Midterm 1

Instructions:

Download the labor.csv data file. You can open an Rmd on your desktop RStudio application or using posit.cloud.

Answer all questions to the best of your ability. ***Do not forget to self-assess your performance after you completed the midterm!***

Consider a study in the 1970's on the factors affecting a woman's participation in the paid labor force (employed or not) in the United States. In particular, it is of interest to determine the relationship between age, number of children a woman has that are under age 6, and paid labor force participation (whether or not a woman has a job). A random sample of n = 753 women was collected and the resulting data can be found in the file named labor.csv.

Importing data tips:

- Using a Project: Download the csv file, put it in your STAT 630 data folder and use the here function (see intro_ggplot2.Rmd from Week 3 for an example if you need help)
- Downloading "manually": If you do not want use an RStudio Project, simply put the data file in the same folder as your midterm Rmd, then go to the Environment window and click on Import Dataset > from Text (readr) > Browse to find your data. You will see code generated to download your dataset.

If you are having trouble importing your data, please ask Dr. Moore to come help right away.

Part 1: Study Design

We do not have information on how this data was collected. Imagine you were asked to consult on this project investigating factors that affect women's participation in the workforce in the U.S.

1. If you had all the resources and time in the world, how would you obtain a sample of women to be a part of your study?
2. Ideally, we want to generalize the results of our study to all women who are working age in the U.S. Based on your chosen sampling method in the previous question, what is the population of women that you can generalize your results to?
3. What is a variable (NOT listed in the dataset) that you think would be a useful factor in determining whether or not a woman participates in the paid labor force? Explain.

Part 2: Exploratory Data Analysis

4. The variable kids_under6 has the values 0, 1, and 2. Do you think this variable should be treated as an integer or a factor variable? Explain your reasoning.

5. Using the R package of your choice (or manually creating in markdown), create the following table of summary statistics. Calculate the **mean** and **standard deviation** for quantitative data and the **counts** and **percentages** for categorical data.

	Does not participate	Participates
Age		
Number of children under 6		
Wife College		
Husband College		
Family Income (excluding wife)		

mean (sd) or n (%)

6. Create a plot to visualize the relationship between participation in the paid labor force and number of children under 6.
7. Using your plot and the summary statistics you calculated in question 5, comment on any similarities or differences between whether or not a woman works and the number of kids she has under the age of 6.
8. Create a plot to visualize the relationship between participation in the paid labor force, number of children under 6, and age.
9. Now that you have added age to the plot, how does that affect the relationship between workforce participation and the number of kids under the age of 6? Comment.

Part 3: Sampling Distributions

10. In this problem, you will create a 95% confidence interval for the true proportion of women who participated in the paid labor force in the 1970's.
    a. First, check that both the necessary conditions for the Central Limit Theorem are met. Show all work/code.
    b. Regardless of whether or not the conditions are met, create and interpret the 95% confidence interval using the context of the problem.
11. According to the internet (so this may or may not be true), 40% of married women were employed by 1970. Based on the confidence interval you calculated, does this percentage seem reasonable? Comment on why or why not.

Reflection Questions

1. Was there anything you found difficult with this exam? What topics (if any) do you feel you still need more work on?

2. Give yourself a rating for this assignment using the EMRN rubric.

E - Excellent

M - Meeting expectations

R - Revision needed

N - Not accessible (mostly blank or did not complete)