Statistical Analysis Project

Due 11:59pm Wednesday 5/26

Please visit the website https://vincentarelbundock.github.io/Rdatasets/datasets.html. Each row of this table corresponds to a dataset. For this project, you will perform some basic statistical analysis on one dataset of your choice. Try to find a dataset that interests you and has at least 30 rows. Browse the titles and read the descriptions in the DOC files to understand what the data is and how it was collected. In the DOC file, read the section labelled Format to see what sort of data is recorded and what units are used. Most datasets have lots of different measurements (columns), and you will need to pick *one* to focus on. Once you have selected a dataset and a column within that data set, click on the CSV file. Open this file with any spreadsheet program/app (Microsoft Excel, Mac Numbers, Google Sheets, etc). The numbers of rows (minus the header row) is the sample size. The column that you selected contains all of the sample measurements. Use these sample measurements to answer the following questions.

Answer the following questions, and give reasons for your answers. Even if you use a spreadsheet program to make a calculation, you must still provide the formula being used. Upload your project to the following Dropbox request folder before 11:59pm Wednesday 5/26: https://www.dropbox.com/request/BcKRa6cZb1lPfTVvxIvJ

1. What is the *exact* title of the dataset you have selected?
   You must be exact so I can find it on the website above.
2. What is the measurement (column) you have selected, and what is the unit for this measurement?
3. What is the sample size n? That is, how many rows does your dataset have (not counting the header row)? Note: you must select a dataset with $n \geq 30$.
4. What is the sample mean $\overline{x}$?[1]
5. What is the sample standard deviation s?[2]
6. Approximately what kind of distribution does the random variable $\overline{x}$ have, and why? What is the name of the theorem that tells you so?
7. Estimate the population mean μ in two ways.
   (a) Give a point estimate for μ and a "margin of error".
   (b) Construct a 99% confidence interval for μ.

[1] Probably your dataset is very large and calculating $\overline{x}$ by hand is not practical. The spreadsheet program you are using can calculate this very easily. Click in any empty cell and type ″=AVERAGE(″, then highlight all the measurements in the sample, i.e. the entire column of measurements you are studying. Once all of the sample measurements are highlighted, type ″)″ and hit ENTER/RETURN. If this does not work, just google it, e.g. ″how to calculate average in excel″.

[2] Again, the spreadsheet program you are using can calculate this very easily. Click in any empty cell and type ″=STDEV(″, then highlight all the measurements in the sample, i.e. the entire column of measurements you are studying. Once all of the sample measurements are highlighted, type ″)″ and hit ENTER/RETURN. If this does not work, just google it, e.g. ″how to calculate sample standard deviation in excel″.