Assignment 1

Due at 11:59pm on September 16.

This assignment is to be submitted individually. Turn in this assignment as an HTML or PDF file to ELMS. Make sure to include the R Markdown or Quarto file that was used to generate it. You should include the questions in your solutions. You may use the qmd file of the assignment provided to insert your answers.

Git and GitHub

1) Provide the link to the GitHub repo that you used to practice git from Week 1. It should have:

- Your name on the README file.
- At least one commit with your name, with a description of what you did in that commit.

https://github.com/zzeng05/Zeng1-Liu2-a1.git

Reading Data

Download both the Angell.dta (Stata data format) dataset and the Angell.txt dataset from this website: https://stats.idre.ucla.edu/stata/examples/ara/applied-regression-analysis-by-fox-data-files/

2) Read in the .dta version and store in an object called `angell_stata`.

```
library(haven)
angell_stata <- read_dta("/Users/zpzzz/Desktop/SURV727/Zeng1-Liu2-a1/angell.dta")
summary(angell_stata)
```

```
    city               morint          ethhet          geomob
Length:43          Min.   : 4.20   Min.   :10.60   Min.   :12.10
Class :character   1st Qu.: 8.70   1st Qu.:16.90   1st Qu.:19.45
Mode  :character   Median :11.10   Median :23.70   Median :25.90
                   Mean   :11.20   Mean   :31.37   Mean   :27.60
                   3rd Qu.:13.95   3rd Qu.:39.00   3rd Qu.:34.80
                   Max.   :19.00   Max.   :84.50   Max.   :49.80
   region
Length:43
Class :character
Mode  :character
```

3) Read in the .txt version and store it in an object called `angell_txt`.

```
angell_txt <- read.table("/Users/zpzzz/Desktop/SURV727/Zeng1-Liu2-a1/angell.txt")
summary(angell_txt)
```

```
     V1                 V2              V3              V4
Length:43          Min.   : 4.20   Min.   :10.60   Min.   :12.10
Class :character   1st Qu.: 8.70   1st Qu.:16.90   1st Qu.:19.45
Mode  :character   Median :11.10   Median :23.70   Median :25.90
                   Mean   :11.20   Mean   :31.37   Mean   :27.60
                   3rd Qu.:13.95   3rd Qu.:39.00   3rd Qu.:34.80
                   Max.   :19.00   Max.   :84.50   Max.   :49.80
     V5
Length:43
Class :character
Mode  :character
```

4) What are the differences between `angell_stata` and `angell_txt`? Are there differences in the classes of the individual columns?

5) Make any updates necessary so that `angell_txt` is the same as `angell_stata`.

6) Describe the Ethnic Heterogeneity variable. Use descriptive statistics such as mean, median, standard deviation, etc. How does it differ by region?

Describing Data

R comes also with many built-in datasets. The "MASS" package, for example, comes with the "Boston" dataset.

7) Install the "MASS" package, load the package. Then, load the Boston dataset.

8) What is the type of the Boston object?

9) What is the class of the Boston object?

10) How many of the suburbs in the Boston data set bound the Charles river?

11) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each variable.

12) Describe the distribution of pupil-teacher ratio among the towns in this data set that have a per capita crime rate larger than 1. How does it differ from towns that have a per capita crime rate smaller than 1?

Writing Functions

13) Write a function that calculates 95% confidence intervals for a point estimate. The function should be called `my_CI`. When called with `my_CI(2, 0.2)`, the function should print out "The 95% CI upper bound of point estimate 2 with standard error 0.2 is 2.392. The lower bound is 1.608."

*Note: The function should take a point estimate and its standard error as arguments. You may use the formula for 95% CI: point estimate +/- 1.96*standard error.*

Hint: Pasting text in R can be done with: `paste()` *and* `paste0()`

14) Create a new function called `my_CI2` that does that same thing as the `my_CI` function but outputs a vector of length 2 with the lower and upper bound of the confidence interval instead of printing out the text. Use this to find the 95% confidence interval for a point estimate of 0 and standard error 0.4.

15) Update the `my_CI2` function to take any confidence level instead of only 95%. Call the new function `my_CI3`. You should add an argument to your function for confidence level.

Hint: Use the `qnorm` *function to find the appropriate z-value. For example, for a 95% confidence interval, using* `qnorm(0.975)` *gives approximately 1.96.*

16) Without hardcoding any numbers in the code, find a 99% confidence interval for Ethnic Heterogeneity in the Angell dataset. Find the standard error by dividing the standard deviation by the square root of the sample size.

17) Write a function that you can `apply` to the Angell dataset to get 95% confidence intervals. The function should take one argument: a vector. Use if-else statements to output NA and avoid error messages if the column in the data frame is not numeric or logical.