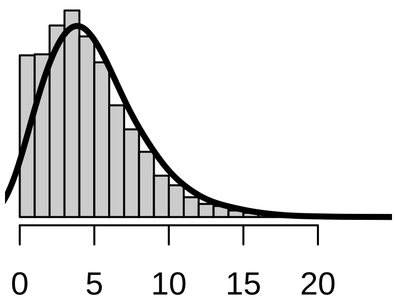
DATA 606 Fall 2016 - Final Exam

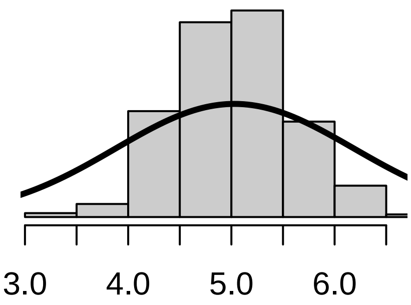
Part I

Figure A below represents the distribution of an observed variable. Figure B below represents the distribution of the mean from 500 random samples of size 30 from A. The mean of A is 5.05 and the mean of B is 5.04. The standard deviations of A and B are 3.22 and 0.58, respectively.

A. Observations



B. Sampling Distribution



a. Describe the two distributions (2 pts).

b. Explain why the means of these two distributions are similar but the standard deviations are not (2 pts).

c. What is the statistical principal that describes this phenomenon (2 pts)?

Part II

Consider the four datasets, each with two columns (x and y), provided below.

```
options(digits=2)
data1 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))
data2 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74))
data3 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))
data4 <- data.frame(x=c(8,8,8,8,8,8,8,19,8,8,8),
                    y=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.89))
```

For each column, calculate (to two decimal places):

a. The mean (for x and y separately; 1 pt).

b. The median (for x and y separately; 1 pt).

c. The standard deviation (for x and y separately; 1 pt).

For each x and y pair, calculate (also to two decimal places; 1 pt):

d. The correlation (1 pt).

e. Linear regression equation (2 pts).

f. R-Squared (2 pts).

For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts)