Stat 311 Homework 2

This assignment has some problems related to Lesson2 and emphasizes exploratory data analysis (EDA)—visualization and numeric summaries for qualitative and quantitative data. We recommend that you create a new folder for this assignment. Download the data files and the Rmd template to this folder before you begin. This template only provides the header and setup code, and the headers for the main problems. You need to add everything else. Use the same formatting structure as you did for Homework 1, using #### a), etc. to label subparts of problems. Check out the two .Rmd files that appear on the Lesson 2 Presentations page—they contain code examples for several types of summaries that were presented in the lectures. Upload your knitted HTML file to Canvas.

Problems 1 – 3 do not require any code. Simply type your answers into the .Rmd file. Problems 4 and 5 require the use of R code.

To reinforce the concepts in the Lesson 2 lectures and for extra practice with R commands, I recommend that you try some of the OpenIntro tutorials that I linked on the Readings page for Lesson 2.

1. For each part, compare the distributions, A and B, based on means/SDs and medians/IQRs. **Do not show any calculations—you do not need them to answer.** Simply by looking at the numbers, state how the means/SDs or medians/IQRs compare. Make sure to explain your reasoning.

   a) Compare the means/SDs for **A:** −20, 0, 0, 0, 15, 25, 30, 30 and **B:** −40, 0, 0, 0, 15, 25, 30, 30

   b) Compare the means/SDs for **A:** 100, 200, 300, 400, 500 and **B:** 0, 50, 300, 550, 600

   c) Compare the medians/IQRs for **A:** 0, 10, 50, 60, 100 and **B:** 0, 100, 500, 600, 1000

   d) Compare the medians/IQRs for **A:** 1, 2, 3, 4, 5 and **B:** 6, 7, 8, 9, 10

2. Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead, he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. To achieve this goal, should he fire employees who have the greatest number of days off, least number of days off, or those who have about the average number of days off?

3. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found a mean score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

   a) Does the new student's score increase or decrease the mean score? Give a qualitative argument.

   b) What is the new mean? Show your work. [Two bonus points if you use the $$ notation to write out your equations that show your work. To get a fraction, use `$$\frac {numerator}{denominator}$$`. To get a bar over a letter, use `$$\bar x$$`. To get a subscript, such as s[x], use `$$s_{x}$$`. These various commands can be combined such as `$$\bar x = \frac {1+2+3}{3} = 2$$.` `$$` puts equations on their own line, whereas use `$` for inline equations.

   c) Does the new student's score increase or decrease the standard deviation of the scores? Explain.

Stat 311 Homework 2

4. This problem uses the same data regarding environmental policy versus economic policy that were presented in Lesson 2, Lecture 1, except the data are categorized by education level or party identity.
    a) Read in GallupByEd.csv and GallupByPI.csv, creating two separate objects that store the data. Convert variables in each object to factors as needed. Reorder/rename factors. [The code for this is provided in the template, so nothing to do but look at the code to understand how factors were reordered/relabeled. This might come in handy for your final group project, where we will not provide the code].
    b) How many observations are there in each data set?
    c) Produce two two-way contingency tables, one for each data set, with education or party ID in rows and the response in the columns. You can just leave the tables as they display in the R output.
    d) What is the joint percentage of people who favor economic policy and have some college?
    e) What are the marginal percentages for party identification? [Hint: you will be reporting three percentages]
    f) Produce two more tables that show row conditional percentages instead of counts. What are the conditional percentages for Response for those participants who identify as democrats?
    g) Pick one of the two data sets and create two bar graphs (your choice of versions) to explore the association between either education or party ID and Response. Make sure the axes are appropriately labeled. Summarize the information you glean from the bar graphs.
    h) For the data set you picked, does the row variable (either education or party ID) appear to be associated with the response. Explain. [Note: this is a qualitative answer based on data visualization.]
5. Complete the following parts using a data set about popular diets (PopularDiets.csv). The data dictionary for the data set is found in the file DietDataDescription.pdf. The journal article that explains the study with results is in the file JournalArticleForDietStudy_joc40214. You will need to browse the journal article to answer parts a) and c).
    a) Was this study an observational study or an experiment? Briefly explain.
    b) What participants were sampled for this study?
    c) What do you believe to be the population of interest? Do you think the results can be generalized to the population of interest or some other population? Explain.
    d) Read in the data. Set variables to factors as needed. [Code for this is provided in the template] There is nothing you need to do.
    e) How many observations are in the data set? How many of the subjects completed the study?
    f) Explore the initial weight variable. Present summary statistics and two different graphs that give you slightly different information (your choice) that provides insight into the distribution of this variable. Summarize your overall findings for the distribution of initial weight by describing the information you get from the summary statistics and graphs. What numeric statistics do you think are best to use to summarize the distribution? Explain.
    g) Explore how initial weight varies by type of diet. [Hint: use comparative box plots or faceted histograms by diet type] Make a qualitative assessment regarding differences in the distribution of initial weight by diet type.
    h) Do you think that the distributions of initial weights across the four diets is important? Explain.