STAT 345 Fall 2018
Homework 8 - Data Exploration Name: ______________________________

Please adhere to the homework rules as given in the Syllabus.

1. Sample Statistics. Timothy and Jimothy run competing lemonade stands. Their sales (in dollars) for this past week are given in the following table. **All calculations here should be done by hand**.

Timothy (x_i)	Jimothy (y_i)
12	10
13	14
9	9
10	11
6	8
11	14
14	13

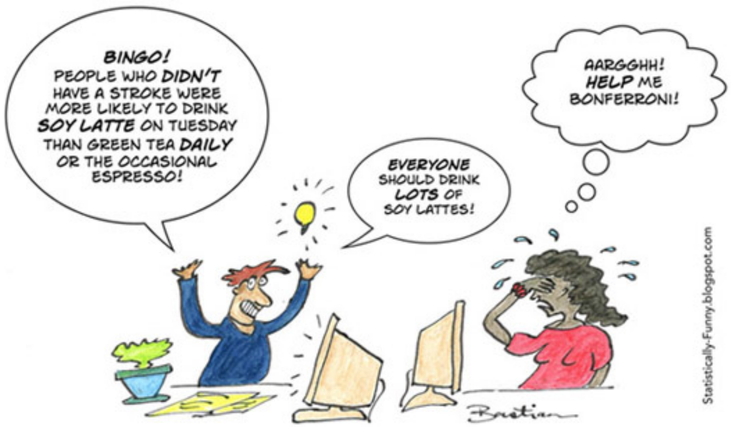
a) Calculate the sample mean, variance and standard deviation for Timothy sales. Do the same for Jimothy.

b) Calculate the sample correlation between Timothy and Jimothy's sales.

c) Combine Timothy and Jimothys sales into a single variable (14 observations total now). Find the 5 number summary of these data. Interpret the 3^{rd} quartile.

2. Guess the Correlation. Go to `guessthecorrelation.com`. Enter your full name as a username. Play the game until you get at least 50 points. Submit a screenshot showing your high-score and user name for credit. The 3 students with the highest score in the class will receive bonus points! (last years winner scored around 200)

3. Bonferonni Correction. The point of this comic is that, often, one can find a "significant" result if you search hard enough. If done incorrectly, this is sometimes known as *data dredging* (or p-hacking, data snooping etc.). Another entertaining comic can be found at `xkcd.com/882`.



Assume we have n hypotheses, and for each hypothesis, we can have an independent test A_i, $i = 1, 2, \cdots n$. If a hypothesis if wrong, the test A_i can give a false positive with probability α (usually α is small, like 0.05 or 0.01.). For this problem, lets assume that ALL of the hypotheses are wrong, therefore $P(A_i) = \alpha$ for $i = 1, 2, \cdots n$.

a) What is the probability that at least 1 test gives a (false) positive result? More importantly, what happens as the number of hypotheses gets large?
Hint: $P(\textit{At least one A occurs}) = P(A_1 \cup A_2 \cup \cdots \cup A_n)$

b) Bonferonni's Inequality states that

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) \leq P(E_1) + P(E_2) + \cdots P(E_n)$$

Prove this inequality for two events, i.e. $n = 2$. *For challenge points, prove the general version of the inequality using induction.*

c) Now, instead of testing each hypothesis with test A_i, let's use a new test $\tilde{A}_i$ which has false positive rate of α/n. *Comment: In reducing the false positive rate, we almost certainly will increase the false negative rate. No such thing as free lunch.* Use Bonferonni's inequality to show that the probability of getting any false positives is now less than α.

4. Data Analysis. The CDI dataset contains demographic information on the 440 largest counties in the US. The dataset can be read into R (from the course webpage) with the following command.

```
read.csv('http://math.unm.edu/ knrumsey/classes/spring17/MiniProjects/data.csv')
```

Extract the column named "PerCapitaIncome" and divide it by 1000 so the units are \$1000. Your answer to this problem should be in the form of a typed report.

a) Create a Histogram of the data. Comment on the skew of the data. Does the data look normally distributed? (*Note: I recommend using the argument* `breaks=12` *to get a better looking histogram*

b) Create a Boxplot of the data.

c) Report the mean, variance, and standard deviation. Also give the five number summary of the data and the IQR.

d) Give the mode of the data. Do this again after rounding the data to 1 decimal place.

e) Use a Kernel Density Estimate (KDE) to estimate the mode of the data. Which of the three estimates of the Mode seems to fit the data the best?

f) Create a QQ-plot of the data. What does this say about the Normality of the data?

g) Use the $1.5 \times IQR$ rule to identify potential outliers.

Challenge Problems

h) Create a Box-Cox plot for the data. Choose a transformation of the data based on the Box-Cox plot. Make histograms and QQ-plots to determine the Normality of the transformed data.

i) Now that the transformed data is (theoretically) more Normal, use the z-score method on the transformed data to identify potential outliers.

j) Repeat the process of outlier identification using a Bonferonni correction. Do the results differ?