LAB 1 INSTRUCTIONS

DISPLAYING AND DESCRIBING DISTRIBUTIONS

These instructions may assist you in learning how to display and describe quantitative/numerical data in Excel. In particular, you will learn how to obtain histograms and boxplots. Moreover, you will study how to summarize numerical data with the **Descriptive Statistics** and **Insert Function** features.

1. Activating the Data Analysis Add-In

In order to use the **Data Analysis** feature in Excel, you have to load the **Analysis ToolPak** first. The **Analysis ToolPak** is a Microsoft Office Excel add-in program (a supplemental program that adds custom features to Excel) that is available when you install Microsoft Excel.

When accessing the **Data** tab of the "Tabs and Commands" ribbon above the spreadsheet, if the **Data Analysis** add-in does not appear in the top right, then follow the instructions below.

1. Click **File** in the top left and click on **Options** in the bottom left.

2. Click **Add-ins** on the left of the opened dialog box.

3. At the bottom of the dialog box, make sure **Excel Add-ins** is selected from the drop-down menu beside "Manage:".

4. Click on **Go…**

5. In the new dialog box, check the box beside **Analysis ToolPak** and then, click **OK**.

The **Data Analysis** feature should now be present at the top right of the **Data** tab as either the phrase "Data Analysis" followed by an icon or simply the icon itself.

2. Charts in Excel: The Basics

In Excel, charts/graphs are made from ranges of numbers in a spreadsheet. Basically, we need to specify (a) the cells that contain the values to be graphed, (b) the type of graph we want drawn, (c) the location where the graph is to be placed. Charts/graphs can be made through the **Data Analysis** feature or through the **Insert** tab.

A chart in Excel consists of several components. Some of these components are displayed by default, others can be added as needed. You can change the display of the chart components by moving them to other locations in the chart, resizing them, or by changing their format. You can also remove those chart components that are not needed to display your data.

There are several charts/graphs available in Excel, yet these instructions will go on to focus only on the ones needed for this course.

3. Histograms

Though it is possible to create a histogram through the **Insert** tab, this course prefers creating histograms from the **Data Analysis** feature. If the feature is not appearing in your version of Excel, please return to point 1 above. Creating a histogram may also depend on the Excel version and operating system you have, yet there is preferred histogram formatting for this course (title, names of axes, no gaps between bars), such that the following data will produce the corresponding histogram. For best results, first try recreating the following histogram in Excel before moving forward with the data in the lab assignments. Usually, you will also be provided particular bins to use for histogram creation, so carefully examine lab assignments for this information. When changing aspects of your histogram, histograms should initially be 17 cells high (each cell height = 13.20 pixels) and 7 cells wide (cell width = 8.11 pixels). You may reduce them (within reason) when pasting into your assignment.

The data given below represent the examination scores of 50 students.

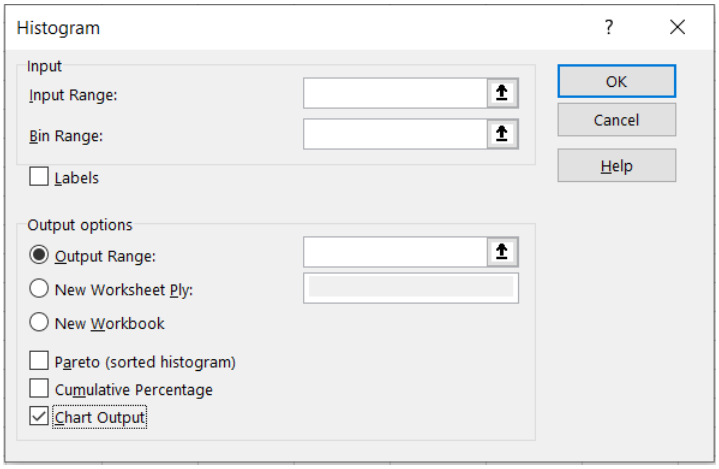
75	43	42	75	84	36	65	59	63	34
78	37	99	66	90	79	80	89	67	57
28	55	79	88	76	60	77	49	92	83
71	78	53	81	77	58	93	85	70	62
80	74	69	90	62	84	64	73	48	72

Enter the data into a single column called *Scores* in the spreadsheet. It may be possible to transpose the data to help create a single column such that this YouTube video may be useful. **Note:** The video provides both static and dynamic methods to tranpose data yet the "dynamic" method has issues, so the static method is preferred.

Create a second column called *Bins* so that the first value is 0, the second value is 10, the third value is 20, and the sequence continues until reaching a value of 100.

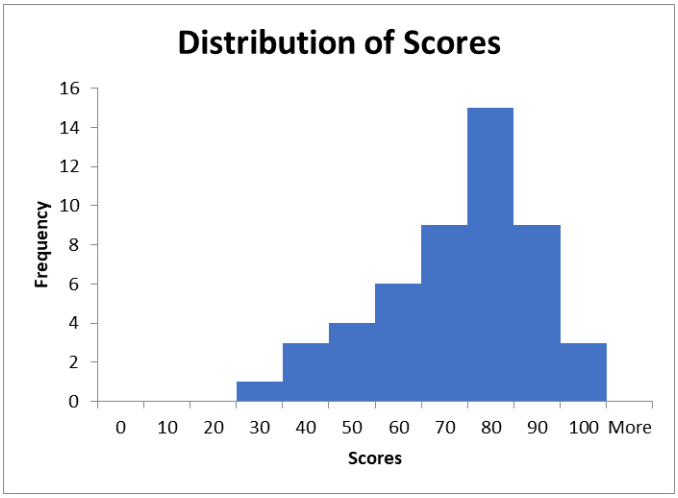
Using the **Histogram** feature within the **Data Analysis** feature and clicking on OK will present a dialog box exactly the same or similar to the following one. The **Input Range** should be the range presented by the cells containing *Scores* while the **Bin Range** should be the range presented by the cells contain *Bins*. The 'up' arrow to the right of each box can be clicked to permit the mouse to select the data directly from the spreadsheet. If including the cells that include the words *Scores* and *Bins*, check the **Labels** box.

For the **Output Options**, it is preferred to select *Output Range* and choose a single cell to present the upper-left corner of all the corresponding output. Also preferred is checking *Chart Output* so that Excel provides a frequency table as well as a histogram when providing output. Then click **OK**.



Bins	*Frequency*
0	0
10	0
20	0
30	1
40	3
50	4
60	6
70	9
80	15
90	9
100	3
More	0

The output on the left will be the result of checking the *Chart Output*, providing frequencies appearing within each bin. For example, 50 represents all observations in *Scores* that are above 40 and less than or equal to 50.



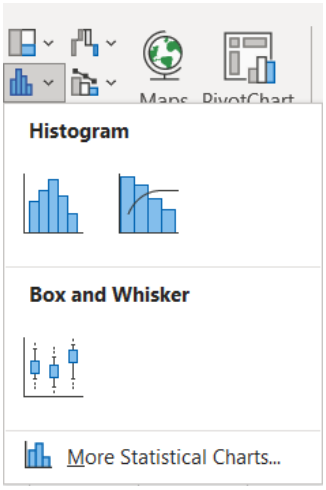
The histogram on the left is the histogram of *Scores*. Note the formatting in terms of the title, the names of the axes, no gaps between bars, and no legend.

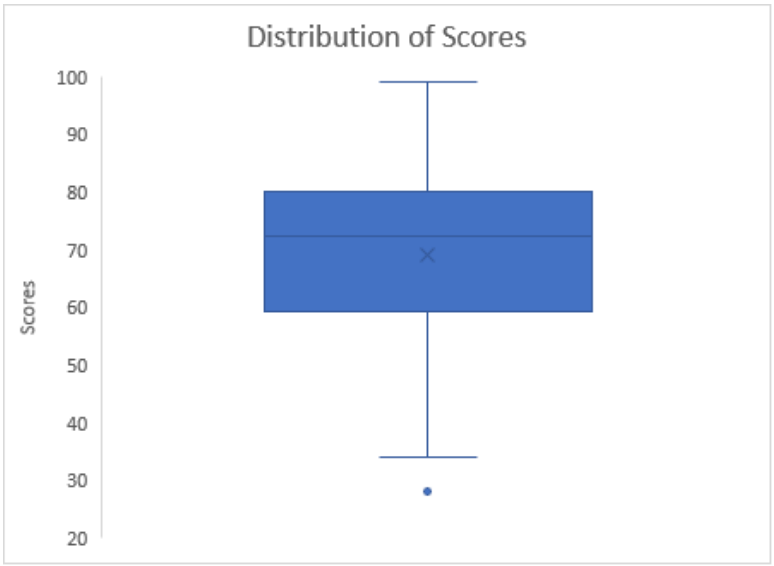
Note: Excel puts the x-axis values at the center of each bin, although they are, in fact, the values that separate the bins.

4. Boxplots

A relatively new feature to Excel is creating a boxplot from the **Insert** tab. The following will provide a brief example with preferred formatting, yet more can be determined from this Microsoft link that focuses on the Windows operating system. **Note:** This course *includes* the median when calculating quartiles, which is an option (*Inclusive median* vs. *Exclusive median*) you may need to change in the **Format Data Series** pane.

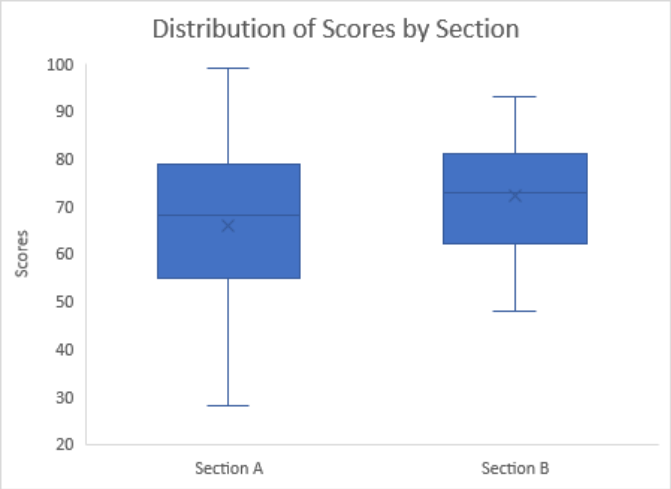
Using the *Scores* variable from above, highlight the entire column (including the label) and then access the **Insert** tab in Excel and clicking the drop-down menu for the histogram (diagram on left) also provides the "Box and Whisker" (or boxplot) option. This leads to a boxplot that can be reformatted to become the boxplot on the right. Note the formatting in terms of the title, the name of the y-axis, the range of the y-axis, the absence of name for the x-axis, and quartile calculation that uses *Inclusive median*. (**Note:** If *Exclusive median* was used, there would be no outlier.) The graph also permits comparison between the mean ('×') and the median (line within the box).





It is also possible to create side-by-side boxplots if a numerical variable (such as *Scores*) connects to a categorical variable (such as *Section*). Suppose the first 25 values of *Scores* came from Section A and the last 25 values came from Section B. Create a new column to the left of *Scores* called *Section* that indicate the first 25 values of *Scores* to be from Section A and the last 25 values of *Scores* to be from Section B.

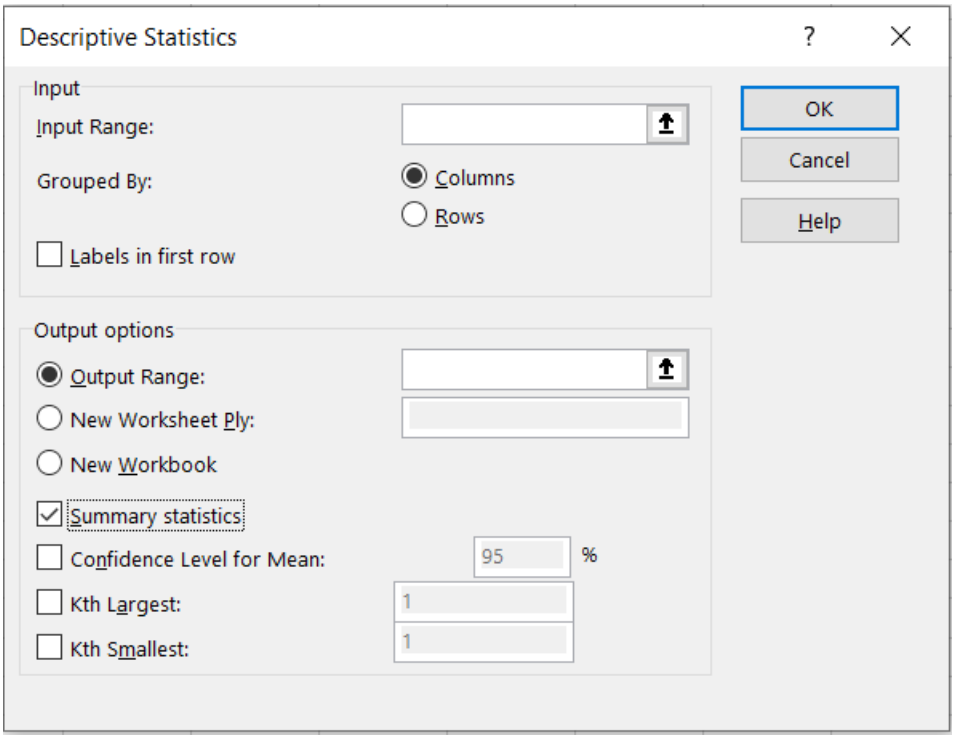
Highlight both the *Section* and *Scores* columns and return to the **Insert** tab to access the "Box and Whisker" option. This leads to a boxplot that can be reformatted to become the boxplot below. Note the formatting in terms of the title, the name of the y-axis, the range of the y-axis, the absence of name for the x-axis, and quartile calculation that uses *Inclusive median*.



5. Descriptive Statistics

The **Descriptive Statistics** feature provides measures of center, spread, and shape. Using the **Descriptive Statistics** feature within the **Data Analysis** feature and clicking on OK will present a dialog box exactly the same or similar to the following one. The **Input Range** should be the range presented by the cells containing *Scores.* The 'up' arrow to the right of the box can be clicked to permit the mouse to select the data directly from the spreadsheet. If including the cell that includes the word *Scores*, check the **Labels in first row** box.

For the **Output Options**, it is preferred to select *Output Range* and choose a single cell to present the upper-left corner of all the corresponding output. Checking *Summary Statistics* is required to produce the necessary output. The remaining checkboxes are not needed at this time yet may be discussed later in the course. Then click **OK**. To adjust the width of a column to fit the longest entry, double-click the column heading border between the column and the next column.



The **Descriptive Statistics** output (available below) contains three measures of center: the mean (68.92, the arithmetical average), the median (72.5, the middle-ranked value), and the mode (75, the most frequently occuring value).

The output table also contains several measures of spread: the range (71, the maximum minus the minimum), the sample standard deviation (16.93425, the typical deviation of a data value from the mean), the sample variance (286.769, the square of the standard deviation). The standard error of the mean (2.394865) equals the sample standard deviation divided by the square root of the sample size. The standard error is a measure of uncertainty about the mean, and it is used for statistical inference.

There are two measures of shape reported in the output: the skewness and kurtosis. The skewness coefficient is a measure of the lack of symmetry of the distribution. Kurtosis measures the heaviness of the tails of the data distribution. As these two measures of shape are not covered in the course material, we do not discuss them here in detail.

Scores	
Mean	68.92
Standard Error	2.394865
Median	72.5
Mode	75
Standard Deviation	16.93425
Sample Variance	286.769
Kurtosis	-0.21749
Skewness	-0.60693
Range	71
Minimum	28
Maximum	99
Sum	3446
Count	50

6. Insert Function

Most of the statistics provided in the **Descriptive Statistics** summary above have corresponding functions in Excel.

Statistic	**Corresponding Excel Function**
Mean	AVERAGE()
Median	MEDIAN()
Mode	MODE()
Standard Deviation	STDEV()
Sample Variance	VAR()
Kurtosis	KURT()
Skewness	SKEW()
Minimum	MIN()
Maximum	MAX()
Sum	SUM()
Count	COUNT()

The remaining **Descriptive Statistics** values could be found by creating functions using these values, such as finding the standard error by dividing the standard deviation by the square root of the count, using SQRT(). The range would be the difference between the maximum and minimum.

Note that STDEV() assumes the data is a sample while STDEVP() assumes the data is a population.

Two other important measures (the first quartile and the third quartile) are not included in Excel's **Descriptive Statistics** output. We will calculate the values using the **Insert Function** feature and add them to the output. The **Insert Function** feature can be accessed by clicking on $\mathbf{f_x}$, clicking on the formula bar, or simply writing an equation in any cell.

↑ ↑
f_x **formula bar**

Using the last option, type the function into any cell. For example, if the *Scores* data was in the cells A2:A51, then you could type "=QUARTILE.INC(A2:A51, 1)" to calcualte the first quartile (59.25). The value of '1' would change to '3' for the third quartile (80). The other values of 0, 2, and 4 would represent the minimum, median, and maximum, respectively, which would be the same values as seen in the **Descriptive Statistics** output.

Note: The QUARTILE.EXC() function calculates quartiles in a slightly different way, as seen with boxplot creation.

7. Inserting Excel Output into a Word Document

In your lab assignments, you will be required to answer some questions about various data sets using Excel. You can paste the results obtained with Excel into a word processing document. Excel has a feature making it possible to export the whole spreadsheet or its parts, charts and/or tables, to a word document.

Start up **Microsoft Word** and click on the **New** command in the **File** menu to create a word document that will contain the answers to the questions in your lab assignment. The assignments must be typewritten.

In order to paste the part of the Excel worksheet containing the results into the word document, copy the results in Excel into the clipboard and then paste into the word document using **Paste** in the **Home** menu (or Ctrl-C to copy the results to the clipboard and Ctrl-V to paste the results into the document).