ENGRD 2700: Basic Engineering Probability and Statistics
Spring 2019

Homework 1

Due Saturday, February 9 at 2 am. Submit to Gradescope by clicking the name of the assignment. See `https://people.orie.cornell.edu/yudong.chen/engrd2700_2019sp.html#homework` for detailed submission instructions.

When completing this assignment (and all subsequent ones), keep in mind the following:

- You must complete the homework individually and independently.
- Provide evidence for each of your answers. If a calculation involves only very minor computation then explain the computation you performed and give the results. If a calculation involves more complicated steps on many many records then hand in the calculations and formulas for the first few records only.
- Write clearly and legibly. You are encouraged to *type* your work although you do not have to. We may deduct points if your answers are difficult to read or disorganized.
- For questions that you answer using R, attach any code that you write, along with the relevant plots. You may use other software, but the same condition applies.
- Submit your homework a single pdf file on Gradescope.

1. The file `quartet.csv` contains four datasets of x and y values, side by side.

   (a) Compute the sample mean, sample median, and sample standard deviation for each column of the dataset.

   (b) Based solely upon the summary statistics you computed in part (a), how do the four datasets compare?

   (c) Construct scatterplots for each of the four datasets. (Hint: In R, you can use the command `par(mfrow=c(2,2))` to combine multiple plots into a single 2-by-2 graph in R. If you do, this command should precede any code that you use to generate plots.)

   (d) Based solely upon the plots you generated in part (c), how do the four datasets compare?

   (e) What's the moral of the story? (That is, what does this example suggest about what should be done when analyzing data?)

2. Answer the questions below about the dataset `CountyData.csv` from the U.S. Census Bureau, performing any data analysis you deem appropriate. The dataset consists of 3143 observations on 53 variables, which are described in the file `CountyData_Info.pdf`.

   (a) Provide a histogram of the per-county percentage of residents who speak a foreign language at home during 2006-2010.

   (b) What was the median per-county amount of federal spending in 2009?

   (c) Create a scatter plot of the percentage of residents below the poverty level (y-axis) versus the percentage of the population with a bachelors degree. Comment on what you see.

   (d) What fraction of counties have a population whose percentage under the age of 18 is above 30%?

3. A subdivision of 24 houses has a mean price of \$500,000, a median of \$440,000, and a standard deviation of \$30,000. A new house is then built in the subdivision that has a price of \$700,000.

   (a) What is the new mean house price?

   (b) What is the new standard deviation?

   (c) Does the median increase, decrease, or stay the same after the new house is built? Or can no conclusion be made? Explain.

4. Consider a data sample $x_1, x_2, \ldots, x_n$. Let $\bar{x}$ and s_x^2 denote its sample mean and sample variance.

   (a) Suppose that you modify these data by adding a constant c to each observation in the sample, and then multiplying by another constant k, to obtain a modified sample $y_1, \ldots, y_n$. (That is, $y_i = (x_i + c) \times k$ for each i.)

   What are $\bar{y}$ and s_y^2, the sample mean and variance of the modified data? Justify your answer mathematically, using the definition of the sample mean and variance. (We saw a similar question in lecture; here you need to write down the proof yourself.)

   (b) Finally, suppose we built another modified data sample $z_1, \ldots, z_n$, where

$$z_i = \frac{x_i - \bar{x}}{s_x},$$

   where s_x is the sample standard deviation of the x–data. This procedure *standardizes* the original data. What are $\bar{z}$ and s_z^2? Justify your answer.