DATA 606 Fall 2020 - Final Exam

Gabriel Campos

Part I

Please put the answers for Part I next to the question number (2pts each):

1. **(b)** *daysDrive*
2. **(b)** $mean = 2.9$, $median = 3.8$
3. **(d)** *Both studies* (a) *and* (b) *can be conducted in order to establish that the treatment does indeed cause improvement*
4. **(c)** *there is an association between natural hair color and eye color.*
5. **(a)** 37.0 *and* 49.8
6. **(d)** *median and interquartile range*; *mean and standard deviation*

7a. Describe the two distributions (2pts).

- A *Observations* is right skewed
- B *SampleDistribution* is nearly normal.

7b. Explain why the means of these two distributions are similar but the standard deviations are not (2 pts).

- B is showing the mean of the means for its samples.
- As the number of samples increases the more nearly normal the data becomes.
- With 500 random samples we are closer to the true mean of our data and the critical values or deviation decreases.

7c. What is the statistical principal that describes this phenomenon (2 pts)?
Central limit theorem

Part II

Consider the four datasets, each with two columns (x and y), provided below. Be sure to replace the `NA` with your answer for each part (e.g. assign the mean of `x` for `data1` to the `data1.x.mean` variable). When you Knit your answer document, a table will be generated with all the answers.

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

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

```
data1.x.mean <- as.numeric(format(round(mean(data1$x),4), nsmall = 4))
data1.y.mean <- as.numeric(format(round(mean(data1$y),4), nsmall = 4))
data2.x.mean <- as.numeric(format(round(mean(data2$x),4), nsmall = 4))
data2.y.mean <- as.numeric(format(round(mean(data2$y),4), nsmall = 4))
data3.x.mean <- as.numeric(format(round(mean(data3$x),4), nsmall = 4))
data3.y.mean <- as.numeric(format(round(mean(data3$y),4), nsmall = 4))
```

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

```
data1.x.median <- as.numeric(format(round(median(data1$x),4), nsmall = 4))
data1.y.median <- as.numeric(format(round(median(data1$y),4), nsmall = 4))
data2.x.median <- as.numeric(format(round(median(data2$x),4), nsmall = 4))
data2.y.median <- as.numeric(format(round(median(data2$y),4), nsmall = 4))
data3.x.median <- as.numeric(format(round(median(data3$x),4), nsmall = 4))
data3.y.median <- as.numeric(format(round(median(data3$y),4), nsmall = 4))
```

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

```
data1.x.sd <- as.numeric(format(round(sd(data1$x),4), nsmall = 4))
data1.y.sd <- as.numeric(format(round(sd(data1$x),4), nsmall = 4))
data2.x.sd <- as.numeric(format(round(sd(data2$x),4), nsmall = 4))
data2.y.sd <- as.numeric(format(round(sd(data2$x),4), nsmall = 4))
data3.x.sd <- as.numeric(format(round(sd(data3$x),4), nsmall = 4))
data3.y.sd <- as.numeric(format(round(sd(data3$x),4), nsmall = 4))
```

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

d. The correlation (1 pt).

```
data1.correlation <- as.numeric(format(round(cor(data1$x,data1$y),4), nsmall = 4))
data2.correlation <- as.numeric(format(round(cor(data2$x,data2$y),4), nsmall = 4))
data3.correlation <- as.numeric(format(round(cor(data3$x,data3$y),4), nsmall = 4))
```

e. Linear regression equation (2 pts).

```
data1.slope <- (data1.y.mean / data1.x.mean) * data1.correlation
data2.slope <- (data2.y.mean / data2.x.mean) * data2.correlation
data3.slope <- (data3.y.mean / data3.x.mean) * data3.correlation

data1.intercept <- data1.y.mean - data1.slope * data1.x.mean
data2.intercept <- data2.y.mean - data2.slope * data2.x.mean
data3.intercept <- data3.y.mean - data3.slope * data3.x.mean
```

f. R-Squared (2 pts).

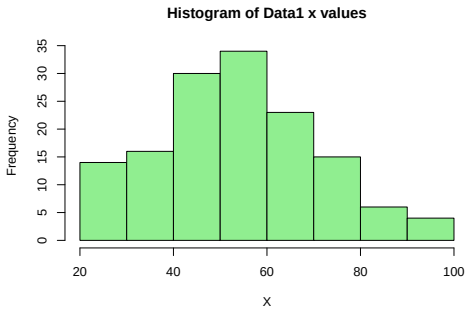
```
data1.rsquared <- data1.correlation^2
data2.rsquared <- data2.correlation^2
data3.rsquared <- data3.correlation^2
```

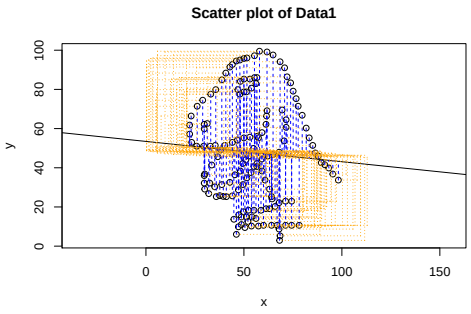
Summary Table

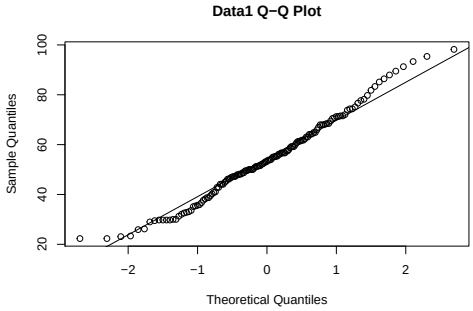
```
## Warning: package 'kableExtra' was built under R version 4.0.3
```

	Data 1		Data 2		Data 3	
	x	y	x	y	x	y
Mean	54.26	47.83	54.27	47.84	54.27	47.83
Median	53.33	46.03	53.14	46.40	53.34	47.54
SD	16.77	16.77	16.77	16.77	16.77	16.77
r	-0.06		-0.07		-0.06	
Intercept	50.92		51.14		50.90	
Slope	-0.06		-0.06		-0.06	
R-Squared	0.00		0.00		0.00	

g. 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)

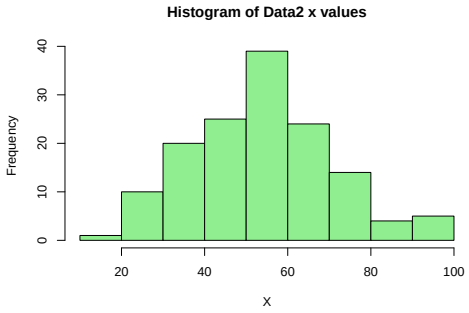


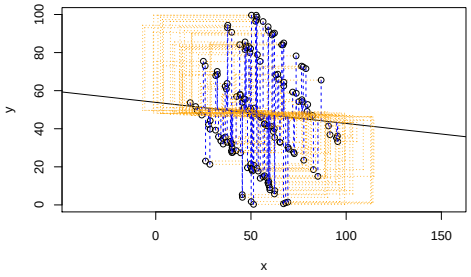


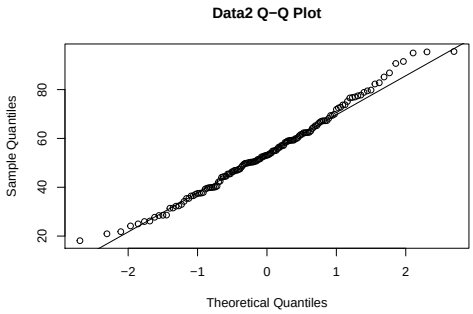


ANSWER - Not all conditions are met, regression method will be rejected.

- **Linearity-** Positive, slightly linear with multiple outliers
- **Nearly Normal Residual-** Left skewed but nearly normal
- **Constant Variability-** Condition not met met based on residual plot
- **Independence Observation-** Condition met

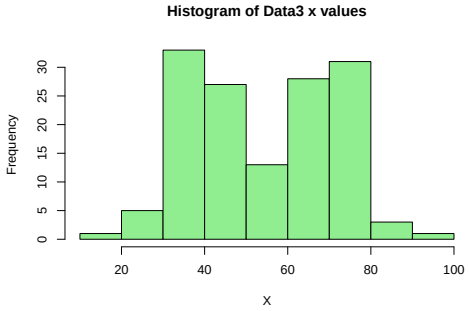


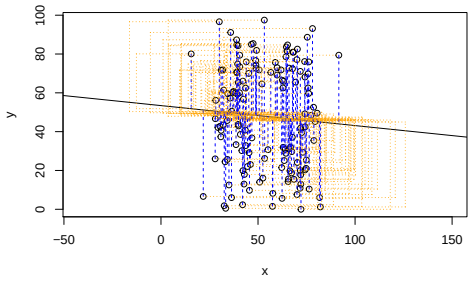


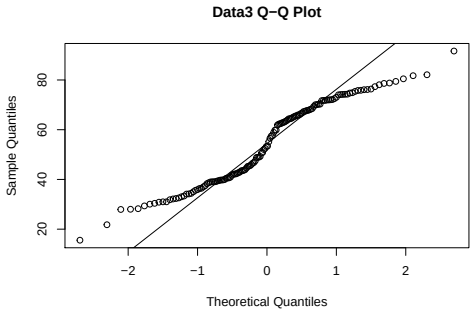


ANSWER - Regression method can be applied

- **Linearity-** Positive, linear
- **Nearly Normal Residual-** Nearly normal condition met
- **Constant Variability-** Condition Met
- **Independence Observation-** Condition met







ANSWER - Not all conditions are met, regression method will be rejected.

- **Linearity-** Condition not met
- **Nearly Normal Residual-** Condition not met
- **Constant Variability-** Condition met
- **Independence Observation-** Condition met

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

ANSWER

- Large chunks of data can be too difficult to analyze without visualizations.
- Visualization tools allow for easy summary of data
- Multiple visualizations helps avoid making assumptions, e.g. some conditions were met when deciding if linear regression model is appropriate, but not all. Deciding based on only one would have led to the wrong assumption