POLI 502 FA20: Homework 1

Due at the start of class on Sept. 16, 2020

Directions. Answer each question completely. Submit your answers along with all graphics and all R code, annotated to note which exercise/sub-exercise it addresses, as well as what it does. For all problems that involve simulation/random number generators, set the random seed equal to the year and month of your birthday: for example, set.seed(198101). You might want to consider using knitr to integrate R code into LaTeX-typeset answers. Please upload your completed assignment as a .pdf file to Blackboard. Reminder: LaTeX formatting is not required until Homework 2.

Exercise 1

A. Following from the instructor-led example regarding potential gender discrimination, use a simulation with one million repetitions and produce a histogram showing the distribution of *difference in proportion* of men vs. women recommended for promotion. Take all numbers/probabilities from the experiment that was discussed in the lecture slides.

Hint 1: There are 24 total subjects of each gender. 21 men and 14 women were recommended for promotion. Assume that, if there were no discrimination, the baseline rate of recommendation is $\frac{35}{48}$ (total recommended divided by total subjects).

Hint 2: The proportion of women recommended $= \frac{\#women\ promoted}{\#women}$.

B. Extract frequencies from the vector of outcomes used in part A above in order to estimate the likelihood that the difference in proportion of men vs. women recommended for promotion could be greater than or equal to 0.3 (with more men recommended). Then, describe whether you think gender discrimination occurred, and why.

Hint 3: You can count the number of events with the following code, in this case looking at instances of disparities favoring men ≥ 5 (in counts):

```
# Create a table that counts the occurrence of each outcome, but only for disparities greater than
# or equal to 5 favoring men
table(Test1M[Test1M >= 5])
```

Hint 4: R can sum all those outcome frequencies and give you their total proportion out of 1 million (i.e., the probability of this occurrence):

```
# Count all occurrences of a 5+ disparity in favor of men, and divide by number of random draws
sum(table(Test1M[Test1M >= 5]))/1000000
```

Exercise 2

Note: Exercise 2 parts A-H are based on Lab 1 from OpenIntro. If you download this lab from the OpenIntro site, you can work through practice that prepares you for the following questions. Although variables can probably be understood from their names, the lab might provide additional useful info.

Load the Behavioral Risk Factor Surveillance System (BRFSS), an annual telephone survey of 350,000 people in the United States. As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of coverage. The data is available here:

http://www.openintro.org/stat/data/cdc.R.
With data in this format, you can use the following code to do so:

```
# Load the BRFSS data
source("http://www.openintro.org/stat/data/cdc.R")
```

A. How many cases are there in this data set? How many variables? For each variable, identify its data type (e.g., numerical & discrete).

B. Create a numerical summary for *height* and *age*, and compute the interquartile range for each. Compute the relative frequency distribution (i.e., create tables) for *gender* and *exerany*. How many males are in the sample? What proportion of the sample reports exercising?

C. Create a **new data frame** called *BRFSS.23.S*, a subset of the original BRFSS data frame, which contains all observations of respondents under the age of 23 that have smoked at least 100 cigarettes in their lifetime. Write the code you used to create the new object as the answer to this exercise.

D. In the **original BRFSS data frame**, create a **new variable** that takes the value of 1 if a respondent is under the age of 23 and has smoked 100 cigarettes in their lifetime, and takes the value of 0 otherwise. Write the code you used to create the new object as the answer to this exercise.

E. Make a scatterplot of weight versus desired weight **using the ggplot2 package**. Be sure to put the presumed explanatory variable on the x-axis, and to label the axes and provide a title for the plot. Describe the relationship between these two variables.

F. Let's consider a new variable: the difference between desired weight, *wtdesire*, and current weight, *weight*. Create this new variable by subtracting the two columns in the data frame and assigning them to a new object called *wdiff*.

G. At what level of measurement is *wdiff*, and what are its units? If an observation has *wdiff* equal to 0, what does this mean about the person's weight and desired weight. What if *wdiff* is positive or negative?

H. Create a histogram of *wdiff* **using the base-R hist() command**, then describe the distribution of *wdiff* in terms of its center, shape, and spread. What does this tell us about how people feel about their current weight?

Exercise 3

A. Download the following two datasets from the Correlates of War website (http://www.correlatesofwar.org/) and load them as R data frames: (1) COW National Trade 4.0 (National_COW_4.0.csv, and (2) the participant-level MID data 4.3 (MIDB_4.3.csv).

B. Merge the two data frames by state and year using an inner join (hint: use *ccode* and *year* variables in the trade data; *ccode* and *StYear* in the MID data). Explain the unit of analysis for the new data frame.

C. Merge the two data frames again, this time using a full join, and aggregating as necessary to create a new data frame where the unit of analysis is the state-year (i.e., where there is one observation per state, per year). Then create a new variable that is equal to 1 if at least one MID originates for a given state in a given year, and 0 otherwise.

D. Assume we want to know the mean trade value in state-years in which no MIDs originated as well as in state-years where one or more MIDs originated. What type of graph would communicate this information well (a written description is sufficient; no need to code it)? And why couldn't we use the data frame from part B above for this task?

From the Text

Note: Do all math in R and submit the (annotated) code. Show all your work!

Diez et al.: Chapter 1 & 2 Exercises: 1.2, 1.10, 1.14, 2.21, 2.33.