STAT 324 Spring 2020: Homework 2

To receive credit, you must submit your assignment to Canvas before **6pm, Thursday, February 6**. The file submission must be a knitted .html or .pdf file, made using RMarkdown. The code you used to answer the questions should be included in your file. You do not need to submit your .rmd file. The assignment is worth 50 points.

Questions:

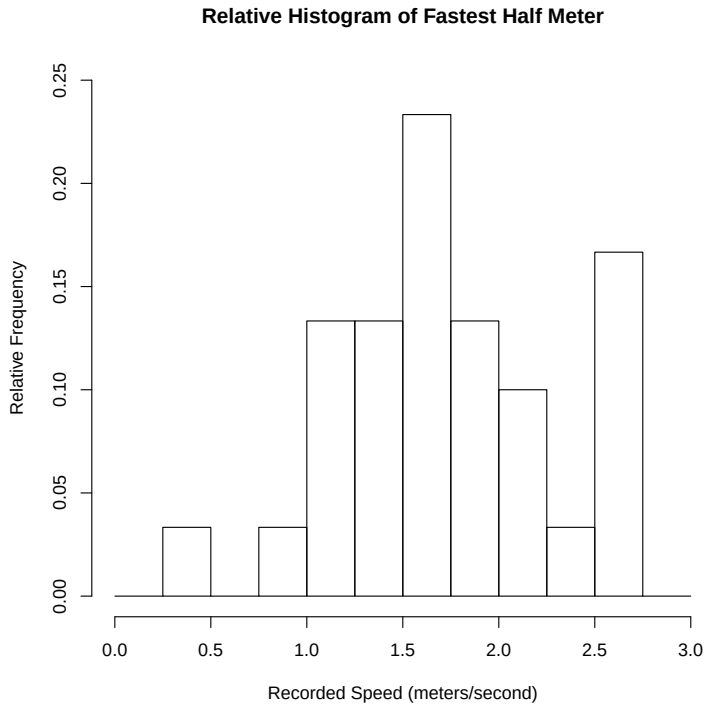
1. There are 12 numbers on a list, and the mean is 24. The smallest number on the list is changed from 11.9 to 1.19.

   (a) Is it possible to determine the direction in which (increase/decrease) the mean changes? Or how much the mean changes? If so, by how much does it change? If not, why not?

   (b) Is it possible to determine the direction in which the median changes? Or how much the median changes? If so, by how much does it change? If not, why not?

   (c) Is it possible to predict the direction in which the standard deviation changes? If so, does it get larger or smaller? If not, why not? Describe why it is difficult to predict by how much the standard deviation will change in this case.

2. A zoologist collected wild lizards in the Southwestern United States. Thirty lizards from the genus *Phrynosoma* were placed on a treatmill and their speed measured. The recorded speeds (meters/second) (the fastest time to run a half meter) for the thirty lizards are summarized in the relative frequency histogram below. (Data Courtesy of K. Bonine *)



(a) Is the percent of lizards with recorded speed below 1.25 closest to: 25%, 50%, or 75%?

(b) In which interval are there more speeds recorded: 1.5-1.75 or 2-2.5?

(c) About how many lizards had recorded speeds above 1 meters/second?

(d) In which bin does the median fall? Show how you know.

3. After manufacture, computer disks are tested for errors. The table below gives the number of errors detected on a random sample of 100 disks. Hint: You can use the rep() function in R to make a vector of repeated numbers.

Number of Defects	0	1	2	3	4
Frequency	41	31	15	8	5

(a) Describe the type of data (ex: nominal) that is being recorded about the sample of 100 disks, being as specific as possible.

(b) Construct a frequency histogram of the information with R.

(c) What is the shape of the histogram for the number of defects observed in this sample?

(d) Calculate the mean and median number of errors detected on the 100 disks by hand and with R. How do the mean and median values compare and is that consistent with what we would guess based on the shape?

(e) Calculate the sample standard deviation with R. Explain what this value means in the context of the problem.

(f) Calculate the first and third quartiles and IQR by hand and with R. Are the values consistent betweeen the two methods? Explain what the three values mean in the context of the problem.

(g) What proportion of the computer disks had a number of errors greater than the mean number of errors?

(h) What range of values for this sample data are not considered outliers using the [Q1-1.5IQR, Q3+1.5IQR] designation (using the IQR you calculated by hand)?.

(i) Make a boxplot of the data using R and compare the lines to the values you calculated by hand.

(j) Compare and contrast (briefly) the information about the data given by the histogram in part b and the boxplot in part i.

4. The file brexit.csv contains the results of 127 polls, including both online polls and telephone polls, carried out from January 2016 to the referendum date on June 23, 2016. Use that dataset to answer the following questions.

   a. Use R to create a histogram for the proportion who answered "Remain" when polled. Describe the shape of the data.

   b. Now construct two separate histograms for the proportion who answered "Remain". Make one histogram for online polls and another histogram for telephone polls. Describe the shape and relative position of the data.

   c. Compute the mean and median proportion voting "Remain" observed for the online and telephone polls. Compare both measures of center across the two groups.

   d. Compute and compare the standard deviation observed in the two groups.

   e. Use R to help you create side by side boxplots of the two sets so they are easily comparable.

f. How many values were identified as outliers? Would these values have been identified as an outlier in the other type of poll? Use the 1.5IQR rule for identifying outliers.

h. What would be the mean and median proportion answering "Remain" if we combined the two poll types together? Show how one of these can be calculated directly from your summary measures in part (c).

i. Next, calculate the mean proportion of respondents that answered "Leave" for both online and telephone polls. What other factor in the data can explain the much smaller gap between means here compared to part c? Explain.