bc09622b69d634b1.html>

## Problem 2: Wayfair Question and Answer Page (20 points)

Suppose Wayfair is interested in ways to increase the conversion rate of their users. Their hypothesis is that by providing a question and answer section on each furniture page, they would be able to induce users to buy more furniture. The problem is inspired by [this](http://people.bu.edu/zg/publications/reviews-qa.pdf) paper but you don’t have to read it.

Consider the following population of ten users whose sales rates we can measure.

|  |  |  |
| --- | --- | --- |
|  | Potential Outcomes (POs): Revenue | |
|  | If User Saw Q&A | If User did not see Q&A |
| User 1 | 1100 | 1100 |
| User 2 | 100 | 600 |
| User 3 | 500 | 500 |
| User 4 | 900 | 900 |
| User 5 | 1600 | 700 |
| User 6 | 2000 | 2000 |
| User 7 | 1200 | 1200 |
| User 8 | 700 | 700 |
| User 9 | 1100 | 1000 |
| User 10 | 140 | 140 |

* 1. What do User 2’s two potential outcomes mean in plain language? **[5 points]**
  2. Give the individual true treatment effect (that is, the true “impact” or “causal effect” of Q&A) for each of the ten users. We are thinking of the “Treatment” as seeing the Q&A. **[5 points]***Notes:*

● This is only possible because we are working with hypothetical potential outcomes; we could never have this much information with real-world data.

● We encourage the use of computing tools on all problems, but please describe your work so that we can determine whether you are using the correct values (e.g.: include code if using R). This will become more important in future problem sets.

* 1. In a single paragraph, tell a story that could explain this distribution of treatment effects. What might cause some users to have different treatment effects than others? **[5 points]**
  2. For these users, what is the true average treatment effect (ATE) of seeing Q&A? **[5 points]**

How long did this assignment take you to do (hours)? How hard was it (easy, reasonable, hard, too hard)?