Homework 1

Due: Friday February 7 at 2:00 pm

See general homework tips and submit your files via the course website.

For all exercises, use the **seeds** data set defined in **HW1Data.sas** in the Homework 1 folder on the course website. The data set is based on the **seeds** data[1] from the UCI Machine Learning Repository[2]. The raw data is contained in **seeds_dataset.txt**, which is available from the UCI Machine Learning Repository and in the course web site. The data contains measurements for seeds from 3 different varieties of wheat, and the measurements are described on http://archive.ics.uci.edu/ml/datasets/seeds.

Exercise 1:

a) Obtain basic descriptive statistics for **area** for all of the data. Comment on typical values, skew, and ranges for the area values of wheat seeds.

b) Repeat the analysis in part **a)** by **variety**. Comment on what this tells us about similarities and differences of the areas of the three varieties' seeds.

Exercise 2:

a) For **area**, visually and quantitatively check if an assumption of normality would be reasonable and state your conclusion.

b) Repeat the analysis in part **a)** by **variety**. Based on the results, comment on which of the wheat varieties' seeds show significant differences from normality in their areas.

Exercise 3:

a) A claim is made that a typical (mean or median) area for wheat seeds is about 14. Test whether the typical (mean or median) wheat seed area for the population this sample came from is significantly different from 14. Consider your test for normality in **Exercise 2** to choose the appropriate location test. State your conclusions.

b) Another claim is made that Rosa wheat seeds have smaller area than Kama wheat seeds (i.e. the **area** value for Rosa wheat seeds is significantly less). Perform a hypothesis test to check this claim, and state your conclusions. Again, consider the hypothesis tests in **Exercise 2** when choosing your test for difference in seed area.

Exercise 4:

a) Perform a correlation analysis for the **area**, **compactness**, and **width** variables for all of the data. State what this tells us about relationships between these three seed measurements in the data sampled, and what we might infer about relationships in the population it was sampled from.

b) Perform the same correlation analysis by **variety**. Comment on how the relationships between the three characteristics differ across wheat varieties, and note any differences with what you found in part **a)** for the combined data.

[1] http://archive.ics.uci.edu/ml/datasets/seeds

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.