

Introduction

In this lab you will use Excel to perform some elemental statistics. These statistical manipulations may be useful in future projects as well as in other Geography classes. I am convinced that knowing how to use Excel is a skill that all geography students should leave Texas A&M with. While learning how to use Excel may be difficult at first, it is well worth it in the end!

First you need to download the data onto your own computer. I have put a hyperlink on the website to a file called project04.xls. This is an Excel **spreadsheet** (or workbook in Microsoft lingo) containing two worksheets called **stats** and **state_population** that will be used for the first and second part of the project, respectively. You can access the two sheets by clicking on the tabs at the bottom.

There are two ways to access the file. The easiest is simply to click on the hyperlink called *Data*. On many computers this will open the file in Excel (or save it to the disk if tell it to). Another thing that you can do is to click on the hyperlink using the right mouse button - this will give up a lot of options, one being to **save Link as....** Once you have the file on your computer you are set to rock and roll.

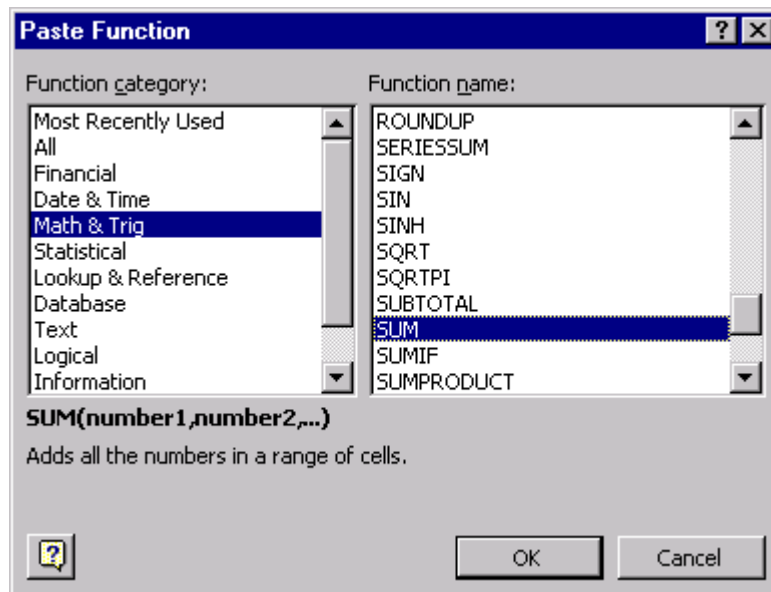
Part 1 - Standard Statistics

One common thing to do with any data set is to calculate a few standard statistical measures. In this lab, we will learn to do this using Excel. Everything you learn here is applicable to the map making projects in the future (plus many other classes you will take).

1. All calculations should be performed in Excel and results rounded off to 3 decimal places (use Format -> Cells -> Number to set the number of decimal places displayed)
2. Begin by selecting the **stats** sheet and copying the data in **Column A** into **Column B**.
3. Arrange the data in **Column B** into ascending order and set the format to 3 decimal places
4. Using the statistical functions in Excel calculate the following statistical properties *of the dataset*.

- **Mean** : _____
- **Median**: _____
- **Mode**: _____
- **Variance** _____
- **Standard Deviation**: _____

I will demonstrate how to sum up the data, and you can figure out how to do the other manipulations. To calculate the sum of the data go to the cell below the column of data you wish to sum (I went to cell B154). In Excel, equations start with an equal sign. So in cell B154 I would type `=sum(B2:B152)`. I could also simply highlight the block of cells I wanted to sum after starting my parentheses. Once my equation is finished, I hit return (enter) and the result appears. I get 10477.000 for my sum, by the way. Once you have successfully calculated a sum you should be able to calculate the remainder of the values. If you are stumped as to a function name, you can go to the menu bar and select **Insert -> Function**, which will provide a Wizard describing all available functions in Excel. An example of the Wizard illustrating the `sum` function is shown below



Part 2 - Histograms

It is often useful to be able to examine the distribution of your data and a common means of accomplishing this is a histogram. Microsoft Excel does an adequate job of creating histograms. It takes a bit of getting used to, but once you become comfortable it really helps. Before you begin you should read the portion of chapter 5 concerning histograms.

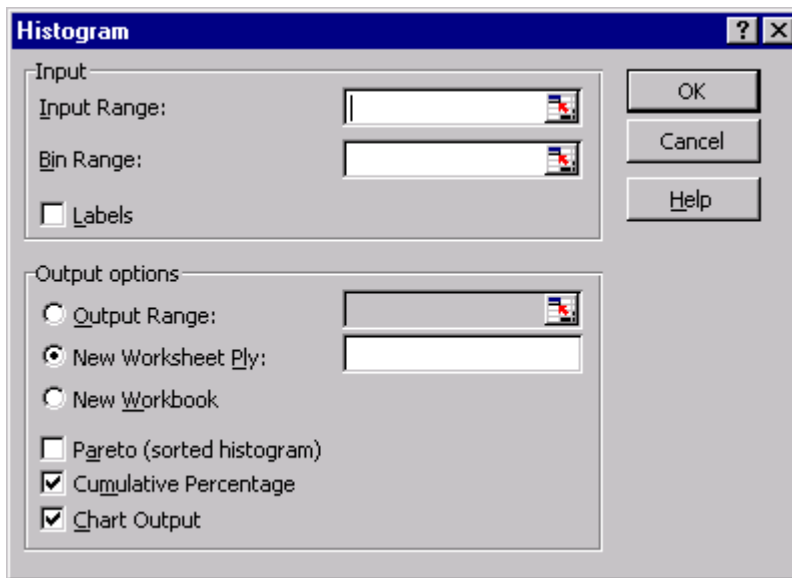
The easiest way to make histograms is to use the **Histogram Wizard**. This is located under the **Tools** menu in *Data Analysis*. If the *Data Analysis* option does not appear you need to go to Add-ins under the **Tools** Menu and select *Analysis Pack* - Once this is selected the *Data Analysis* option should appear.

The Histogram tool is fairly easy to use. It asks for several things. The first thing it wants is the **Input Range**. You simply need to highlight the block of cells you want to make a histogram of, or type the range into the box.

The Histogram tool asks for an optional **Bin Range**. The bin range is a set of cells that contain the endpoints of each histogram bin. For instance, if I wanted to make a

histogram with bins that went 0-9,10-19,20-29...90-99. I would create in column D for example a series of cells with the numbers, 0,9,19,29...99. The numbers must be in ascending order. If no bin range is specified Excel will create a set of evenly distributed bins between the minimum and maximum data values.

You can also specify what output you want. For the purposes of this lab, you should check the **chart output** and the **cumulative percentage** options. If the labels option is selected then your data columns (or rows) should have a label describing the data in the cell above (or to the left) of your first data point. An example of data suitable for use in a histogram, including labels, is show below.



Number of ISPs by county	Bins for Histogram
1	0
19	9
4	19
10	29
3	39
0	49
56	59
23	
7	
6	
12	
19	
34	
26	
8	

Once all the histogram options are to your liking, click on OK and Excel will create a histogram. You can then format the histogram by clicking on the various elements. I suggest you tinker with each element of the graph, just to get some idea of Excel's flexibility. Once you are comfortable with the histogram tool, do the following things.

1. Create a histogram and cumulative Frequency Histogram of the data as chart output.
2. Describe the data distribution based on the Histogram and the calculations made in part and answer the following questions.
 - a. Does the data fit a normal distribution
 - b. is it unimodal, bimodal?
 - c. Is the distribution skewed, if so in what direction?
3. Assume that this data represents some quantify concerning counties in a state (like

number of internet service providers). Your job as the state cartographer of Texas is to make a map showing the distribution of these values in the state. Your boss wants you to make a map breaking down these values into several categories (for instance show counties that have many providers, moderate number of providers, few providers, and hey, this county doesn't even have a telephone). To make this map you must decide several things:

- How many categories will you use?
- What will be the minimum and maximum values for each category
- How will you choose the categories?
- Equal size bins?
- Equal number of values in each bin?

Create a histogram representing your final selection and write up a justification for your choices. To do this you should create a set of meaningful bins

For the histogram portion of this exercise, you should turn in two histogram charts, one with the Excel default bins and one with the bins of your making as well as you description of the data distribution described in question 2.

Part 3 - Areal Statistics

As a cartographer, you will work with spatial statistical measures as well as standard ones. Unlike standard statistical measures, most software packages do not have built in spatial statistical functions which means it is up to you the user to undertake the calculations. Chapter 5 in Dent discusses spatial statistical measures. Some of these calculations (like the coefficient of areal correspondence) are best accomplished using a Geographic Information System. Others, however, can be done easily using Excel. Using Excel and the following equations, determine the areal mean and the standard distance for the following data points: (4,3), (2,5), (6,1), (1,3), (4,9), (10,8), (3,9), (7,2), (5,3), (8,8)

Areal Means

At its simplest, the areal mean is simply the average X and Y positions of a data set. More generally, areal means can reflect weighted values, that is to say one point may carry more weight than another in the calculations.

$$\bar{X} = \frac{\sum fx}{\sum f} \text{ and } \bar{Y} = \frac{\sum fy}{\sum f}$$

where f is the weight at each x,y location. For this exercise all points are the same weight (e.g. f=1), so in effect we are doing a simple averaging.

1. What is the **Areal Mean** of the data points?

X _____ Y _____

2. Compute the **Standard Distance** for the data points

Standard Distance _____

Standard Distance

Standard distance is basically the 2-D equivalent of the standard deviation. It is a measure of the dispersion of the values around the mean. To calculate a standard distance it is first necessary to determine the Euclidean distance from each point to the mean (d). This is accomplished using simple geometry. Once these distances are determined standard distance is calculated as follows. **Please note that equation in Dent is incorrect.**

$$SD = \sqrt{\frac{\sum d^2}{N}}$$

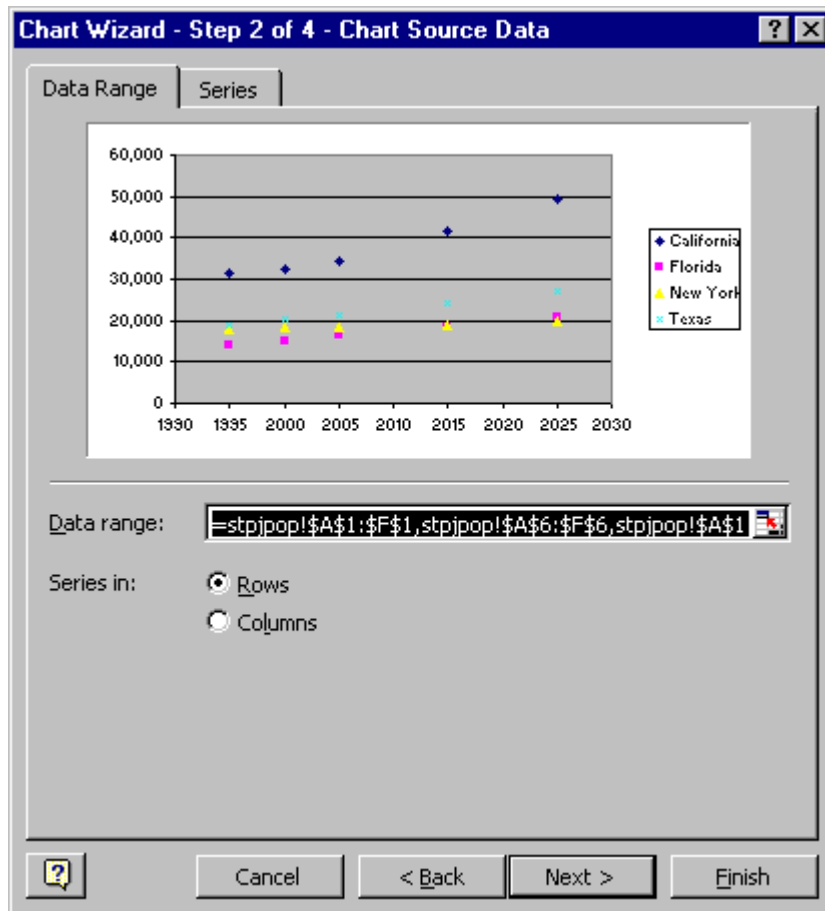
where N is the number of points.

Part 4 - Graphing

In this part of the lab, you will use Excel to examine the changes in projected population 3 states. The information you will need to use is located in *state_population* worksheet in the Excel spreadsheet that you can download from the web site.

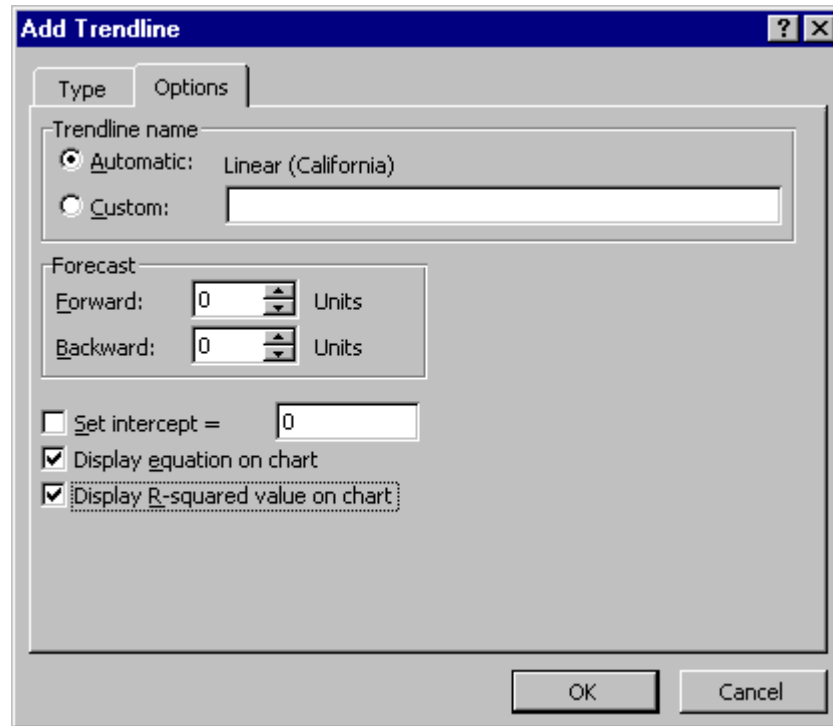
You should consider at the following states - California, Florida, New York and Texas.

- 1) You need to create a simple **XY scatter plot** of the data that shows the population of each state plotted by year using the Chart Wizard as *a series of symbols without connecting lines*. I don't want to lead you step by step through the process, but here are a few hints to help you along. The first is that you can hold down the control key (Ctrl) and highlight the various rows in the spreadsheet you want to plot up. This will allow you to put all the states on the same graph. At step 2 using the chart wizard you may need to tell Excel that your data is in rows as opposed to columns. At this point you should have produced a chart that looks something like this....

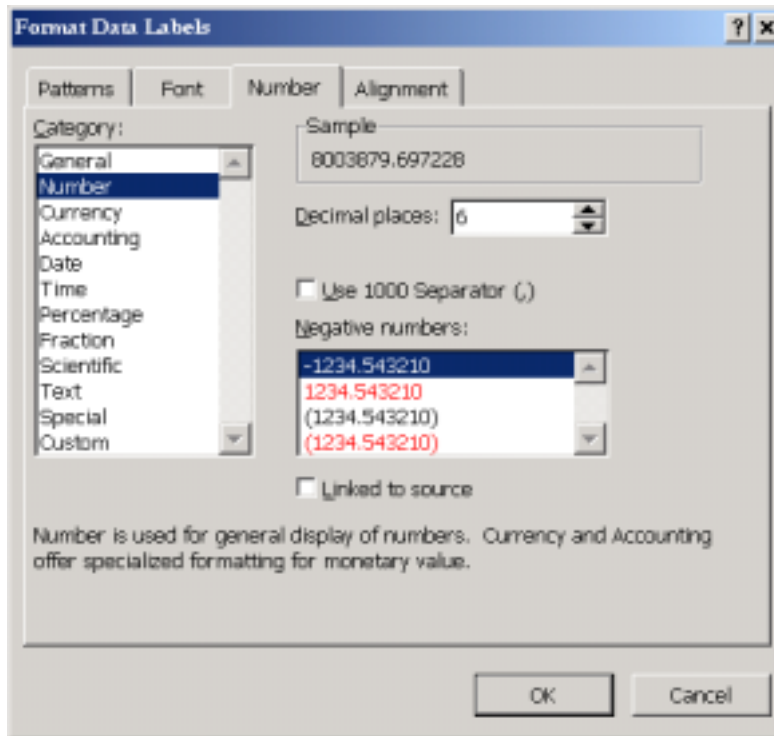


- 2) You should add appropriate titles and label the axes correctly and when given the option save it as a **new sheet**.

- 3) Once you have created your chart, I want you to add a trendline (a regression line) to each state. This is easily done by clicking on each state's value, then going to the menu bar and from chart selecting **Add Trendline**. This will give you a window with two tabs (Type and Options). For all the following work go to the **Options** and make sure you **Display equation on chart** and **Display R-squared value on chart**.



Once you created the trendline and added the equation, you will have to format the number in the equation. This is the same process that you can do to change how the values in cells are changed, such as increasing the number of decimal places. You can do this by selecting the equation and then right clicking. This will bring up the following dialog box, which you can use to change the properties. You should change the formula of the equation to have between 6 and 8 decimal places. This is necessary to have the precision required to estimate the states population in 2050.



Now the Fun Begins....

- 1) Add a **linear** trendline to each state's population curve and print out a copy of the completed map (it doesn't have to be in color and you may wish to rearrange things so I can tell which trendline equation goes to which line)
- 2) Complete the following table for each of the states

State	Trendline Equation	R^2
California		
Florida		
New York		
Texas		

- 3) Add a **2nd order polynomial** trendline to each state's population curve and print out a copy of the completed map (it doesn't have to be in color and you may wish to rearrange things so I can tell which trendline equation goes to which line)
- 4) Complete the following table for each of the states

State	Trendline Equation	R²
California		
Florida		
New York		
Texas		

- 5) Based on the two Trendline Equations determined in (2) and (4), compute each state's estimated population for the year 2050

State	Estimated 2050 population using linear trendline	Estimated 2050 population using 2nd order
California		
Florida		
New York		
Texas		

- 6) Describe why the two estimated populations differ and which type of trendline you is best to estimate these future populations and why you say so.

If you have any problems do not hesitate to come see me