DATA 606 Spring 2017 - Final Exam

Part I

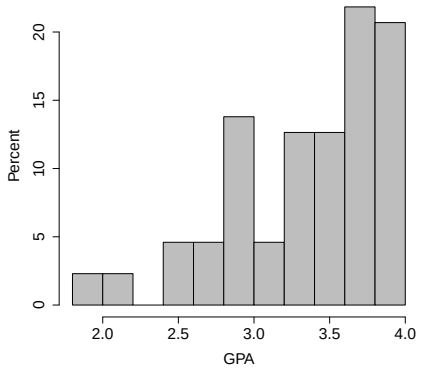
1. A student is gathering data on the driving experiences of other college students. A description of the data car color is presented below. Which of the variables are quantitative and discrete?

car	1 = compact, 2 = standard size, 3 = mini van, 4 = SUV, and 5 = truck
color	red, blue, green, black, white
daysDrive	number of days per week the student drives
gasMonth	the amount of money the student spends on gas per month

a. car
b. daysDrive
c. daysDrive, car
d. daysDrive, gasMonth
e. car, daysDrive, gasMonth

---

2. A histogram of the GPA of 132 students from this course in Fall 2012 class is presented below. Which estimates of the mean and median are most plausible?



a. mean = 3.3, median = 3.5
b. mean = 3.5, median = 3.3
c. mean = 2.9, median = 3.8
d. mean = 3.8, median = 2.9
e. mean = 2.5, median = 3.8

3. A researcher wants to determine if a new treatment is effective for reducing Ebola related fever. What type of study should be conducted in order to establish that the treatment does indeed cause improvement in Ebola patients?

a. Randomly assign Ebola patients to one of two groups, either the treatment or placebo group, and then compare the fever of the two groups.
b. Identify Ebola patients who received the new treatment and those who did not, and then compare the fever of those two groups.
c. Identify clusters of villages and then stratify them by gender and compare the fevers of male and female groups.
d. Both studies (a) and (b) can be conducted in order to establish that the treatment does indeed cause improvement with regards to fever in Ebola patients.

---

4. A study is designed to test whether there is a relationship between natural hair color (brunette, blond, red) and eye color (blue, green, brown). If a large χ^2 test statistic is obtained, this suggests that:

a. there is a difference between average eye color and average hair color.
b. a person's hair color is determined by his or her eye color.
c. there is an association between natural hair color and eye color.
d. eye color and natural hair color are independent

---

5. A researcher studying how monkeys remember is interested in examining the distribution of the score on a standard memory task. The researcher wants to produce a boxplot to examine this distribution. Below are summary statistics from the memory task. What values should the researcher use to determine if a particular score is a potential outlier in the boxplot?

min	Q1	median	Q3	max	mean	sd	n
26	37	45	49.8	65	44.4	8.4	50

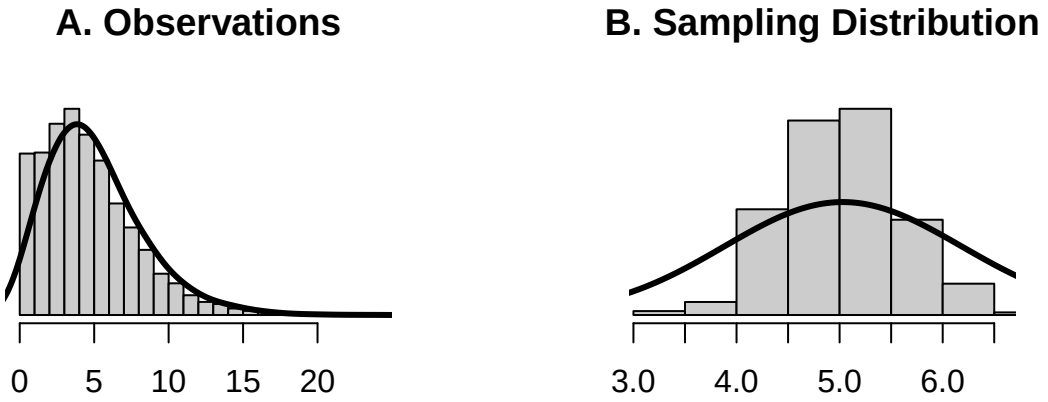
a. 37.0 and 49.8
b. 17.8 and 69.0
c. 36.0 and 52.8
d. 26.0 and 50.0
e. 19.2 and 69.9

---

6. The _____ are resistant to outliers, whereas the _____ are not.

a. mean and median; standard deviation and interquartile range
b. mean and standard deviation; median and interquartile range
c. standard deviation and interquartile range; mean and median
d. median and interquartile range; mean and standard deviation
e. median and standard deviation; mean and interquartile range

7. 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. 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)