CHAPTER 1 - LAB SESSION
INTRODUCTION TO EXCEL

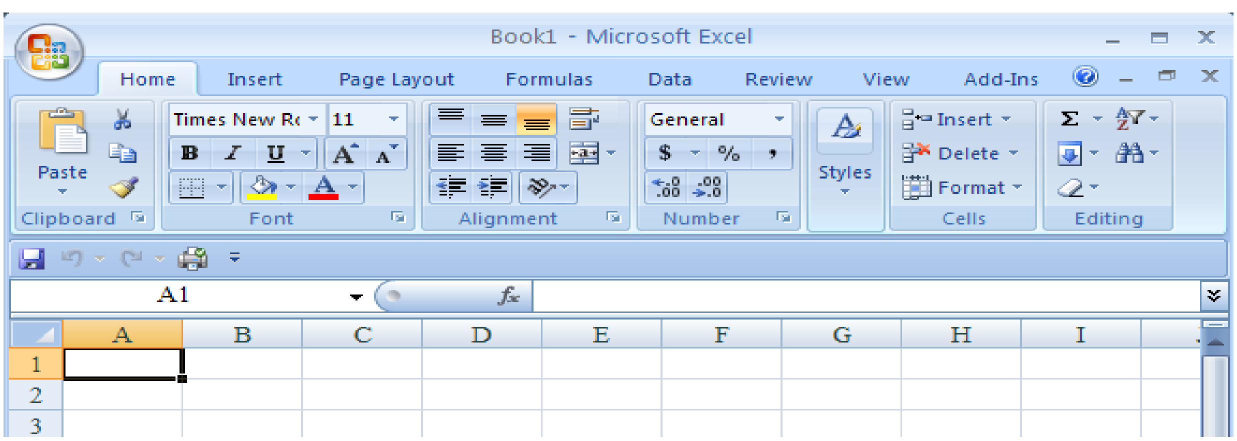
INTRODUCTION: This lab session is designed to introduce you to the statistical aspects of Microsoft Excel. During this session you will learn how to enter and exit Excel, how to enter data and commands, how to print information, and how to save your work for use in subsequent sessions. As with any new skill, using this software will require practice and patience. Excel is a spreadsheet used for organizing data in columns and rows. It is an integrated part of Microsoft Office, and so data can be easily imported and exported into word processing documents, databases, graphics programs, etc. It offers a wide range of statistical functions and graphs and so is an alternative to specific statistical software.

BEGINNING AND ENDING AN EXCEL SESSION

To start Excel: Click on the Start button and choose Programs/Excel. If you have the Office shortcut bar installed, simply click on the Excel icon.

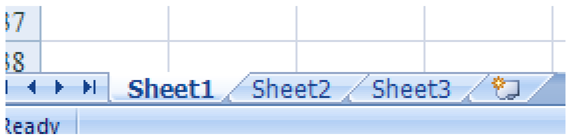
To exit Excel:
To end a Excel session and exit the program, choose **File** from the menu bar and then choose **Exit**. A dialog box will appear, asking if you want to save the changes made to this worksheet. Click **Yes** or **No**. You can also exit Excel by clicking the X in the upper right corner of the window.

THE EXCEL WINDOWS
When you enter Excel, there are actually two windows open. The outer window is the Excel application window, which contains all of the buttons and menus that control the functionality of the program. The inner window contains the workbook with all of its sheets and controls. The tabs across the top take the place of the drop down menus of the previous version of Excel.

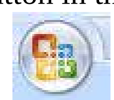


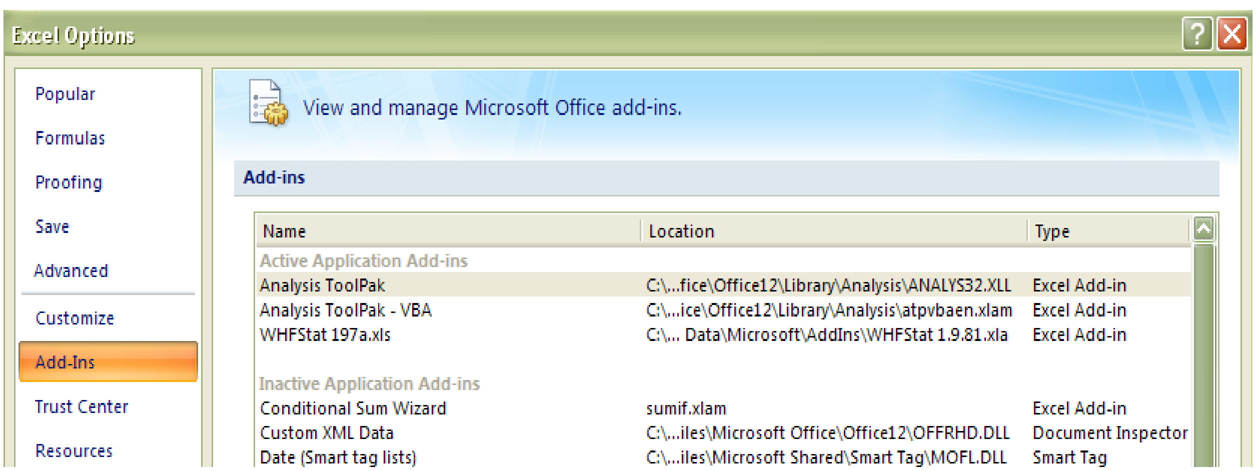
The Document (sheet) Window:
When you first start Excel you will be in a window titled “Microsoft Excel - Book 1” Excel organizes itself in workbooks, each of which is made up of worksheets that are 65,536 rows by 256 columns. You can enter and edit data on several worksheets simultaneously and perform

calculations based on data from multiple worksheets. When you create a chart, you can place the chart on the worksheet with its related data or on a separate chart sheet. Each of the cells within the sheet is identified by the intersection of its row and column, for example A2, or B7. Note the three tabs and the bottom of the screen, called “sheet1”,”sheet2”, and “sheet3”. The default is a workbook with three sheets, but the number of sheets in a workbook is limited only by available memory. To add a single worksheet, click Worksheet on the Insert menu. To delete sheets from a workbook, select the sheets you want to delete. Then on the Edit menu, click Delete Sheet. To Rename a sheet, double-click the sheet tab, and type a new name over the current name.

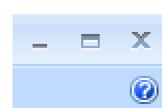


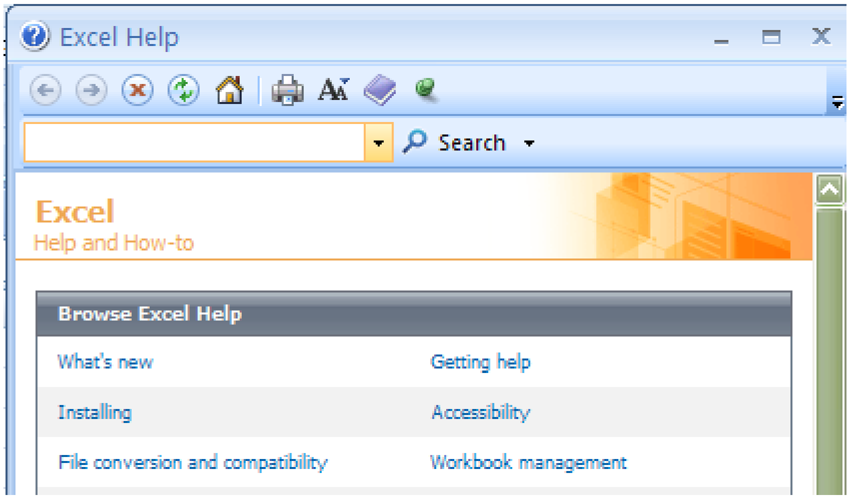
Analysis ToolPak : Microsoft Excel provides a set of data analysis tools — called the Analysis ToolPak — that you can use to save steps when you develop complex statistical or engineering analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables. If the **Data Analysis** command is not on the DATA tab, you need to install the Analysis ToolPak. To do this, go to the Office Button in the upper left corner and select Excel Options in the lower right corner and select **Add-ins**. When the dialog box appears, check **Analysis Toolpak** and **Analysis VBA**. Then click on **OK**





The Help Window in Excel
Information about Excel is stored in the program. If you forget how to use a command or need general information, you can ask Excel for help. From the Menu Bar choose **Help**. A drop down menu will appear that gives you several choices. You can select from a list of topics or enter a particular question in the search bar.





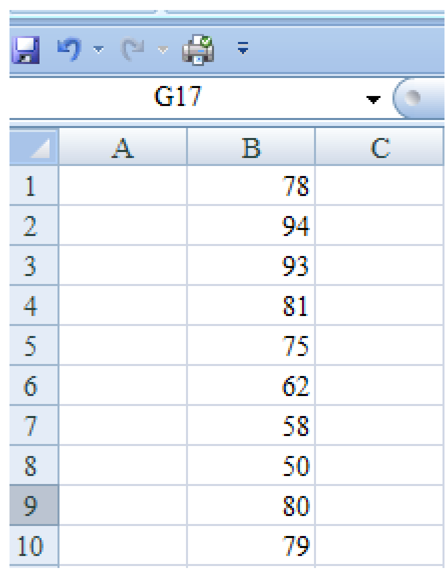
ENTERING DATA

When a workbook is first opened, the cell A1 is outlined in black. This indicates the active cell. Move your cursor around the sheet, clicking into different cells to activate them. Note that the address changes in the box above A1. The address (row and column) of the active cell always appears here.

Let's enter data in the second column:

78 94 93 81 75 62 58 50 80 79

To do this press the down arrow key or Enter key to move to the next entry position.



Let’s fill the first column with the numbers 1 through 10. We can do it the same way, or we can let Excel do it for us. Enter a 1 in cell A1. And a 2 in A2. Highlight these two cells and grab the lower right corner and drag it down until you have the numbers 1 through 10 in the column.

Column 1 should now contain the integers 1 through 10. While you are in the sheet window, fill columns 3 and 4 with a set of ten test scores each. You should now have four columns of data.

Changing a value entered

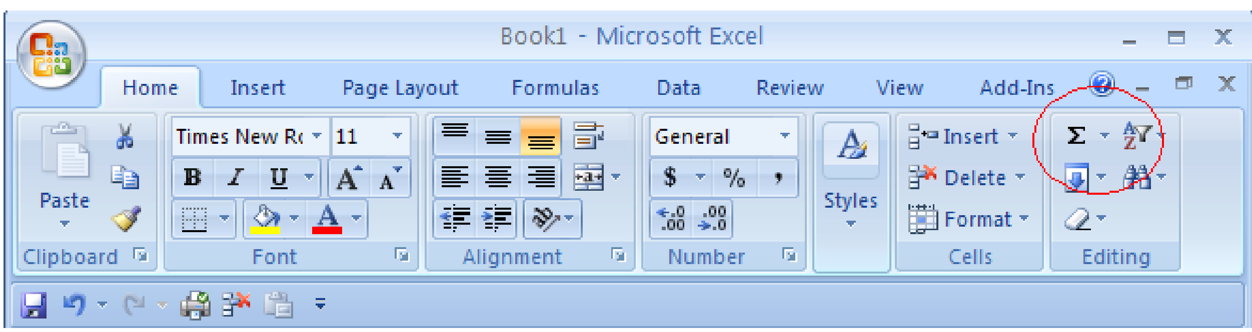
We can edit data directly in the cell or from the formula bar at the top of the sheet. If you have not hit the Enter key yet, you can simply back space and correct your mistake. If you have entered the data, click on the cell you wish to edit to make it active. You can either retype to overwrite the data, or click into the formula bar and edit the entry.

Suppose we had inadvertently left out a value and we wish to enter it in a particular position. Place the cursor in the cell in which you wish to insert the new value. Click the Insert Cells button on the toolbar. A dialog box will appear, asking which way you wish to move the cells. A blank cell is created and the missing value can be entered. Entire rows and columns can be added the same way. You can take a short cut to this by using **Control +.**

A cell can be deleted by making the cell active, then Choose: **Edit > Delete Cells** or by using **Control -**

Copying Data
To copy the contents of one cell to another, simply activate the cell, use **Control C.** Activate the cell that you want to paste the value into and **Control V.** This can also be done for a range of cells. Activate the upper left cell of the range. Press shift and click the lower right corner of the range. This should highlight the entire range. You can then copy and paste as above. You can also put the copy and paste controls in the quick access bar and use them.

Cell References:
Previously, you entered four columns of data. Click on cell B11. On the Home tab you will see a summation sign ˜. Click on it and the ten values above it will be enclosed in a box. Press enter and the sum of the ten values will be in cell B11 .



Now activate cell B11, press **Control C**, highlight cells C11 and D11, and press **Control V.** This should give you the sums of columns C and D. Note what happened in the formula when you copied it. The references were changed to reflect the new column. This is called a relative reference.

A	B	C	D
1	78	82	88
2	94	74	89
3	93	99	81
4	81	100	77
5	75	76	72
6	62	54	56
7	58	53	65
8	50	82	92
9	80	87	67
10	79	65	80
	=SUM(B1:B10)		
	SUM(**number1**, [number2], ...)		

A	B	C	D
1	78	82	88
2	94	74	89
3	93	99	81
4	81	100	77
5	75	76	72
6	62	54	56
7	58	53	65
8	50	82	92
9	80	87	67
10	79	65	80
	750	772	767

If you need to preserve the value of a certain cell when copying a formula, you will have to use absolute referencing. This is accomplished by placing $ within the address. (A6 would keep the value in cell A6 wherever it was copied to within the worksheet.)

SAVING YOUR WORK

An Excel workbook contains all your work; the data, graphs, and all the sheets within the workbook. When you save a project, you save all of your work at once. When you open a project, you can pick up right where you left off. The contents of each sheet can be saved and printed separately from the project, in a variety of formats. You can also delete a worksheet or graph, which removes the item from the project.

RETRIEVING A FILE

You can open a wide variety of files with Excel. Choose **File Open** to select the appropriate one. There is an **Import Wizard** that will guide you through the process.

A CD ROM accompanies Johnson/Kuby's Elementary Statistics, 11/e This disk has data in Excel format for many of the problems in the text. Follow the instructions that accompany the disk for use on your computer.

PRINTING:

You have many options when it comes to printing from Excel. Go to the standard toolbar and choose the **File** drop down menu. The Set Print Area choice allows you to select the range of cells you wish to print.

The **Page SetUp** dialog box has four tabs that will help you customize your output. You can also access this dialog box through **Print Preview**. This is a good choice because it allows you to play with your selections to get the best layout for your output before you commit it to paper.

ASSIGNMENT:

1. Create a data file on your disk that consists of the heights of 15 of your classmates (in column 1) and their weights (in column 2).

2. Retrieve the data file created in #1 above, and produce a paper copy (commonly called 'hard-copy') to hand in.

3. Retrieve the data file for Exercise 2.23 from the Student Suite CD that accompanies the Johnson/Kuby text, and print a hard copy to hand in.

CHAPTER 2 - LAB SESSION 1
GRAPHIC PRESENTATION OF UNIVARIATE DATA

INTRODUCTION: Graphically representing data is one of the most helpful ways to become acquainted with the sample data. In this lab you will use Excel to present data graphically. You will be analyzing data using four types of graphs: Circle graphs, Bar graphs, histograms, and cumulative (relative) frequency plots (ogives).

GRAPHIC PRESENTATIONS OF DATA

There are several ways to display a picture of the data. These graphical displays help us get acquainted with the data and to begin to get a feel for how the data is distributed and arranged. In attempting to get a pictorial representation of data, we must decide what type of graphic display would best present the data and their distribution. The type of display used depends, in large part, on the type of data and the idea to be presented.

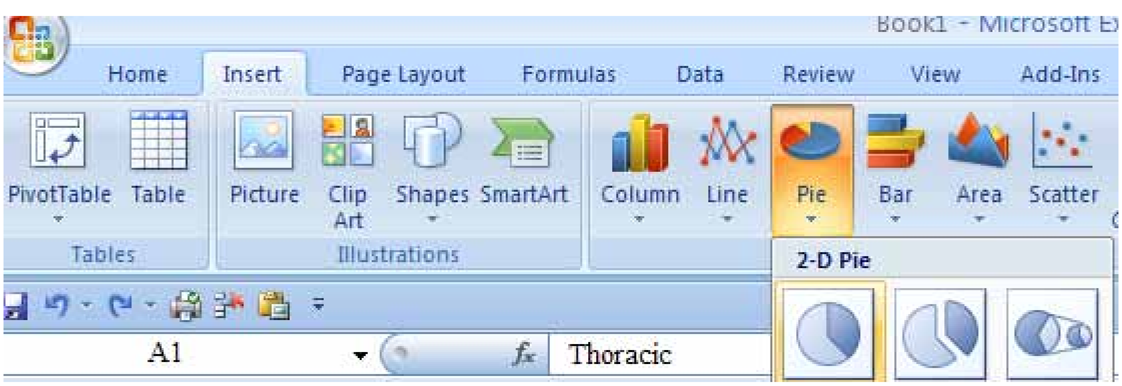
GRAPHIC DISPLAYS FOR QUALITATIVE (CATEGORICAL) DATA

CIRCLE GRAPHS

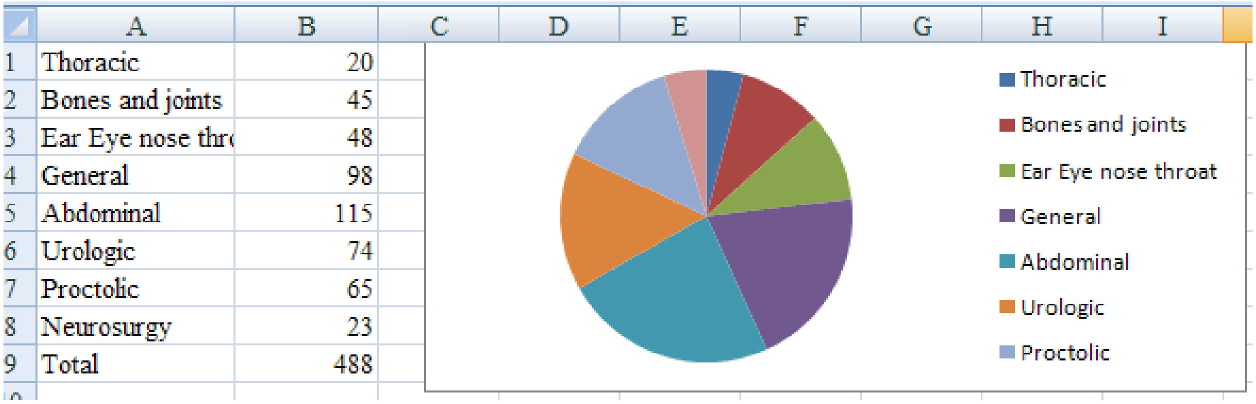
A circle graph shows the amount of data that belongs to each category as a proportional part of a circle. Consider Example 2.1. We are instructed to construct a circle graph, with data presented as a frequency distribution. Enter the data (either by hand, or opening the data file.)

	A	B
1	Thoracic	20
2	Bones and joints	45
3	Ear Eye nose thr	48
4	General	98
5	Abdominal	115
6	Urologic	74
7	Proctolic	65
8	Neurosurgy	23
9	Total	488

Highlight the data. From the ribbon, select the **Insert tab > Pie > 2D pie**

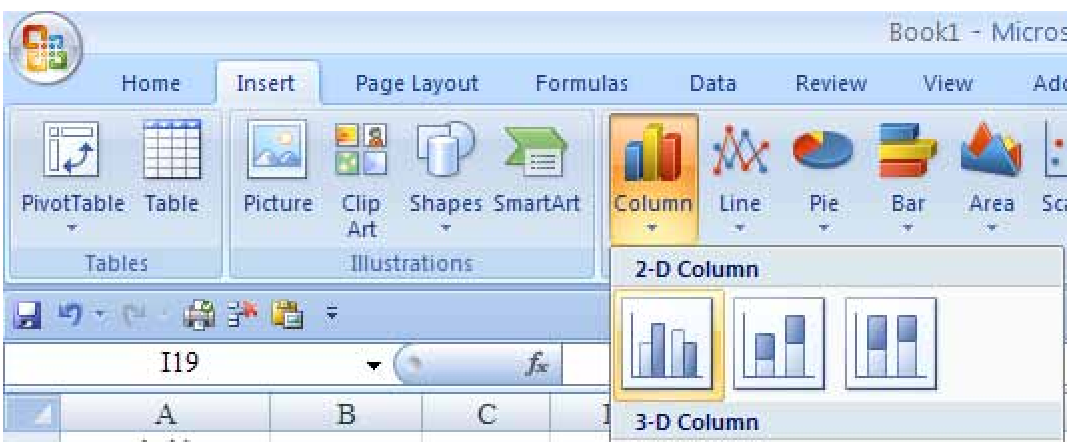


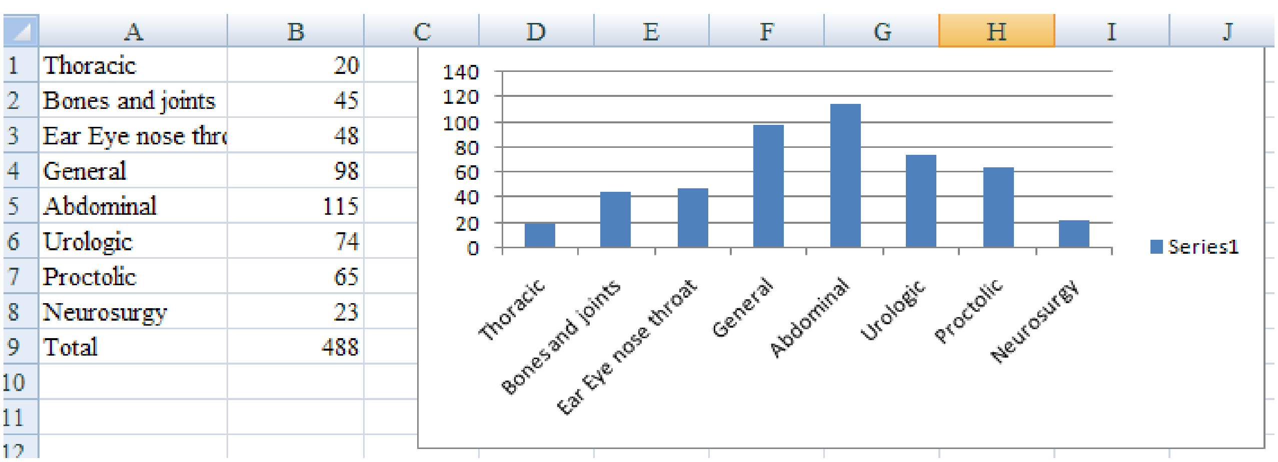
The chart will appear in the worksheet. By right clicking on the chart, you may format the chart as you wish, adding titles, changing colors, etc. Be careful not to include the total line.



BAR GRAPHS

A bar graph shows the amount of data that belongs to each category as proportionally sized rectangular areas. Let's continue to use the data from Exercise 2.1, and present this data as a bar graph. Since we already have the data entered we can go right to the commands to create the bar graph:

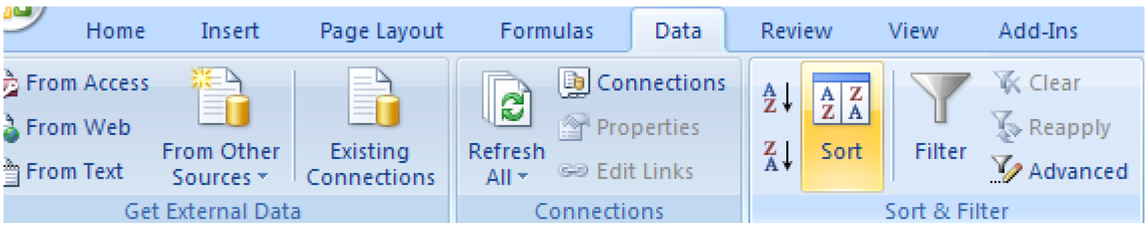




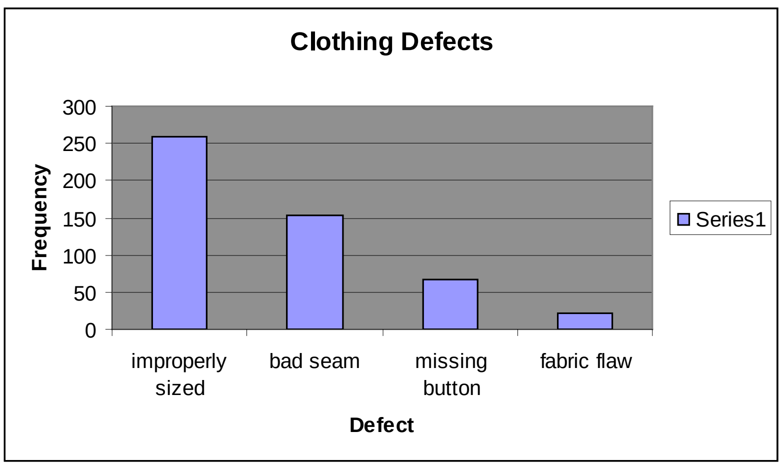
PARETO DIAGRAMS – A Special Type of bar graph

Consider Exercise 2.11. We are instructed to construct a Pareto diagram in this instance since this a quality control application. In constructing a Pareto diagram for Exercise 2.11, basically we are doing a bar graph, but sorting the data first. After you have input the categories into column A and the corresponding frequencies into column B, then continue by selecting the data, and choosing the DATA tab from the ribbon. Then select SORT.

	A	B
1	Missing Button	67
2	Bad seam	153
3	Improper size	258
4	Fabric flaw	22



Then continue with the commands necessary to create the bar graph.



Note: Excel does not include the line graph

GRAPHICAL DISPLAYS FOR QUANTITATIVE DATA

DOTPLOTS

Dotplots are a quick and efficient way to get a preliminary understanding of the distribution of your data. The dotplot display is not available, but the initial step of ranking the data can be done. Input the data into a column,

Choose: **Data** > **Sort**
Enter: Sort by: **Column A** (or whatever column the data is in
Select: **Ascending** > My list has: **Header row** or **No Header row**

Use the sorted data to finish constructing the dotplot.

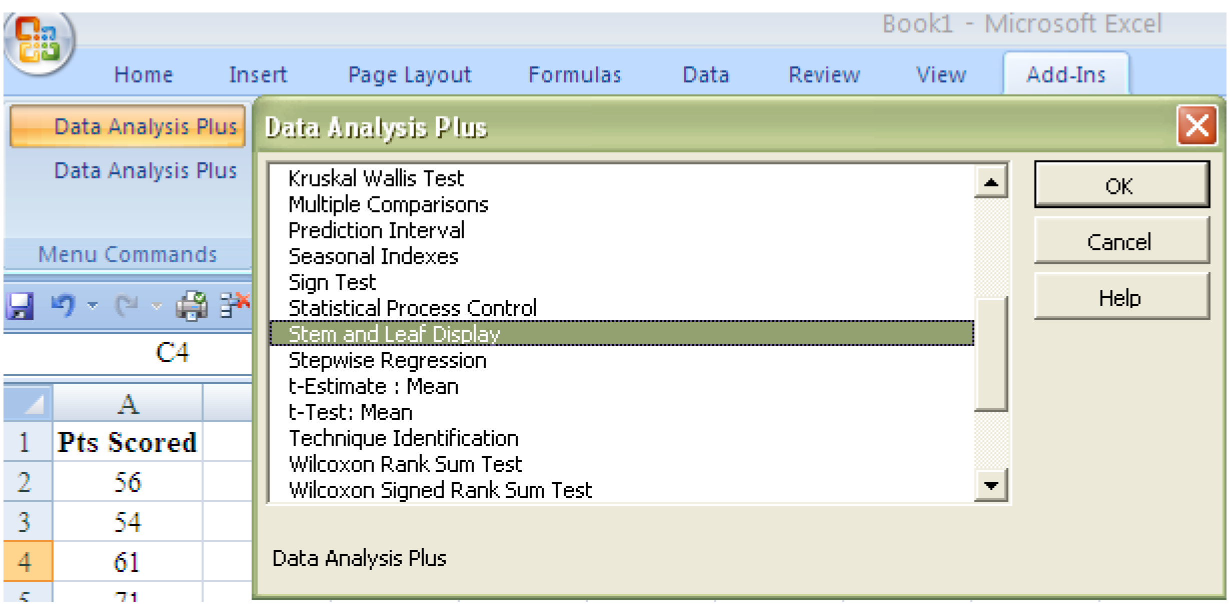
STEM AND LEAF DISPLAY

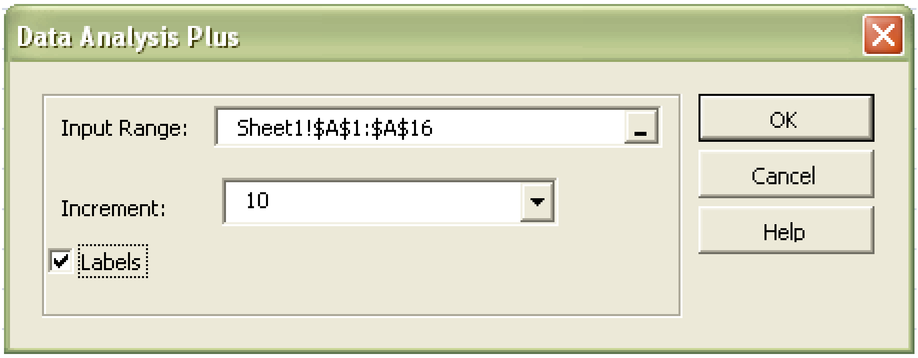
The stem-and-leaf diagram is not available with a standard version of Excel. However, Data Analysis Plus (a collection of statistical macros for Excel) can be downloaded onto your computer from your Student Suite CD.

To illustrate the commands necessary to construct a stem-and-leaf display, let's use the data from Exercise 2.19 (points scored). Enter the data into column A with a heading in cell A1, then continue with:

	A
1	Pts Scored
2	56
3	54
4	61
5	71
6	46
7	61
8	55
9	68
10	60
11	66
12	54
13	61
14	52
15	36
16	64
17	51

Choose: **Add-ins** > **Data Analysis Plus** > **Stem and Leaf Display** > **OK**
Enter: Input Range : **(A2:A17 or select cells)**
Increment: **(the stem increment you wish to use)**





Stem & Leaf Display

Stems	Leaves
	3->6
	4->6
	5->124456
	6->0111468
	7->1

Notice, originally the macro chose an increment of 10. If you click the down arrow for the increment box, note the different increment options. None of the other increments make sense for this particular data set.

HISTOGRAMS

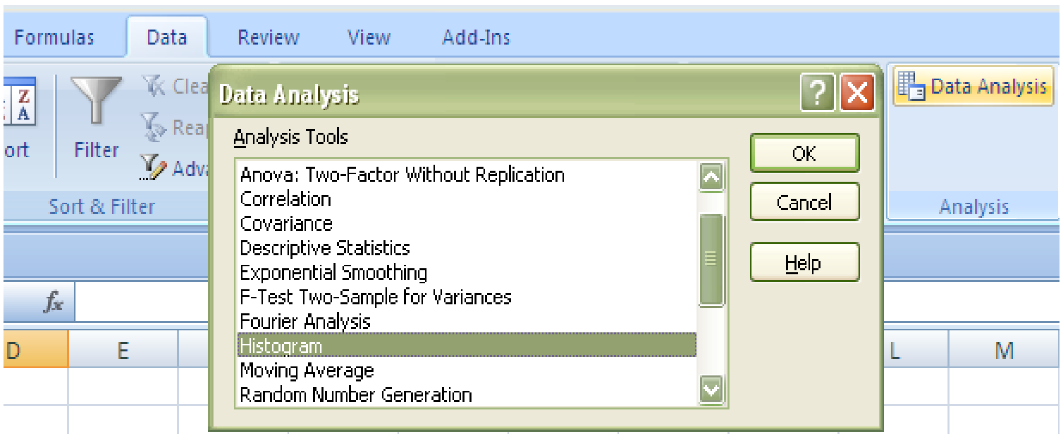
Histograms are more useful for large sets of data. We expect the histogram of a sample to be similar to that of the population. To illustrate the many options under the **HISTOGRAM** command, let's use the data in Exercise 2.39 (on the Student Suite CD). The **HISTOGRAM** command separates the data into intervals on the x-axis and draws a bar for each interval whose height, by default, is the number of observations (or frequency) in the interval.

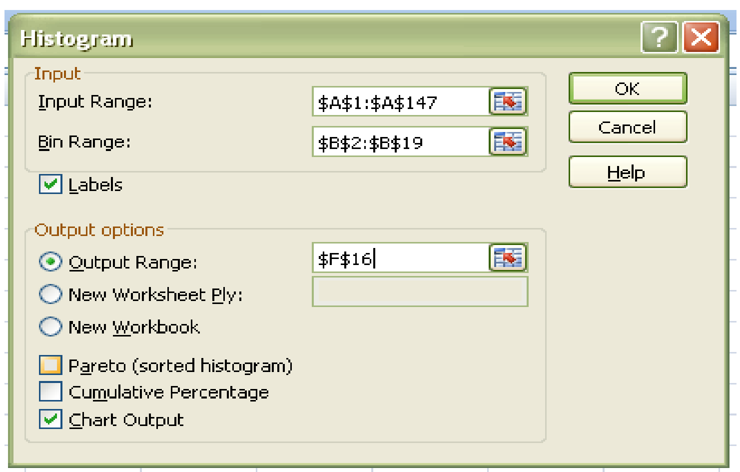
Input the data into column A (or retrieve data worksheet from the Student Suite CD). Input the upper class limits into column B (this is optional, but recommended).

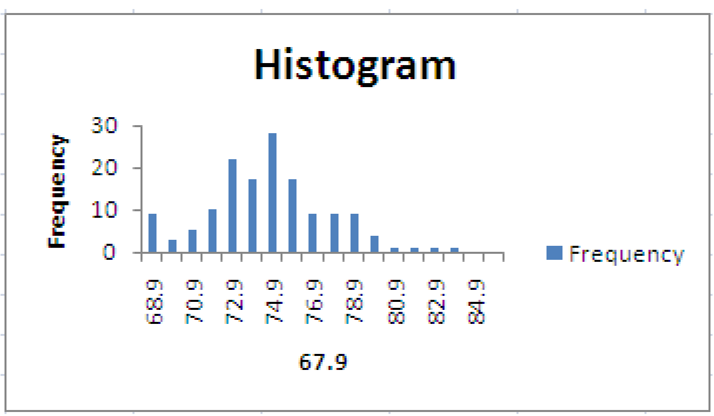
	A	B
1	GolfScor	
2	69	67.9
3	73	68.9
4	72	69.9
5	74	70.9
6	77	71.9
7	80	72.9
8	75	73.9
9	74	74.9
10	72	75.9
11	83	76.9

Note: Data continues down the column.

Choose: **Data > Data Analysis** > Histogram > OK**
Enter: Input Range: **(A2:A147 or select cells)**
Bin Range: **(B2:B19 or select cells)**
Select: **Output Range**
Enter: area for freq. distribution, & graph: **(C1 or select cell)**
Select: **Chart Output**





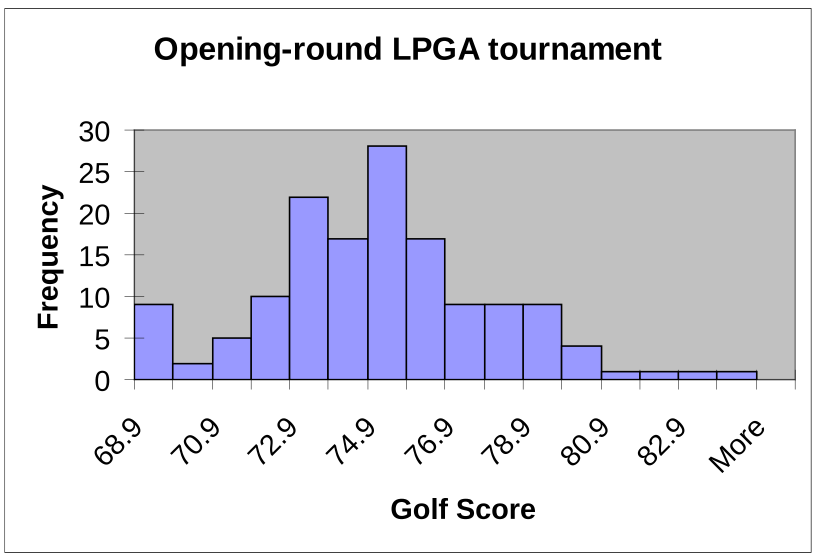


**If Data Analysis does not show on the Tools menu:
Choose: Tools > Add-Ins
Select: Analysis ToolPak
Analysis ToolPak-VBA

To remove gaps between bars
Click on: **Any bar on graph**
Click on: **Right mouse button**
Choose: **Format Data Series > Options**
Enter: Gap width: **0**

To edit histogram:
Click on: Anywhere clear on the chart
-use handles to size
Any title or axis name to change

67.9	*Frequency*
68.9	9
69.9	2
70.9	5
71.9	10
72.9	22
73.9	17
74.9	28
75.9	17
76.9	9
77.9	9
78.9	9
79.9	4
80.9	1
81.9	1
82.9	1
83.9	1
More	0



Note that the upper class limits appear in the center of the bars. Replace with class midpoints.

Also note that if the data is already tabled, the commands are different. See your text.

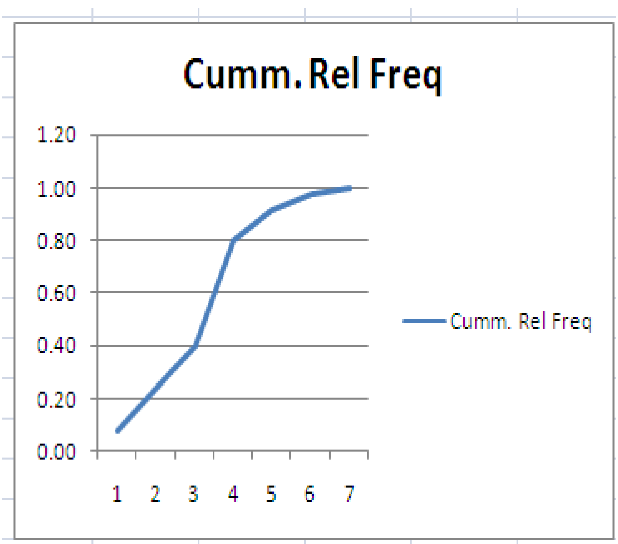
OGIVES
To construct an ogive, the class boundaries must be in listed in column A and the cumulative percentages listed in column B. Let's use Exercise 2.55 in your text as an example. We are presented with a grouped frequency distribution.

Now you need this same information presented as a cumulative relative frequency distribution:

	A	B	C	D
1	KSW Test Score	Frequency	Cumm. Freq.	Cumm. Rel Freq
2	0 - 4	4	4	0.08
3	4 - 8	8	12	0.24
4	8 - 12	8	20	0.40
5	12 - 16	20	40	0.80
6	16 - 20	6	46	0.92
7	20 - 24	3	49	0.98
8	24 - 28	1	50	1.00

Highlight the cumulative relative frequency column.
Choose: **Insert > Line > 4th picture (usually)**

You may then highlight the chart and format it any way you choose by clicking on the Design or Layout tabs.



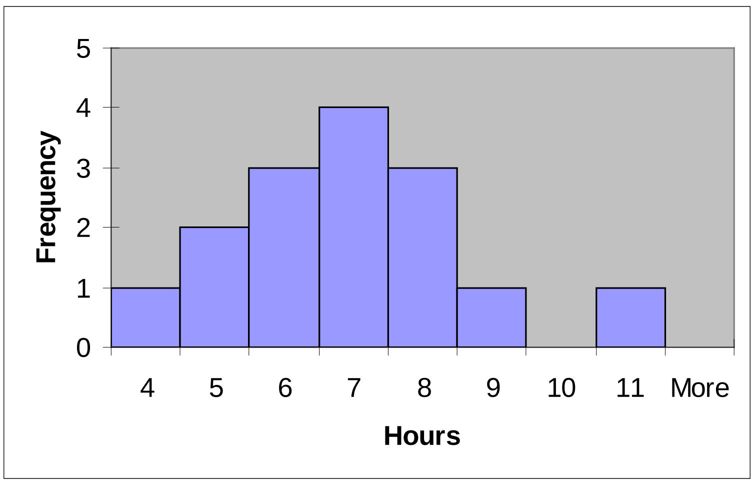
ASSIGNMENT: Do Exercises 2.7, 2.19, 2.43, 2.48, and 2.54in your text.

CHAPTER 2 - LAB SESSION 2
NUMERICAL PRESENTATION OF UNIVARIATE DATA

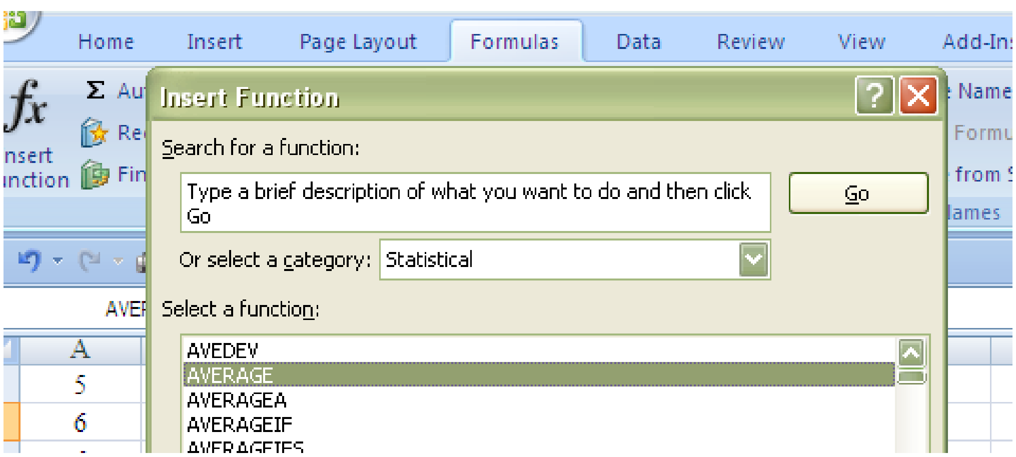
INTRODUCTION: The basic idea of descriptive statistics is to describe a set of data in a variety of abbreviated ways. In this lab you will investigate measures of central tendency and dispersion. The box-and-whiskers display, a graphical display of the 5-number summary of a set of data, will also be introduced.

MEASURES OF CENTRAL TENDENCY AND DISPERSION

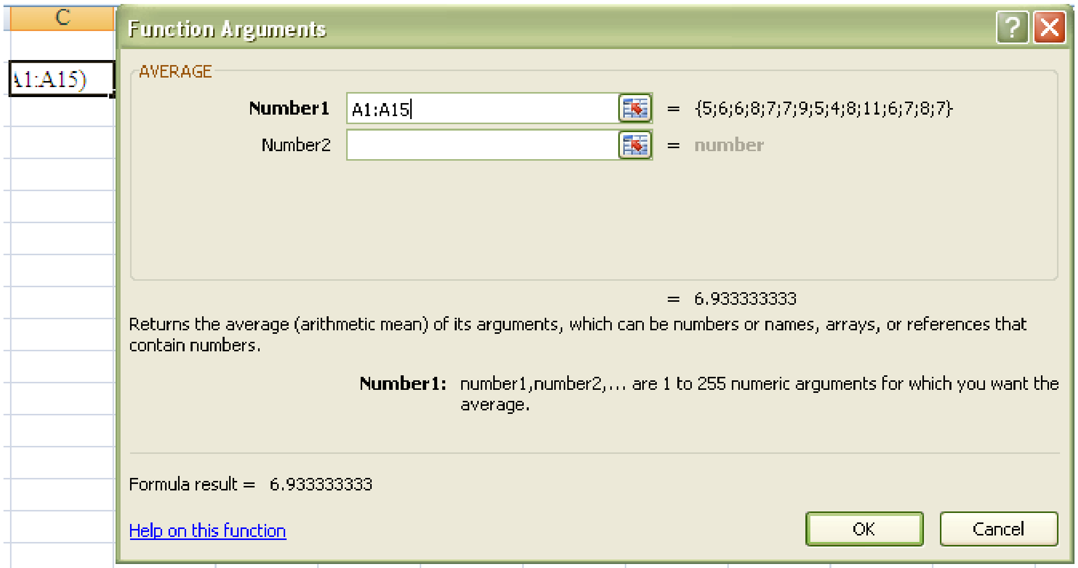
Measures of central tendency and variation are the foundation of descriptive statistics but most of these formulas are quite tedious to compute, even with a calculator. Fortunately, we can find a number of commonly used descriptive statistics using just a single command. Enter the data in Exercise 2.76 into column A.
Get a histogram of your data and visually approximate the "center".



Calculate the mean (and median) using the following commands.
Activate a cell where you want the answer to remain.
Choose: **Formulas** > **Insert function, f_x** > **Statistical** > **AVERAGE (or MEDIAN)> OK**



Enter: Number 1: **(A1:A15 or select cells)**



The answers will be placed in the selected cells.

C	D
Mean	Median
6.933333	7

We can also compute the midrange by using the statistical functions MAX and MIN as follows:

select a cell to hold the result, then click in the formula box and type (selecting the appropriate statistical function - shown in bold)

.5*(**MAX**(A1:A15)+**MIN**(A1:A15))

E	F	G
Max	Min	Midrange
11	4	7

Visually locate the three calculated centers on the histogram. Notice the three measures of central tendency are approximately the same. How well did you visually approximate the center?

Now, place the values of hours of sleep (column A) plus 4 into column B, do a histogram, visually locate the 'center', then determine the mean, median and midrange.

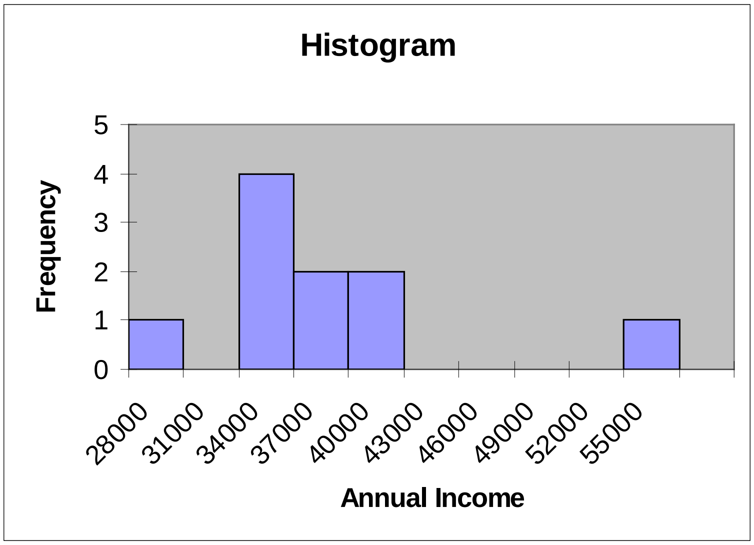
How did the three measures of central tendency (mean, median, and midrange) change?

Next, place the values of column A times 3 into column C, and follow the procedure above.

mean	
	20.8
median	
	21
midrange	
	22.5

Compare the three measures of central tendency for the columns of data A, B and C. How and why did a change in the measures occur? If a different transformation was performed (such as dividing each entry in A by 2) could you make an educated guess about the effect on these three measures?

Consider Applied Example 2.11 in the text. Retrieve the data and do a histogram and calculate the mean, median and midrange. What is there about the distribution of these ten data values that causes these three averages to be so different?



mean	
	35400
median	
	33375
midrange	
	39750
mode	
	31500

Compare the standard deviations for each of the previous four examples, along with how similar or how different the three measures of central tendency were. Can we use the standard deviation to predict whether we expect these three measures of central tendency to be quite similar or quite different?

FREQUENCY DISTRIBUTIONS

When the sample data are in the form of a frequency distribution, we can still use Excel to describe the distribution. The class marks need to be listed in one column with the corresponding frequencies in another. Start a new Excel workbook. (Choose: File > New > Workbook), and enter the following information, where X represents the number of radios in a household and Frequency is the number of households having X radios:

X	Freq
1	20
2	35
3	100
4	90
5	65
6	40
7	5

Name column A as Radios, and B as Frequency. Create column C to be x*f and D to be x^2*f as follows:

Activate C2
Enter: **=A2*B2**
Drag: **Bottom right corner of C2 down to give other products**

Activate D2 and repeat above commands replacing the formula with **=A2*C2**

Activate the data in columns B, C and D.
Choose: **AutoSum**

To find mean, activate E2, then continue with:
Enter: **=(C9/B9)**
To find the variance, activate E3, then continue with:
Enter: **=(D9-(C9^2/B9))/(B9-1)**
To find the standard deviation, activate E4, the continue with:
Enter: **=SQRT(E3)**

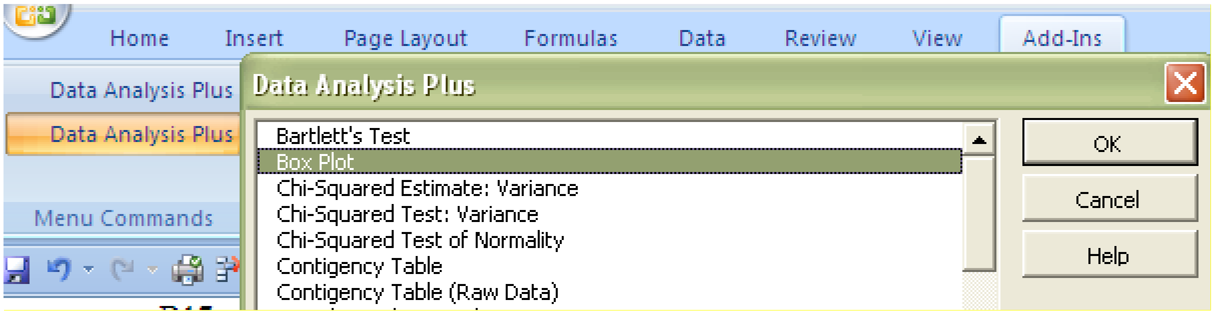
*Reminder: in the case of a grouped frequency distribution enter the class marks in one column and the corresponding frequencies in another.

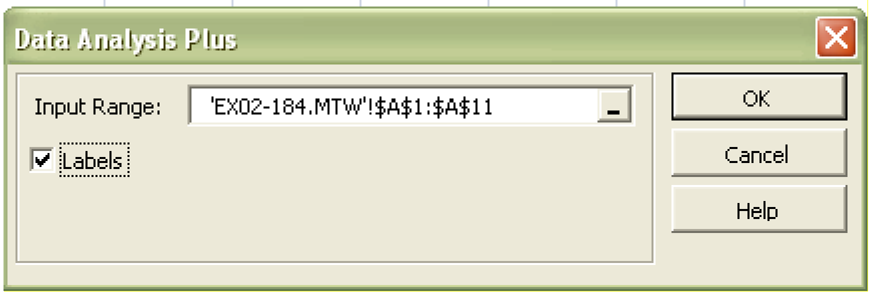
BOX-AND-WHISKER DISPLAY

The boxplot (Excel's name for the box-and-whisker display) is a simple graph that gives a graphic 5-number summary. Information about the center, dispersion, and skewness of a data set will be illustrated. Retrieve the data for Exercise 2.184 and construct a boxplot for each of columns A and B.

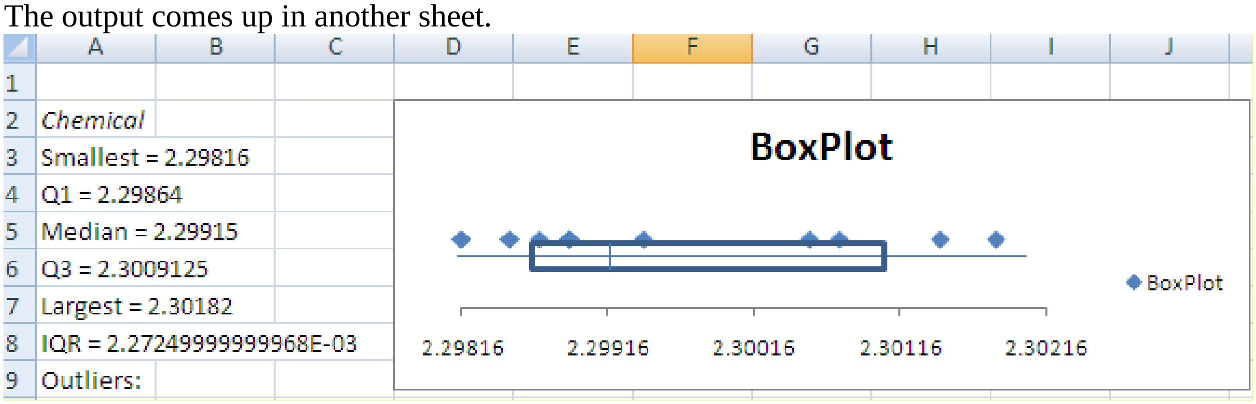
	A	B
1	Chemical	Atmospheric
2	2.30143	2.31017
3	2.29890	2.30986
4	2.29816	2.31010
5	2.30182	2.31001
6	2.29869	2.31024
7	2.29940	2.31010
8	2.29849	2.31028
9	2.29889	2.31163
10	2.30074	2.30956
11	2.30054	

Choose: **Add-Ins > Data Analysis Plus > BoxPlot > OK**
Enter: **A2:A24 or select cells > OK**

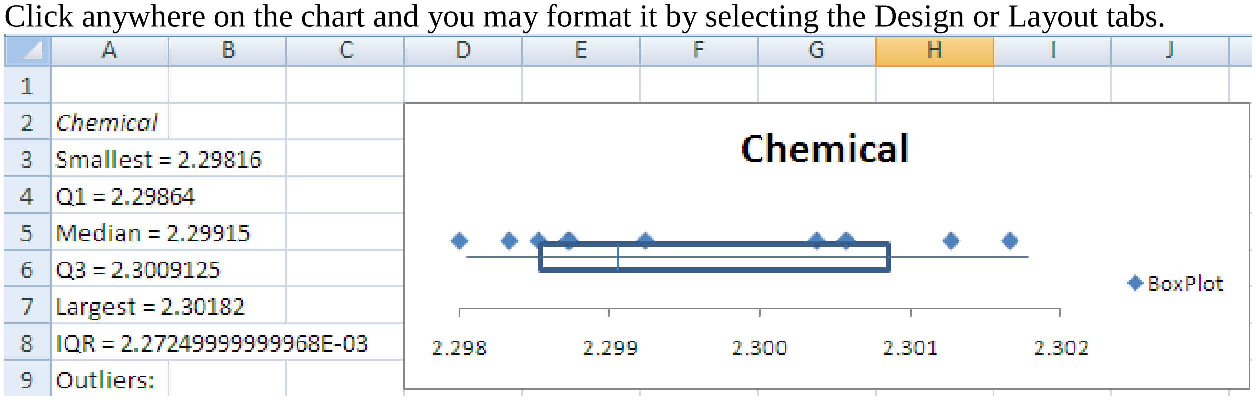




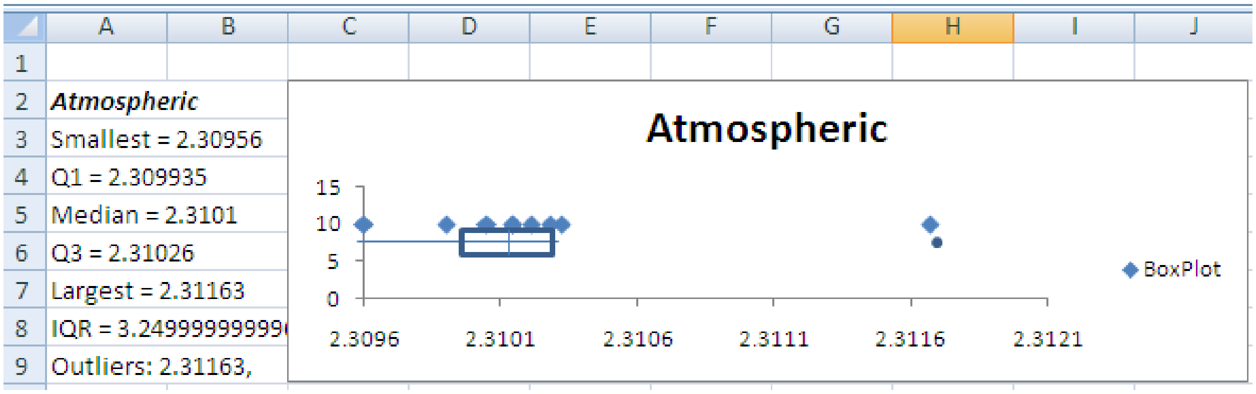
The output comes up in another sheet.



Click anywhere on the chart and you may format it by selecting the Design or Layout tabs.



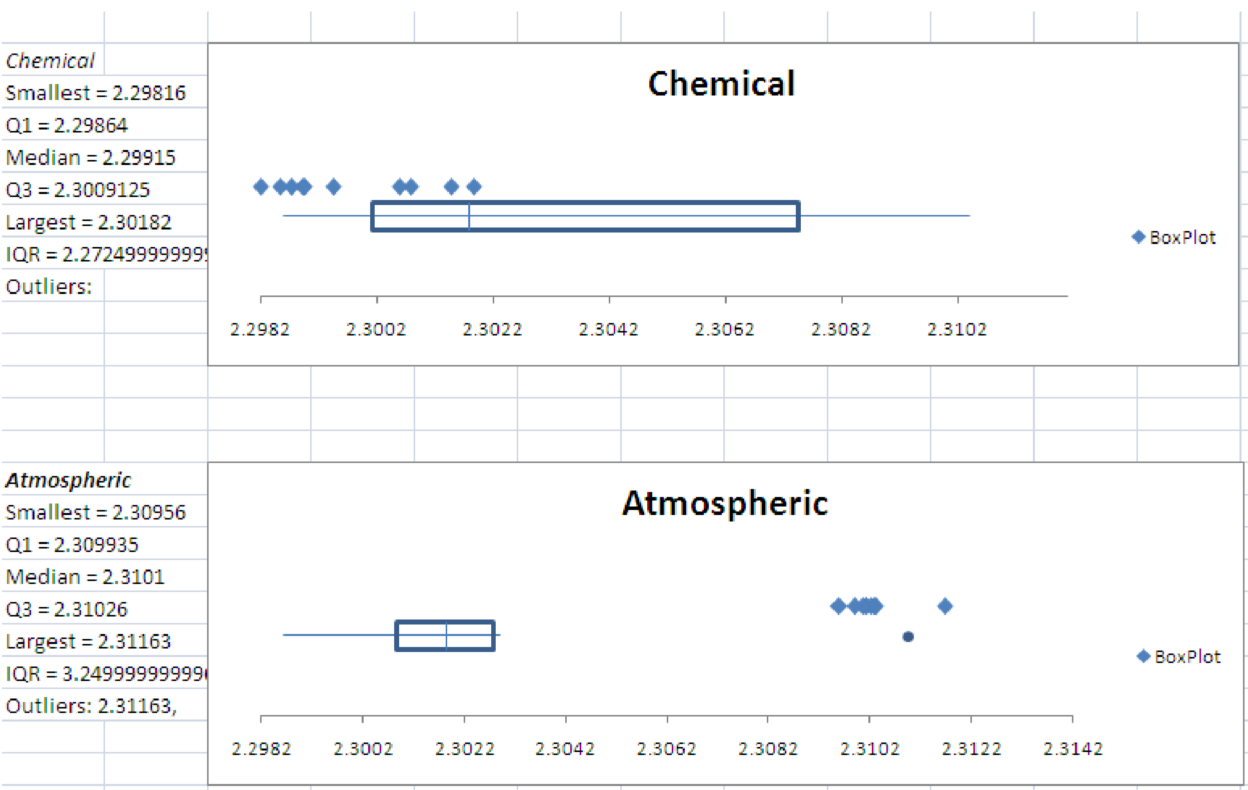
Repeating the procedure for Column B, Atmospheric data we get the following chart.



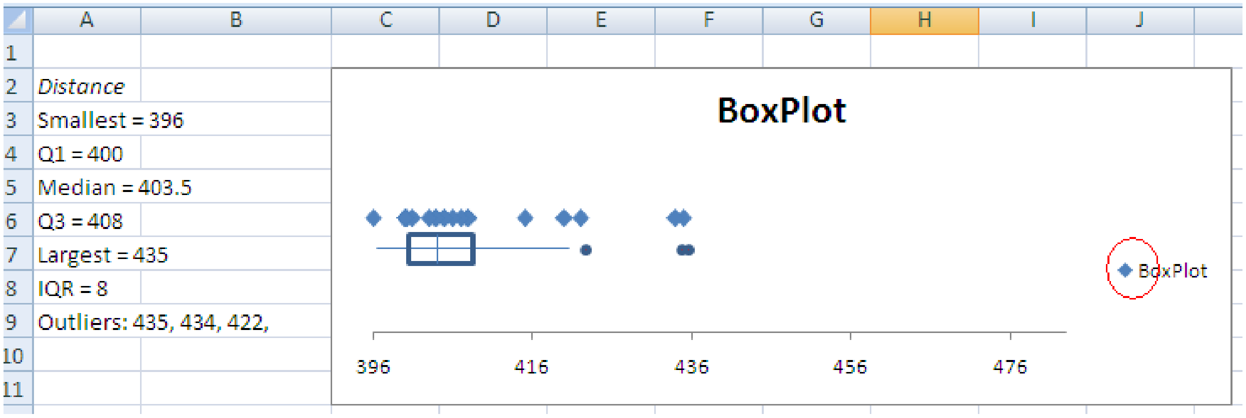
A rectangle is constructed between the two quartiles, with a line across the box indicating the location of the median. The box encloses the middle half of the data. The whiskers extend in either direction to indicate the maximum and minimum values.

Although “side-by-side” BoxPlots cannot be constructed in Excel, we can generate BoxPlots with the same scale for better comparison of the distributions. Activate all of columns A, and B, then

Choose: **Add-Ins > Data Analysis Plus > Box Plot > OK**



Consider Exercise 2.126. Retrieve the data and perform a BoxPlot of the data in column A.



The red oval in the boxplot indicates an outlier- a data value that is far removed from the rest of the data.

ASSIGNMENT: Do Exercises 2.76, 2.118, 2.125 and 2.128 in your text.

CHAPTER 3 - LAB SESSION 1
PRESENTATION OF BIVARIATE DATA

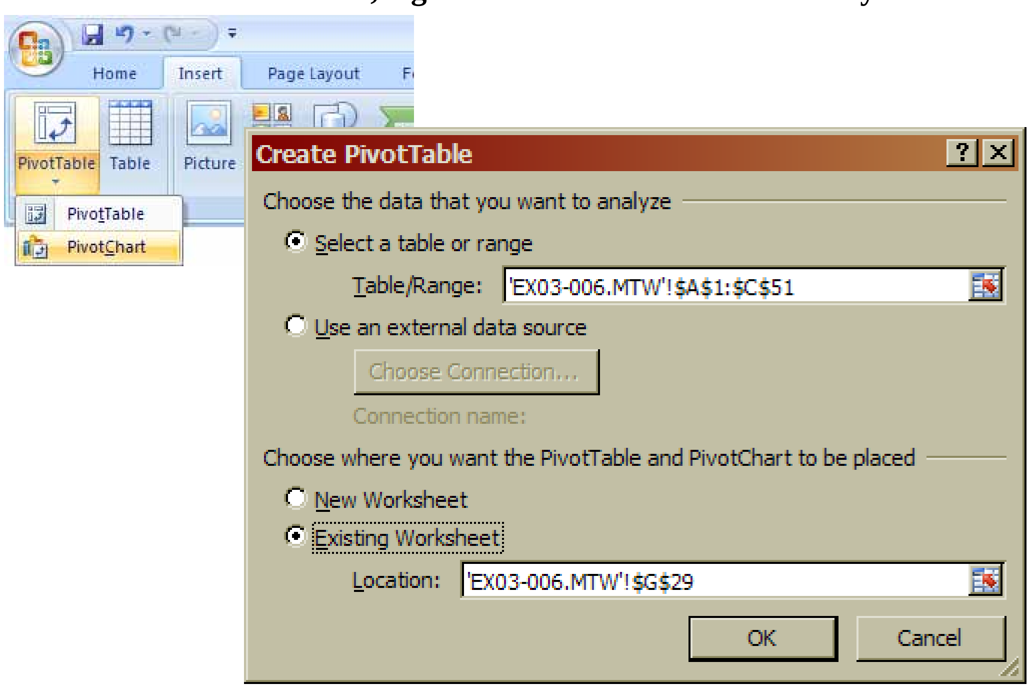
INTRODUCTION: It is frequently interesting to view the relationship of two variables. In this lab we will see how Excel can help us plot bivariate data and discover some trends in the relationship. We can set up the data as ordered pairs, with the independent variable as the x and the dependent variable as the y.

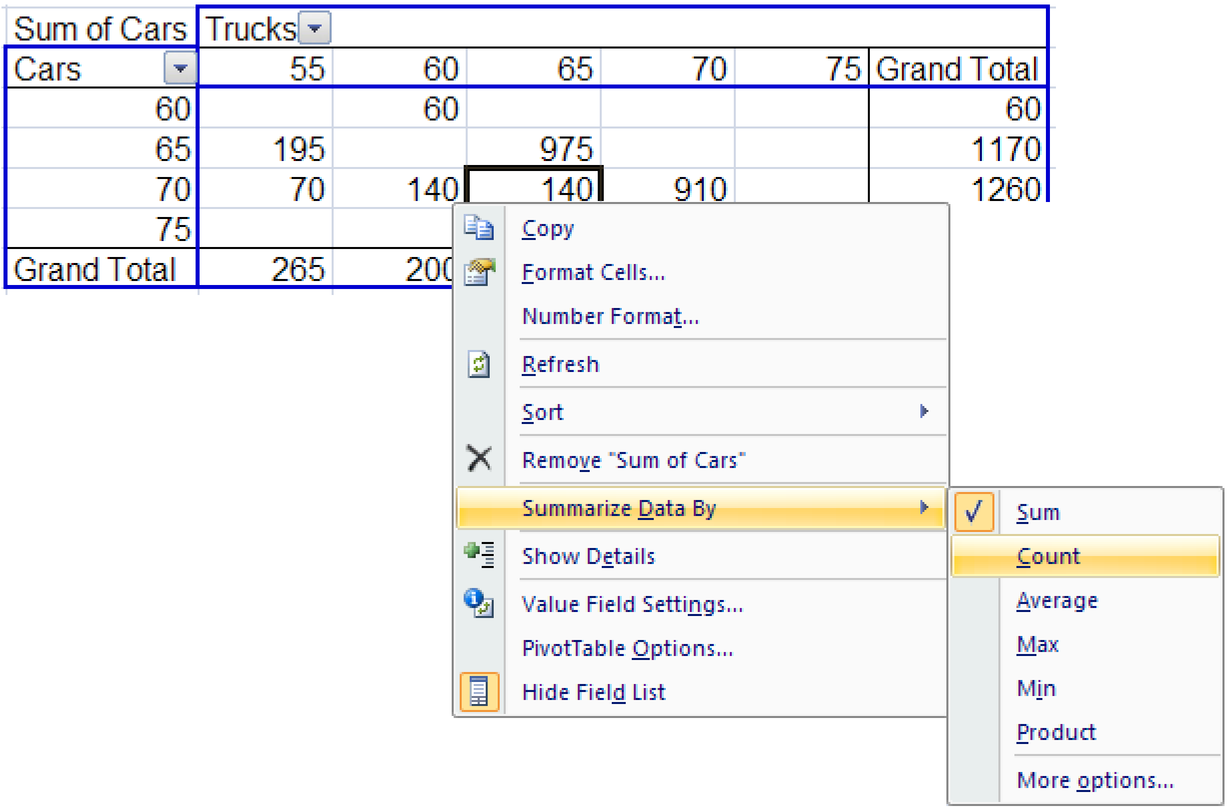
TABULAR PRESENTATION OF BIVARIATE DATA

We can arrange the data resulting from two qualitative variables in a cross tabulation or contingency table. These tables often show relative frequencies (percentages) that can be based on the entire sample, or on the subsample classification (either a row or a column).

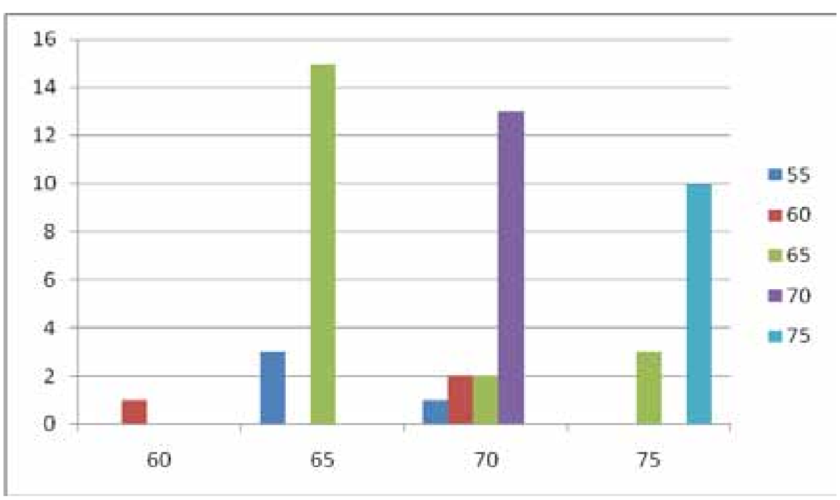
Let's use the data in the Highway Speed Limits table in Exercise 3.6. Retrieve the data (EX03-06). Note the data is arranged as follows: Column A is titled State. Column B is titled Cars, and column C is titled Trucks. We need to associate the vehicle type with each value in columns B and C.

To construct a cross-tabulation table of the two variables, vehicle type and maximum speed limit:
Choose: **Insert >Tables > Pivot Table pulldown > Pivot Chart**
Select: **Select a table or range**
Enter: **select appropriate cells of columns B and C**
Select: **Existing Worksheet**
Enter: **select a cell > OK**
Drag: **Car Label to row and Truck Label to column**
Either label into data area; right click and select Summarize by Count > OK



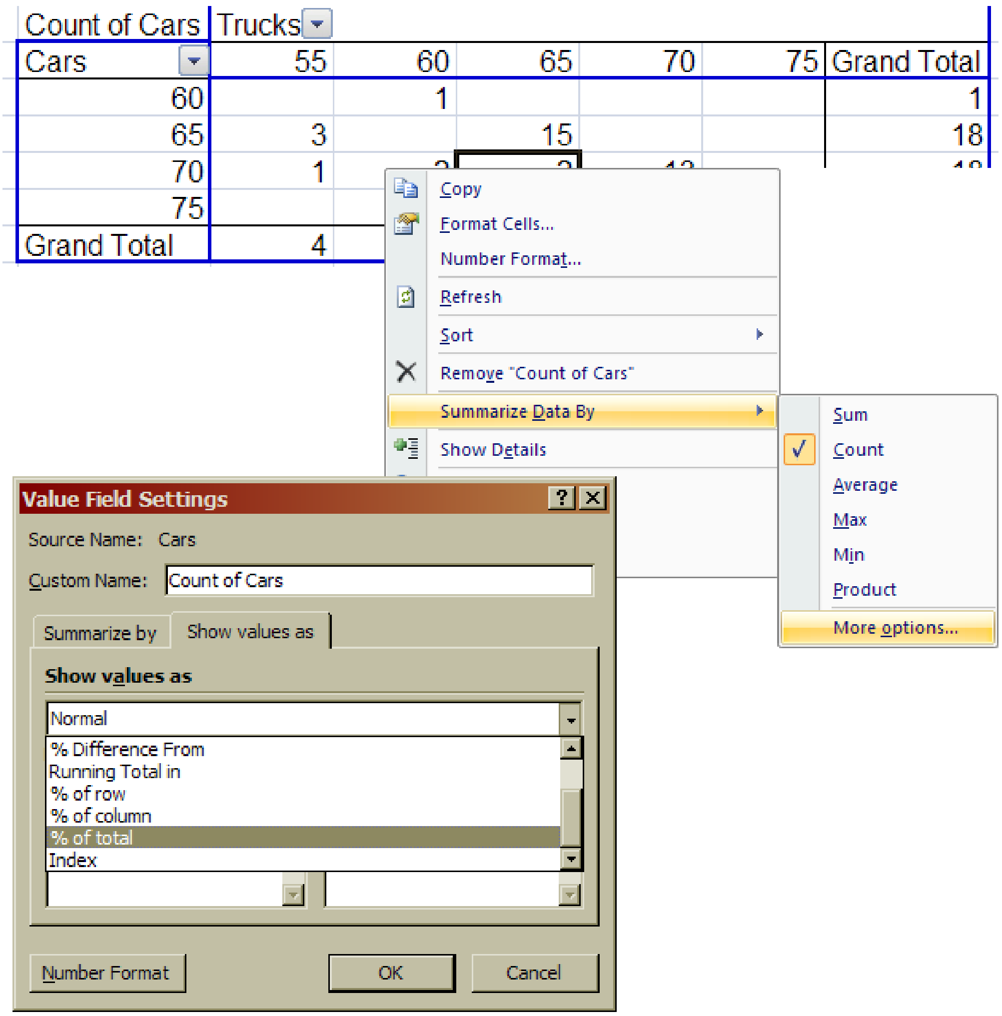


Count of Cars	Trucks					
Cars	55	60	65	70	75	Grand Total
60		1				1
65	3		15			18
70	1	2	2	13		18
75			3		10	13
Grand Total	4	3	20	13	10	50



Now let's do the same thing, only this time select the summarize by total percent.

Right Click: in data area box;
Choose: **Summarize Data by > More options**
Select: **Show values as: % of total > OK**



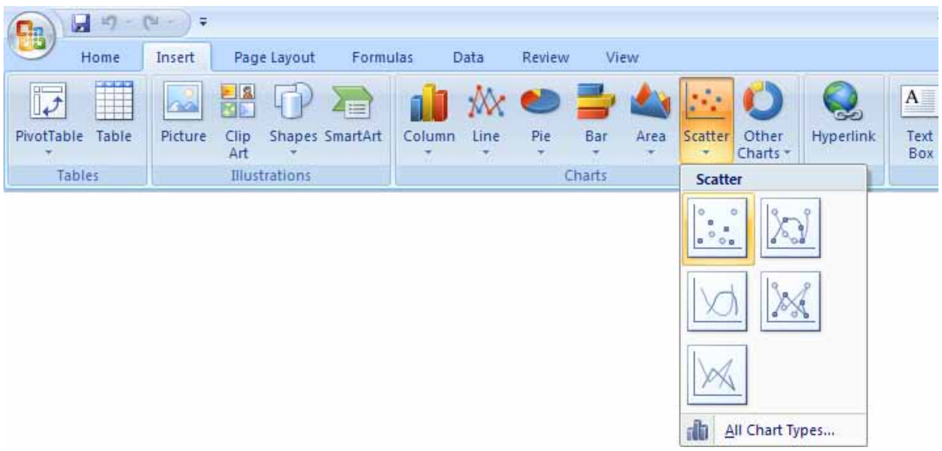
Count of Cars	Trucks					
Cars	55	60	65	70	75	Grand Total
60	0.00%	2.00%	0.00%	0.00%	0.00%	2.00%
65	6.00%	0.00%	30.00%	0.00%	0.00%	36.00%
70	2.00%	4.00%	4.00%	26.00%	0.00%	36.00%
75	0.00%	0.00%	6.00%	0.00%	20.00%	26.00%
Grand Total	8.00%	6.00%	40.00%	26.00%	20.00%	100.00%

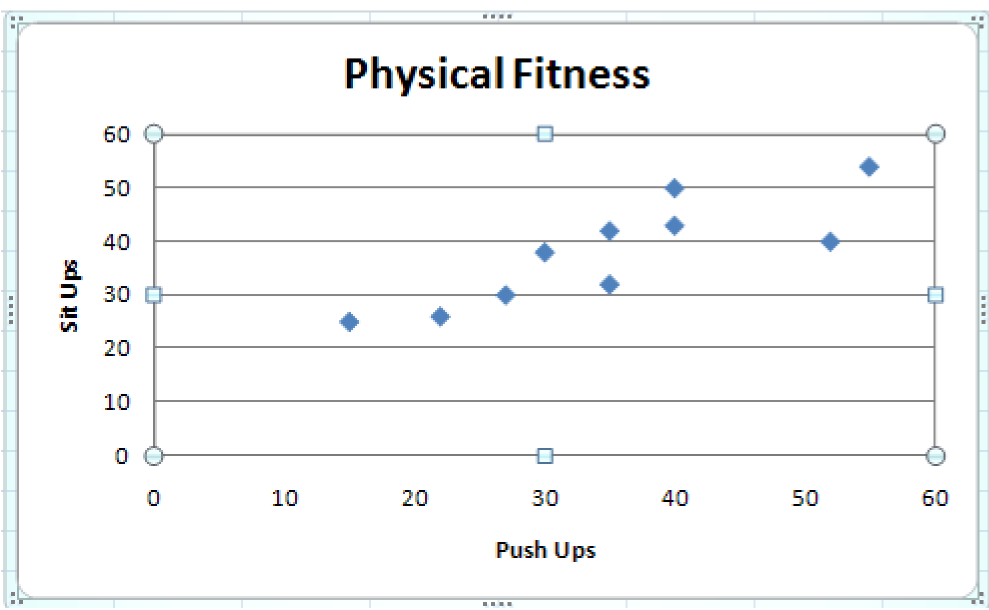
SCATTER DIAGRAMS

To do a scatter diagram illustrating the relationship between two quantitative variables we will enter the data into two columns. For this illustration, the data from Table 3-10 will be used (TA03-10).

The x-variable (push-ups) is in column B, and the y-variable (sit-ups) is in column C. Select the data in columns B and C and continue with:

Choose: **Insert > Scatter > 1st picture**
Choose: **Chart Layouts > Layout 1**
Enter: Chart title: **Physical Fitness**
Value (x) axis: **Push ups**
Value (y) axis: **Sit ups**

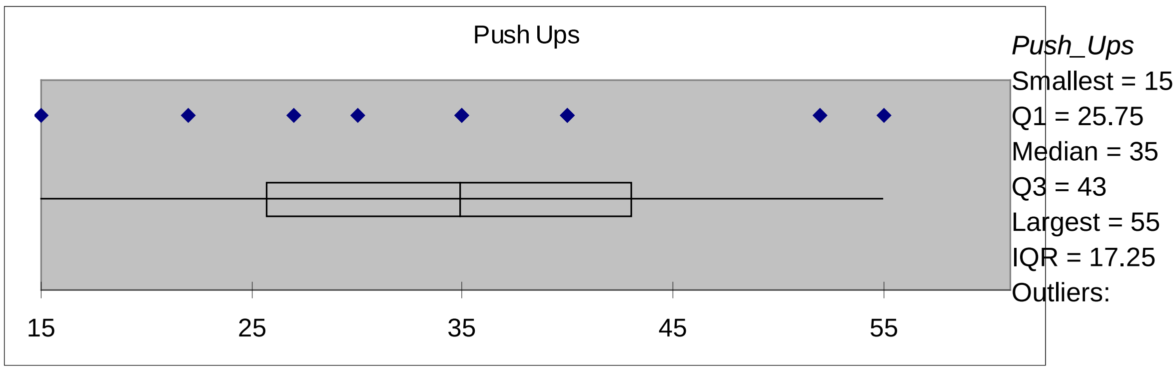


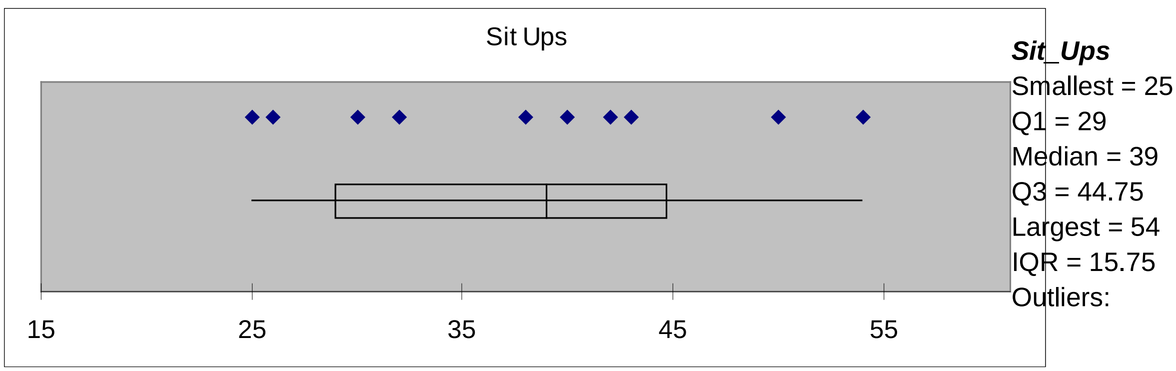


For the person(s) that did 35 push-ups, how many sit-ups were they able to do?

How many push-ups and sit-ups were done by the person represented by the dot in the upper right corner?

To compare these two variables in a different way, lets do a box-and-whisker display with common scale:





Compare the two types of exercises. Which indicates greater range of ability? Which exercise do most of those sampled find more difficult to do (as measured by number done)?

ASSIGNMENT: Do Exercises 3.11, 3.25 in your text

CHAPTER 3 - LAB SESSION 2
CORRELATION AND REGRESSION

INTRODUCTION: Not only is it important to analyze single variables, but frequently one needs to determine if and how two variables are related. The correlation coefficient is a measure of the strength of the linear relationship between two variables. In these exercises you will use Excel to analyze this statistic, and these exercises will also give you a very brief introduction to linear regression.

INVESTIGATIONS OF THE CORRELATION COEFFICIENT
The data set below is a sample of weight and waist size for 11 women. You will use that data to estimate the correlation between a woman's weight and her waist size. Once that value has been determined you will show that this value is independent of the scale of the two variables.

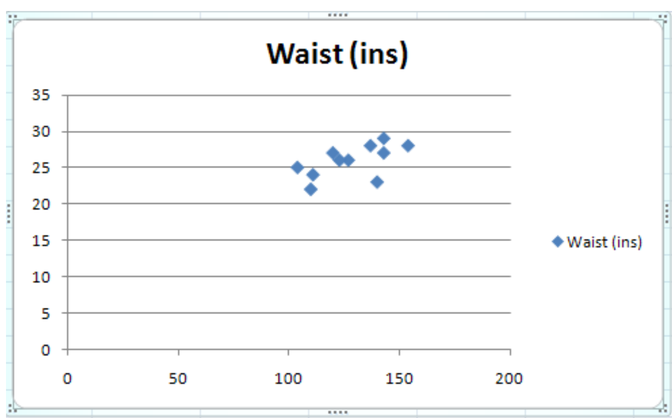
Weights and Waist Sizes

weight(lbs):	110	143	120	127	143	111	137	154	123	104	140
waist (ins):	22	29	27	26	27	24	28	28	26	25	23

Enter the data into the worksheet and name the two variables.

Get a scatter diagram of the bivariate data set. The variable "Weight" should be on the x-axis and "Waist" on the y-axis. Select the data and continue with:

Choose: **Insert > Scatter > 1st picture**
Choose: **Chart Layouts > Layout 1**

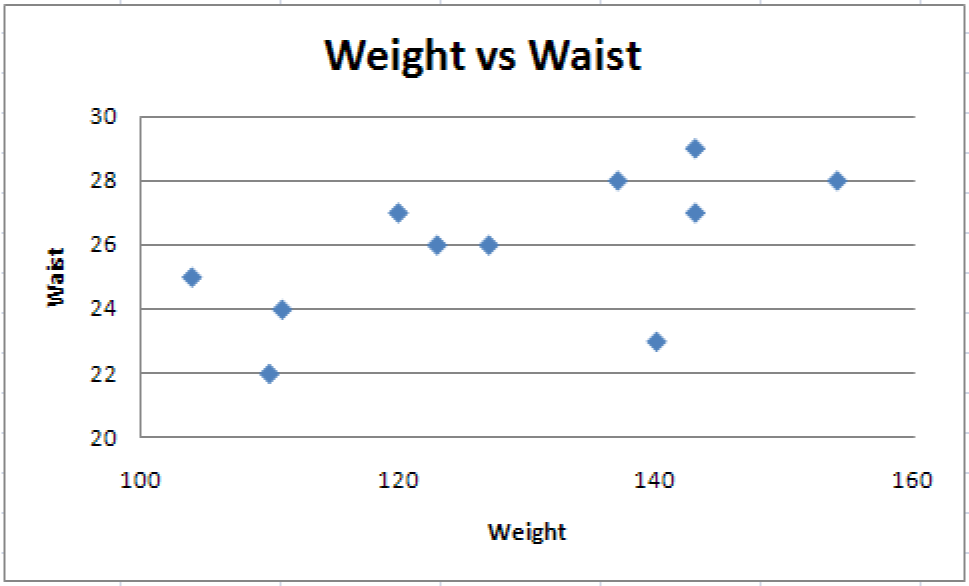


We can edit the scatter plot (since all the points are in a corner), and rescale the axes to reflect the data range.

Right click on the X-axis
Select: **Format Axis**
Click: radio buttons for **Fixed** on each of the options and enter

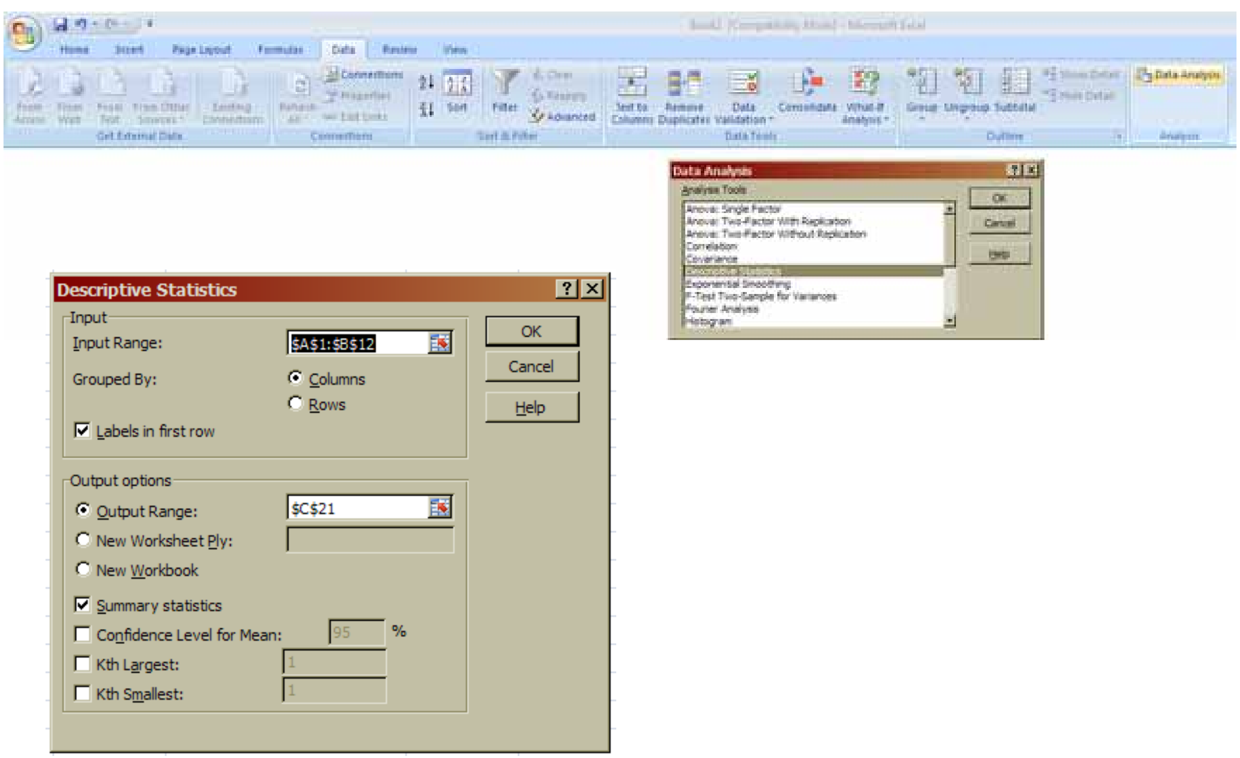
Minimum: **100**
Maximum: **160**
Major unit: **20**
Minor unit: **5**

Also make appropriate changes to the y-axis scale



Let's also generate descriptive statistics for each of these variables:

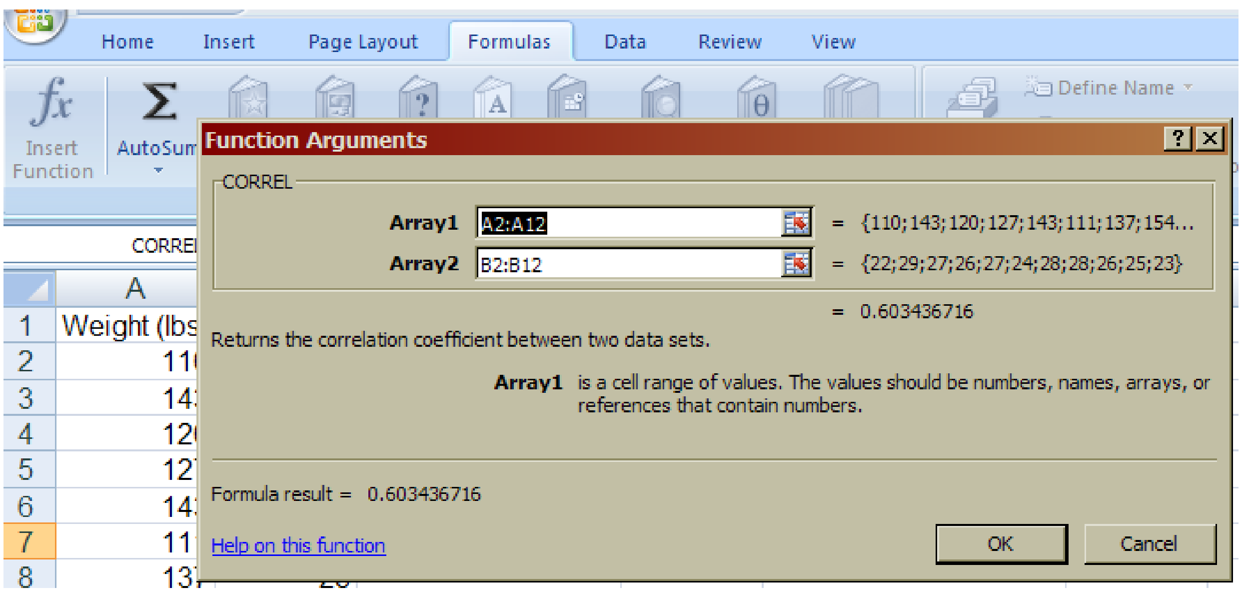
Choose: **Data > Data Analysis > Descriptive Statistics > OK**
Enter: Input Range: **select cells**
Select: **Labels in First Row** (if necessary)
Output Range: **select cell**
Select: **Summary statistics > OK**



* the output shown below is edited for clarity

Weight		*waist (ins)*	
Mean	128.3636	Mean	25.90909
Median	127	Median	26
Mode	143	Mode	27
Standard Deviation	16.21279	Standard Deviation	2.21154
Range	50	Range	7
Minimum	104	Minimum	22
Maximum	154	Maximum	29
Count	11	Count	11

Calculate the correlation coefficient, r.
Select the Formulas Tab, then
Choose: **Insert function, f_x > Statistical > Correl > OK**
Enter: Array 1: **x data range**
Array 2: **y data range > OK**



correlation
0.603436716

QUESTIONS:

1. Would you say that the variables were positively or negatively correlated? Is there a strong or weak correlation?

2. If you were to add an equal amount of weight to each woman (assume no change in waist size), would the value of r, the correlation coefficient, change? Test your conjecture by adding 25 lbs. to each woman's weight and recalculate r. The necessary commands are:

   Activate cell C2 and type “= A2 + 25”, then drag right corner down to perform the same calculation on all of column A. Redo the correlation using column C for Array 1.

3. If you were to change the scale of the variables: weight to kg and waist size to meters, would the value of r change? Test your conjecture by multiplying 'WEIGHT' by 0.453 and 'WAIST' by .0254 and recalculate r. How will the scatter diagram change when you change the scales?

4. The last observation in your data set was for a model known for her especially thin figure. If you eliminated it from the data set, how much would r change? Would you say that the statistic, r, is sensitive to extreme observations? Explain.

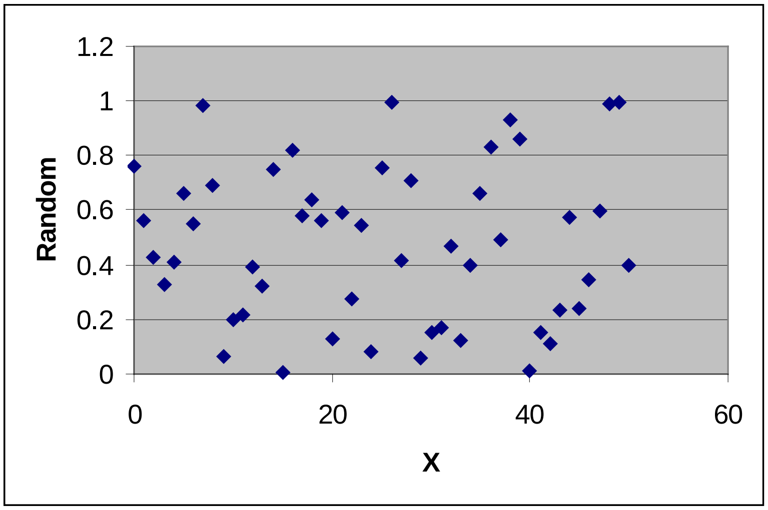
INTERPRETATION OF THE CORRELATION COEFFICIENT

In this next section, we will be examining some scatter diagrams of computer-generated data to gain a more thorough understanding of just what the value of the correlation coefficient means. For each pair of variables, you will calculate r and look at the corresponding scatter diagram.

Enter the values from 0 to 50 for your first variable and name your variable “x”.
Enter x in A1, and enter 0 in A2, enter 2 in A3, select A2 and A3, then right click on lower right corner and drag to A52

In cell B1 enter the name **Random,** then activate cell B2, and continue with:
Enter: **=rand( )**
Click and drag: **lower right corner of B2 cell to row 52**

Get a scatter diagram of the two variables and calculate r.

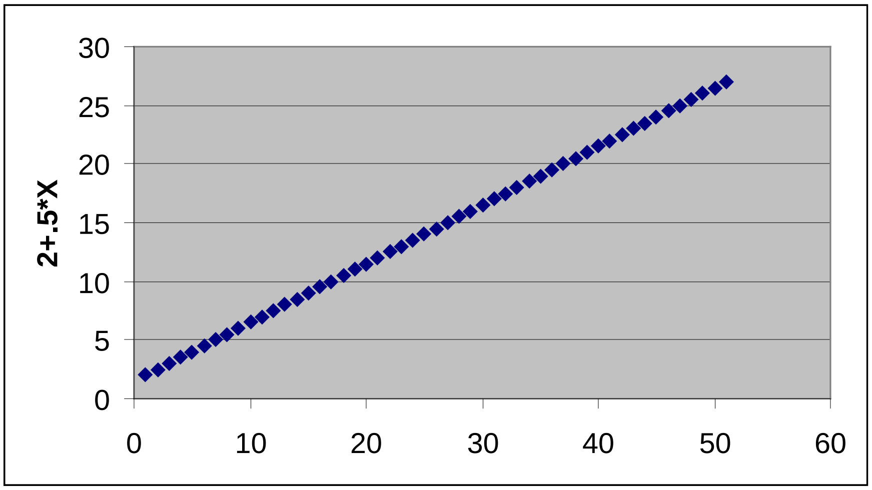


correlation
0.047592

When comparing your output to that presented here, remember you are working with random data and there will be variation in results.

Next, generate a set of y values which has no random component:

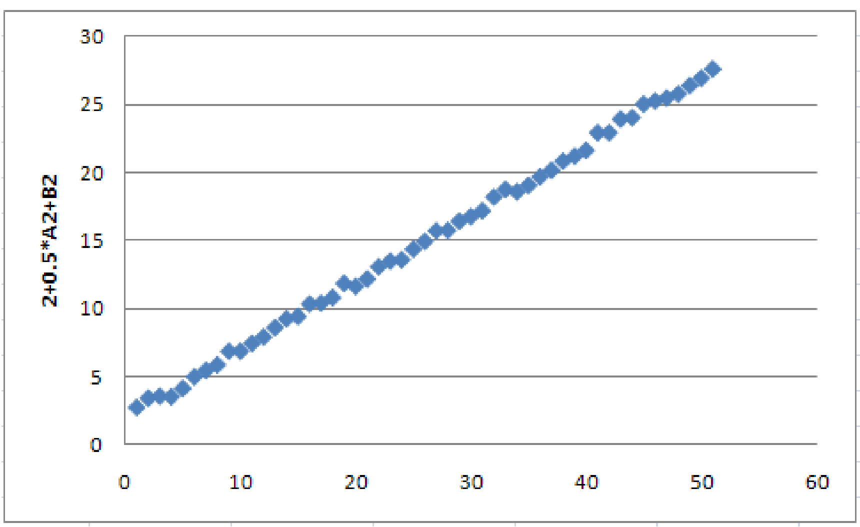
Activate cell C2, type **=2+A2*.5, click and drag lower right corner of cell down through row 52.**



correlation
1

Generate a set of y values that have a small random component and repeat above procedure.

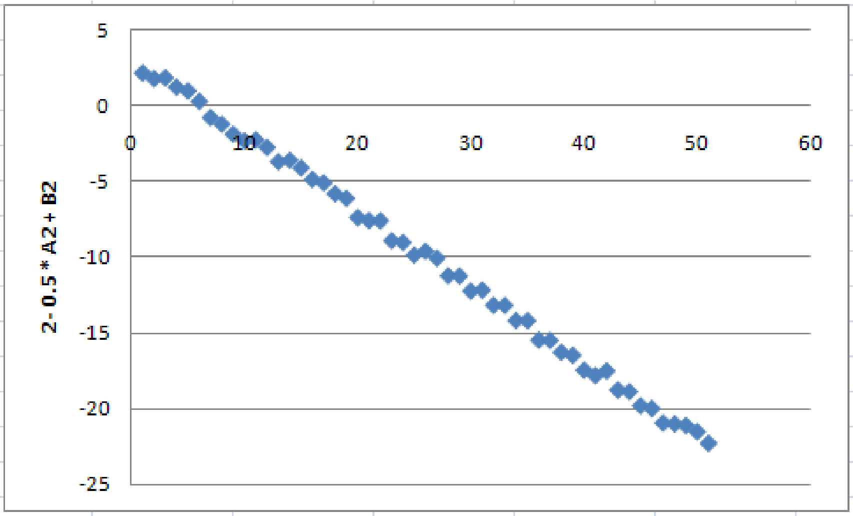
fill column D with **= 2 + 0.5 * A2 + B2**



correlation
0.999255

Generate a set of y values that are negatively correlated, and repeat above procedure.

fill column E with **= 2 - 0.5 * A2 + B2**

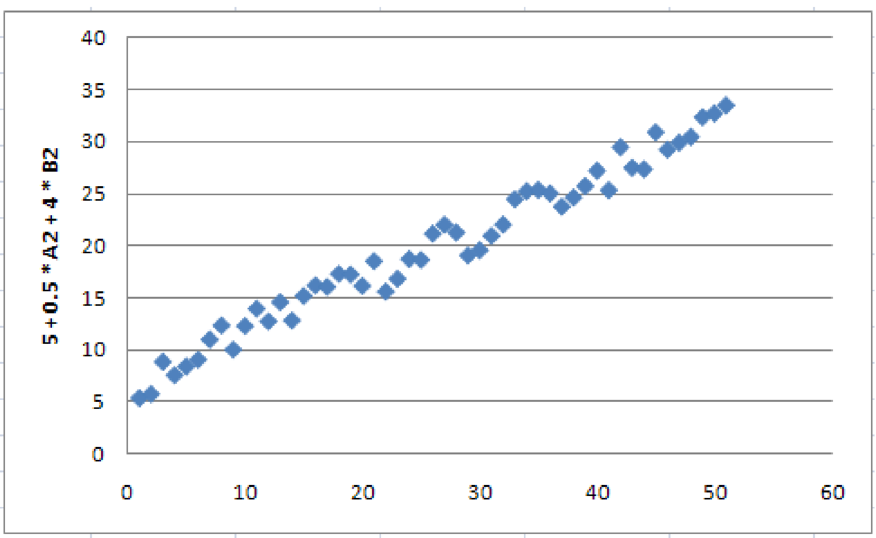


correlation

-.99911

Generate a set of y values that have a large random component and repeat previous procedure.

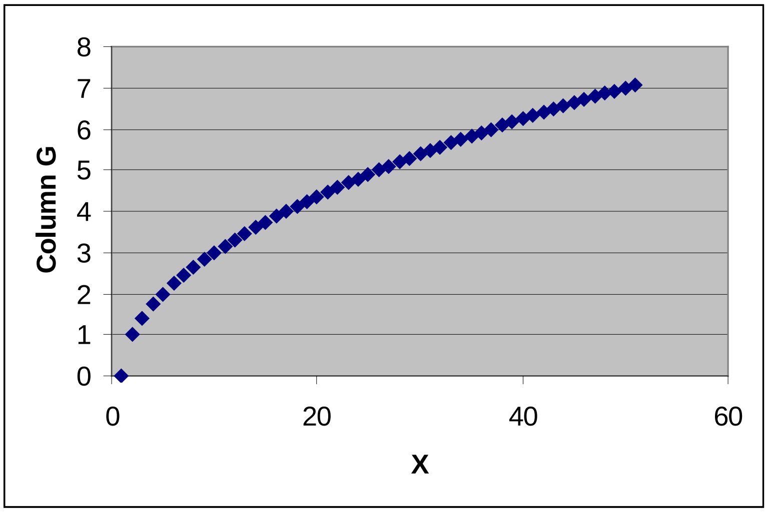
fill column F with **= 5 + 0.5 * A2 + 5 * B2**



correlation
0.989753

Generate a set of y values that are non-linearly related to x.

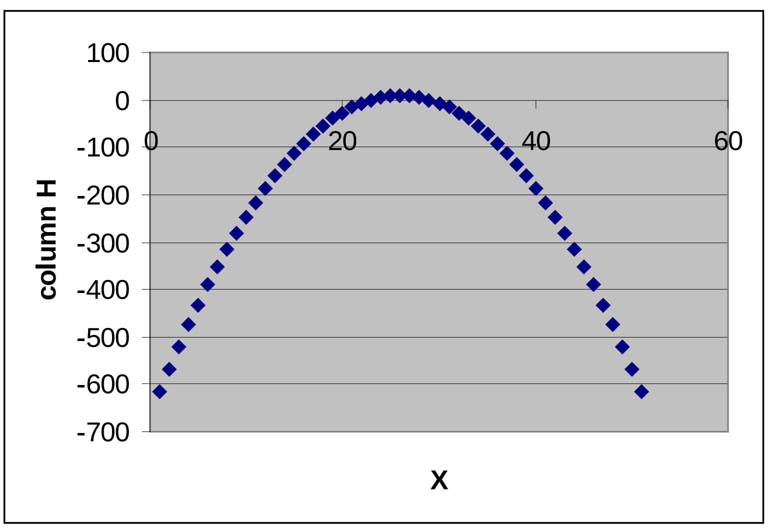
fill column G with **= SQRT(0.1*A2)**



correlation
0.973675

Generate a second set of y values which are related but not linearly related to x and repeat previous procedure.

fill column H with **= 9 - (A2 - 25)**2**



correlation
-1.25106E-17

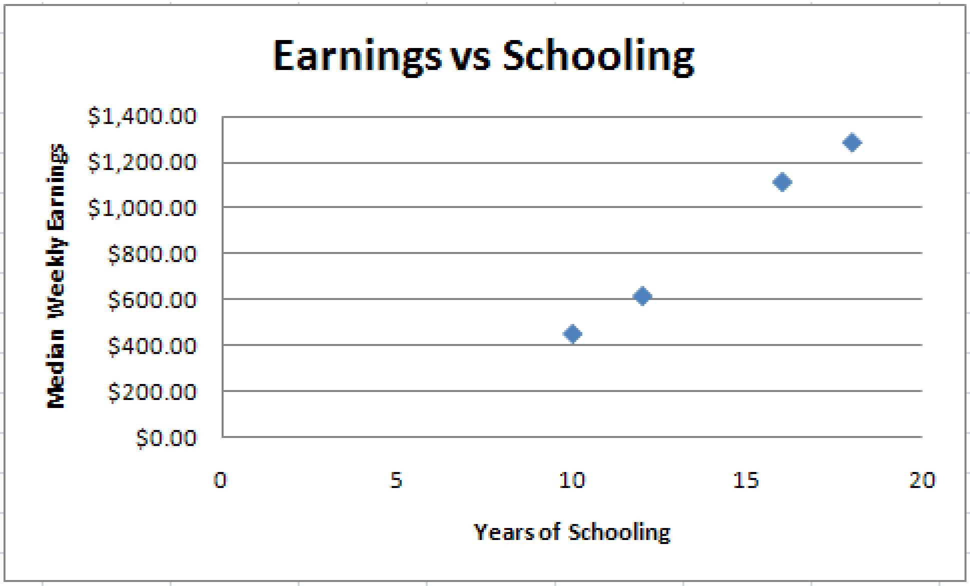
QUESTIONS:

1. Using the results from above, what type of relationship can you determine between the correlation coefficient and the scatterplot? What type of pattern do you see in the scatter diagram when r is close to zero? When r is close to one? What is the pattern like when r is negative?

2. Does r being close to zero imply that the two variables are unrelated? Check column H versus column A before answering this question.

LINEAR REGRESSION

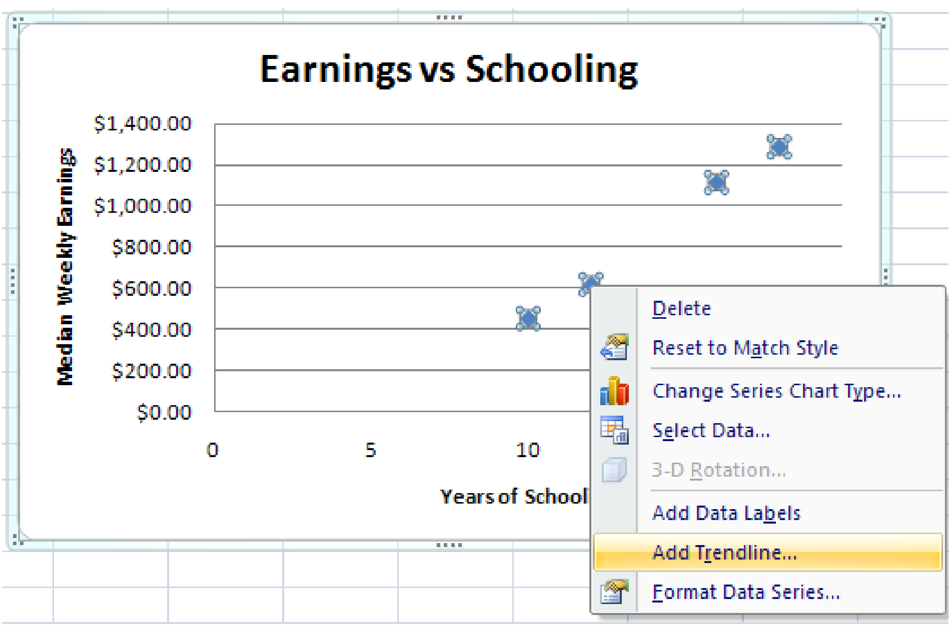
Retrieve the data from text Exercise 3.75.(EX03-075) Get a feeling for whether years of schooling and median usual weekly earnings are correlated by doing a scatterplot.

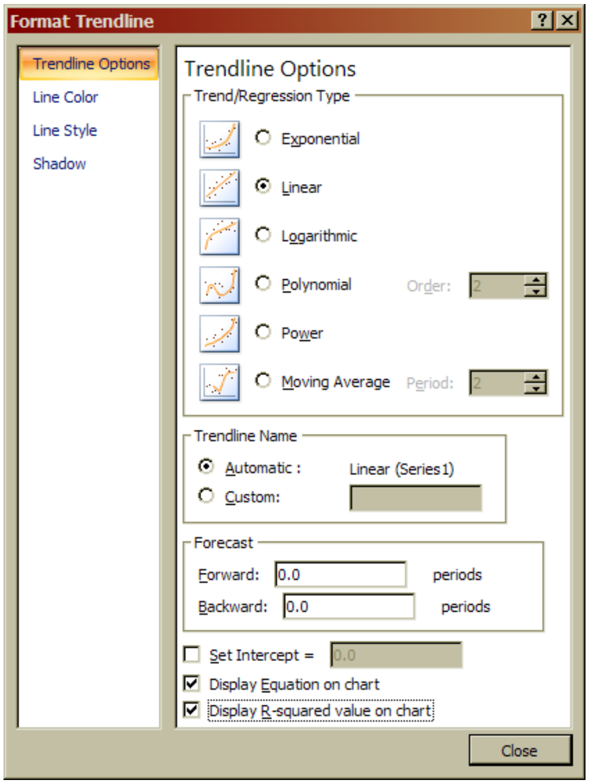


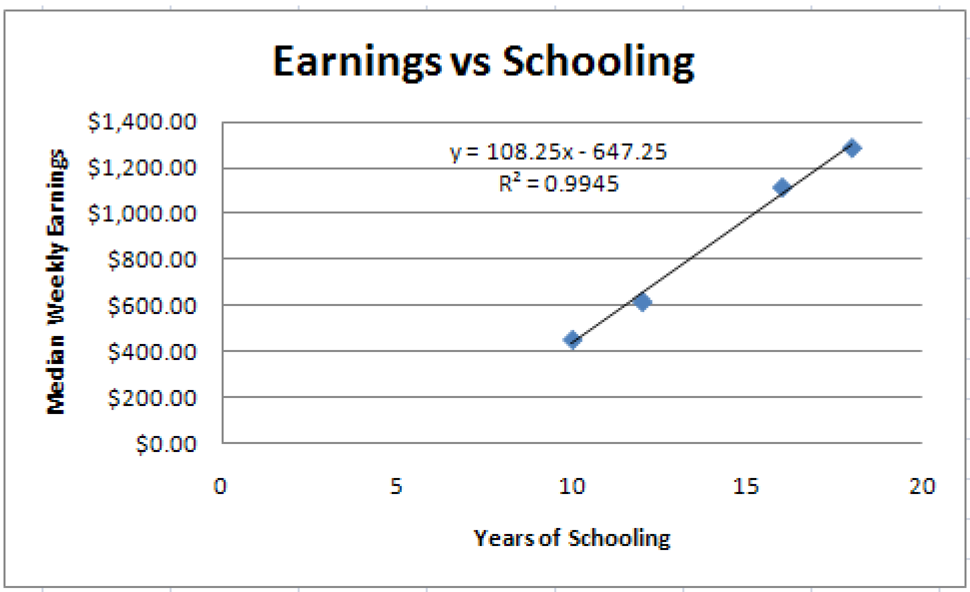
correlation
0.9972673

The least squares line can be added to the plot, along with its equation and the value of r^2. Right click on one of the data points shown in the scatter plot. A drop-down menu will appear.

Select: **Add Trendline**
Select: Type : **Linear**
Select: Options,
then check **Display equation of chart and Display R-squared on chart > OK.**

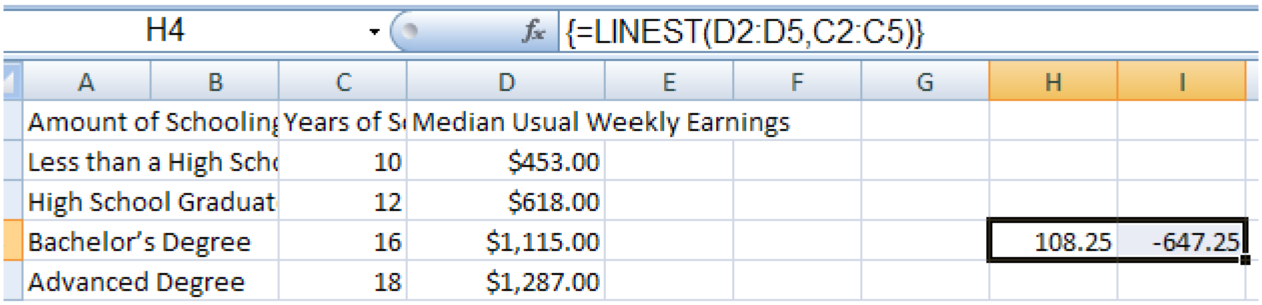



Format Trendline
Trendline Options
Line Color
Line Style
Shadow
Trendline Options
Trend/Regression Type
Exponential
Linear
Logarithmic
Polynomial
Order:
Power
Moving Average
Period:
Trendline Name
Automatic :
Linear (Series1)
Custom:
Forecast
Forward:
0.0
periods
Backward:
0.0
periods
Set Intercept =
0.0
Display Equation on chart
Display R-squared value on chart
Close

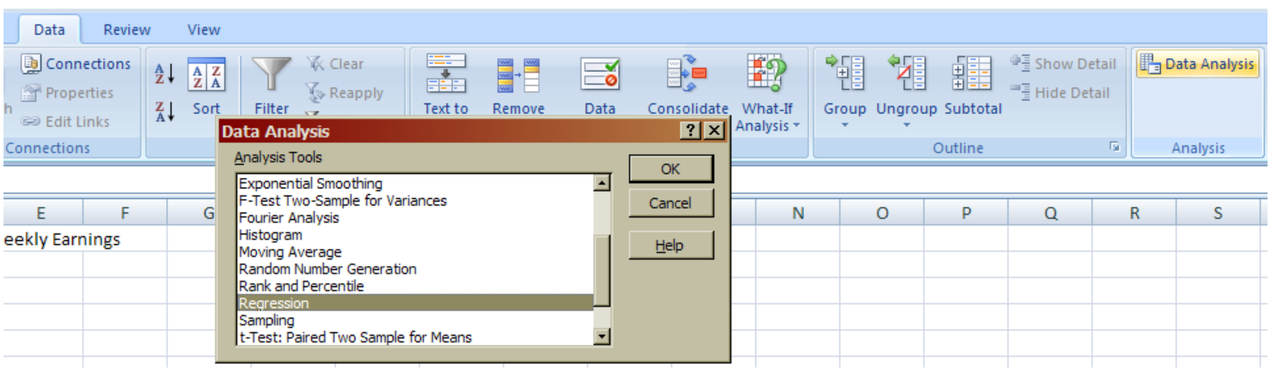

Earnings vs Schooling
y = 108.25x - 647.25
R² = 0.9945
Median Weekly Earnings
$1,400.00
$1,200.00
$1,000.00
$800.00
$600.00
$400.00
$200.00
$0.00
0
5
10
15
20
Years of Schooling

If we want to obtain the values of slope and intercept without using the doing the scatter diagram and adding the trendline, we can use LINEST(y-range, x-range).

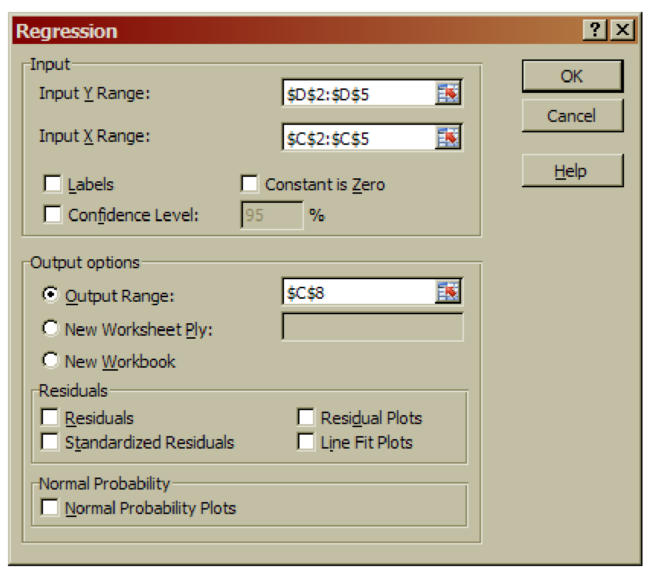
Activate two horizontally adjacent cells on the worksheet, type **=LINEST(D2:D5, C2:C5)** in the formula bar and press **CTRL+Shift+Enter** to generate the values of both slope and intercept.



To do the regression: Choose: **Data>Data Analysis >Regression >OK**



Indicate the location of the data, as appropriate and click OK.



Here is the default output generated by the **Regression** command for Exercise 3.75. Notice that a great deal of information is generated, but at this point we would need only the coefficients highlighted in red.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.997267348
R Square	0.994542163
Adjusted R Square	0.991813244
Standard Error	35.86258496
Observations	4

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	468722.5	468722.5	364.4455	0.002733
Residual	2	2572.25	1286.125		
Total	3	471294.75			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-647.25	81.38515682	-7.95292	0.015445	-997.422	-297.078	-997.422	-297.078
X Variable 1	108.25	5.670372563	19.09046	0.002733	83.85236	132.6476	83.85236	132.6476

ASSIGNMENT: Do Exercises 3.20, 3.38, 3.45, 3.59, 3.99 in your text.

CHAPTER 5 - LAB SESSION
RANDOM NUMBERS AND PROBABILITY

INTRODUCTION: This lab session is designed to introduce you to random numbers and their use in simulating experiments. The outcomes of events in normal life cannot be predicted, but it is possible to have an idea of what outcomes are possible. The theory of probability was developed to help analyze experiments whose outcomes are uncertain. We can use Excel to simulate certain experiments such as flipping a coin or rolling a die.

RANDOM NUMBERS

You were introduced to the RANDOM command in Chapter 3 – Lab Session 2. There we used the RAND worksheet function, to return an evenly distributed random number greater than or equal to 0 and less than 1 every time the worksheet is calculated. Now we'll look at the Random Number Generation analysis tool. This tool is part of the Analysis ToolPak. This tool fills a range with independent random numbers drawn from one of several distributions. You can characterize subjects in a population with a probability distribution. For example, you might use a normal distribution to characterize the population of individuals' heights, or you might use a Bernoulli distribution of two possible outcomes to characterize the population of coin toss results.

Suppose we want to simulate the outcomes for tossing a coin 100 times.

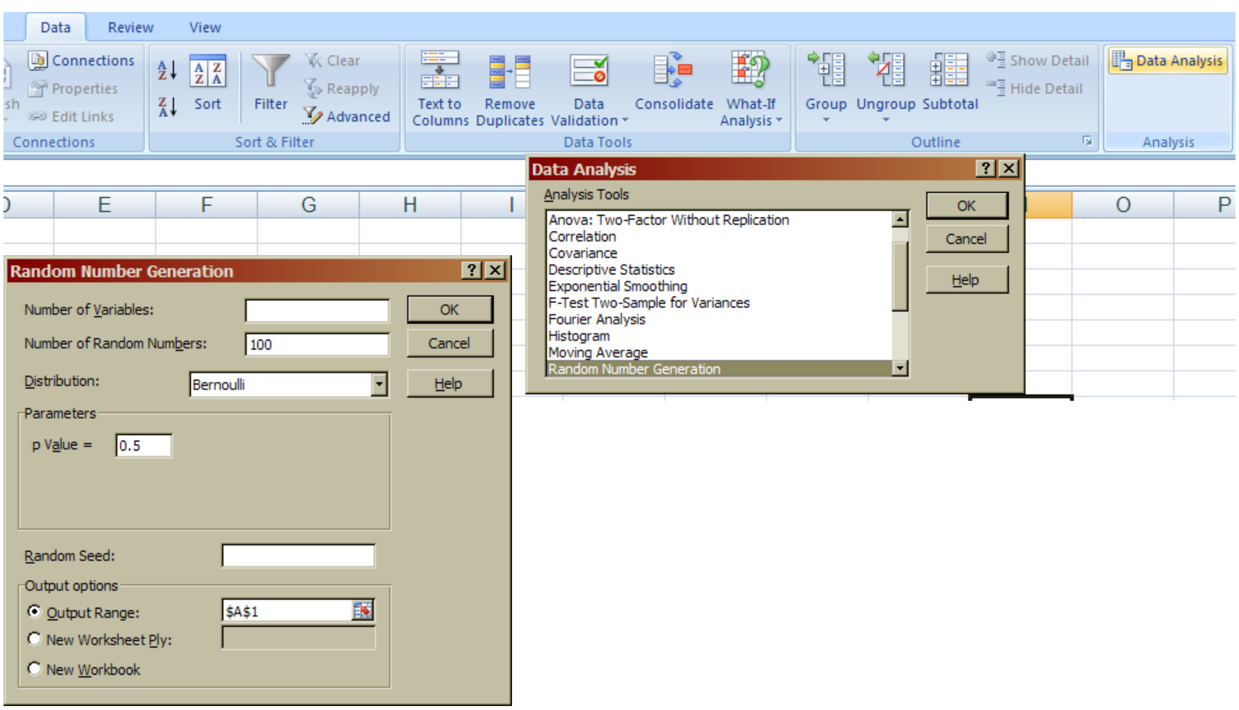
Choose: **Data > Data Analysis > Random Number Generation > OK**

Enter: Number of Random Numbers: **100**

Distribution: **Bernoulli**

p-value: **0.5**

Select: Output Range: **select cell > OK**

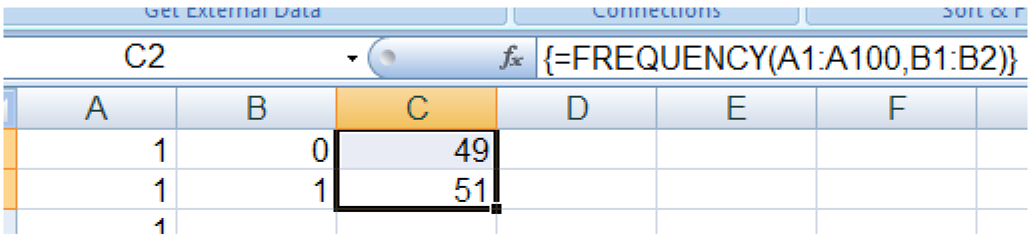


Give the relative frequency for a head (1) and a tail (2) based on the Excel output.
Certainly, let the computer do the work:

Enter a 0 and a 1 into the first two cells of column B. (This is done to indicate the classes of values to be tallied.) Select cells C1:C2 (to store the tallies) and then continue by typing

=Frequency(A1:A100, B1:B2)

Since this is an array formula you must press Ctrl + Shift + Enter



Questions:

1. What commands would be used for simulating the rolling of a die 50 times?
2. Create a new Excel workbook and place 50 simulated rolls into columns A and B. Give the relative frequency for the outcomes 1, 2, 3, 4, 5, and 6 based on the Excel output.

THE LAW OF LARGE NUMBERS

To see how the law of large numbers works, we need to create a third column with the sums of two dice rolls simulated by columns A and B. First, since Excel generates real decimal values, place the INT (the integer function) of column A values in column C, and the INT of column B values into column D.

in cell E1 enter **= C1 + D1,** click and drag the fill handle to cell E50

To determine the relative frequency of each outcome:
Enter: the possible outcomes **2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12** into column F
Select cells G1 through G11, and type **=FREQUENCY(E1:E50,F1:F11)** and press Ctrl + Shift + Enter. To get the relative frequencies, in cell H1 enter **=G1/50** and click and drag the fill handle to cell H11.

F	G	H
2	1	0.02
3	4	0.08
4	6	0.12
5	9	0.18
6	6	0.12
7	8	0.16
8	4	0.08
9	7	0.14
10	2	0.04
11	2	0.04
12	1	0.02

Remember, these sums were randomly generated, so your output may differ.

Note: You may also use **Insert > Table > PivotChart . . .** to generate a table containing the outcomes and their frequencies. See lab Chapter 3 Lab 1 for more information.

Interpreting the results:

1) What is the observed probability of obtaining a sum of 2 on the dice?
2) What is the observed probability of obtaining a sum of 7 on the dice?
3) What is the observed probability of obtaining a sum of 11 on the dice?

Using similar commands, create two additional columns containing 500 simulated rolls of a single die and a third column containing the sums of these 500 simulated rolls of 2 dice.

4) Answer the above three questions about this simulation. How do the answers compare to the theoretical probability? (Use both numerical and graphic evidence.)

THE BINOMIAL PROBABILITY DISTRIBUTION

Consider the following situation: Suppose you bought four light bulbs. The manufacturers claim that 85% of their bulbs will last at least 700 hours. If the manufacturer is right, what are the chances that all four of your bulbs will last at least 700 hours? That three will last 700 hours, but one will fail before that?

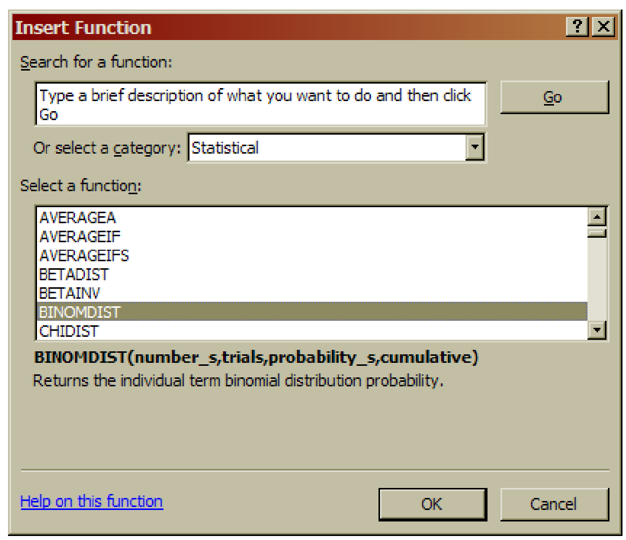
Consider another situation. You've somehow gotten enrolled in a class in advanced Greek Mythology. You don't know anything about mythology but you're to take a pop quiz. You'll have to guess on every question. It's a multiple-choice test; each of the 20 questions has 3 possible answers. To pass you must get at least 12 correct. What are the chances that you'll pass?

How would you answer the above questions? Excel can help us with this by using the BINOMDIST (The Binomial Probability Distribution Function) to generate binomial probabilities. (Remember what a binomial distribution requires.)

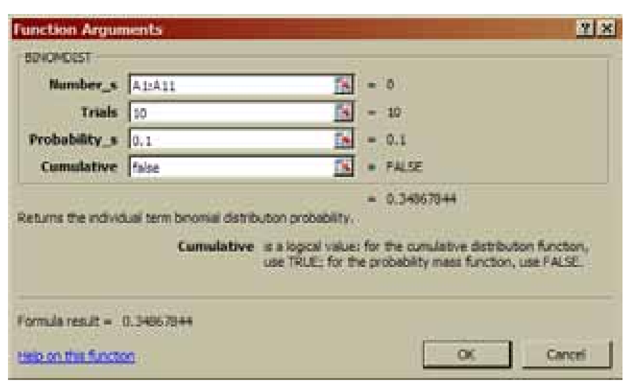
Calculating Binomial Probabilities with BINOMDIST

To obtain the probability of each possible outcome for a binomial distribution with n = 10 and p = 0.1, you will use the following commands. You must first create a column with the values for which you wish to find the corresponding probabilities. Input the values 0 to 10 into column A. Activate B1, then continue with:

Select the Formulas Tab, then
Choose: **Insert function, f_x > Statistical > BINOMDIST > OK**



Enter: Number_s: **A1:A11, or select cells**
Trials: **10**
Probability_s: **0.1**
Cumulative: **false > OK**
Drag: **fill handle down to give other probabilities**



This results in the following values:

	A	B
1	0	0.348678
2	1	0.38742
3	2	0.19371
4	3	0.057396
5	4	0.01116
6	5	0.001488
7	6	0.000138
8	7	8.75E-06
9	8	3.65E-07
10	9	9E-09
11	10	1E-10

Looking back to our original questions, to find the probability that three of your four light bulbs will be successes (last more than 700 hours) and one will fail we use:

"x-values" into column C (0, 1, 2, 3, 4) then activate cell D1 and select the Formulas tab:

Choose: **Insert function, f_x > Statistical > BINOMDIST > OK**

Enter: Number_s: **C1:C5, or select cells**

Trials: **4**

Probability_s: **0.85**

Cumulative: **false > OK**

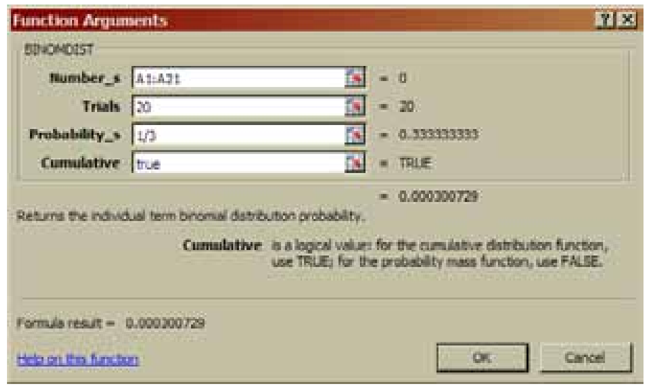
Drag: **fill handle down to give other probabilities**

	C	D
1	0	0.000506
2	1	0.011475
3	2	0.097538
4	3	0.368475
5	4	0.522006

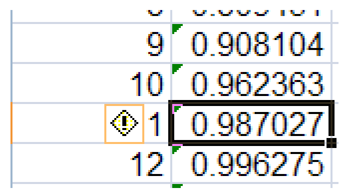
So we see that the probability of exactly 3 of the 4 bulbs being successes (lasting more than 700 hours) is .368475.

Cumulative Probabilities

The same statistical function BINODIST can be used to generate cumulative probabilities. A cumulative probability is the probability that your result will be less than or equal to a particular value. As an example, suppose we calculate the probability you will fail the test in advanced Greek Mythology. Here n = 20 and p = .3333. You will fail the test if you get less than or equal to 11 questions correct. (You will pass if you get 12 or more right.) The following commands can be used to calculate this probability:



Choose: **Insert function, f_x > Statistical > BINOMDIST > OK**
Enter: Number_s: **A1:A21, or select cells**
Trials: **20**
Probability_s: **1/3**
Cumulative: **true > OK**
Drag: **fill handle down to give other probabilities**



The cell directly next to the x-value 11 is the probability you will fail (answer 11 or fewer correct). So, what is the probability that you will pass? (1 - .9870 = .013)

Mean and Standard Deviation of the Binomial Distribution

Excel can be used like a calculator to determine the mean and standard deviation for the binomial distribution. Let's continue using the Greek Mythology test example, assuming we have X in column A, and binomial probabilities in column B (not the cumulative probabilities):

Activate cell C1, then
Enter: **=A1*B1**
Click and drag: **fill handle down to complete the column calculation**
Activate cell D1, then
Enter: **=A1*C1**
Click and drag: **fill handle down to complete the column calculation**
Activate cell C22 and click AutoSum and press enter (this is the mean)
Activate cell D22 and click AutoSum and press enter
(this is the sum of x^2P(x))
Finish the calculation for standard deviation as follows:
Activate any cell and enter **=SQRT(D22-C22^2)**

mean	std dev'n
6.666667	2.108185

6.67 is the mean number of questions expected to be answered correctly and 2.108 is the standard deviation expected among the number of questions answered correctly per test consisting of 20 questions.

ASSIGNMENT: Do Exercises 4.32, 5.36, 5.68, 5.69 in your text

CHAPTER 6 - LAB SESSION
NORMAL APPROXIMATION OF THE BINOMIAL

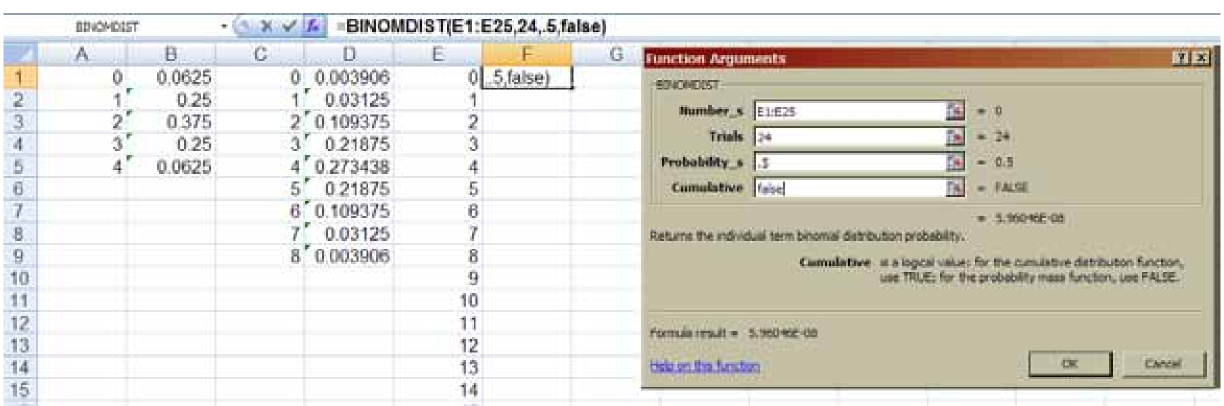
INTRODUCTION: The normal distribution is one of the most important distribution functions in statistics. We will now see how the binomial probabilities can be reasonably estimated by using the normal probability distribution. Later we will need to determine whether normality is a reasonable assumption. We will start our investigation with a few specific binomial distributions.

Step 1:Entering the data. For this demonstration we will use columns A, D, and G to hold a series of numbers. The corresponding probabilities will be placed into B, E and H. Enter the numbers 0, 1, 2, 3, and 4 into column A. Similarly, set D to the numbers 0, 1, ..., 8 and to set G to the numbers 0, 1, 2, ..., 24. These three columns will be used for three specific situations: n = 4, n = 8, and n = 24.

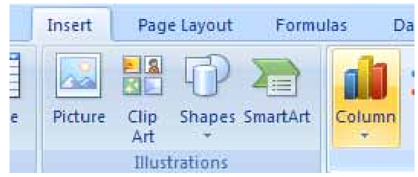
Step 2:Calculating and Storing the Probabilties. We will now place the binomial probabilities for A into B using BINOMDIST with n = 4 and p = .5.

Reminder on how to do this from Chapter 5 – Lab Session 1:
Activate cell B1, select the Formulas tab and continue with:
Choose: **Insert function, f_x > Statistical > BINOMDIST > OK**
Enter: Number_s: **A1:A5, or select cells**
Trials: **4**
Probability_s: **.5**
Cumulative: **false** > **OK**
Drag: **fill handle down to give other probabilities**

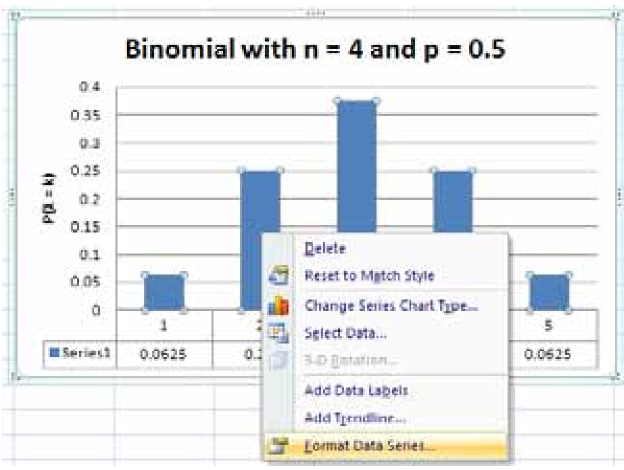
Place the binomial probabilities for D into E and G into H, being sure to use n = 8 and n = 24, respectively (keep p = 0.5.)

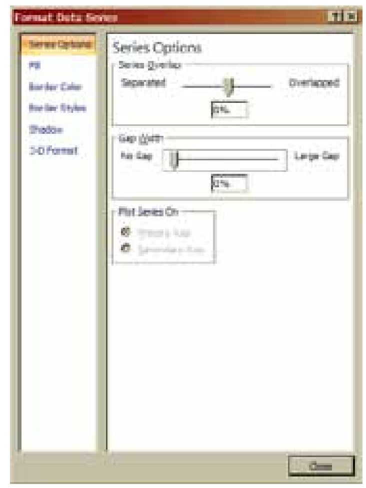


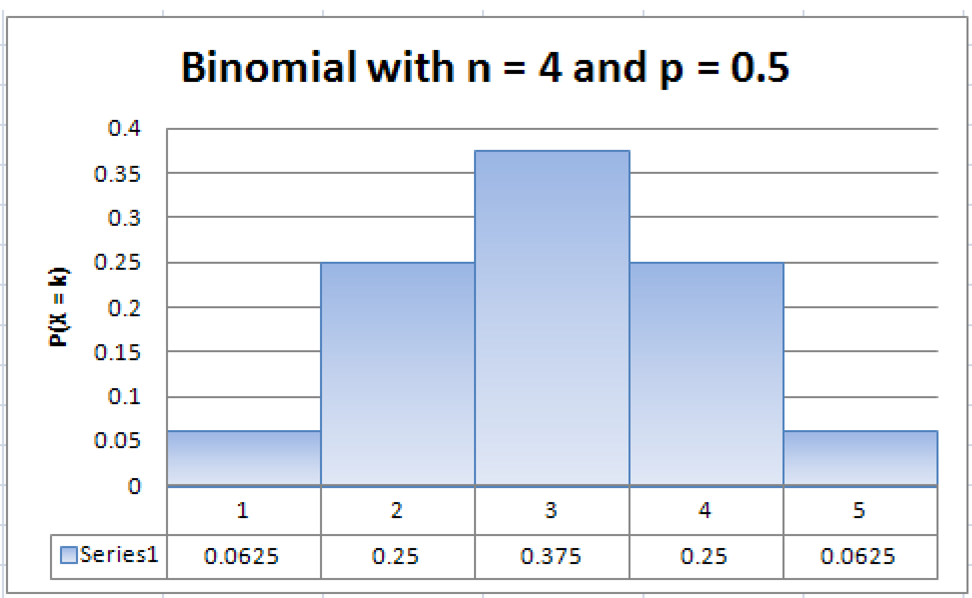
Step 3:Plotting the Probabilities Now we will plot each of the probabilities of x for 0 to n for n = 4 by using procedures identical to earlier constructions of charts. We will have to use a bar chart, because the Histogram option under Data Analysis does not allow construction of a histogram based on a probability distribution.



Highlight column B, then continue with
Select: **Insert > Column > 1st picture**
Enter: **appropriate titles > Next > Finish**
Chart can then be modified to remove the gaps
Right click on one of the bars of the chart,
select Format Data Series, and slide the gap width to 0%.







Repeat this procedure for plotting E versus D and H versus G. What can we say about the distribution as n becomes larger?

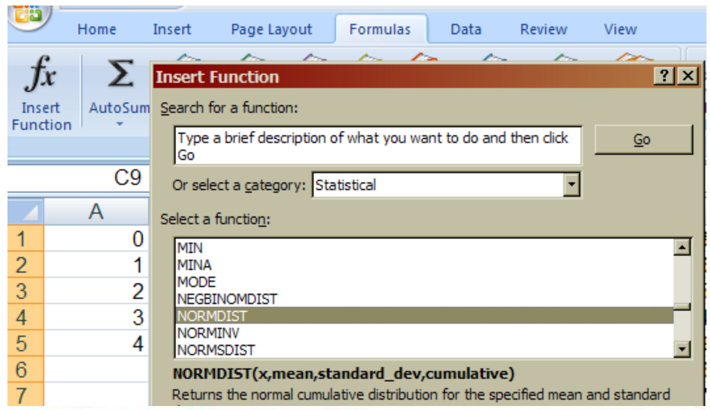
Step 4:Interpreting the results.

Let's see how the normal distribution approximates a binomial with p = .5 and n = 8. The approximating normal distribution has mean mu = 8(.5) = 4 and standard deviation sigma = sqrt((8)(.5)(.5)) = 1.414

First, we need to place the normal probabilities for each x (column C) into another column, say column F.

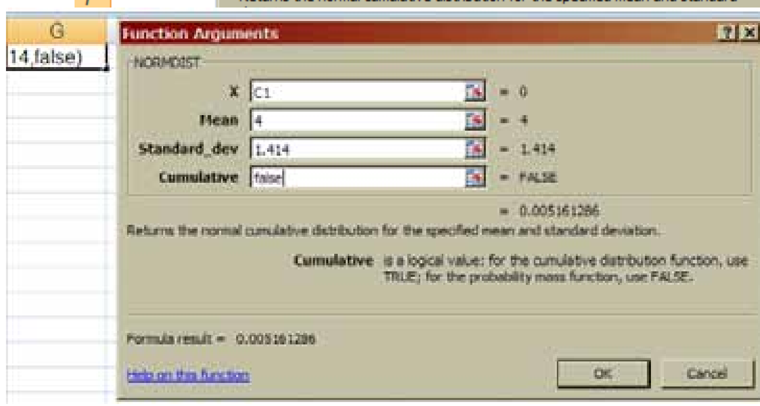
Activate cell F1, select the Formulas tab and continue with:

Choose: **Insert function, f_x >**
Statistical > NORMDIST>OK



Enter: X: **C1, or select cell**
Mean: **4**
Standard_dev: **1.414**
Cumulative: **FALSE**
OK

Click and drag: **fill handle to generate normal probability for each x**

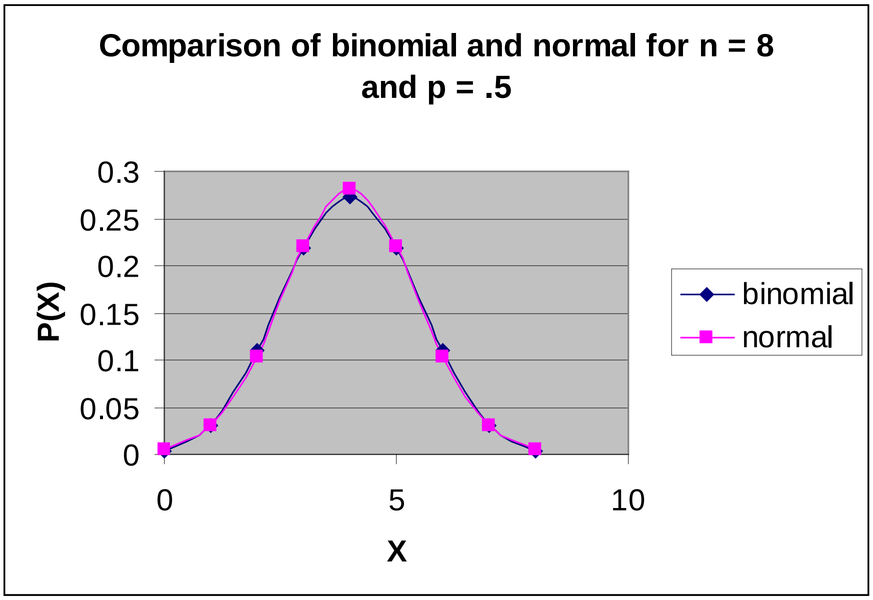


F
0.005161
0.029717
0.103761
0.219712
0.282137
0.219712
0.103761
0.029717
0.005161

To draw the graph of a the normal probability curve along with a binomial probability curve, activate cells C1 through F9, then continue with

Choose: **Insert** > **Scatter** > **2nd picture**

And edit titles appropriately



The chart just executed plotted the probability distribution function for the binomial and for the normal approximation on the same axes. This will help us see why we can approximate a binomial by a normal and how to do the appropriate calculations.

You should visualize the histogram corresponding to the binomial probabilities. The height of a bar is the probability the binomial variable is equal to the corresponding value. For example, the height of the bar centered at 5 is the probability the binomial is equal to 5. The base of a bar is 1 unit wide. Therefore, the area of a bar is equal to its height, and is thus equal to the corresponding probability.

Also visualize the normal curve.

Here are some calculations that will help the explanation. Suppose we want the probability that the binomial variable has a value from 5 to 7. This probability is the sum of the probabilities at 5, 6, and 7. (Look in Rows 6, 7 and 8 in column E: the sum is 0.359375) The area under the normal curve that goes from 4.5 to 7.5 approximates the area of the three binomial bars. How could we determine this area?

The probability the binomial variable has a value from 5 to 7 is .359375 . The approximation obtained from the normal probability distribution is .353205 without continuity correction, which is very close to the true probability. If we were to use a normal approximation for a binomial with p = .5 and n = 24 (like in columns G and H), the approximation would look even better. In the exercises, we'll look at other values of p.

ASSIGNMENT:

1. (a) Make plots as in the first part of the lab, but use p = .4 instead of p = .5. Use n = 4, 8 and 24.
   (b) Repeat part (a) using p = .2.
   (c) What can you say about the normal approximation to the binomial? For what values of n and p does it seem to work best?

2. Suppose X has a binomial distribution with p = .8 and n = 25. Use Excel to calculate each of the probabilities below exactly. Also compute the normal approximation to these probabilities. Compare the binomial results with the normal approximations.
   (a) $P(X = 21)$
   (b) $P(X \leq 21)$
   (c) $P(X > 24)$
   (d) $P(21 \leq X \leq 24)$

3. Do Exercises 6.103 and 6.133 in your text

CHAPTER 7 - LAB SESSION
SAMPLE VARIABILITY

INTRODUCTION: In an effort to predict population parameters, we need to investigate the variability in the sample means obtained from repeated sampling. The Central Limit Theorem tells us that the sampling distribution of sample means, $\bar{x}$, is approximately normally distributed. In the following lab you will test the results of the Central Limit Theorem.

GENERATING THE DISTRIBUTIONS OF SAMPLE MEANS

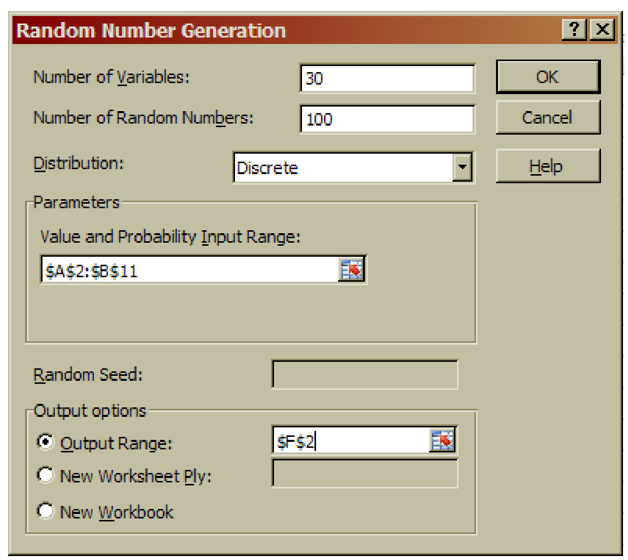
Uniform Distribution

Enter the values 0 through 9 into column A and name column A 'X': Enter the probabilities into column B. For the uniform distribution assign probabilities of .1 to the x-values 0 through 9. Name column 2 'UNIFORM':

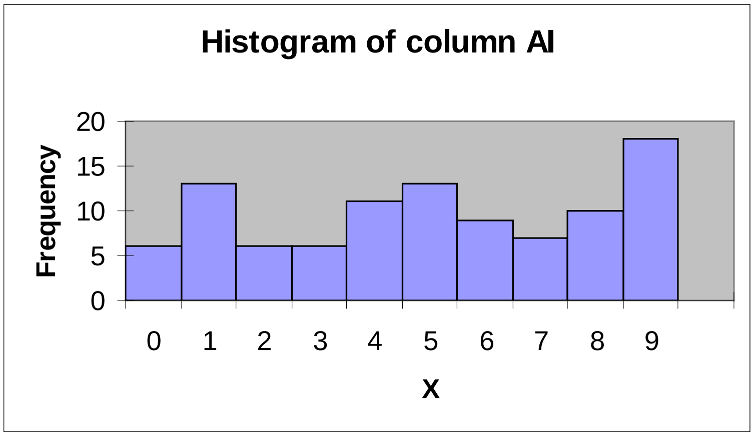
A	B
X	UNIFORM
0	0.1
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1

Generate 30 sets of 100 uniform deviates (random numbers with a uniform distribution) and store them in columns F through AI. (Reminder:
Data > Data Analysis > Random Number Generation > OK

Number of Variables: **30**
Number of Random Numbers: **100**
Distribution: **Discrete**
Value and Probability Input Range: **A2:B11**
Output Range: F2
OK



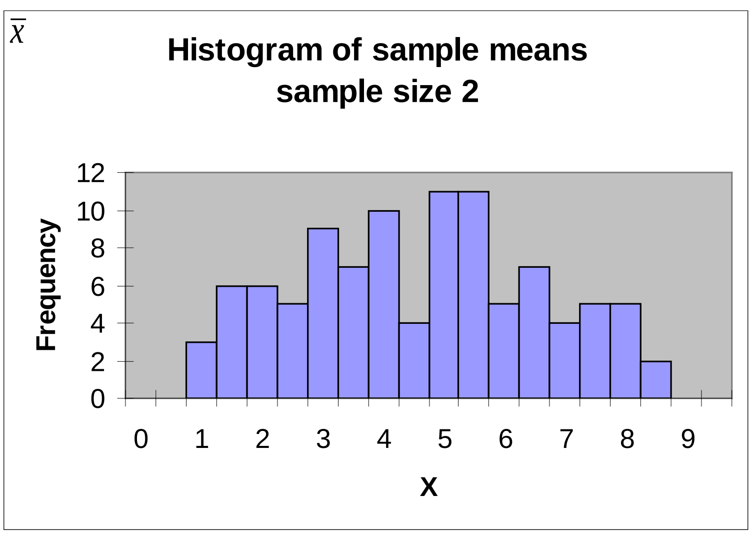
Observe the distribution of the data in AI.



To illustrate the concept of a sampling distribution we're considering the finite population {0, 1, 2, ..., 9}. We shall generate values from three very different distributions and investigate, empirically, sampling distributions of the sample means, $\overline{x}$, for samples of size n=2, n=5, and n=30 for each of the different distributions.

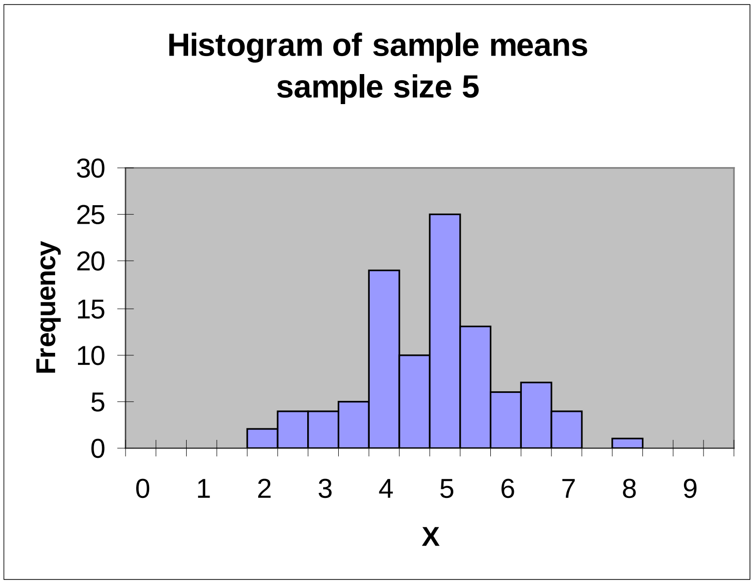
(N=2) Calculate the sample mean, $\overline{x}$, for each pair of values given in columns F and G and store in column AJ:

Observe the distribution of the sample means in column AJ:



Notice that this distribution of sample means does not look like the population.

(N=5) Calculate $\bar{x}$ for the values in H through L, storing your results in AK, then observe the distribution of the sample means in AK.



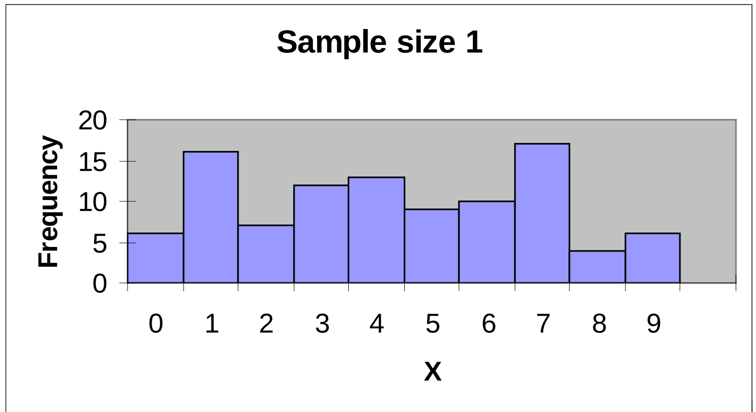
(N=30) Repeat the above procedure for the values in columns F through AJ, storing your results in AL.

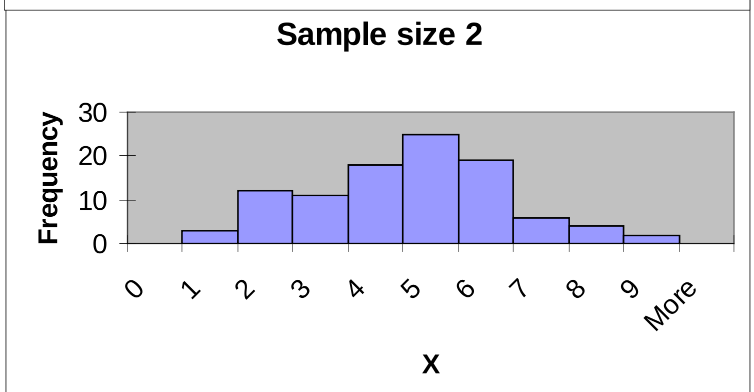
Compare the descriptive statistics and distributions for each of the calculated means

Distribution of sample mean

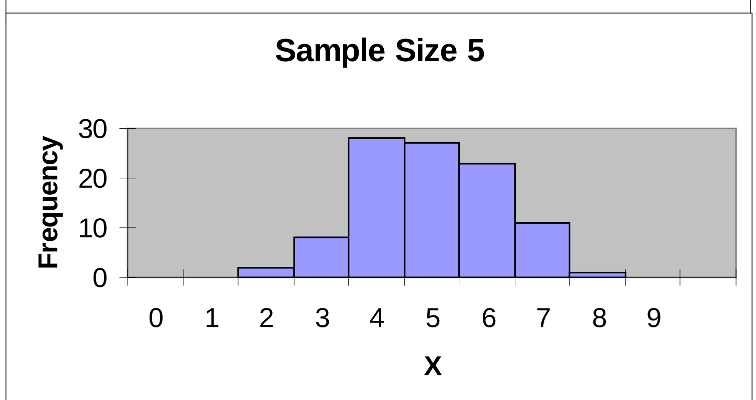
sample size 2		*Sample size 5*		*Sample size 30*	
Mean	4.375	Mean	4.604	Mean	4.498667
Standard Deviation	1.73987	Standard Deviation	1.255285	Standard Deviation	0.460905
Range	8	Range	6.2	Range	1.966667
Minimum	0.5	Minimum	1.2	Minimum	3.6
Maximum	8.5	Maximum	7.4	Maximum	5.566667

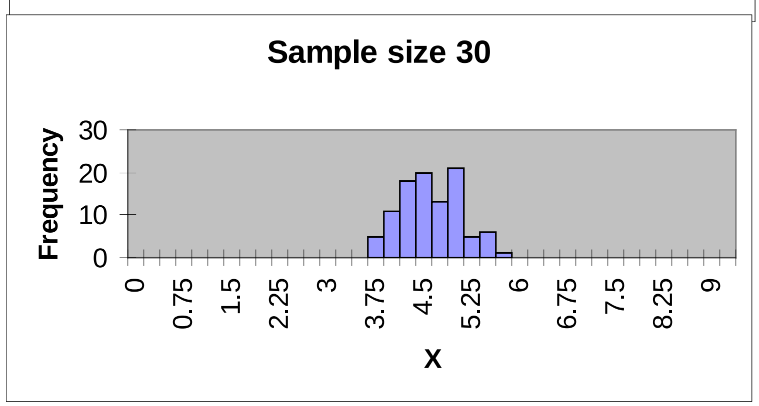
Now, look at the distribution of sample means for samples of size 1(column F), size 2 (column AJ), size 5 (column AK), and size 30 (column AL) graphically (using the same scale for each):





Note the shape of each of the distributions of the sample means. These distributions don't look like the original data (F), but they do have a shape we're familiar with.





J-Shaped Distribution

Enter the following probabilities into column C: .39 .26 .22 .18 .15 .13 .12 .10 .05 .02 and repeat the previous procedure.

U-Shaped Distribution

Enter the following probabilities into column D: .18 .15 .09 .06 .02 .02 .06 .09 .15 .18 and repeat the previous procedure.

Questions:

1. What are the parameter values for each of the three distributions?
2. What happened to the means and standard deviations of the $\bar{x}$'s as n got larger?
3. How did the distributions of $\bar{x}$'s compare to the normal distribution as n got larger? Were the results similar for the different distributions?
4. Do Exercises 7.9, 7.15, 7.40, 7.45 and 7.46 in your text.

CHAPTER 8 - LAB SESSION
ESTIMATION AND HYPOTHESIS TESTING

INTRODUCTION: Two indispensable statistical decision-making tools for a single parameter are (i)confidence intervals, and (ii) hypothesis tests to investigate theories about parameters. In this lab you will learn how to calculate confidence intervals and perform hypothesis tests (assuming we know sigma) using Excel.

CONFIDENCE INTERVALS
As an introduction, let's follow Example 8.4 in your text.

Begin a new worksheet and generate 40 random integers the range 0 to 9 in column A. You can use either the Random Numbers Table (Table 1) or the Random Number Generation tool of Excel and then use the INT (integer) function to transform to integers in the range 0 to 9:

Choose: **Data > Data Analysis Plus > Random Number Generation > OK**
Enter: Number of Variables: **1**
Number of Random Numbers: **40**
Select: Distribution: **Uniform**
Enter: Parameters, Between **0** and **10**
Output Range: **B1** > **OK**
Enter: in cell A1: **=INT(B1)**
Click and drag: lower right corner to cell A40

To see the mean, standard deviation and maximum and minimum values for the data set use:
Select: **Data > Data Analysis > Descriptive Statistics > OK**
enter input and output range as appropriate, and select Summary Statistics

Column1	
Mean	5
Standard Error	0.457137
Median	5
Mode	8
Standard Deviation	2.891189
Sample Variance	8.358974
Kurtosis	-1.35942
Skewness	-0.12062
Range	9
Minimum	0
Maximum	9
Sum	200
Count	40

(Your results may be slightly different, since we are using random data.)

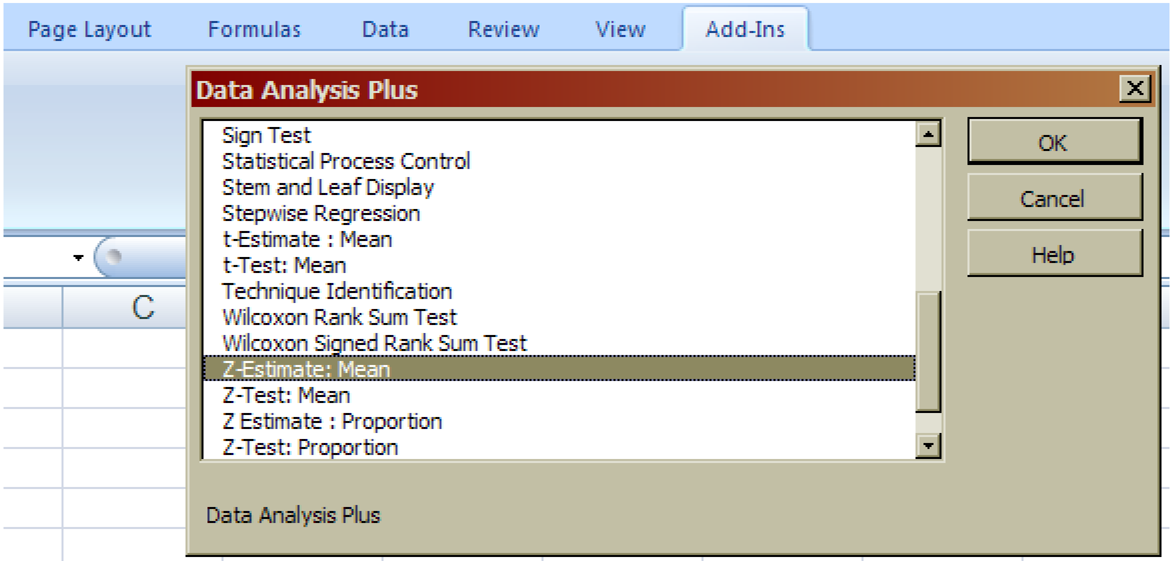
Find the 90% confidence interval for the mean of these values we generate in column A:

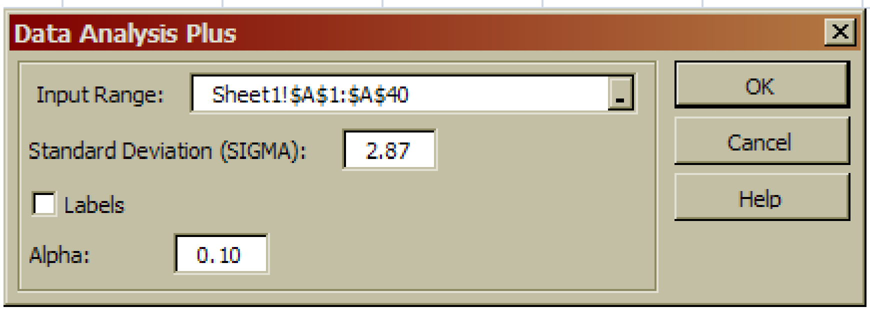
Choose: **Data** > **Data Analysis Plus** > **Z-Estimate: Mean** > **OK**

Enter: Input Range: **A1:A40**

Standard Deviation (SIGMA): **2.87** > **OK**

Alpha : **.10** > **OK**





z-Estimate: Mean

	Column 1
Mean	5
Standard Deviation	2.8912
Observations	40
SIGMA	2.87
LCL	4.253587
UCL	5.746413

So the 90% confidence interval for the mean is 4.25 to 5.74.

Find the 95% and 99% confidence intervals for the mean of this same set of data and record the results.

Looking at these three intervals

1. Consider the means obtained from 100 samples of size 40. If these means were used to construct 100 confidence intervals, determine the expected number of times the population mean would be included in one of these intervals.

2. In the 99% confidence interval that you found, the level of significance is 99%. What is the value of a ? What does a represent?

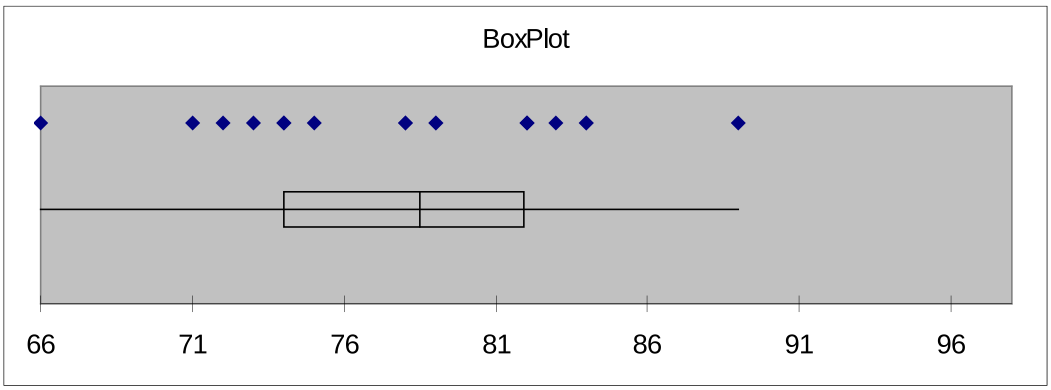
3. In which of these intervals is the maximum error, E, the smallest? What does this mean? In which of these intervals are you being more certain to include the population mean?

HYPOTHESIS TESTING

A standard final examination in an elementary statistics course is designed to produce a mean score of 75 and a standard deviation of 12. The hypothesis you will try to verify is: "This particular statistics class is above average." At the .05 level of significance, test the claim that the following sample scores reflect an above-average class (assuming sigma = 12):

79	79	78	74	82	89	74	75	78	73
74	84	82	66	84	82	82	71	72	83

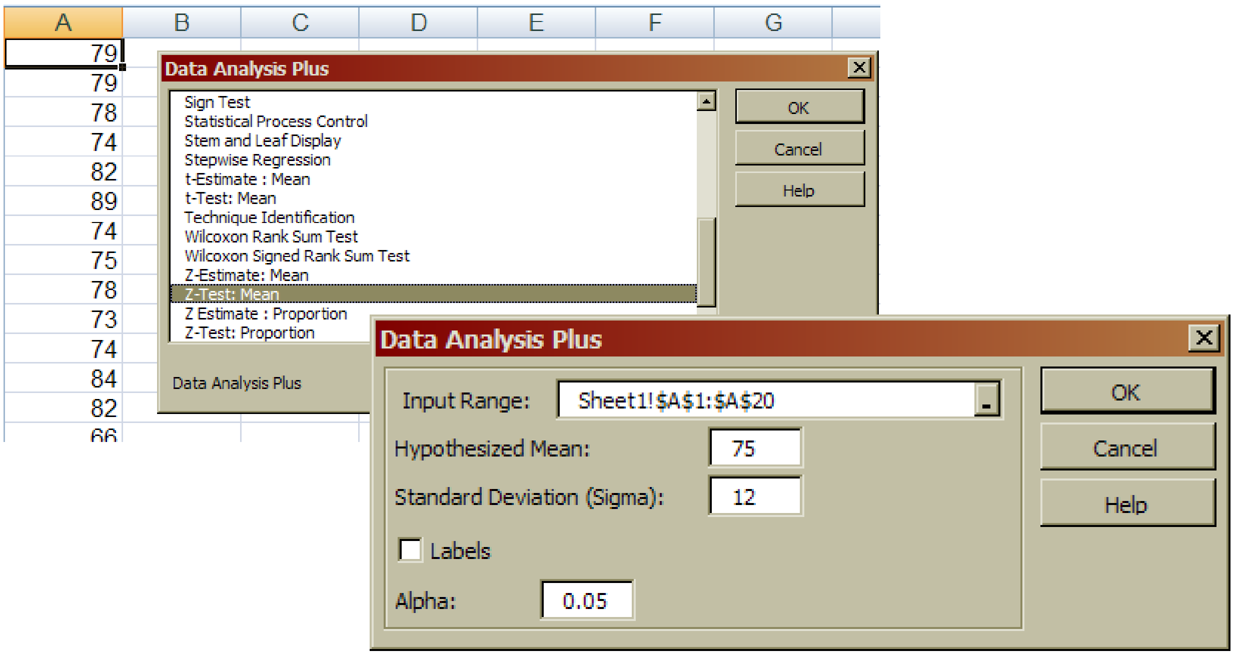
Enter the data and get a preliminary graphical analysis.



Column1	
Mean	78.05
Standard Error	1.251263
Median	78.5
Mode	82
Standard Deviation	5.595816
Sample Variance	31.31316
Range	23
Minimum	66
Maximum	89
Sum	1561
Count	20

Test the hypothesis, "The mean test grade for this class is greater than 75."

Choose: **Add-Ins** > **Data Analysis Plus** > **Z-Test: Mean** > **OK**
Enter: Input Range: **A1:A20 or select cells** > **OK**
Hypothesized mean: **75**
Standard Deviation (SIGMA): **12** > **OK**
Alpha: **.05** > **OK**



The results are as follows:

Z-Test: Mean

	Column 1
Mean	78.05
Standard Deviation	5.5958
Observations	20
Hypothesized Mean	75
SIGMA	12
z Stat	1.1367
P(Z<=z) one-tail	0.1278
z Critical one-tail	1.6449
P(Z<=z) two-tail	0.2556
z Critical two-tail	1.96

Note that the p-values and critical values for both one-tail and two-tail tests are given.

Questions:

1. What are the formal null and alternative hypotheses?

2. What is the value of the test statistic, and what is your decision? Is the mean of this class above “average”?

ASSIGNMENT: Do Exercises 8.41, and 8.115 in your text, and the following two problems.

1. In one region of a city, a random survey of households includes a question about the number of people in the household. The results are given in the accompanying frequency table. Construct the 90% confidence interval for the mean size of all such households. Assume that the sample standard deviation can be used as an estimate of the population standard deviation.

Household size	1	2	3	4	5	6	7
Frequency	15	20	37	23	14	4	2

2. An aeronautical research team collects data on the stall speeds (in knots) of ultralight aircraft. The results are summarized in the accompanying stem-and-leaf plot. Construct the 95% confidence interval for the mean stall speed of all such aircraft. Assume sigma = 1.

MTB > Stem-and-Leaf c1.

Stem-and-leaf of C1   N = 16
Leaf Unit = 0.10

21. | 7 8
22. | 3 4 4 6
23. | 2 2 5 8 9 9
24. | 0 1 3
25. | 2

CHAPTER 9 - LAB SESSION 1
ANALYZING MEAN (SIGMA UNKNOWN)

INTRODUCTION: The t-statistic is used when making inferences concerning the population mean when sigma is an unknown quantity. We will introduce the t-test and compare the z and t distributions.

THE CONFIDENCE INTERVAL

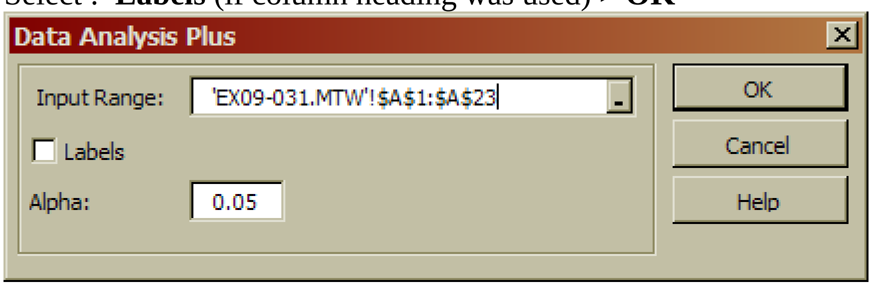
To generate a confidence interval using the t-statistic we use Inference About a Mean command, specifying the level of confidence and the column of data for which the estimation is being made.

Consider the data presented in exercise 9.31[EX09-031] of your text. Open the data file. Before we complete a 95% confidence interval estimate for the mean length of lunch breaks at Giant Mart, we check the normal probability plot and boxplot to verify the normality assumption. Excel uses a test for normality, not the probability plot.

Choose: **Add-Ins** > **Data Analysis Plus** > **Chi-Squared Test of Normality** > **OK**

Enter: Input Range: **select cells**

Select : **Labels** (if column heading was used) > **OK**



Chi-Squared Test of Normality

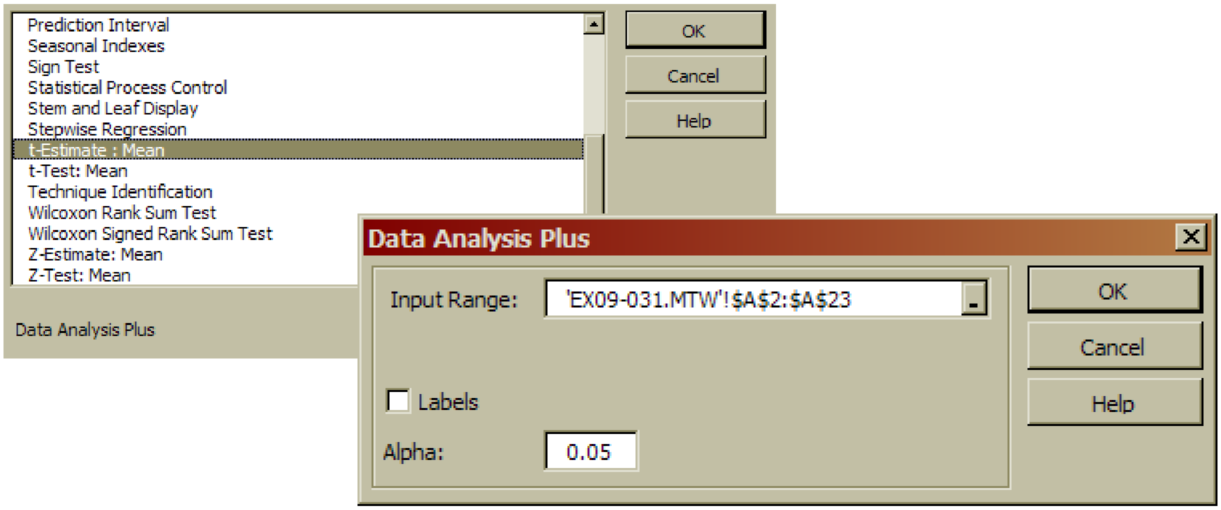
	Time (min)
Mean	29.31818182
Standard deviation	4.9221
Observations	22

Intervals	Probability	Expected	Observed
(z <= -1)	0.158655	3.49041	4
(-1 < z <= 0)	0.341345	7.50959	10
(0 < z <= 1)	0.341345	7.50959	4
(z > 1)	0.158655	3.49041	4

chi-squared Stat	2.6149	
df	1	
p-value	0.1059	*p*-value greater the .05, given distribution approximately normal
chi-squared Critical	3.8415	

To complete a 95% confidence interval estimate for the mean length of lunch breaks at Giant Mart complete the following steps:

Choose: **Add-Ins > Data Analysis Plus > t-Estimate:Mean > OK**
Enter: Input Range: **A1:A22**
Enter: Alpha: **.05 > OK**



This results in the following output, which appears in a new worksheet.

t-Estimate: Mean

	Column 1
Mean	29.3182
Standard Deviation	4.9221
LCL	27.13584
UCL	31.50053

So we have:
With 95% confidence we estimate the mean length of lunch breaks at Giant Mart to be between 27.14 and 31.50 minutes.

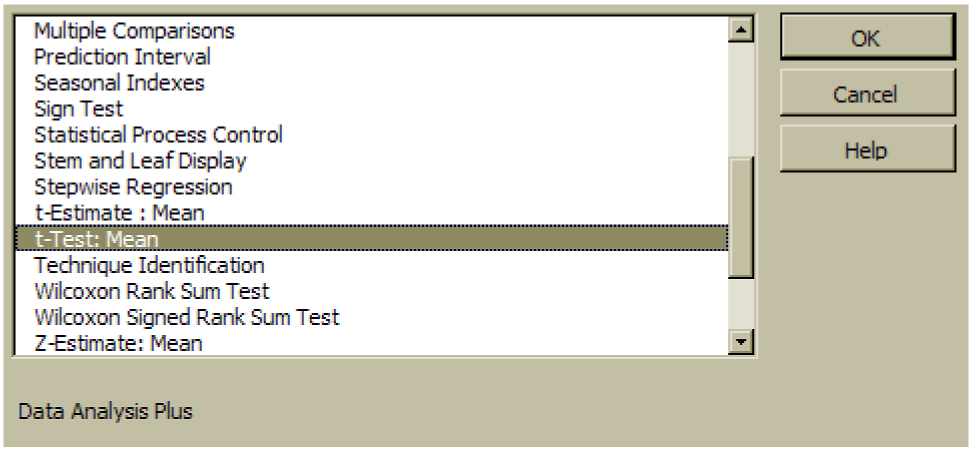
THE TTEST

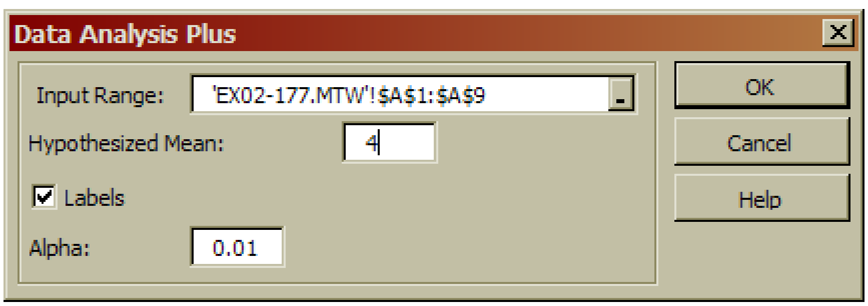
Using text exercise 9.29[EX02-177] as the basis of our discussion, open the data file. Suppose we have been asked to determine whether this accelerator has decreased the drying time by significantly more then 4% at the 0.01 level. The hypotheses to be tested are:

H_0: $\mu = 4.0$
H_a: $\mu > 4.0$

To perform the test, use the following commands:

Choose: **Add-Ins > Data Analysis Plus > t-Test Mean > OK**
Enter: Input Range: **A2:A9 > OK**
Hypothesized mean: **4**
Alpha: **0.01 > OK**





The output appears on a new worksheet as follows:

t-Test: Mean

	Column 1
Mean	4.5625
Standard Deviation	1.3405
Hypothesized Mean	4
Df	7
t Stat	1.1869
P(T<=t) one-tail	0.137
t Critical one-tail	2.9979
P(T<=t) two-tail	0.274
t Critical two-tail	3.4995

Is there sufficient evidence to show that this accelerator has decreased the drying time significantly more than 4% at the .01 level?

As another example consider the point spread between opposing teams in the 1996 bowl games :
5 20 19 33 6 10 7 18 29 41 6 32 9 36.

Enter the data into Column A of a new worksheet.

Test the hypothesis, "The average spread between the scores of the winning and the losing teams in a college bowl game is less than 20." Assume sigma is unknown.

Use the same commands as above to get the following output:

t-Test: Mean

	Column 1
Mean	19.3571
Standard Deviation	12.7013
Hypothesized Mean	20
df	13
t Stat	-0.1894
P(T<=t) one-tail	0.4264
t Critical one-tail	2.6503
P(T<=t) two-tail	0.8528
t Critical two-tail	3.0123

Questions:

1 What are the formal null and alternative hypotheses?

2. What is the value of the test statistic, and what is your decision if $\alpha = .10$? Is the final point spread of a bowl game less than 20?

3. What does the size of the p-value tell us?

ASSIGNMENT: Do Exercises 9.56, 9.60 in your text

COMPARISON OF THE Z AND T DISTRIBUTION

Why do you use two different distributions depending on the availability of the standard deviation, σ ? What basic assumptions are necessary to use the t-statistic? Is the basic assumption that the parent population is normally distributed a necessary one? Why? If the parent population is not known to be normally distributed, when can we use the t-statistic? In this exercise you will generate both types of statistics from the same 100 samples and be able to compare the two empirical distributions.

In a new workbook, generate 100 samples of size 5 from a normal distribution with μ =15 and σ =10, and store the mean and standard deviation of each of the 100 samples.

Choose: **Data** > **Data Analysis** > **Random Number generation** > **OK**
Enter: Number of Variables: **5**
Number of Random Numbers: **100**
Distribution: **Normal**
Mean: **15**
Standard Deviation: **10**
Select: Output Options: **Output Range** > **A1** > **OK**

This will make 5 columns of 100 random numbers each.

Calculate the Mean and Standard Deviation of each row and place them in columns F and G. (Do this for row 1, and click and drag to fill the remainder.)

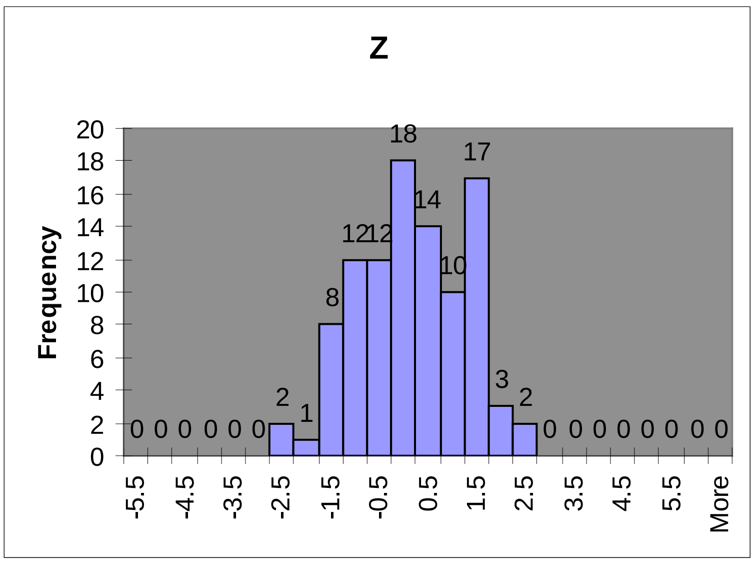
Calculate both *z* and *t* statistics of each row and place them in H and I.

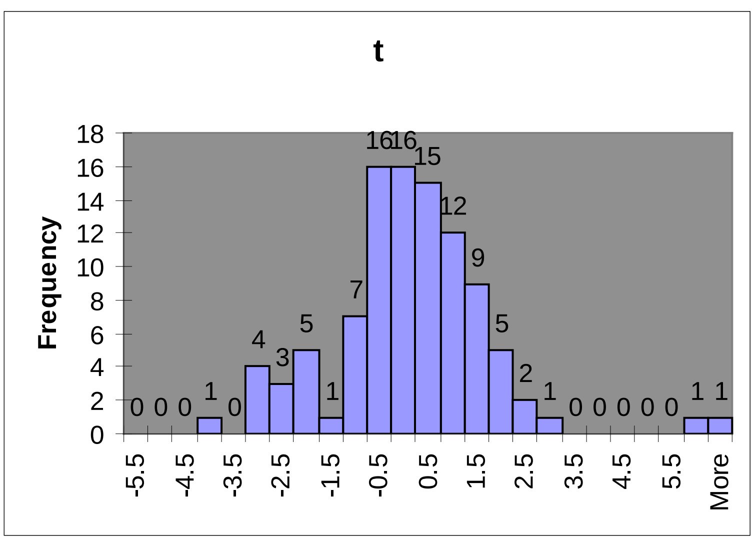
Recall: $z = \dfrac{\overline{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ and $t = \dfrac{\overline{x} - \mu}{\dfrac{s}{\sqrt{n}}}$

Replicate these for all 100 rows by highlighting and dragging the lower right corner.

For each of the two statistics, z and t, count the number of times their value is more than 2 units away from the origin.

Compare the two distributions graphically by using histograms (recall the method from Lab 2)





QUESTIONS:

1. How many of the calculated *z*-statistics were more than two units away from the origin? How many of the *t*-statistics?

2. What did the distributions for the two statistics look like? Compare their centers, spread, and overall shape.

3. Would you describe the *t*-distribution as bell-shaped? If so, would you say it is approximately normal?

4. If you were to increase n, would you expect the difference between the two distributions to increase or decrease?

ASSIGNMENT: Do Exercise 9.64 in your text.

CHAPTER 9 - LAB SESSION 2
ANALYZING THE POPULATION PROPORTION

INTRODUCTION: In this lab we will investigate the inferences that can be made about the binomial parameter *p*. Inferences concerning the population binomial parameter *p* are made using procedures that closely parallel the inference procedures for the population mean μ (see Chapter 9 – Lab Session 1).

CONFIDENCE INTERVALS

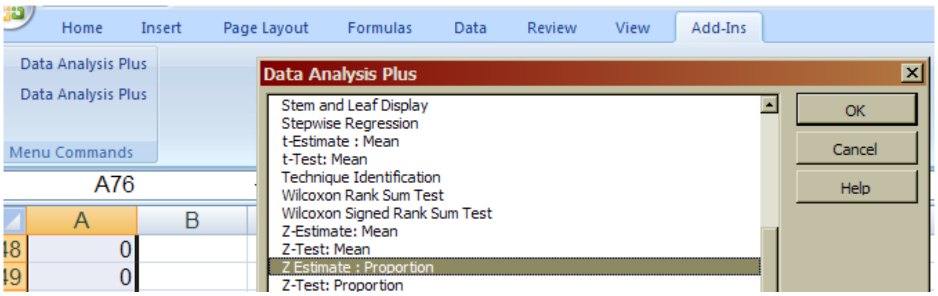
Consider the following sample problem.

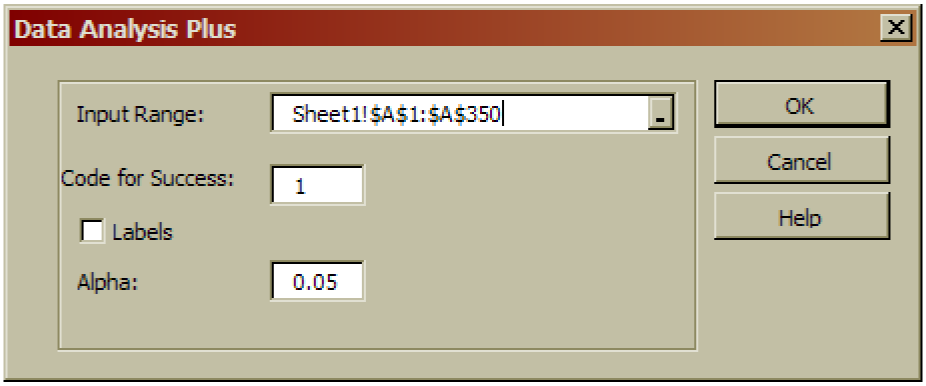
A telephone survey was conducted to estimate the proportion of households with a personal computer. Of the 350 households surveyed, 75 had a personal computer. Give a point estimate for the portion of the population that had a personal computer. Give the 95% confidence interval.

The data to be entered will be a series of 0's and 1's, each number designating one of two categories. Since the parameter of concern is the proportion of households with a personal computer, we use 1 to represent 'has a personal computer' and use 0 to represent 'does not have a personal computer'.

To enter the data:
Enter: **1** in Cell A1
Drag: **Lower right corner down** to Cell A75
Enter **0** in Cell A76
Drag: **Lower right corner down** to Cell A350

Finally, determine a 95% confidence interval for *p*:
Choose: **Add-Ins > Data Analysis Plus> Z-Estimate: Proportion > OK**
Enter: Input Range: **A1:A350 > OK**
Code for Success: **1**
Alpha: **.05 > OK**





The output looks like this:

z-Estimate: Proportion

	Column 1
Sample Proportion	0.2143
Observations	350
LCL	0.1713
UCL	0.2573

So our 95% confidence interval estimate for proportion of households that have a personal computer is 17.13% to 25.73%

HYPOTHESIS TESTING

This sample problem will take you through the steps of entering the data and performing a hypothesis test for exercise 9.105 in your textbook.

Since the parameter of concern is the proportion of claims settled within 30 days, we'll let 1 represent 'claim settled within 30 days' and 0 represent 'claim not settled within 30 days'.

Enter the data as before:

Enter: **1** in Cell A1
Drag: **Lower right corner down** to Cell A55
Enter **0** in Cell A56
Drag: **Lower right corner down** to Cell A75

The hypotheses for this test are H_0: $p = .9$ vs H_a: $p < .9$

To test the hypothesis :

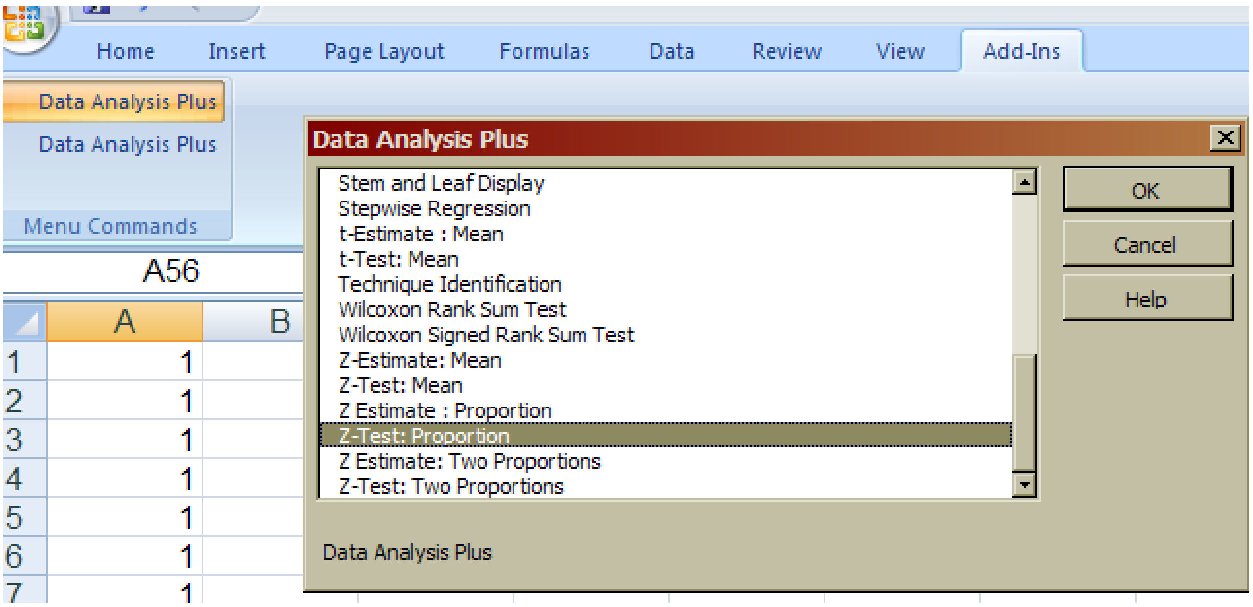
Choose: **Add-Ins > Data Analysis Plus> Z-Test: Proportion > OK**

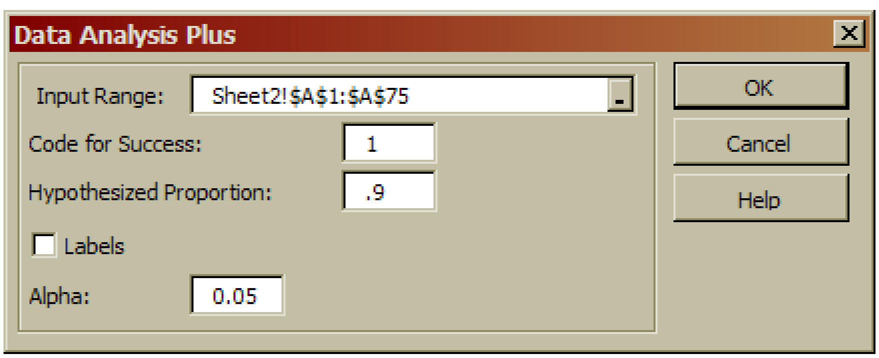
Enter: Input Range: **A1:A75 > OK**

Code for Success: **1**

Hypothesized Proportion: **.9**

Alpha: **.05** > **OK**





This creates the following output on a new sheet.

z-Test: Proportion

	Column 1
Sample Proportion	0.7333
Observations	75
Hypothesized Proportion	0.9
z Stat	-4.8113
P(Z<=z) one-tail	0
z Critical one-tail	1.6449
P(Z<=z) two-tail	0
z Critical two-tail	1.96

1. What decision should be made based on these results?
2. What does p value = 0.0 tell us?

ASSIGNMENT: Do Exercises 9.107, 9.109 in your text.

CHAPTER 9 - LAB SESSION 3
ANALYZING THE POPULATION VARIANCE

INTRODUCTION: In this lab we will present the hypothesis test for the standard deviation for a normal population. When sample data are skewed, just one outlier can greatly affect the standard deviation. It is very important, especially when using small samples, that the sampled population be normal; otherwise the procedures are not reliable. However, unlike the analysis for the mean you will not have convenient computer commands to help you.

To use Example 9-19 as an example of using Excel to aid in completion of the hypothesis test, let's assume the 12 samples tested yielded the following data:

165	172	180	189	181	173
167	192	212	169	198	171

Enter the data into Column A.
Determine the descriptive statistics by the following:
Choose: **Data** > **Data Analysis** > **Descriptive Statistics**

This gives you the following:

	Column1	
165		
172		
180	Mean	180.75
189	Standard Error	4.152627865
181	Median	176.5
173	Mode	#N/A
167	Standard Deviation	14.38512489
192	Sample Variance	206.9318182
212	Kurtosis	0.37412902
169	Skewness	1.005775368
198	Range	47
171	Minimum	165
	Maximum	212
	Sum	2169
	Count	12
	Confidence Level(95.0%)	9.139876928

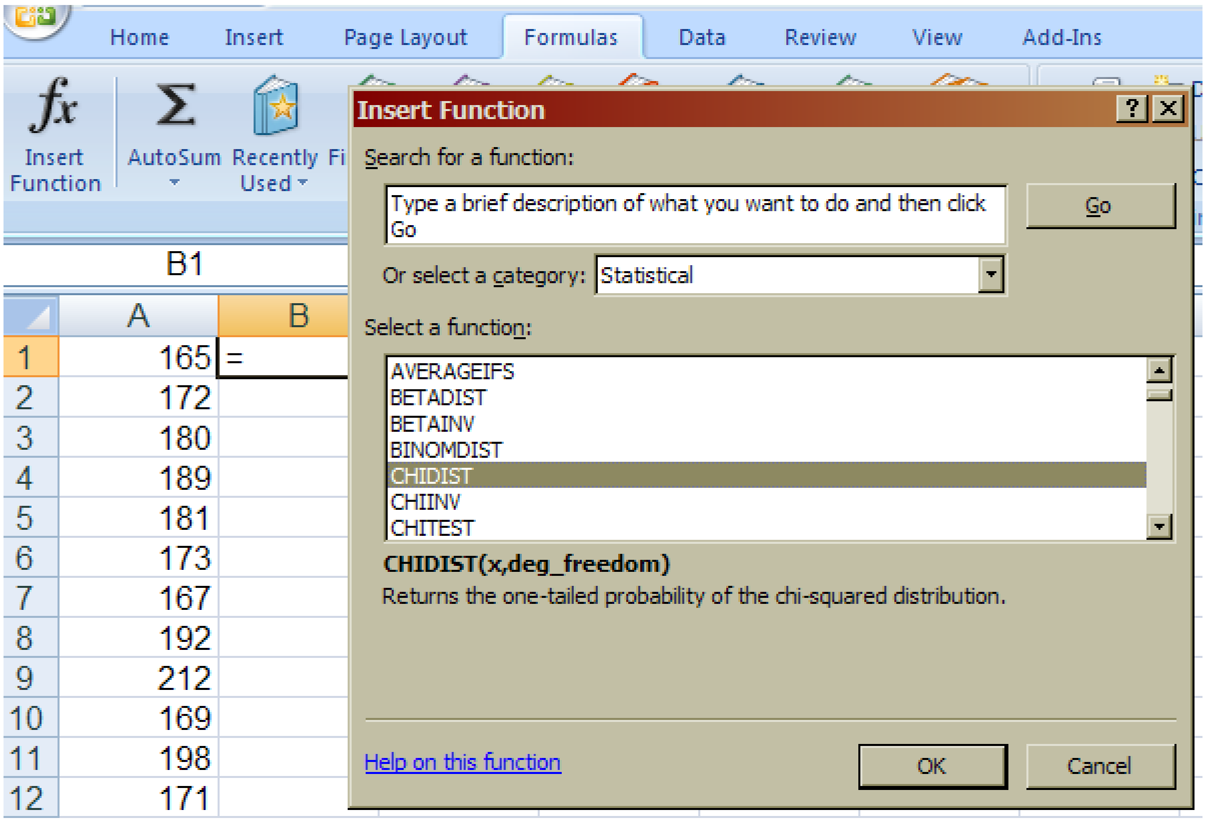
From the table we see that n = 12, s = 14 and we calculate $X^{2*} = 21.56$

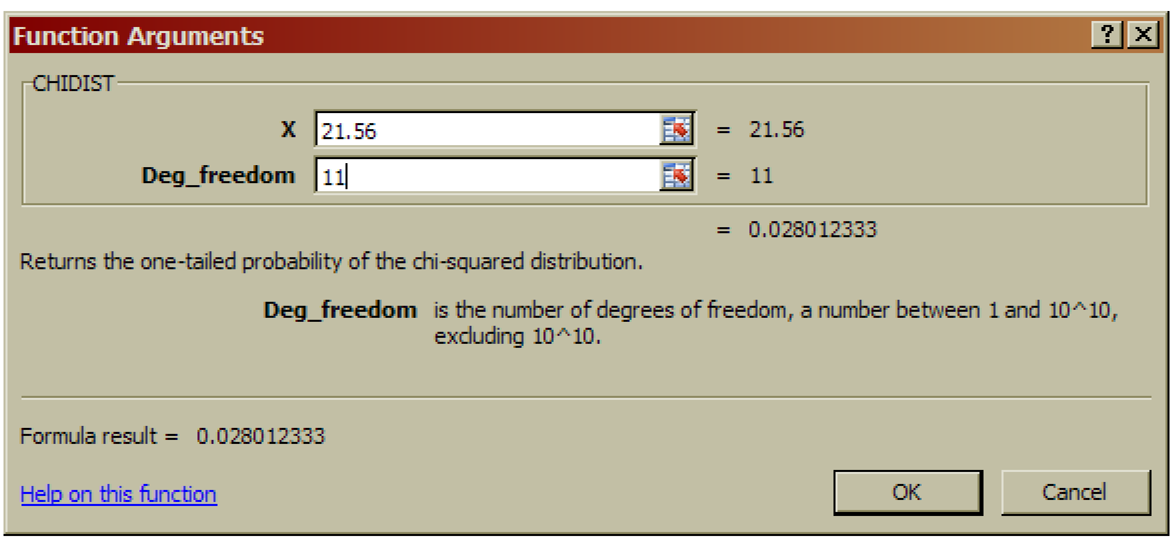
To calculate the p-value, activate Cell B1, select the Formulas Tab and continue with

Choose: ***fx* Insert Function > Statistical > CHIDIST > OK**

Enter: X^2*: **21.56**

Df: **11 > OK**





This gives you the value 0.0280.

Recall that the manufacturer claims "shelf life" is normally distributed.
Why is this important?

What decision should be made?
Does your conclusion match that for Example 9-16?

ASSIGNMENT: Do Exercises 9.137 and 9.144 in your text.

Use the following data for 9.137

31.6	31.9	32.6	31.9	31.5	32.5	32.0	32.2	31.9	32.0
32.2	31.8	31.8	32.3	31.1	31.8	31.5	31.7	31.8	31.8

CHAPTER 10 LAB SESSION
INFERENCES INVOLVING TWO POPULATIONS

INTRODUCTION: When comparing two populations we need two samples, one from each population. Two kinds of samples can be used: dependent or independent, determined by the source of the data. The methods of comparison are quite different.

CASE 1. DEPENDENT SAMPLE (PAIRED DATA):

The two data values, one from each set, that come from the same source are called paired data. They are compared by using the difference in their values, called the paired difference, d. Because the distribution of the paired difference, $d = x_1 - x_2$, will be approximately normally distributed when paired observations are randomly selected from normal populations, we will use the t-test. We wish to make inferences about μ_d where the random variable (d) involved has an approximately normal distribution with an unknown standard deviation (σ_d).

Confidence Interval

Consider the data presented in exercise 10.16 of your text. Use Excel to generate the 95% confidence interval for the mean improvement in memory resulting from taking the memory course. (d = after - before).

Retrieve the data file for Exercise 10-016 or enter it yourself in columns A and B.

Form the paired difference and put it in Column C.

	A	B	C
1	**Before**	**After**	**Difference**
2	93	98	5
3	86	92	6
4	72	80	8
5	54	62	8
6	92	91	-1
7	65	78	13
8	80	89	9
9	81	78	-3
10	62	71	9
11	73	80	7

To generate the interval:

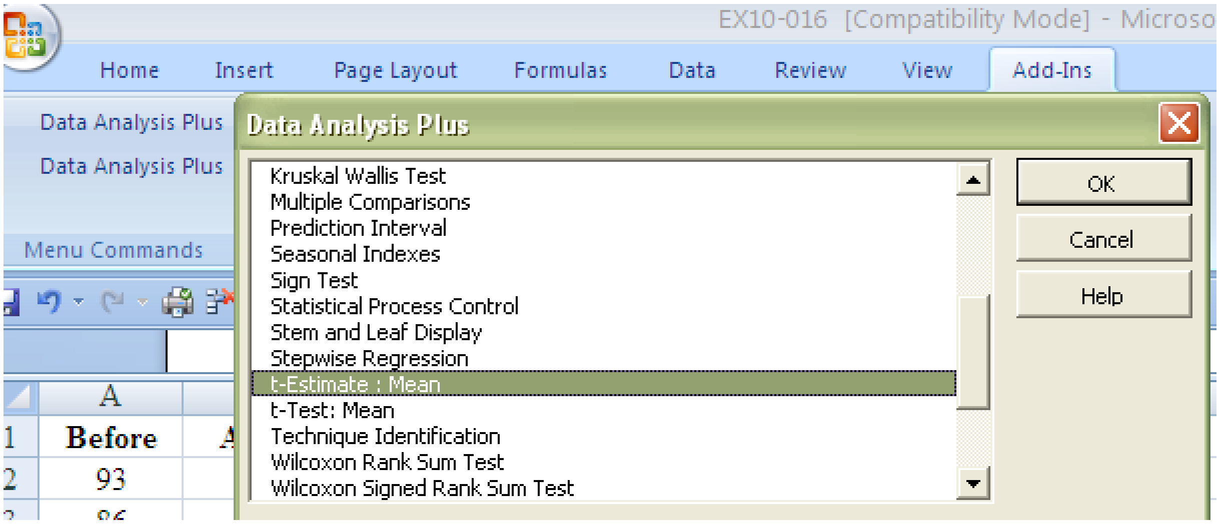
Click into any empty cell.

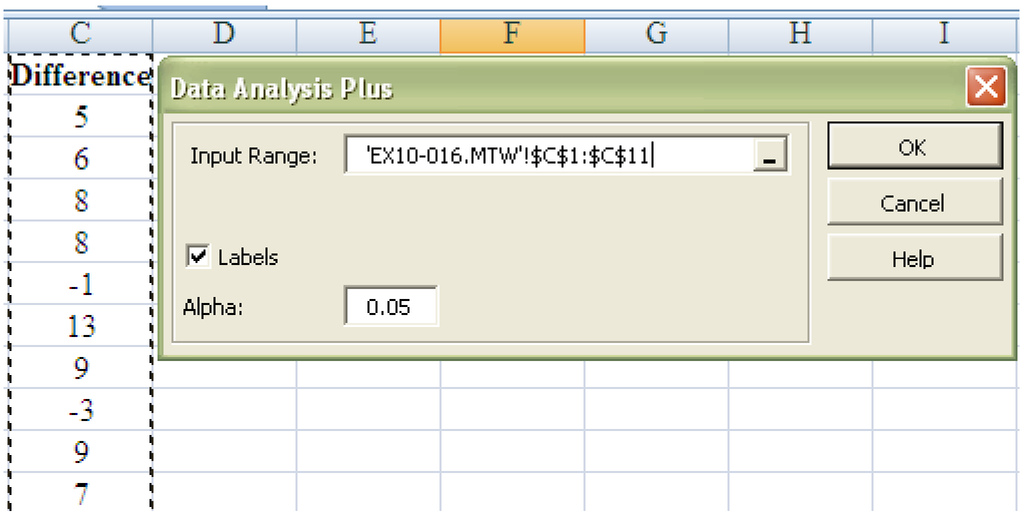
Choose: **Add-Ins > Data Analysis Plus > t-Estimate: Mean**

Enter: Input range: **C2:C11**

Select: Labels (if necessary)

Enter: Alpha: **(or 0.05)**





The output is placed in a separate sheet.

	A	B	C
1	**t-Estimate: Mean**		
2			*Difference*
3	Mean		6.1
4	Standard Deviation		4.7947
5	LCL		2.6701173
6	UCL		9.5298827

Hypothesis Testing

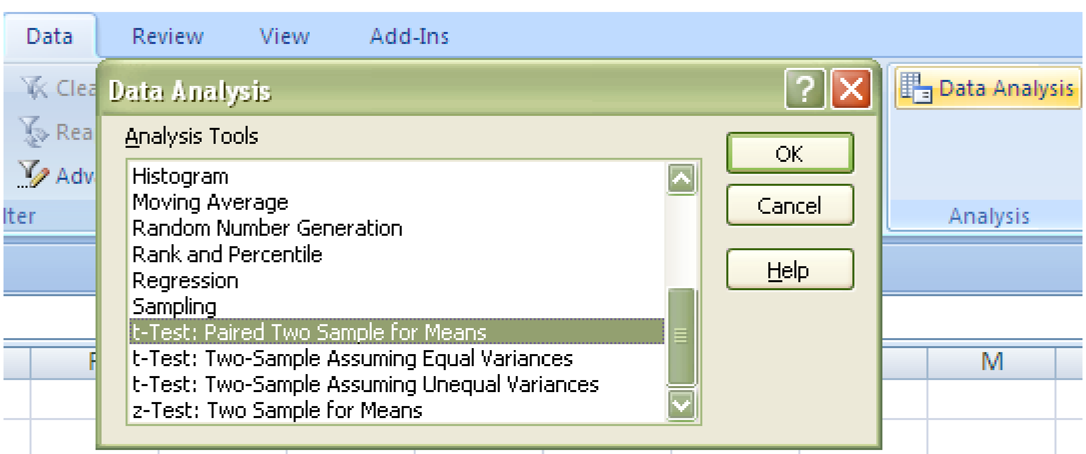
To demonstrate the procedure for a hypothesis test on the mean difference we will do Exercise 10.38.

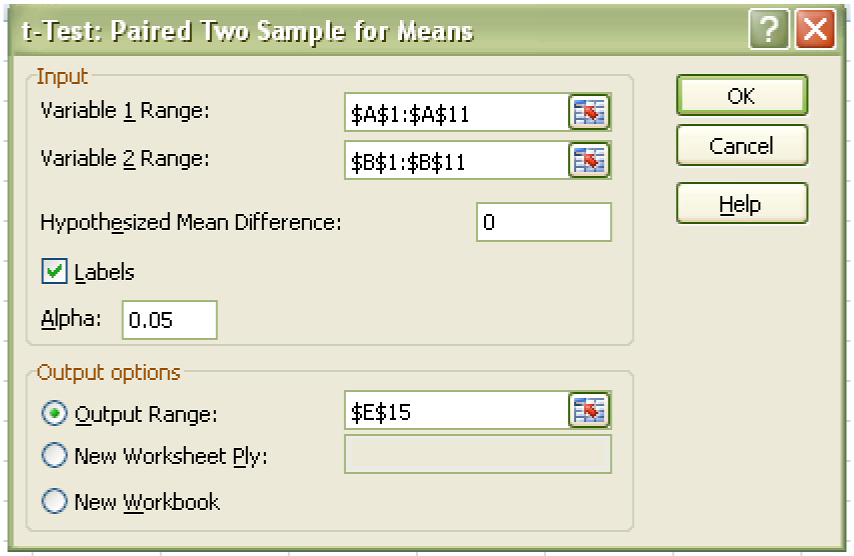
Enter the data for Before in column A and for After in column B or by retrieving it from the Student Suite CD (ex10-038) and calculate the paired differences.

	A	B	C
1	**Before**	**After**	**Diff**
2	29	30	1
3	22	26	4
4	25	25	0
5	29	35	6
6	26	33	7
7	24	36	12
8	31	32	1
9	46	54	8
10	34	50	16
11	28	43	15

Then perform a t-test on the paired differences (After – Before).

Choose: **Add-Ins > Data Analysis > t-Test: Paired Two Sample for Means**
Enter: Variable 1 Range**: B4:B14**
Enter: Variable 2 Range: **A4:A14**
Select: Labels
Enter: α (example 0.05)
Select: Output Range
Enter: **A15** (or any empty cell)
Click: **OK**





The results you get look like this:

t-Test: Paired Two Sample for Means		
	Before	*After*
Mean	29.4	36.4
Variance	46.26666667	94.48888889
Observations	10	10
Pearson Correlatio	0.810662928	
Hypothesized Me	0	
df	9	
t Stat	-3.821341258	
P(T<=t) one-tail	0.002040758	
t Critical one-tail	1.833112923	
P(T<=t) two-tail	0.004081516	
t Critical two-tail	2.262157158	

Note: t statistic = 3.82 and the p-value = 0.0041. How would you interpret these results?

ASSIGNMENT: Do Exercises 10.19, 10.20, 10.22, 10.34 in your text.

<u>CASE 2. INDEPENDENT SAMPLES:</u>

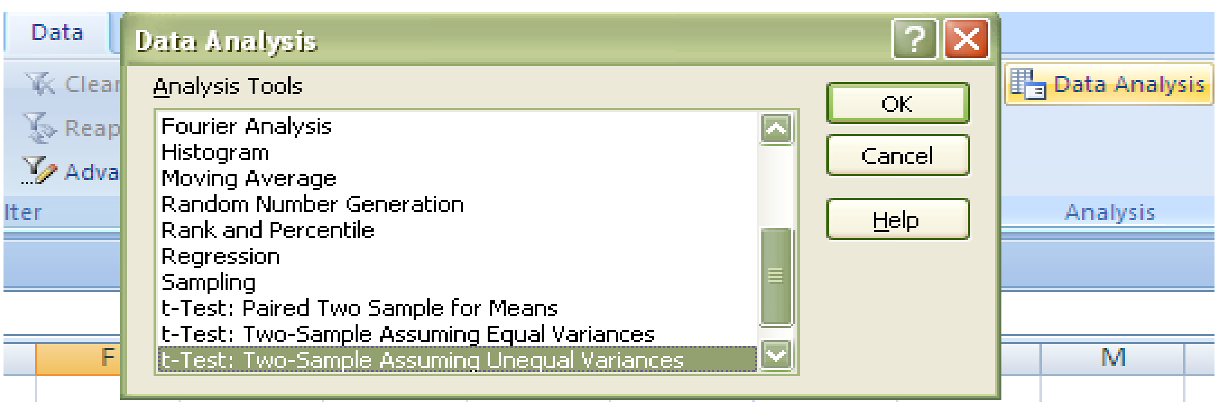
If two samples are selected, one from each of the populations, the two samples are independent if the selection of objects from one population is unrelated to the selection of objects from the other population. Since the samples provide the information for determining the standard error, the t distribution will be used as the test statistic, and the degrees of freedom will be calculated by Excel.

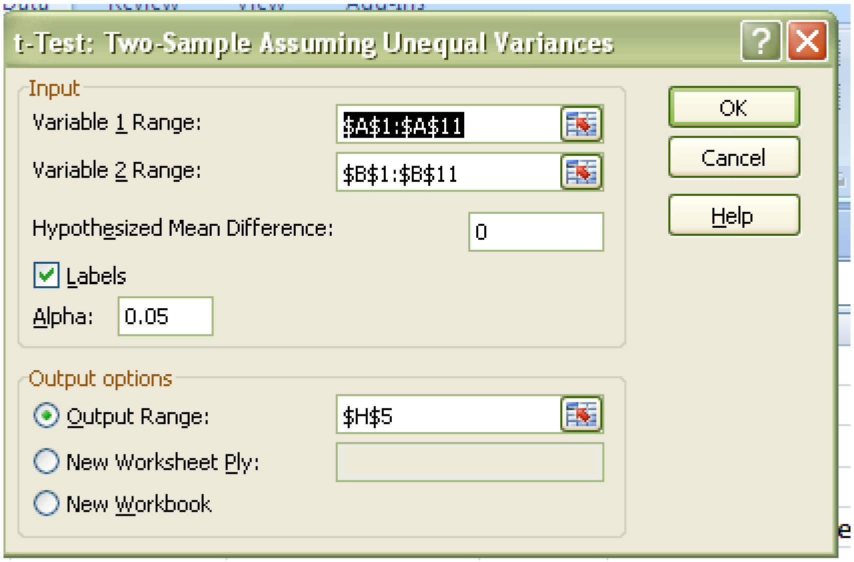
a) Complete the hypothesis test presented in Exercise 10.72 of your text. Retrieve the data from the Student Suite CD and note that the data for Diet A is in Column A and Diet B is in Column B.

	A	B
1	**DietA**	**DietB**
2	5	5
3	14	21
4	7	16
5	9	23
6	11	4
7	7	16
8	13	13
9	14	19
10	12	9
11	8	21

Perform a t-test as follows:

Choose: **Add-Ins > Data Analysis > t-Test: Two Sample Assuming Unequal Variances**
Enter: Variable 1 Range**: B1:B11**
Enter: Variable 2 Range: **A1:A11**
Hypothesized Difference**: 0.0**
Select: Labels
Enter: α (example 0.10)
Select: Output Range
Enter: **A15** (or any empty cell)
Click: **OK**





We then get the following output:

t-Test: Two-Sample Assuming Unequal Variances

	DietA	***DietB***
Mean	10	14.7
Variance	10.44444444	46.0111111
Observations	10	10
Hypothesized Mean Difference	0	
df	13	
t Stat	-1.97808302	
P(T<=t) one-tail	0.034755712	
t Critical one-tail	1.770933383	
P(T<=t) two-tail	0.069511424	
t Critical two-tail	2.160368652	

Do the data justify the conclusion that the mean weight gained on diet B was greater than the mean weight gained on diet A, at the $\alpha = .05$ level of significance?

Now that we have concluded that there is a difference, let us consider giving a 90% confidence interval estimate for this difference. The ToolPak does not print a confidence interval directly, but the output from the t-test provides us with the information to construct one. To complete the interval you must compute the formula for the confidence interval.

$$\overline{X}_1 - \overline{X}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Difference of the Means (Diet A - Diet B)	-4.7
SE =SQRT((s^2/n_1) + (s^2/n_2)	2.376037785
t*	1.770931704
ME= (t*)(SE)	4.207800642
lower = mean Diff - ME	-8.907800642
upper = Mean Diff + ME	-0.492199358

So the 90% interval for the difference of means is: (-8.91, -0.49)

b) Consider Exercise 10.49 in your text. Retrieve the data from the Student Suite CD: the data for the males is in Column A and the females is in Column B.

	A	B
1	males	females
2	76	76
3	76	70
4	74	82
5	70	90
6	80	68
7	68	60
8	90	62
9	70	68
10	90	80
11	72	74
12	76	60
13	80	62
14	68	72
15	72	
16	96	
17	80	
18		

Doing a t-test as above gives the following output:

	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*Variable 1*	*Variable 2*
4	Mean	77.375	71.07692
5	Variance	69.71666667	85.07692
6	Observations	16	13
7	Hypothesized Mean Difference	0	
8	df	25	
9	t Stat	1.907486345	
10	P(T<=t) one-tail	0.034004546	
11	t Critical one-tail	2.485103323	
12	P(T<=t) two-tail	0.068009092	
13	t Critical two-tail	2.787437552	
14			
15			
16	Difference of the Means	6.19808	
17	SE	3.301767764	
18	t*	2.787	
19	ME	9.202026757	
20	lower = difference of means - ME	-3.00394676	
21	Upper = difference of means - ME	15.40010676	

So the interval is (-3.004, 15.400). What does this imply? (Note the interval includes 0).

ASSIGNMENT: Do Exercises 10.76, and 10.79in your text. Both sets of data are found on the Student Suite CD.

Enrichment Assignment: Do Exercise 10.80 or 10.81. Turn in a typed paper detailing your procedures and results. Include the session commands you used and a printed copy of your output to substantiate your conclusions.

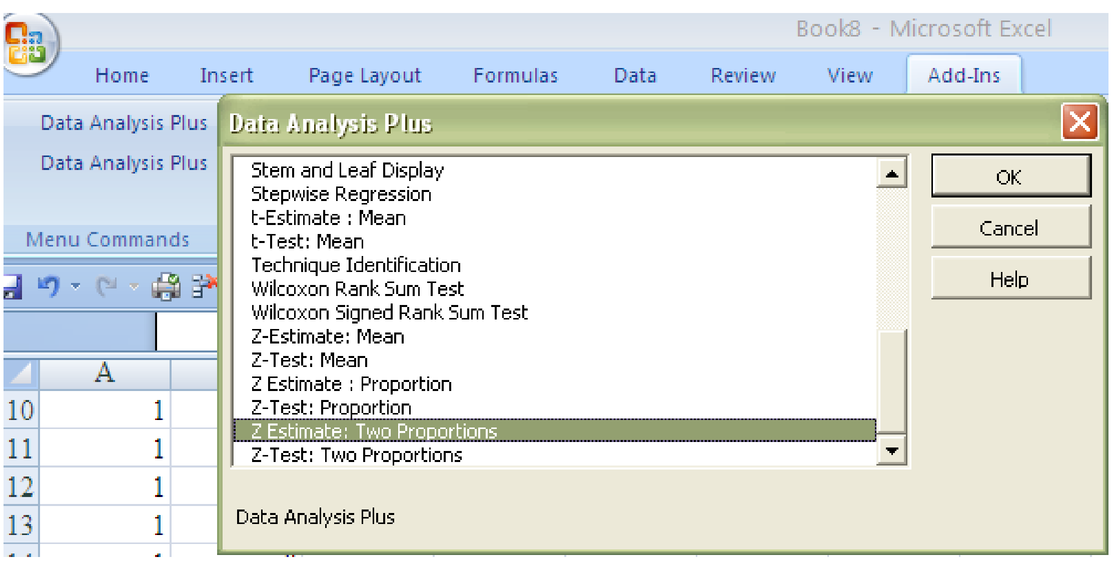
COMPARING TWO PROPORTIONS USING TWO INDEPENDENT SAMPLES

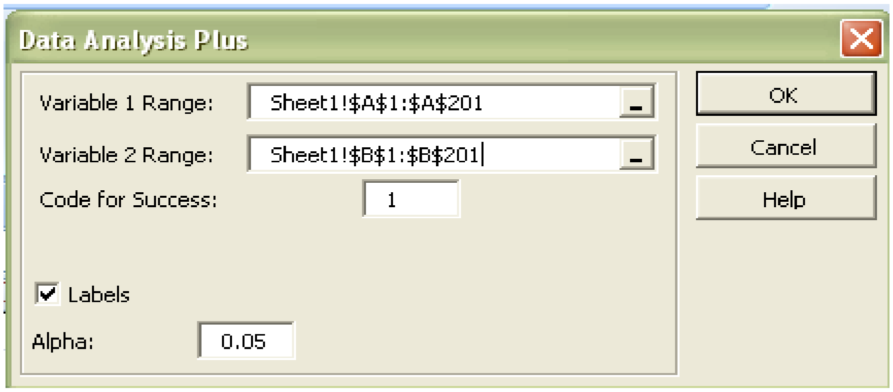
Confidence Interval

Consider the following problem: We are interested in estimating the difference in the proportion of male and female teenagers who have ever gambled. The sample evidence given is that 66% of the 200 males (x = 132) and 37% of the 199 females (x = 74) have "ever gambled".

The Data Analysis Plus Add-in is unable to do inference regarding proportions from summary data. We must create the data to fit the summary statistics we are given.
After entering 132 ones and 68 zeros into column A to represent the males, and 74 ones followed by 125 zeros to represent the females.

Choose: **Add-Ins > Data Analysis Plus > Z Estimate : Two Proportions**
Enter: Variable 1 Range: (**select cells from column A**)
Variable 2 Range: (**select cells from column B**)
Code for Success : **1**
Select: Labels (if necessary)
Enter: Alpha : **.05 > OK**



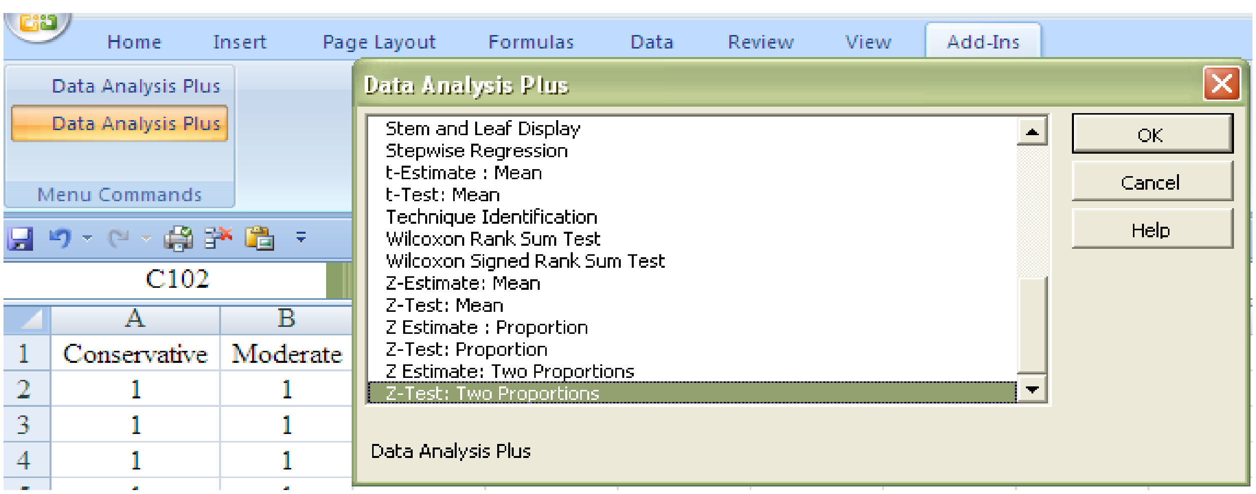


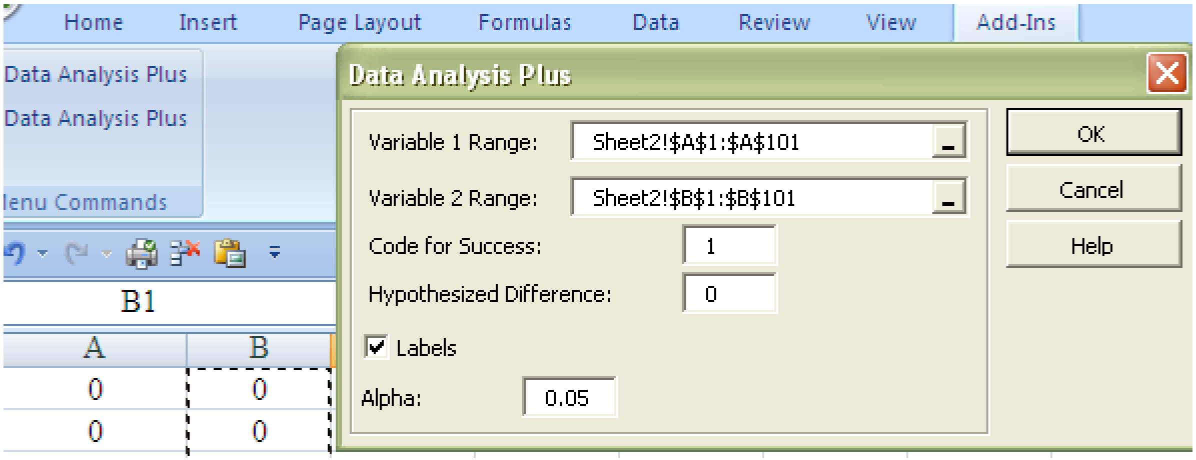
We obtain the following output:

	A	B	C	D
1	**z-Estimate: Two Proportions**			
2				
3			*Males*	*Females*
4	Sample Proportions		0.66	0.37
5	Observations		200	200
6				
7	LCL		0.196314	
8	UCL		0.383686	

Hypothesis Test

Consider exercise 10.101. To complete the test, create the data entering 0s and 1s as above.
Choose: **Add-Ins > Data Analysis Plus > Z-Test : Two Proportions**
Enter: Variable 1 Range: (**select cells from column A**)
Variable 2 Range: (**select cells from column B**)
Code for Success : **1**
Select: **Labels** (if necessary)
Enter: Alpha : **.05** > **OK**





The results are shown below.

	A	B	C	D
1	**z-Test: Two Proportions**			
2				
3			*Conservative*	*Moderate*
4	Sample Proportions		0.4	0.5
5	Observations		100	100
6	Hypothesized Difference		0	
7	z Stat		-1.4213	
8	P(Z<=z) one tail		0.0776	
9	z Critical one-tail		1.6449	
10	P(Z<=z) two-tail		0.1552	
11	z Critical two-tail		1.96	

So, what conclusion should be reached?

ASSIGNMENT: Do Exercises 10.90, and 10.100

CHAPTER 11 LAB SESSION
ANALYZING ENUMERATIVE DATA

INTRODUCTION: The data used in this lab is enumerative -- that is, the data is placed in categories and counted. The observed frequencies list exactly what happened in the sample. The expected frequencies represent the theoretical expected outcomes (what is expected to happen "on the average"). These expected values must always add up to n.

When we perform a hypothesis test on these two sets of values, we are really asking, "how different are they"? If the difference is small, we may attribute it to the chance variation in the samples. However, if the difference is large there may be a difference in the proportions in the population and we may reject the null hypothesis. We can use the χ^2 distribution in our test. We will first make inferences concerning multinomial experiments and then extend that to contingency tables.

MULTINOMIAL EXPERIMENTS

A multinomial experiment consists of n independent trials, whose outcome fits into only one of k possible cells. The probabilities of each of these cells remains constant and the sum of all the probabilities = 1. For multinomial experiments, we will always use a right tail critical region of the distribution. The expected frequency for each cell is obtained by multiplying the probability for that cell by the total number of trials, n.

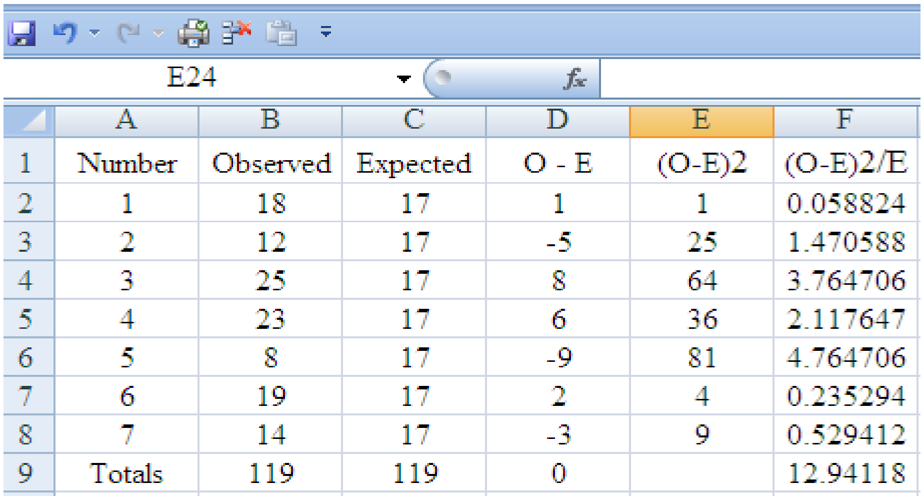
We can use Excel to calculate the Chi-Square statistic by entering the data, and the probability for each cell, calculating the expected values for each cell, and the Chi-Square value for each cell. We then need to sum each of these columns. Let us do Example 11.1 from the text, implementing Excel to do the calculations.

Enter the column headings in Row 1
Enter the seven observed values into column B.
Since there are seven sections, we can assume the probability of choosing any one of them would be 1/7 of the 119 students. Therefore, we will enter 17 in seven rows of column C. These are the Expected values.

Next, calculate the sums of each column by using the Σ from the toolbar.

In Cell D2 enter the formula = B2 – C2, and copy it down the column.
In Cell E2 enter the formula = D2*D2, and copy it down the column.
In Cell F2 enter the formula = E2/ C2, and copy it down the column.

Calculate the sums of columns D and F. The sum in column D should be 0, to give you a check on your data. The sum in Column F is the value of X^2.
Compare your results to the text.



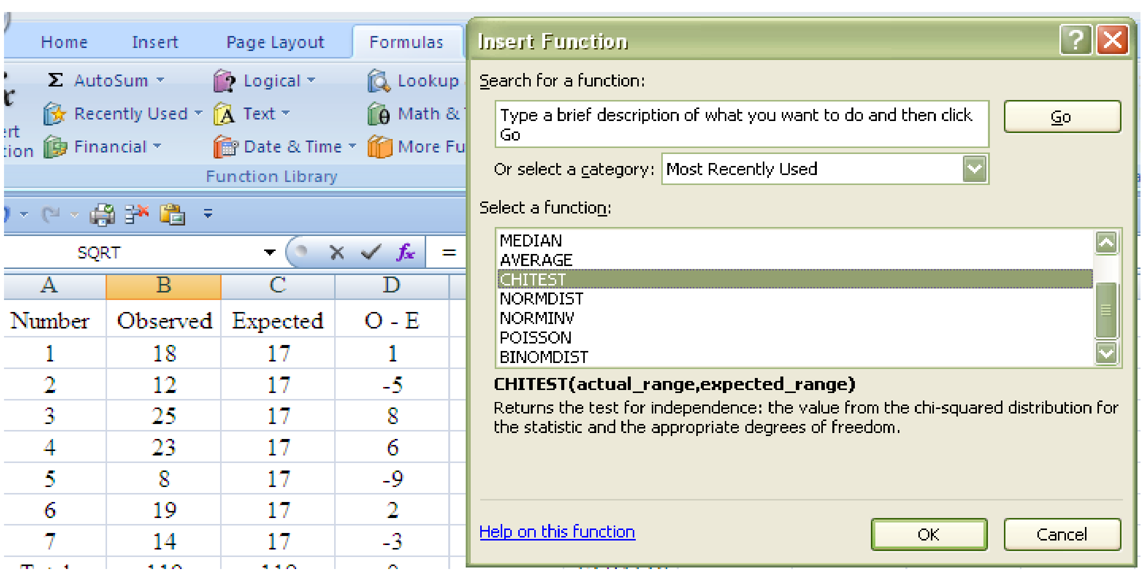
	A	B	C	D	E	F
1	Number	Observed	Expected	O - E	(O-E)2	(O-E)2/E
2	1	18	17	1	1	0.058824
3	2	12	17	-5	25	1.470588
4	3	25	17	8	64	3.764706
5	4	23	17	6	36	2.117647
6	5	8	17	-9	81	4.764706
7	6	19	17	2	4	0.235294
8	7	14	17	-3	9	0.529412
9	Totals	119	119	0		12.94118

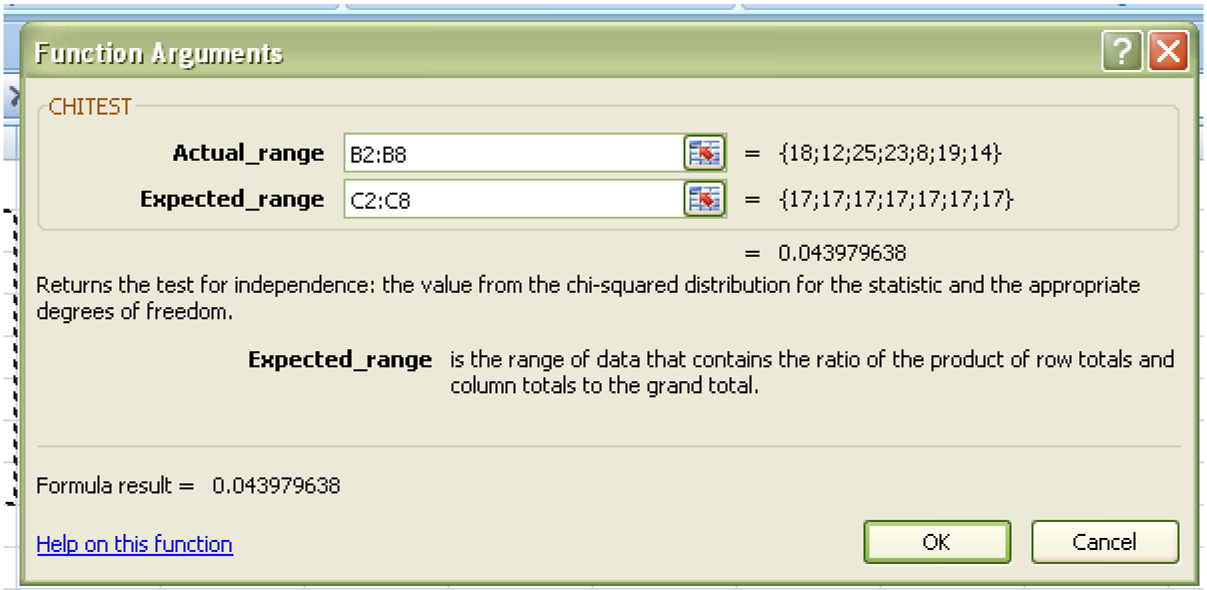
Let's enter some data to make our chart complete.

In Cell A11 enter **α = 0.05**
In Cell A12 enter **df = 6**
In Cell A13 enter **X^2 = 12.94**
In Cell A14 enter **p-value =**

To calculate your p-value click on Cell B14.

Click: **Formulas** > **Insert f_x** > **Statistical** > **CHITEST**
Enter: Actual Range: **B2:B8**
Expected Range: **C2:C8**
OK





11	alpha =	0.05
12	df =	6
13	X2=	12.94
14	p value =	0.04398

You now have to finish the test and state your conclusion.

ASSIGNMENT: Do Exercises 11.15, 11.21, and 11.22 in your text.

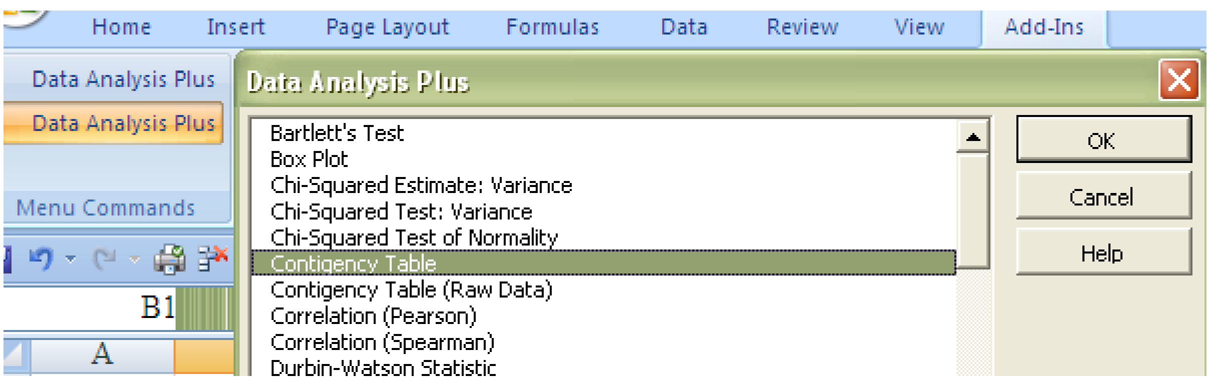
INFERENCES ABOUT CONTINGENCY TABLES

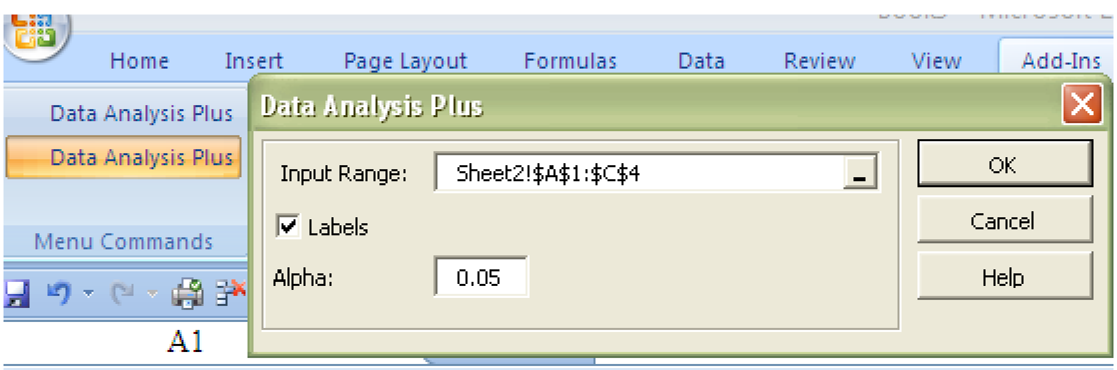
Contingency tables arrange data into a two-way classification. It involves two variables, and the first question we need to ask is are they independent or dependant. The two tests that use contingency tables are the Test of Independence and the Test for Homogeneity.

In a new sheet, enter the data from Example 11-6 including appropriate titles.

	A	B	C	D
1	**Residence**	**Favor**	**Oppose**	**Total**
2	Urban	143	57	200
3	Suburban	98	102	200
4	Rural	13	87	400
5	Total	254	246	500

Click on an empty cell.
Choose: **Add-Ins > Data Analysis Plus > Contingency Table > OK**
Enter: Input range: **B2:D4** > OK
Select: Labels (if necessary)
Enter: Alpha α (.05)





A new sheet will be created (note the tab name) that will contain the following:

	A	B	C	D
1	**Contingency Table**			
2				
3		*Favor*	*Oppose*	TOTAL
4	*Urban*	143	57	200
5	*Suburban*	98	102	200
6	*Rural*	13	87	100
7	TOTAL	254	246	500
8				
9	chi-squared Stat			91.7155
10	df			2
11	p-value			0
12	chi-squared Critical			5.9915

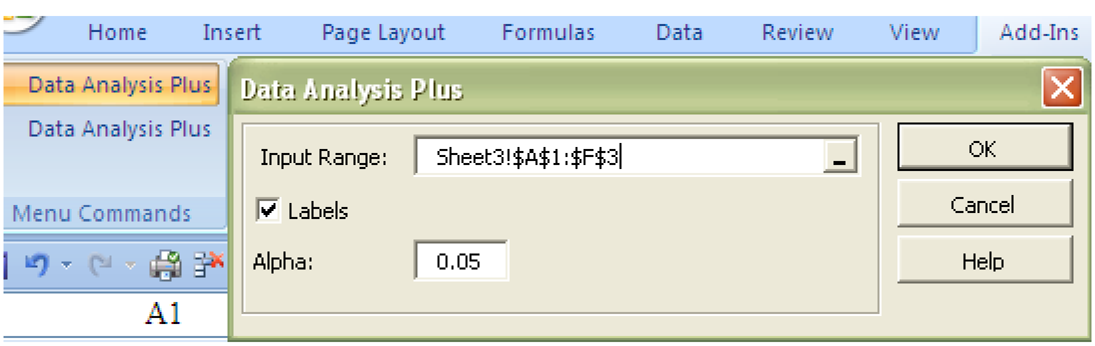
You now have to complete the test, noting that your df = 2, and state your conclusion.

Let us perform the procedure using the data from Exercise 11-45. First, label your columns and rows. Enter your data.

	A	B	C	D	E	F
1		**M**	**T**	**W**	**T**	**F**
2	**Defective**	85	90	95	95	90
3	**Non-Defective**	15	10	5	5	10

Click on an empty cell.

Choose: **Add-Ins > Data Analysis Plus > Contingency Table > OK**
Enter: Input range: **A1 – F3**
Select: Labels (if necessary)
Enter: Alpha (.05)



A new sheet will be created (note the tab name) that will contain the following:

	A	B	C	D	E	F	G
1	**Contingency Table**						
2							
3		*M*	*T*	*W*	*T*	*F*	TOTAL
4	*Defective*	85	90	95	95	90	455
5	*Non-Defective*	15	10	5	5	10	45
6	TOTAL	100	100	100	100	100	500
7							
8	chi-squared Stat			8.547			
9	df			4			
10	p-value			0.0735			
11	chi-squared Critical			9.4877			

You will still need to frame the null and alternative hypothesis; set the criteria, and then, using the above results, draw your conclusion.

ASSIGNMENT: Do the following Exercises 11.50, 11.51, 11.68, 11.74 in your text.

CHAPTER 12 LAB SESSION
ANALYSIS OF VARIANCE

INTRODUCTION: In earlier sessions you have examined and compared means from two samples. We will now practice a technique that tests hypothesis about several means. While we could compare the means in pairs as we have done before, the process could become too unwieldy to be of any use. Analysis of variance (ANOVA) allows us to test all the means at the same time to see if there is any significant difference between them.

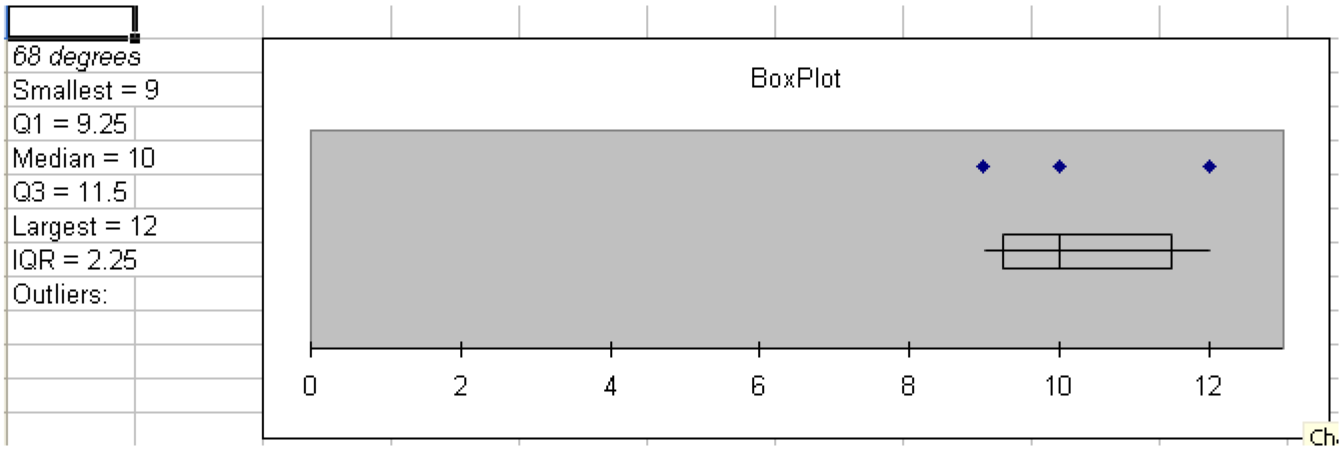
The Logic Underlying The Anova Technique

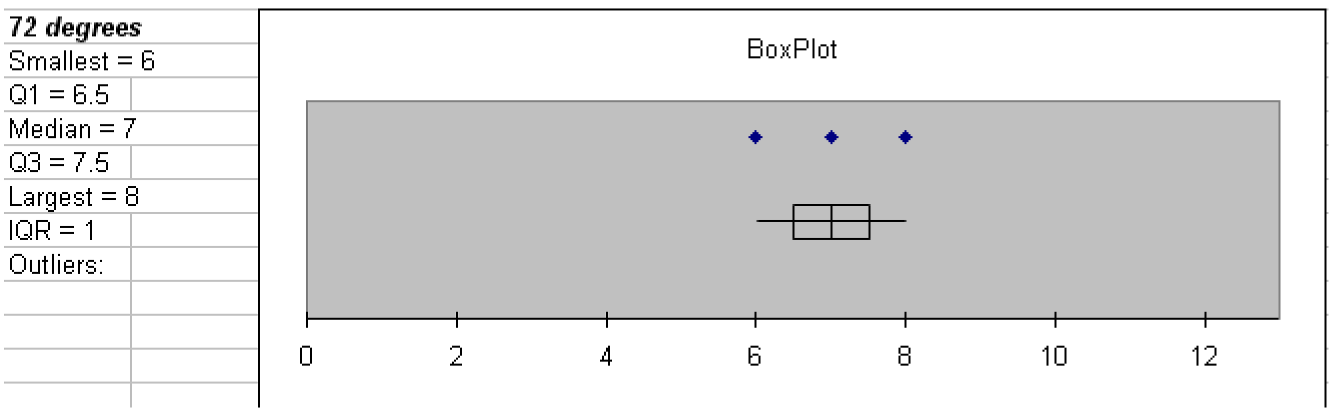
We will be forming a comparison between two estimates of the population variance: one based on the variance within each set of data and the other between the sets of data. We will use the F distribution for this comparison. If there is relatively little difference within each group and a large difference between the sample means, we will reject the null hypothesis. (Remember we always word the null hypothesis to say "there is no difference..."). If there is a lot of variance within a group and little between groups, we cannot conclude that the population means are different. We also need to know that the groups under investigation are approximately normally distributed and independent. ANOVA is presented as a table, and we need to define our terms in order to understand what the table is telling us. The Factor is the variable whose means we are interested in studying. When we first set up our data charts in Excel, each column will represent different Levels of the Factor we are examining. Each row will be a data value from repeated samplings, called a Replicate. The ANOVA table will give a summary of the data with the different levels of the Factor in the first column, followed by each levels' count, sum average and variance in subsequent columns across the rows. It then gives you a chart describing the sources of variation both Between Groups and Within Groups

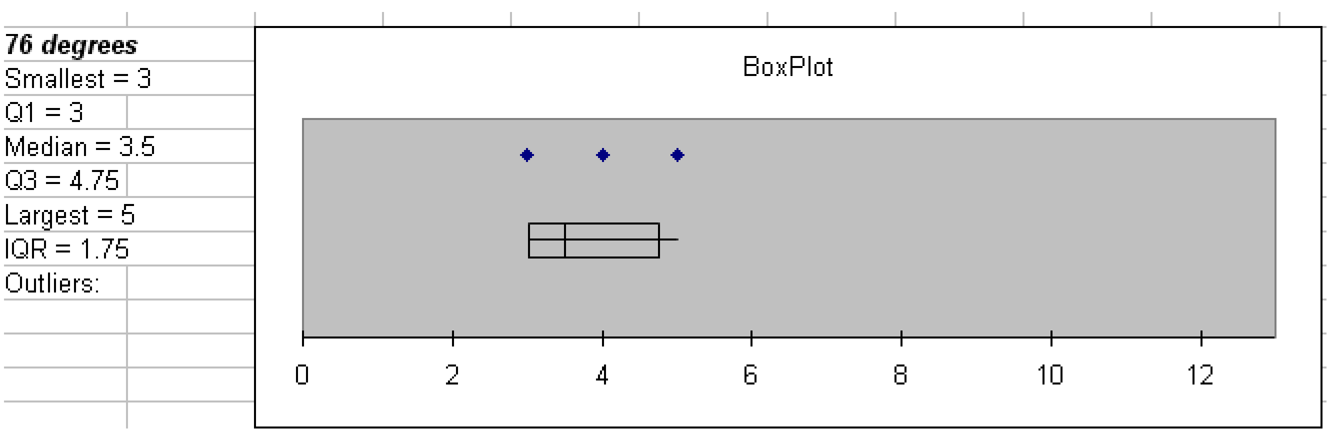
PERFORMING AN ANOVA ANALYSIS

This sample problem will take you through the steps of entering the data and generating the ANOVA table for Example 12-1 in your textbook. The FACTOR we are looking at is temperature and whether it has any effect on production. We will examine production at three different temperature levels: 68°, 72°, 76°. These levels form our columns. The production amounts are the replicates and form the rows of the data table. You can name the columns and enter the data directly into the worksheet.

If we did a Box Plot of the three columns some interesting things are shown.

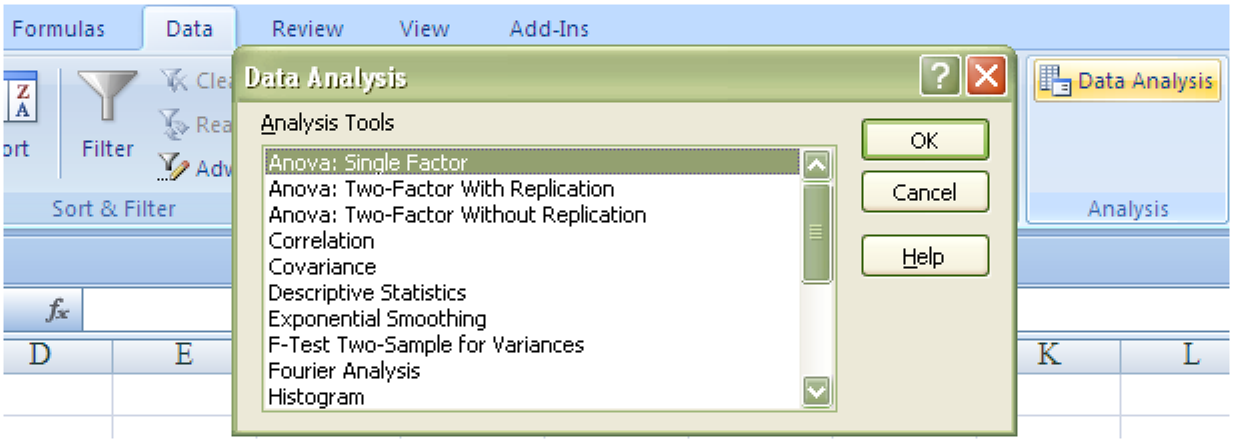


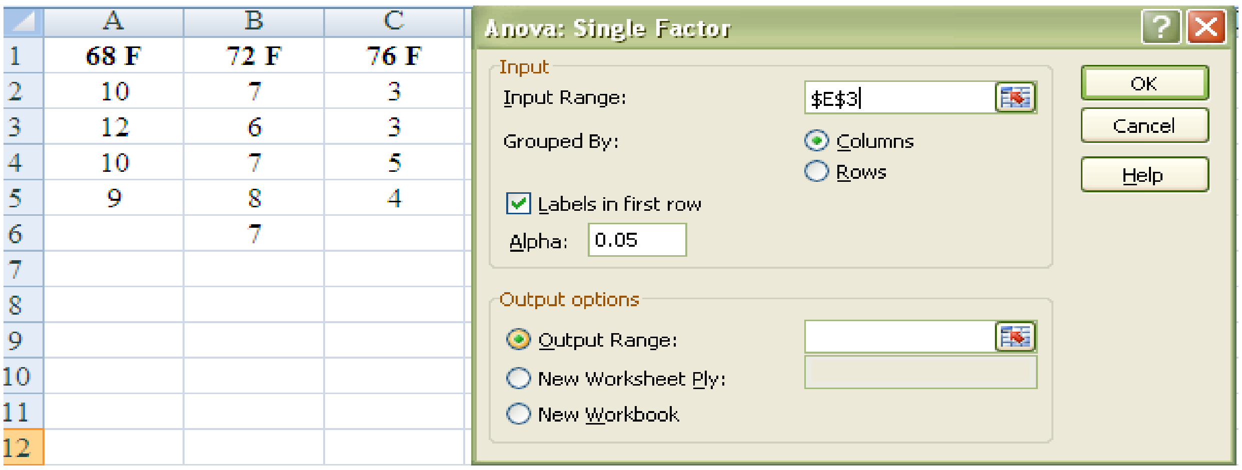




Note that the points within each level are fairly close, but the three levels hardly overlap at all. The commands for generating the ANOVA table is as follows:

Choose:	**Data > Data Analysis > ANOVA: Single Factor > OK**	
Enter:	Input Range: **A3:C8**	
Select:	**Output Range:**	
Enter:	**A10**	(or the upper left corner where you want it)





The worksheet will look like this:

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
68 F	4	41	10.25	1.583333		
72 F	5	35	7	0.5		
76 F	4	15	3.75	0.916667		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	84.5	2	42.25	44.47368	1.05E-05	4.102821
Within Groups	9.5	10	0.95			
Total	94	12				

Compare the output to the calculations in Illustration 12-1 in the text. Note in particular that the calculated value for F* = 44.47. To make our decision, we need to compare this to the critical value F(2,10,.05) = 4.10. We can therefore conclude that at least one of the temperatures has an effect on the production level. The p-value given in the chart can also be used to determine the conclusion. How would you interpret it?

Exercise 12.53 in the chapter exercises compares the stopping distances for four brands of tires. Using the data given there, is there sufficient evidence to conclude that there is a difference in the mean stopping distances at the $\alpha = .05$ level? This data may be found on the Student Suite CD as ex12-53.

a) State your null and alternative hypotheses.
b) Find your critical region and value for F.
c) 1) Enter your data in columns 1 - 4, naming them A, B, C, D respectively.
2) Do a box plot to get a feel for how the data interact.
3) Perform an ANOVA to calculate F*. What does the p value tell you? Explain.
d) Draw your conclusion about the null hypothesis and explain what it means to you. How would your conclusion change if α changed?

ASSIGNMENT: Do Exercises 12.28, 12.29, 12.51 and 12.55 in your text.

CHAPTER 13 LAB SESSION
LINEAR REGRESSION ANALYSIS

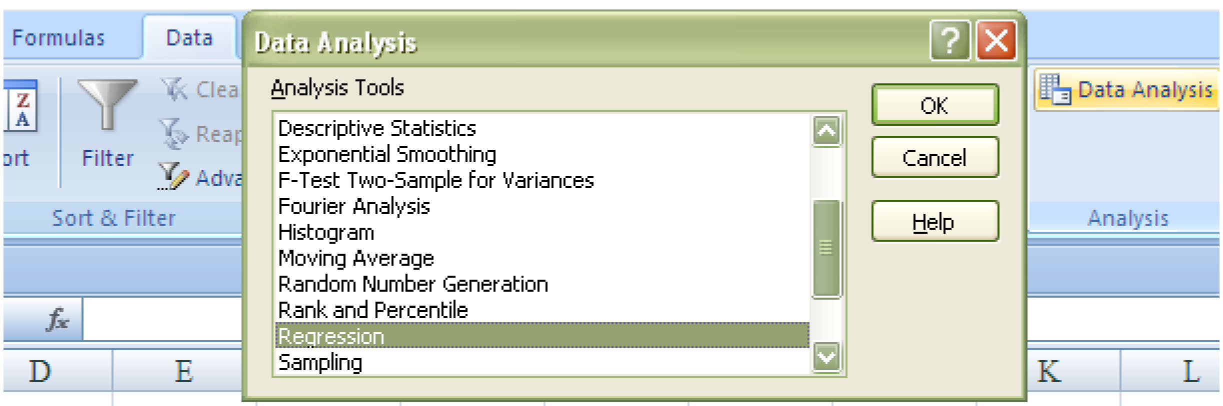
INTRODUCTION: In an earlier lab, we looked at bivariate data, and used the linear correlation coefficient to see if there was a relationship between the two variables. You also looked at a method of developing a line of best fit. In this lab we will look at a method of deciding whether the equation of that line is of any use to us in making point predictions and developing confidence intervals.

Before beginning this lab, you should review the commands for performing a regression analysis in Chapter 3 Lab Session 2. Use the data in Exercise 13.43 just to refresh your memory.

Enter x values in column A (independent variable) and the y values in column B (dependant variable).

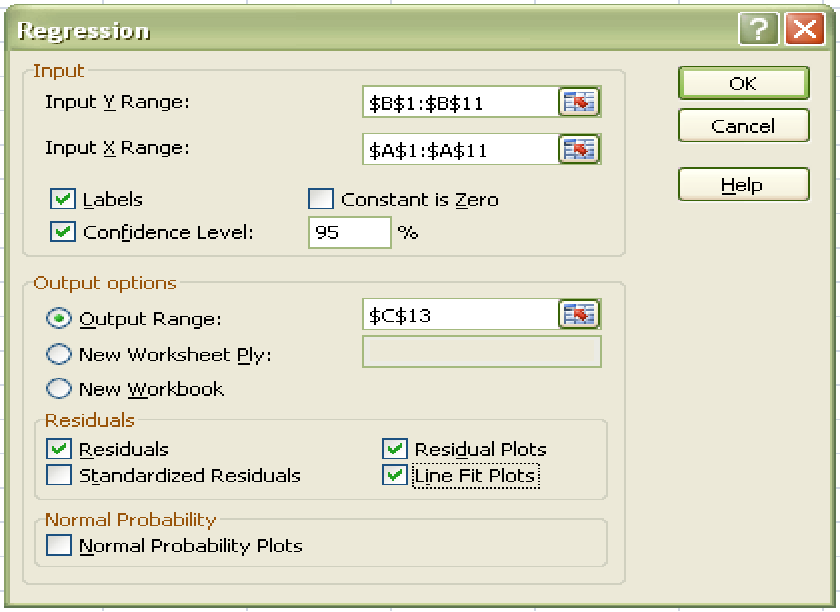
	A	B
1	x	y
2	12	15
3	14	25
4	16	30
5	20	30
6	23	30
7	46	80
8	50	90
9	48	95
10	50	110
11	55	130

Choose: **Data > Data Analysis > Regression > OK**
Enter: Input y range**: B1: B11 (or select cells)**
Input X range**; A1:A11 (or select cells)**
Select: Labels (if you labeled your columns)
Confidence level: **95% (or desired level)**
Select: Output Range**: C3** (or upper left corner of where you want output to appear)



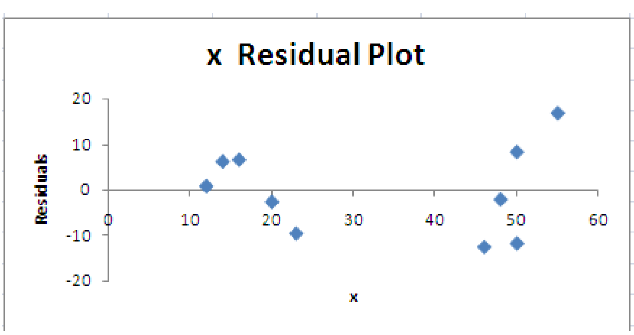
Check the following boxes, depending on desired output:

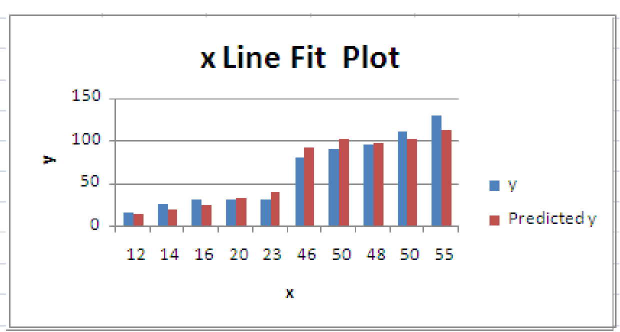
Residuals: to obtain predicted values and their residuals
Residual Plots: scatterplot of residuals against their x values
Line Fit Plots: scatterplot of y against x



Note: You can make the output nicer looking if you choose Format>Column>Autofit

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.973361953							
R Square	0.947433491							
Adjusted R Square	0.940862677							
Standard Error	10.17382633							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	14924.44606	14924.45	144.188154	2.13323E-06			
Residual	8	828.0539369	103.5067					
Total	9	15752.5						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-13.4135162	7.167861608	-1.87134	0.09820431	-29.9426347	3.115602292	-29.9426347	3.11560229
x	2.302799886	0.191774743	12.00784	2.1332E-06	1.860566537	2.745033235	1.860566537	2.74503324

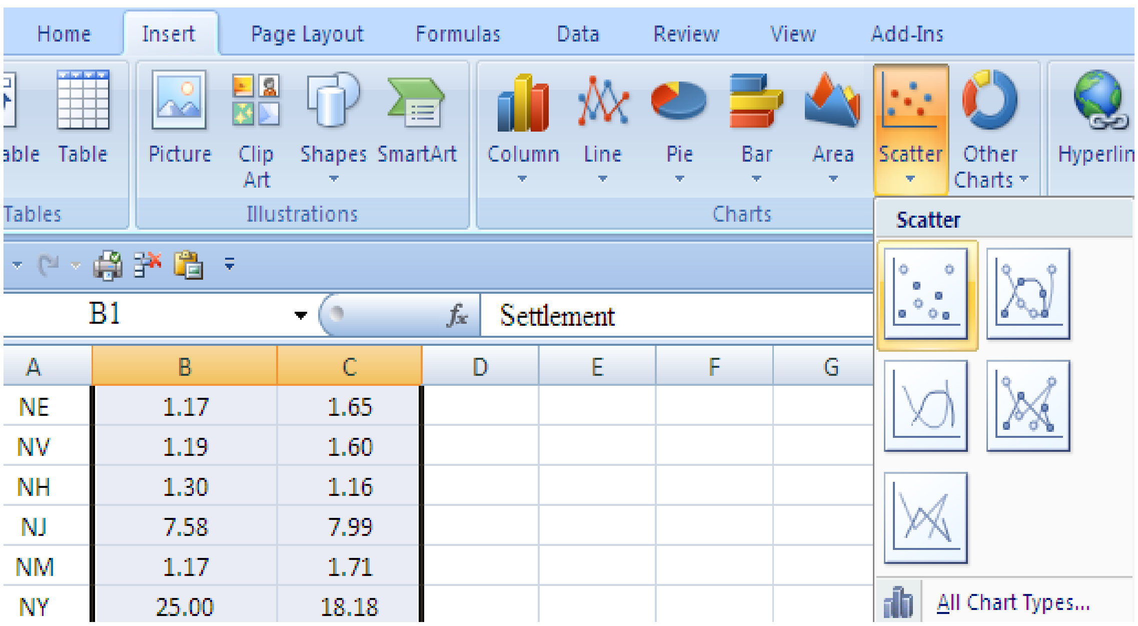


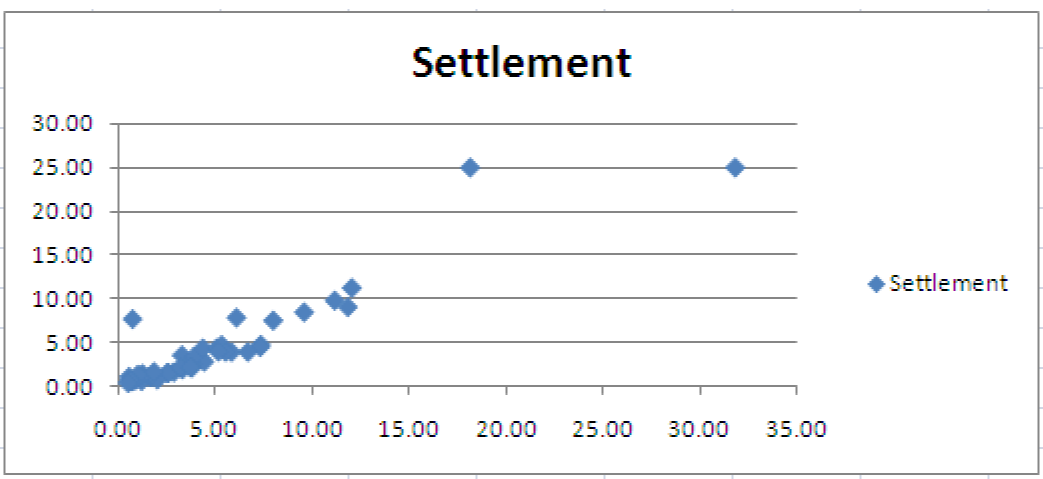


There are several steps in doing a linear regression analysis. First we obtain a least squares estimate for the model equation $y = \beta_0 + \beta_1 x + \varepsilon$, using the REGRESSION command as practiced above. Next, we need to check our assumptions about the random error component, ε. (The mean value of the experimental error is zero. We must also assume that the distribution of the y's is approximately normal and the variances σ^2 of the distribution of random errors is a constant.) Lastly, we can construct a confidence interval for our predictions.

We will use Exercise 13.88 to demonstrate the procedure. First, we do a scatterplot:

a)





b) We then calculate the regression equation. Note the coefficients on the output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.928353486
R Square	0.861840195
Adjusted R Square	0.858836721
Standard Error	1.932327304
Observations	48

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1071.431096	1071.431096	286.9477776	2.1273E-21
Residual	46	171.7588853	3.73388881		
Total	47	1243.189981			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.046638208	0.372357834	0.125251045	0.900870604	-0.702879227	0.796155642	-0.702879227	0.796155642
Population	0.879362439	0.051911847	16.93953298	2.1273E-21	0.774869313	0.983855565	0.774869313	0.983855565

Note the coefficients : The intercept is 0.046638 and the slope is 0.879362 so we have the following equation:

Settlement Amount = 0.046638 + 0.879362 * (Population)

To complete this exercise, follow the formula for the prediction interval, given in the text, and using the values generated by Excel.

ASSIGNMENT: Do Exercises 13.79, 13.87, 13.90 in your text.

CHAPTER 14 LAB SESSION
ELEMENTS OF NON-PARAMETRIC STATISTICS

INTRODUCTION: All the previous methods we have studied are parametric statistics - based on a population that has a certain distribution and can be applied only when special criteria are met. Non-parametric statistical methods can be applied when these criteria are not able to be met and assumptions about the parent population (such as normality) cannot be made, since these techniques do not rely on the distribution of the parent population. Non-parametric methods tend, unfortunately, to waste information and are less sensitive than their parametric counterparts. This, however, can be compensated for very nicely by increasing the sample size. Non-parametric techniques are generally easier to apply and are only slightly less efficient than parametric techniques.

THE SIGN TEST

The Sign test is one of the easiest tests to use, since it reduces the data to plus and minus signs. It can be used in hypothesis test for a single median or for two dependent samples using a paired difference. The basic concept is that because the median is the middle piece of data, with 50% of the data above it (represented by +) and 50% below (represented by -), then P(+) = .5 and P(-) = .5 . The method is fairly simple: all zeroes are rejected and the rest of the data is assigned positive and negative signs. The test statistic is the number of the less frequent sign. This is actually a binomial random variable (outcome either + or -) with a probability of 1/2. Z is calculated by the formula

$$z = (x' - n/2) / [(1/2) \sqrt{n}]$$

We will use the data from Exercise 14.3 fist to form a 95% Confidence Interval around the hypothesized median of 48. Enter the data into column A. To form the interval, we look in Table 12, which shows a critical value of k = 5 for n= 20 and $\alpha = .05$. Sort the data, and drop the last 5 values from each end to get an interval (39, 47).

We will using the same data to perform a sign test of whether the median high temperature is 48.

a) State the hypotheses:

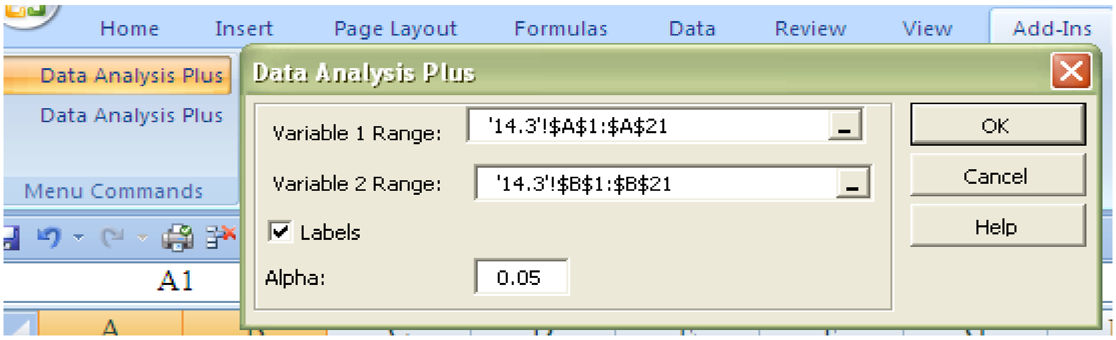
H_0: The median high temperature = 48

H_a: The median high temperature ≠ 48.

b) Set test criteria: This process will compute the differences between the data values and the hypothesized median of 48. In cell B1 enter 48 and drag it down the column to simulate a second sample. We do not count those values that equal the median, so we have a sample size of 19 and $\alpha = .05$.

Choose **Add-Ins > Data Analysis Plus >Sign Test > OK**

Fill in the data ranges as shown.



The results appear in a new sheet,

	A	B	C	D	E
1	**Sign Test**				
2					
3	Difference			*Temps - Median*	
4					
5	Positive Differences			3	
6	Negative Differences			16	
7	Zero Differences			1	
8	z Stat			-2.9824	
9	P(Z<=z) one-tail			0.0014	
10	z Critical one-tail			1.6449	
11	P(Z<=z) two-tail			0.0028	
12	z Critical two-tail			1.96	

c) Notice we have only 3 temperatures above the stated median and 16 below. The actual median of the sample is 45.5.

d) We have a p-value = 0.0028. We therefore reject the H_0 in favor of the H_a.

ASSIGNMENT: Do Exercise 14.14 in your text.
The Sign test can also be used for paired differences with two dependent samples.
Do Exercise 14.15 in your text.

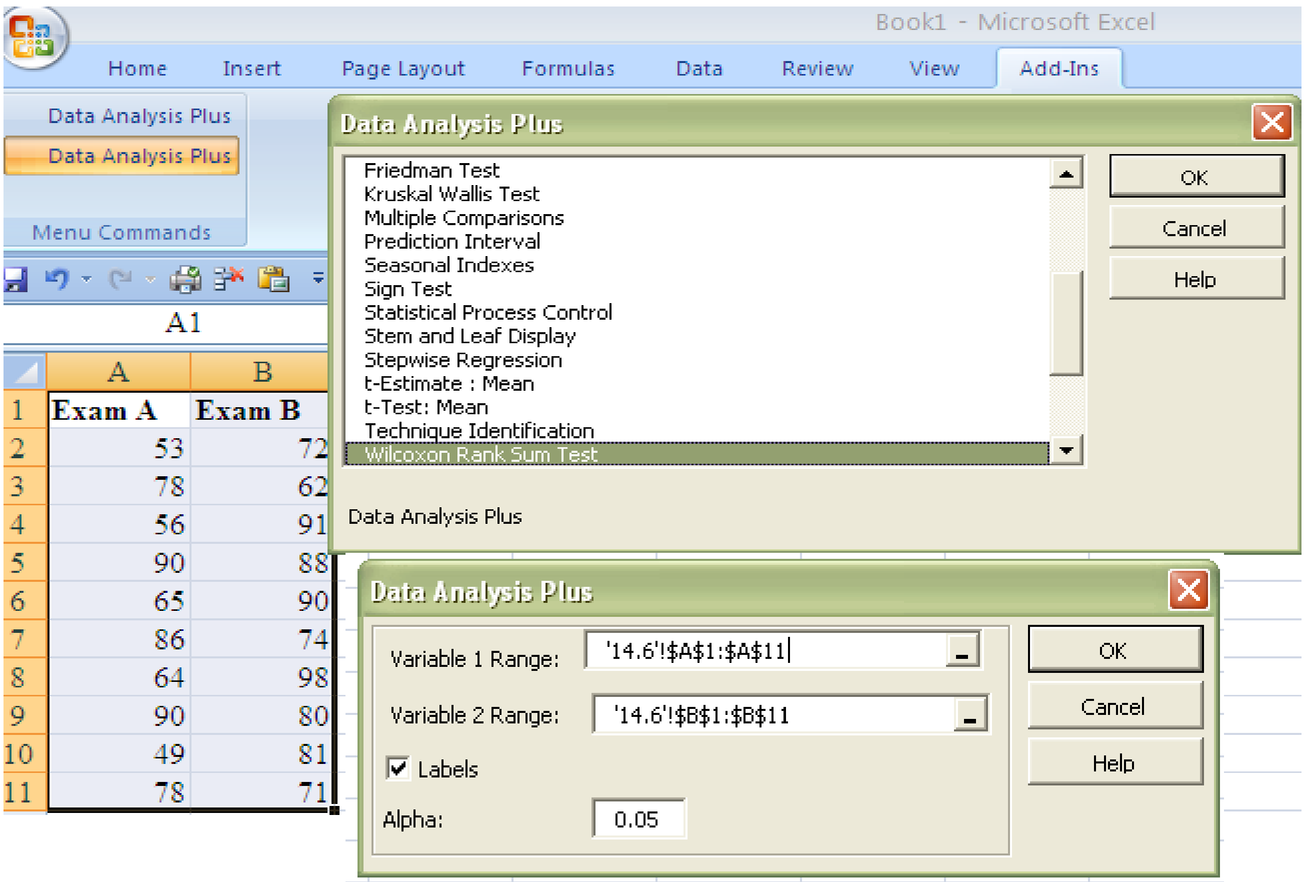
THE MANN WHITNEY TEST

This is an alternative method for the t-test on two independent random samples in which the random variable is continuous (also called Mann-Whitney-Wilcoxon test). By default, a two-sided test is performed. To do one-sided tests, select the test you want from the Alternative dialogue box. The test is carried out as follows: First, the two samples are ranked together, with the smallest observation given rank 1, the next largest given rank 2, and so on. Then the sum of the ranks of the first sample is calculated. If the sum is small, it indicates the observations from the first sample are smaller than those from the second sample, etc. The attained significance level of the test is calculated using a normal approximation (with a continuity correction factor).

The following problem demonstrates Example 14-6 in your text:

.

Choose: **Add-Ins > Data Analysis Plus > Wilcoxon Rank Sum Test***
Enter: variable 1 range:: **A1: A20**
variable 2 range:: **B1: B20**
Select: Labels (if necessary)
Choose: Alpha: α (.05)



The results appear in a new sheet.

Note the p-value = 0.151 > .05.
We therefore Fail to Reject H_0.

	A	B	C	D
1	**Wilcoxon Rank Sum Test**			
2				
3			Rank Sum	Observations
4	*Exam A*		86	10
5	*Exam B*		124	10
6	z Stat		-1.4363	
7	P(Z<=z) one-tail		0.0755	
8	z Critical one-tail		1.6449	
9	P(Z<=z) two-tail		0.151	
10	z Critical two-tail		1.96	

ASSIGNMENT: Do Exercises 14.29, 14.30, and 14.33 in your text.

RUNS TEST FOR RANDOMNESS

How do we really know when a set of outcomes is truly random? It cannot be in just counting the number of outcomes, but also in looking at the order in which those outcomes arise -- their arrangement. A particular run is a sequence of outcomes that have a common property. When that property changes the current run ends and a new one begins with the new property. The random variable to be considered is V, the number of runs. Its critical value is found in Table 14.

Example 14-10 is used to demonstrate the EXCEL technique:

a) State your hypotheses: H_0: The numbers are random
$H_{a:}$ the numbers are not random

b) State criteria: A two tail test with $\alpha = .05$

c) Perform the test: Note: Excel will only compute the differences between the data values and the median. To complete the test, you will need to count the number of runs, V, created by the + and – signs.

First enter the data in column A. Calculate the Median of the 30 data elements.

Enter**: A1 – MEDIAN(A1:A30)**
Drag the lower right corner of B1 to B10 to copy the formula for each data entry.

	A	B	C	D
1	Data	Data - Median		
2	2	-1.5		
3	5	1.5	Median =	3.5
4	3	-0.5		
5	8	4.5		
6	4	0.5		
7	2	-1.5		
8	9	5.5		

Determine the number of runs. We observe $n_a = 15$, $n_b = 15$ and V = 24 runs.

d) Calculate the P – value by using Table 14 and determine if it is smaller than α Since the value for V in the table = 10, less than our observed 24, the p-value is less than .05

e) Conclusion: State your results (Accept or Reject). We therefor reject the H_o.

ASSIGNMENT: Do Exercises 14.31, 14.41, 14.44 and 14.47 in your text.

RANK CORRELATION

This test is a nonparametric alternative to the linear correlation coefficient. The test is used to determine if there is a correlation between two rankings. Let's consider exercise 14.60.

Retrieve the data from the data file.

	A	B	C
1	**Soup**	**Sodium**	**Fiber**
2	A	480	12
3	B	830	0
4	C	510	1
5	D	460	5
6	E	490	3
7	F	580	7
8	G	420	2
9	H	290	4
10	I	450	10
11	J	430	6
12	K	390	9

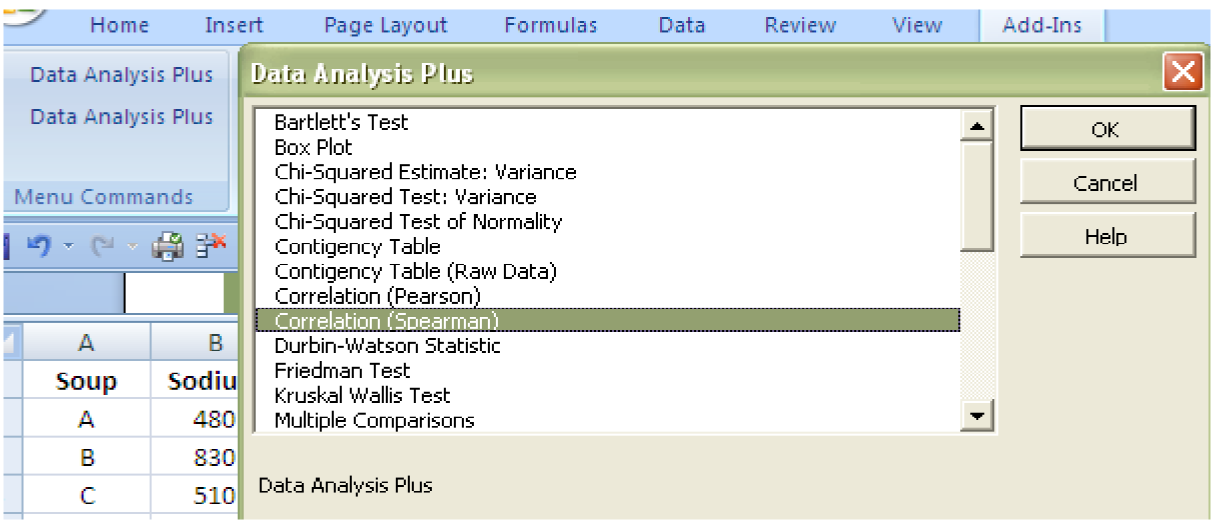
Choose: **Add-Ins > Data Analysis Plus > Correlation (Spearman)**

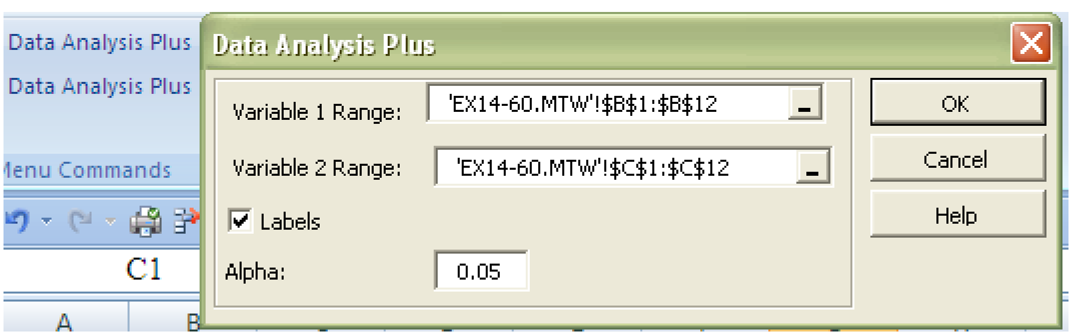
Enter: **Variable 1 Range: select appropriate cells**

Variable 2 Range: select appropriate cells

Select: **Labels (if necessary)**

Enter: **Alpha: 0.05**





The results appear in a new sheet.

	A	B	C	D
1	**Spearman Rank Correlation**			
2				
3	*Sodium and Fiber*			
4	Spearman Rank Correlation			-0.2909
5	z Stat			-0.9199
6	P(Z<=z) one tail			0.1788
7	z Critical one tail			1.6449
8	P(Z<=z) two tail			0.3576
9	z Critical two tail			1.96

What would your conclusion be?

ASSIGNMENT: Do Exercises 14.61, 14.63 in your text.