**Excel Project #2: Descriptive Statistics**

**Due Date:** 10/18/18

Introduction: In this project, you will be using Excel to analyze a dataset with descriptive statistics, estimate quantities of interest and identify possible outliers. You will be provided with a dataset as an Excel spreadsheet. Go to *D2L > Contents > Excel Projects > Project #2* to find the folder containing the datasets. You will find a file called '*EP2\_yourname.xls*'. Replace *yourname* with your own name and save the workbook. The workbook contain a spreadsheets that can be accessed through the tabs on the bottom of the workbook. Follow the instructions for each part and then upload the finished product to D2L before the due date. If you have any questions, email me at: [grant.moore@frontrange.edu](mailto:grant.moore@frontrange.edu) .

**Part 1 : Estimation**

Instructions: Follow the instructions for the first three problems to estimate all quantities of interest regarding the given data set. Label your any cells containing answers. Format, label and title all graphs appropriately. You will be provided with data concerning the number of deaths caused by Manatee attacks per year. The data set is provided as a time series, but you may ignore the year for our purposes. We want to treat this as a *univariate* dataset.

1. **Histograms.** Create the following summaries of the data

a. A relative frequency distribution. Use 7 evenly spaced (*equal class width*) classes.

i. Create a histogram for the relative frequency distribution.

b. The cumulative relative frequency distribution.

ii. Create a histogram for the cumulative relative frequency.

c. Based on the results from the first two parts of this problem, estimate the following quantities (Save your estimates in a cell and label the cell):

i. The Median

ii. The 75th Percentile

2. **Statistics**. Calculate the following descriptive statistics for the data set:

a. The Mean

b. The Median

c. The Mode

d. The Minimum

e. The Maximum

f. The Range

g. The 25th Percentile

h. The 75th Percentile

i. The Interquartile Range

j. The Variance

k. The Standard Deviation

3. **Box Plot.** Use the appropriate results from #2 to create a box plot of this distribution. You will need to use a stacked bar chart. The range of the graph will be the *minimum,* the *25th percentile*, the *median*, the *75th percentile* and the *maximum.*

4. Based on your answers to #1 - #3, what is the shape of the distribution in this dataset? Skewed right? Skewed left? Normal? Uniform? Bimodal? Justify your answer numerically and graphically. Save and label your answer in a cell.

**Part 2: Statistical Theorems**

Instructions: Using our first statistical theorems, make some predictions about thedistribution. Follow the instructions for the next problems to apply the *Empirical Rule* and *Chebyshev's Theorem* to the Manatee dataset. Label all cells containing any answer.

5. **Empirical Rule.** Estimate the 97.5th percentile of the distribution using the empirical rule. You will need the mean and standard deviation from #2 in order to complete this problem. Label the cell containing your estimate.

a. Using the percentile function, compute the actual 97.5th percentile of this distribution. How close is your estimate to the actual value? What might have caused any discrepancies? What does the empirical rule assume about the shape of the distribution? Write your answer in a cell and label the cell.

**EXTRA CREDIT FOR EXAM 2: 2% Points on Exam!**

6. **Chebyshev's Theorem**. Use Chebyshev's Theorem to identify the 94th percentile of this distrbution. How close is your estimate to the actual value? What might have caused any discrepancies? What does Chebyshev's theorem assume about the shape of the distribution? Writer your answer in a cell and label the cell.

7. Looking at the shape of the distribution (i.e. your relative frequency distribution from #1), which rule is more appropriate to apply to this distribution? Why? Writer your answer in a cell and label the cell.

**Part 3: Detecting Outliers**

Instructions: Conclude your statistical analysis by identifying possible outlying observations in the given dataset. Apply both the Empirical Rule and the Interquartile Range Rule in order to determine if an observation fits the conditions for an outliers.

8. **Empirical Rule Detection**.

a. Use the Empirical Rule to determine the boundaries for detecting outliers within 95% certainty. Save the upper and lower boundary in separate cell and label them.

i. Using the boundaries found in part a, highlight any possible outlying observations in dataset in red (or label clearly in some other way.)

b. Use the Empirical Rule to determine the boundaries for detecting outliers within 99% certainty. Save the upper and lower boundary in separate cells and label them.

i. Using the boundaries found in part b, highlight any possible outlying observations in dataset in yellow (or label clearly and distinctly from *part a* in some other way)

9. **Interquartile Range Detection**. Use the Interquartile Range Rule to determine the boundaries for detecting outliers. Save the upper and lower boundary in separate cells and label them. Using these boundaries, highlight any possible outlying observations in blue (or label clearly and distinctly from *#7a* and *#7b* some other way.)