**Conference 1: Introduction to Distributions**

Andrew Lockhart, [abl73@drexel.edu](mailto:abl73@drexel.edu)

**Notes:**

Include all calculations done, graphs made, and the answers to the questions in a word document.

Feel free to either use this word document or just number your questions in another document.

**Coding Nomenclature:** For an overview of formulas see this [link](https://support.microsoft.com/en-us/office/overview-of-formulas-in-excel-ecfdc708-9162-49e8-b993-c311f47ca173) (only look if really interested)

Starting An Equation: putting an equal sign ‘=’ in a cell will start a formula call for that cell. BONUS TIP: hitting the tab key will complete the function you are calling/have selected in the formula menu

Absolute References: Utilizing a ‘$’ in an equation keeps the cell call the same for example $A$1 will ensure all equations use the value from A1 instead of moving down the line to A2

**Useful Functions:**

=AVERAGE(): Returns the average or arithmetic mean of the selected data

=STDEV.P(): Calculates the standard deviation of the selected data

=MEDIAN(): Calculates the median of the selected data

**Introduction**

Data binning is a data pre-processing technique used to reduce the effects of minor observation errors. It can help you visualize your data prior to further statistical analysis. The bin sizes can be used to identify characteristics of your dataset. To bin data, you must consider the **range** of your data set. You want to try different bin sizes before choosing the one that best reflects your data. To make a histogram you need to use the **Analysis ToolPak add-in** in Excel (This is only for older excel versions like before 2016 old but the toolpak can be used to calculate other stats stuff, we just won’t be doing that here). *To install it, consult the following instructions, which can also be found in Help.*

**Instructions to Load the Analysis ToolPak (**this may vary depending on your version of Excel)

The Analysis ToolPak is a Microsoft Excel **add-in** program that is available when you install Microsoft Office or Excel. To use it in Excel, however, you need to install it first.

1. Click the File tab, and then click Options.
2. Click Add-Ins, and the in the Manage box, select Excel Add-Ins.
3. Click Go.
4. In the Add-Ins available box, select the Analysis ToolPak check box, and then click OK.
   1. Tip: If analysis ToolPak is not listed in the Add-Ins available box, click Browse to locate it.
5. If you get prompted that the Analysis ToolPak is not currently installed on your computer, click Yes to install it.
6. After you load the Analysis ToolPak, the Data Analysis command is available in the Analysis group on the Data tab.

**OR**

1. Click the Tools tab, and then click Excel Add-Ins.
2. In the Add-Ins available box, select the Analysis ToolPak check box, and then click OK.
   1. Tip: If analysis ToolPak is not listed in the Add-Ins available box, click Browse to locate it.
3. After you load the Analysis ToolPak, the Data Analysis command is available in the Analysis group on the Data tab.

*In the past, there have questions about how to create a histogram in excel. Please see the following link for assistance:* [*https://www.excel-easy.com/examples/histogram.html*](https://www.excel-easy.com/examples/histogram.html)

(For newer excel versions just follow the steps at the end)

**Question 1: Analyzing White Blood Cell Counts**

Instructions:

Excel:

1. White blood cell (WBC) counts were collected from 20 patients. Calculate the **mean, median, and standard deviation**.
2. Create a **histogram** and bin the data as you see fit. Play around with the bin sizes until you find one that helps you best visualize the data (this will not be apparent until you generate your histogram; you can play around with bin sizes from there). **Bin size matters!**

SPSS:

1. Compare your data to a normal curve.
2. Import the data into SPSS.
3. Graphs, Histogram, Insert data into variable slot, Check the box that says “Display normal curve”.

Questions:

1. Based on the histogram generated, what is the distribution of the data? Is it unimodal? Bimodal? What facet of the histogram indicates this?
2. Given your mean and median, is the data skewed? If so, in what direction?
3. How does the data compare to the normal curve? Does the data follow a normal distribution?

**Question 1, Part B: Central Limit Theorem**

Under the second excel sheet you will find averages of WBC from the same population as before except this time we have a lot more data.

1. Using SPSS repeat the above SPSS instructions to plot a histogram with a normal curve. Does the data appear normal now?
2. Now that we have all of the data in SPSS we can now investigate how close to a normal distribution each set of data is. Test both data sets with Shapiro-Wilk test and the Kolmogorov-Smirnov Test. (Both of these tests will explore the probability that the distributions are normally distributed. The null hypothesis for both is that the data is from a normal distribution)
3. How are the skewness and kurtosis values?

**Question 2: Analyzing Rat Cognition**

Instructions:

1. Rats were tested on a new cognitive task being piloted in the laboratory. The dataset includes memory scores for rats that were scored by two students. Student A scored animals with IDs A100-A116, and Student B scored rats with IDS A117-A133. You expect all animals to show a similar performance. Calculate the **mean, standard deviation,** and **media**n for the complete dataset, and generate a **histogram** of the data.

Questions:

1. Based on the histogram you generated, how would you describe the distribution of the data? Is it unimodal? Bimodal? What aspect of the histogram indicates this? (It may also be helpful to calculate the mean for each student and see what this tells you)
2. Does this distribution of the data indicate anything concerning about the scoring between the two students?
3. In SPSS plot the two histograms above each other and discuss what represents the power and specificity. (Just guestimate where a significance level of 0.05 is and go from there)
4. If you wanted to determine if there was a statistical difference between the scoring done by the two students, what would you do?