DCCS 221 – Probability and Statistics
Lab # 6

1. Open the data set *drivers* in Python.
   a. Go to the DCCS 221 Blackboard.
   b. Download the following files to your DCCS221/Lab6 folder
      i. drivers.csv
      ii. Labnote_6.pdf
   c. Open Python.
   d. In Python, open the "drivers.csv" file.
   e. Create the vector *x*, which is the first column of the matrix drivers. It refers to the number of drivers per household. Create the vector *y*, which is the second column of the matrix drivers, and refers to the household size. All data was obtained from the Bureau of Transportation Statistics and is from 1995.
      a. x=drivers(:,1)
      b. y=drivers(:,2)
2. Create a vector *x20* that includes only the first 20 observations of vector *x*.
   x20=x(1:20,1)
3. Draw a histogram for *x20*.
   a. Around what value do you expect the mean to be? And the median?
   b. Use Python to compute the actual values for the mean and the median.
      i. Mean(x20)
      ii. Median(x20)
   c. Change the last value of *x20* to *7* (reference an observation in a vector and set it to some value; i.e. *x20(20)=7* references the 20th observation of x20 and sets it to 7).
      i. Recalculate the mean and the median. What do you see?
      ii. Which one is more sensitive to the data, the mean or the median?
4. Create a vector x100. x100 should include the first 100 observations of vector x and a vector x4000 that includes the first 4000 observations of vector x. Repeat steps (b) and (c) for each case in the previous Step. What can you say about the effects that the number of observations might have on the calculation of the mean and the median?
5. We discussed dispersion qualitatively. The mean absolute deviation (MAD), sample standard deviation, and sample coefficient of variation (COV) are quantitative measures of dispersion. Calculate these for both *x* and *y*. Which measure would you want to use to compare the dispersion of the two samples?
   a. mad(x) –
   b. std(x)
   c. std(x)/mean(x)
6. Plot a histogram for *x* (using 7 bins) and a histogram for *y* (using 10 bins) separately (you can use the *figure* command to create a new figure for each plot).
   hist(x,7), hist(y,10)
   a. Is the data skewed to the right or to the left for both *x* and *y*? Why?
   b. Use Python to compute the skewness and kurtosis (e.g. *skewness(x), kurtosis(y)*)

c. **Challenge Question (Extra pts.) -** Provide your explanation on skewness and kurtosis.

d. Use Python to compute the covariance and the correlation coefficient.

   - *cov(x,y)*
   - *corrcoef(x,y)*