Instructions

Complete the following simulation and data analysis tasks and upload your results to your course Github repository. You may submit your results as an R file, an R markdown/notebook file, or as a pdf. Please include all relevant R code and outputs from running the code in your answers. Make sure to comment your code.

Simulation

Using R, demonstrate that treatment and control groups are comparable when the treatment is randomly assigned. To help get you started, consider this interactive graph: https://ellaudet.github.io/graphs/random_assignment.html

You do not need to create an interactive graph or use the same variables as shown in this example. However, your response should be a simulation that does the following:

- Randomly samples *n* observations from a population with some distribution of traits
- Randomly assigns each observation to the treatment or control group with an equal probability
- Repeats this process many times
- Calculates the proportion of traits for the entire sample, treatment, and control groups for each iteration

Using this simulation, show the following:

- As *n* increases, the distribution of traits in the sample has similar proportions to the distribution in the population.
- As *n* increases, the distribution of traits in the treatment and control groups have similar proportions

You do not need to conduct any statistical tests do demonstrate that the proportions are similar. Simple tables or plots that pass an eyeball test are sufficient.

Data Analysis

You will be analyzing the data we discussed in class from Gerber et al.'s 2008 paper "Social Pressure and Voter Turnout: Evidence form a Large-Scale Field Experiment." The experiment sent a random selection of voters a message that pressured people to vote by promising to tell their neighbors if they voted in the upcoming election. The dataset is named voting.csv and is in the data folder of the course github repo. There are three variables in the dataset:

birth: Year of birth of registered voter

message: Whether the voter received the social pressure message

voted: Whether the voter voted in the 2006 election

Use this data complete the following tasks/questions.

1. What is the treatment variable? Is it a discrete or continuous variable? What is the variable's data type?
2. Create a new treatment variable in your data frame that is a binary version of the existing treatment variable. Your new variable should equal 1 if the observation was treated, and 0 otherwise.
3. Compute the average outcome for the treatment group and the average outcome for the control group. Interpret the results by writing 1-2 sentences about what these numbers mean substantively.
4. Use brackets to subset the data frame and create two new data frames, one for the treatment group and one for the control group.
5. What is the average birth year for the treatment and control groups?
6. What is the estimated average causal effect for this experiment? Provide the calculated average effect and a substantive interpretation.
7. Suppose we wanted to claim that the estimated causal effect is an estimated effect for the entire U.S. population. What assumption would need to hold for us to make this claim?