2014

Excel Study Guide & Help/User Manual for DA 3410 Students

Fox, William Dr.

Monterey, California. Naval Postgraduate School

http://hdl.handle.net/10945/70575

EXCEL STUDY GUIDE & HELP/USER MANUAL for DA 3410 Students

Dr. William P. Fox

Summer

Table of Contents

Using Excel for Statistical Data Analysis - Caveats

At A Glance

We used Excel to do some basic data analysis tasks to see whether it is a reasonable alternative to using a statistical package for the same tasks. We concluded that Excel is a **poor** choice for statistical analysis beyond textbook examples, the simplest descriptive statistics, or for more than a very few columns. The problems we encountered that led to this conclusion are in four general areas:

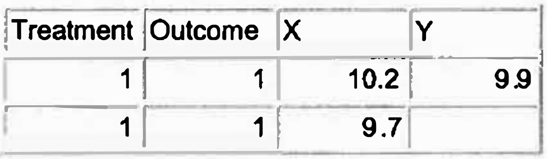
- **Missing values are handled inconsistently, and sometimes incorrectly.**
- **Data organization differs according to analysis, forcing you to reorganize your data in many ways if you want to do many different analyses.**
- **Many analyses can only be done on one column at a time, making it inconvenient to do the same analysis on many columns at the same time.**
- **Output is poorly organized, sometimes inadequately labeled, and there is no record of how an analysis was accomplished.**

Excel is convenient for data entry, and for quickly manipulating rows and columns prior to statistical analysis. However when you are ready to do the statistical analysis, we recommend the use of a statistical package such as SAS, SPSS, Stata, Systat or Minitab.

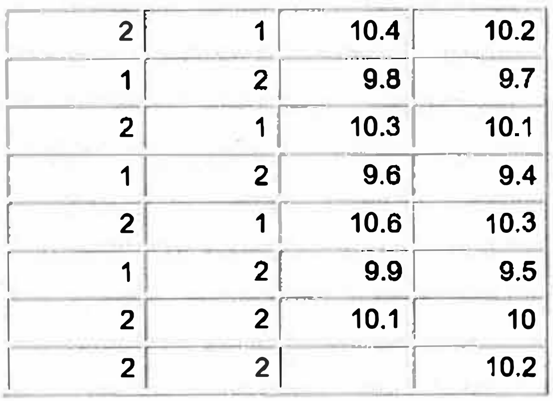
Introduction

Excel is probably the most commonly used spreadsheet for PCs. Newly purchased computers often arrive with Excel already loaded. It is easily used to do a variety of calculations, includes a collection of statistical functions, and a Data Analysis ToolPak. As a result, if you suddenly find you need to do some statistical analysis, you may turn to it as the obvious choice. We decided to do some testing to see how well Excel would serve as a Data Analysis application.

To present the results, we will use a small example. The data for this example is fictitious. It was chosen to have two categorical and two continuous variables, so that we could test a variety of basic statistical techniques. Since almost all real data sets have at least a few missing data points, and since the ability to deal with missing data correctly is one of the features that we take for granted in a statistical analysis package, we introduced two empty cells in the data:



Treatment	Outcome	X	Y
1	1	10.2	9.9
1	1	9.7	



2	1	10.4	10.2
1	2	9.8	9.7
2	1	10.3	10.1
1	2	9.6	9.4
2	1	10.6	10.3
1	2	9.9	9.5
2	2	10.1	10
2	2		10.2

Each row of the spreadsheet represents a subject. The first subject received Treatment 1, and had Outcome 1. X and Y are the values of two measurements on each subject. We were unable to get a measurement for Y on the second subject, or on X for the last subject, so these cells are blank. The subjects are entered in the order that the data became available, so the data is not ordered in any particular way.

We used this data to do some simple analyses and compared the results with a standard statistical package. The comparison considered the accuracy of the results as well as the ease with which the interface could be used for bigger data sets - i.e. more columns. We used SPSS as the standard, though any of the statistical packages OIT supports would do equally well for this purpose. In this article when we say "a statistical package," we mean SPSS, SAS, STATA, SYSTAT, or Minitab.

Most of Excel's statistical procedures are part of the Data Analysis tool pack, which is in the Tools menu. It includes a variety of choices including simple descriptive statistics, t-tests, correlations, 1 or 2-way analysis of variance, regression, etc. If you do not have a Data Analysis item on the Tools menu, you need to install the Data Analysis ToolPak. Search in Help for "Data Analysis Tools" for instructions on loading the ToolPak.

Two other Excel features are useful for certain analyses, but the Data Analysis tool pack is the only one that provides reasonably complete tests of statistical significance. Pivot Table in the Data menu can be used to generate summary tables of means, standard deviations, counts, etc. Also, you could use functions to generate some statistical measures, such as a correlation coefficient. Functions generate a single number, so using functions you will likely have to combine bits and pieces to get what you want. Even so, you may not be able to generate all the parts you need for a complete analysis.

Unless otherwise stated, all statistical tests using Excel were done with the Data Analysis ToolPak. In order to check a variety of statistical tests, we chose the following tasks:

- Get means and standard deviations of X and Y for the entire group, and for each treatment group.

- Get the correlation between X and Y.
- Do a two sample t-test to test whether the two treatment groups differ on X and Y.
- Do a paired t-test to test whether X and Y are statistically different from each other.
- Compare the number of subjects with each outcome by treatment group, using a chi-squared test.

All of these tasks are routine for a data set of this nature, and all of them could be easily done using any of the above listed statistical packages.

General Issues

Enable the Analysis ToolPak

The Data Analysis ToolPak is not installed with the standard Excel setup. Look in the Tools menu. If you do not have a Data Analysis item, you will need to install the Data Analysis tools. Search Help for "Data Analysis Tools" for instructions.

Missing Values

A blank cell is the only way for Excel to deal with missing data. If you have any other missing value codes, you will need to change them to blanks.

Data Arrangement

Different analyses require the data to be arranged in various ways. If you plan on a variety of different tests, there may not be a single arrangement that will work. You will probably need to rearrange the data several ways to get everything you need.

Dialog Boxes

Choose Tools/Data Analysis, and select the kind of analysis you want to do. The typical dialog box will have the following items:

Input Range: Type the upper left and lower right corner cells. e.g. A1:B100. You can only choose adjacent rows and columns. Unless there is a checkbox for grouping data by rows or columns (and there usually is not), all the data is considered as one glop.

Labels - There is sometimes a box you can check off to indicate that the first row of your sheet contains labels. If you have labels in the first row, check this box, and your output MAY be labeled with your label. Then again, it may not.

Output location - New Sheet is the default. Or, type in the cell address of the upper left corner of where you want to place the output in the current sheet. New Worksheet is another option, which I have not tried. Ramifications of this choice are discussed below.

Other items, depending on the analysis.

Output location

The output from each analysis can go to a new sheet within your current Excel file (this is the default), or you can place it within the current sheet by specifying the upper left corner cell where you want it placed. Either way is a bit of a nuisance. If each output is in a new sheet, you end up with lots of sheets, each with a small bit of output. If you place them in the current sheet, you need to place them appropriately; leave room for adding comments and labels; changes you need to make to format one output properly may affect another output adversely. Example: Output from Descriptive has a column of labels such as Standard Deviation, Standard Error, etc. You will want to make this column wide in order to be able to read the labels. But if a simple Frequency output is right underneath, then the column displaying the values being counted, which may just contain small integers, will also be wide.

Results of Analyses

Descriptive Statistics

The quickest way to get means and standard deviations for an entire group is using Descriptive in the Data Analysis tools. You can choose several adjacent columns for the Input Range (in this case the X and Y columns), and each column is analyzed separately. The labels in the first row are used to label the output, and the empty cells are ignored. If you have more, non-adjacent columns you need to analyze, you will have to repeat the process for each group of contiguous columns. The procedure is straightforward, can manage many columns reasonably efficiently, and empty cells are treated properly.

To get the means and standard deviations of X and Y for each treatment group requires the use of Pivot Tables (unless you want to rearrange the data sheet to separate the two groups). After selecting the (contiguous) data range, in the Pivot Table Wizard's Layout option, drag Treatment to the Row variable area, and X to the Data area. Double click on "Count of X" in the Data area, and change it to Average. Drag X into the Data box again, and this time change Count to StdDev. Finally, drag X in one more time, leaving it as Count of X. This will give us the Average, standard deviation and number of observations in each treatment group for X. Do the same for Y, so we will get the average, standard deviation and number of observations for Y also. This will put a total of six items in the Data box (three for X and three for Y). As you can see, if you want to get a variety of descriptive statistics for several variables, the process will get tedious.

A statistical package lets you choose as many variables as you wish for descriptive statistics, whether or not they are contiguous. You can get the descriptive statistics for all the subjects together, or broken down by a categorical variable such as treatment. You can select the statistics you want to see once, and it will apply to all variables chosen.

Correlations

Using the Data Analysis tools, the dialog for correlations is much like the one for descriptive - you can choose several contiguous columns, and get an output matrix of all pairs of correlations. Empty cells are ignored appropriately. The output does NOT include the number of pairs of data points used to compute each correlation (which can vary, depending on where you have missing data), and does not indicate whether any of the correlations are statistically significant. If you want correlations on non-contiguous columns, you would either have to include the intervening columns, or copy the desired columns to a contiguous location.

A statistical package would permit you to choose non-contiguous columns for your correlations. The output would tell you how many pairs of data points were used to compute each correlation, and which correlations are statistically significant.

Two-Sample T-test

This test can be used to check whether the two treatment groups differ on the values of either X or Y. In order to do the test you need to enter a cell range for each group. Since the data were not entered by treatment group, we first need to sort the rows by treatment. **Be sure to take all the other columns along with treatment, so that the data for each subject remains intact**. After the data is sorted, you can enter the range of cells containing the X measurements for each treatment. Do not include the row with the labels, because the second group does not have a label row. Therefore your output will not be labeled to indicate that this output is for X. If you want the output labeled, you have to copy the cells corresponding to the second group to a separate column, and enter a row with a label for the second group. If you also want to do the t-test for the Y measurements, you'll need to repeat the process. The empty cells are ignored, and other than the problems with labeling the output, the results are correct.

A statistical package would do this task without any need to sort the data or copy it to another column, and the output would always be properly labeled to the extent that you provide labels for your variables and treatment groups. It would also allow you to choose more than one variable at a time for the t-test (e.g. X and Y).

Paired t-test

The paired t-test is a method for testing whether the difference between two measurements on the same subject is significantly different from 0. In this example, we wish to test the difference between X and Y measured on the same subject. The important feature of this test is that it compares the measurements within each subject. If you scan the X and Y columns separately, they do not look obviously different. But if you look at each X-Y pair, you will notice that in every case, X is greater than Y. The paired t-test should be sensitive to this difference. In the two cases where either X or Y is missing, it is not possible to compare the two measures on a subject. Hence, only 8 rows are usable for the paired t-test.

When you run the paired t-test on this data, you get a t-statistic of 0.09, with a 2-tail probability of 0.93. The test does not find any significant difference between X and Y. Looking at the output more carefully, we notice that it says there are 9 observations. As noted above, there should only be 8. It appears that Excel has failed to exclude the observations that did not have both X and Y measurements. To get the correct results copy X and Y to two new columns and remove the data in the cells that has no value for the other measure. Now re-run the paired t-test. This time the t-statistic is 6.14817 with a 2-tail probability of 0.000468. The conclusion is completely different!

Of course, this is an extreme example. But the point is that Excel does not calculate the paired t-test correctly when some observations have one of the measurements but not the other. Although it is possible to get the correct result, you would have no reason to suspect the results you get unless you are sufficiently alert to notice that the number of observations is wrong. There is nothing in online help that would warn you about this issue.

Interestingly, there is also a TTEST function, which gives the correct results for this example. Apparently the functions and the Data Analysis tools are not consistent in how they deal with missing cells. Nevertheless, I cannot recommend the use of functions in preference to the Data Analysis tools, because the result of using a function is a single number - in this case, the 2-tail probability of the t-statistic. The function does not give you the t-statistic itself, the degrees of freedom, or any number of other items that you would want to see if you were doing a statistical test.

A statistical package will correctly exclude the cases with one of the measurements missing, and will provide all the supporting statistics you need to interpret the output.

Cross tabulation and Chi-Squared Test of Independence

Our final task is to count the two outcomes in each treatment group, and use a chi-square test of independence to test for a relationship between treatment and outcome. In order to count the outcomes by treatment group, you need to use Pivot Tables. In the Pivot Table Wizard's Layout option, drag Treatment to Row, Outcome to Column and also to Data. The Data area should say "Count of Outcome" – if not, double-click on it and select "Count". If you want percent, double-click "Count of Outcome", and click Options; in the “Show Data As” box which appears, select "% of row". If you want both counts and percent, you can drag the same variable into the Data area twice, and use it once for counts and once for percent.

Getting the chi-square test is not so simple, however. It is only available as a function, and the input needed for the function is the observed counts in each combination of treatment and outcome (which you have in your pivot table), and the expected counts in each combination. Expected counts? What are they? How do you get them? If you have sufficient statistical background to know how to calculate the expected counts, and can do Excel calculations using relative and absolute cell addresses, you should be able to navigate through this. If not, you’re out of luck.

Assuming that you surmounted the problem of expected counts, you can use the Chi test function to get the probability of observing a chi-square value bigger than the one for this table. Again, since we are using functions, you do not get many other necessary pieces of the calculation, notably the value of the chi-square statistic or its degrees of freedom.

No statistical package would require you to provide the expected values before computing a chi-square test of independence. Further, the results would always include the chi-square statistic and its degrees of freedom, as well as its probability. Often you will get some additional statistics as well.

Additional Analyses

The remaining analyses were not done on this data set, but some comments about them are included for completeness.

Simple Frequencies

You can use Pivot Tables to get simple frequencies. (see Cross tabulations for more about how to get Pivot Tables.) Using Pivot Tables, each column is considered a separate variable, and labels in row 1 will appear on the output. You can only do one variable at a time.

Another possibility is to use the Frequencies function. The main advantage of this method is that once you have defined the frequencies function for one column, you can use Copy/Paste to get it for other columns. First, you will need to enter a column with the values you want counted (bins). If you intend to do the frequencies for many columns, be sure to enter values for the column with the most categories. e.g., if 3 columns have values of 1 or 2, and the fourth has values of 1,2,3,4, you will need to enter the bin values as 1,2,3,4. Now select enough empty cells in one column to store the results - 4 in this example, even if the current column only has 2 values. Next choose Insert/Function/Statistical/Frequencies on the menu. Fill in the input range for the first column you want to count using relative addresses (e.g. A1:A100). Fill in the Bin Range using the absolute addresses of the locations where you entered the values to be counted (e.g. $M$1:$M$4). Click Finish. Note the box above the column headings of the sheet, where the formula is displayed. It start with "= FREQUENCIES(". Place the cursor to the left of the = sign in the formula, and press Ctrl-Shift-Enter. The frequency counts now appear in the cells you selected.

To get the frequency counts of other columns, select the cells with the frequencies in them, and choose Edit/Copy on the menu. If the next column you want to count is one column to the right of the previous one, select the cell to the right of the first frequency cell, and choose Edit/Paste (ctrl-V). Continue moving to the right and pasting for each column you want to count. Each time you move one column to the right of the original frequency cells, the column to be counted is shifted right from the first column you counted.

If you want percent as well, you'll have to use the Sum function to compute the sum of the frequencies, and define the formula to get the percent for one cell. Select the cell to store the first percent, and type the formula into the formula box at the top of the sheet - e.g. = N1*100/N$5 - where N1 is the cell with the frequency for the first category, and N5 is the cell with the sum of the frequencies. Use Copy/Paste to get the formula for the remaining cells of the first column. Once you have the percent for one column, you can Copy/Paste them to the other columns. You'll need to be careful about the use of relative and absolute addresses! In the example above, we used N$5 for the denominator, so when we copy the formula down to the next frequency on the same column, it will still look for the sum in row 5; but when we copy the formula right to another column, it will shift to the frequencies in the next column.

Finally, you can use Histogram on the Data Analysis menu. You can only do one variable at a time. As with the Frequencies function, you must enter a column with "bin" boundaries. To count the number of occurrences of 1 and 2, you need to enter 0,1,2 in three adjacent cells, and give the range of these three cells as the **Bins** on the dialog box. The output is not labeled with any labels you may have in row 1, nor even with the column letter. If you do frequencies on lots of variables, you will have difficulty knowing which frequency belongs to which column of data.

Linear Regression

Since regression is one of the more frequently used statistical analyses, we tried it out even though we did not do a regression analysis for this example. The Regression procedure in the Data Analysis tools lets you choose one column as the dependent variable, and a set of contiguous columns for the independents. However, it does not tolerate any empty cells anywhere in the input ranges, and you are limited to 16 independent variables. Therefore, if you have any empty cells, you will need to copy all the columns involved in the regression to new columns, and delete any rows that contain any empty cells. Large models, with more than 16 predictors, cannot be done at all.

Analysis of Variance

In general, the Excel's ANOVA features are limited to a few special cases rarely found outside textbooks, and require lots of data re-arrangements.

One-way ANOVA

Data must be arranged in **separate and adjacent** columns (or rows) for each group. Clearly, this is not conducive to doing 1-ways on more than one grouping. If you have labels in row 1, the output will use the labels.

Two-Factor ANOVA Without Replication

This only does the case with **one observation per cell** (i.e. no Within Cell error term). The input range is a rectangular arrangement of cells, with rows representing levels of one factor, columns the levels of the other factor, and the cell contents the one value in that cell.

Two-Factor ANOVA with Replicates

This does a two-way ANOVA with **equal cell sizes**. Input must be a rectangular region with columns representing the levels of one factor, and rows representing replicates within levels of the other factor. The input range MUST also include an additional row at the top, and column on the left, with labels indicating the factors. However, these labels are not used to label the resulting ANOVA table. Click Help on the ANOVA dialog for a picture of what the input range must look like.

Requesting Many Analyses

If you had a variety of different statistical procedures that you wanted to perform on your data, you would almost certainly find yourself doing a lot of sorting, rearranging, copying and pasting of your data. This is because each procedure requires that the data be arranged in a particular way, often different from the way another procedure wants the data arranged. In our small test, we had to sort the rows in order to do the t-test, and copy some cells in order to get labels for the output. We had to clear the contents of some cells in order to get the correct paired t-test, but did not want those cells cleared for some other test. And we were only doing five tasks. It does not get better when you try to do more. There is no single arrangement of the data that would allow you to do many different analyses without making many different copies of the data. The need to manipulate the data in many ways greatly increases the chance of introducing errors.

Using a statistical program, the data would normally be arranged with the rows representing the subjects, and the columns representing variables (as they are in our sample data). With this arrangement you can do any of the analyses discussed here, and many others as well, without having to sort or rearrange your data in any way. Only much more complex analyses, beyond the capabilities of Excel and the scope of this article would require data rearrangement.

Working with Many Columns

What if your data had not 4, but 40 columns, with a mix of categorical and continuous measures? How easily do the above procedures scale to a larger problem?

At best, some of the statistical procedures can accept multiple contiguous columns for input, and interpret each column as a different measure. The descriptive and correlations procedures are of this type, so you can request descriptive statistics or correlations for a large number of continuous variables, as long as they are entered in adjacent columns. If they are not adjacent, you need to rearrange columns or use copy and paste to make them adjacent.

Many procedures, however, can only be applied to one column at a time. T-tests (either independent or paired), simple frequency counts, the chi-square test of independence, and many other procedures are in this class. This would become a serious drawback if you had more than a handful of columns, even if you use cut and paste or macros to reduce the work. In addition to having to repeat the request many times, you have to decide where to store the results of each, and make sure it is properly labeled so you can easily locate and identify each output.

Finally, Excel does not give you a log or other record to track what you have done. This can be a serious drawback if you want to be able to repeat the same (or similar) analysis in the future, or even if you've simply forgotten what you've already done.

Using a statistical package, you can request a test for as many variables as you need at once. Each one will be properly labeled and arranged in the output, so there is no confusion as to what's what. You can also expect to get a log, and often a set of commands as well, which can be used to document your work or to repeat an analysis without having to go through all the steps again.

Summary

Although Excel is a fine spreadsheet, it is not a statistical data analysis package. In all fairness, it was never intended to be one. Keep in mind that the Data Analysis ToolPak is an "add-in" - an extra feature that enables you to do a few quick calculations. So it should not be surprising that that is just what it is good for - a few quick calculations. If you attempt to use it for more extensive analyses, you will encounter difficulties due to any or all of the following limitations:

- Potential problems with analyses involving missing data. These can be insidious, in that the unwary user is unlikely to realize that anything is wrong.
- Lack of flexibility in analyses that can be done due to its expectations regarding the arrangement of data. This results in the need to cut/paste/sort/ and otherwise rearrange the data sheet in various ways, increasing the likelihood of errors.
- Output scattered in many different worksheets, or all over one worksheet, which you must take responsibility for arranging in a sensible way.
- Output may be incomplete or may not be properly labeled, increasing possibility of misidentifying output.
- Need to repeat requests for the some analyses multiple times in order to run it for multiple variables, or to request multiple options.
- Need to do some things by defining your own functions/formulae, with its attendant risk of errors.
- No record of what you did to generate your results, making it difficult to document your analysis, or to repeat it at a later time, should that be necessary.

If you have more than about 10 or 12 columns, and/or want to do anything beyond descriptive statistics and perhaps correlations, you should be using a statistical package. There are several suitable ones available by site license through OIT, or you can use them in any of the OIT PC labs. If you have Excel on your own PC, and don't want to pay for a statistical program, by all means use Excel to enter the data (with rows representing the subjects, and columns for the variables). All the mentioned statistical packages can read Excel files, so you can do the (time-consuming) data entry at home, and go to the labs to do the analysis.

We also note the using Excel with Apple Mac is prohibitive. We suggest using the NPS lab's MS compatible computers.

However, since we are using Excel, let's begin.

Introduction

This Excel manual provides illustrative experience in the use of Excel for data summary, presentation, and for other basic statistical analysis. We believe the popular use of Excel is on the areas where Excel really can excel. This includes organizing data, i.e. basic data management, tabulation and graphics. For real statistical analysis on must learn using the professional commercial statistical packages such as Minitab, SAS, and SPSS.

We have created templates to allow for ease of conducting hypothesis testing (simple 1 and 2 sample mean or proportions), nonlinear regression for power functions and sin functions, logistics regression, and Poisson regression, as well as AHP for up to a max of 8 alternatives and 8 criterion. All other commands that we desire are part of Excel's Analysis ToolPak or Excel's commands.

Microsoft Excel 1993-2003, 2007, 2008 beyond provides a set of data analysis tools called the **Analysis ToolPak** which you can use to save steps when you perform statistical analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Using Add-IN add the Analysis ToolPak to your Excel. Not being an Apple (MAC) user, I am not sure you can do this with an Apple product. I do know Apple users have had issues in the past.

Add-IN Analysis ToolPak

(we can do this on day one –bring your laptop to class).

If the Data Analysis command is selectable on the Tools menu, then the Analysis ToolPak is installed on your system. However, if the Data Analysis command is not on the Tools menu, you need to install the Analysis ToolPak. For older version of MS-Office the following will work. For later versions, follow along in the first lab.

Step 1: On the Tools menu, click Add-Ins.... If Analysis ToolPak is not listed in the Add-Ins dialog box, click Browse and locate the drive, folder name, and file name for the Analysis ToolPak Add-in — Analys32.xll — usually located in the Program Files\Microsoft Office\Office\Library\Analysis folder. Once you find the file, select it and click OK.

Step 2: If you don't find the Analys32.xll file, then you must install it.

1. Insert your Microsoft Office 2000 Disk 1 into the CD ROM drive.
2. Select Run from the Windows Start menu.
3. Browse and select the drive for your CD. Select Setup.exe, click Open, and click OK.
4. Click the Add or Remove Features button.

5. Click the + next to Microsoft Excel for Windows.
6. Click the + next to Add-ins.
7. Click the down arrow next to Analysis ToolPak.
8. Select Run from My Computer.
9. Select the Update Now button.
10. Excel will now update your system to include Analysis ToolPak.
11. Launch Excel.
12. On the Tools menu, click Add-Ins... – and select the Analysis ToolPak check box.

Step 3: The Analysis ToolPak Add-In is now installed and Data Analysis... will now be selectable on the Tools menu.

Microsoft Excel is a powerful spreadsheet package available for Microsoft Windows and the Apple Macintosh. Spreadsheet software is used to store information in columns and rows which can then be organized and/or processed. Spreadsheets are designed to work well with numbers but often include text. Excel organizes your work into workbooks; each workbook can contain many worksheets; worksheets are used to list and analyze data .

Excel is available on all public-access PCs (i.e., those, e.g., in the Library and PC Labs). It can be opened either by selecting Start – Programs – Microsoft Excel or by clicking on the Excel Short Cut which is either on your desktop, or on any PC, or on the Office Tool bar.

THE BASICS

Opening a Document:

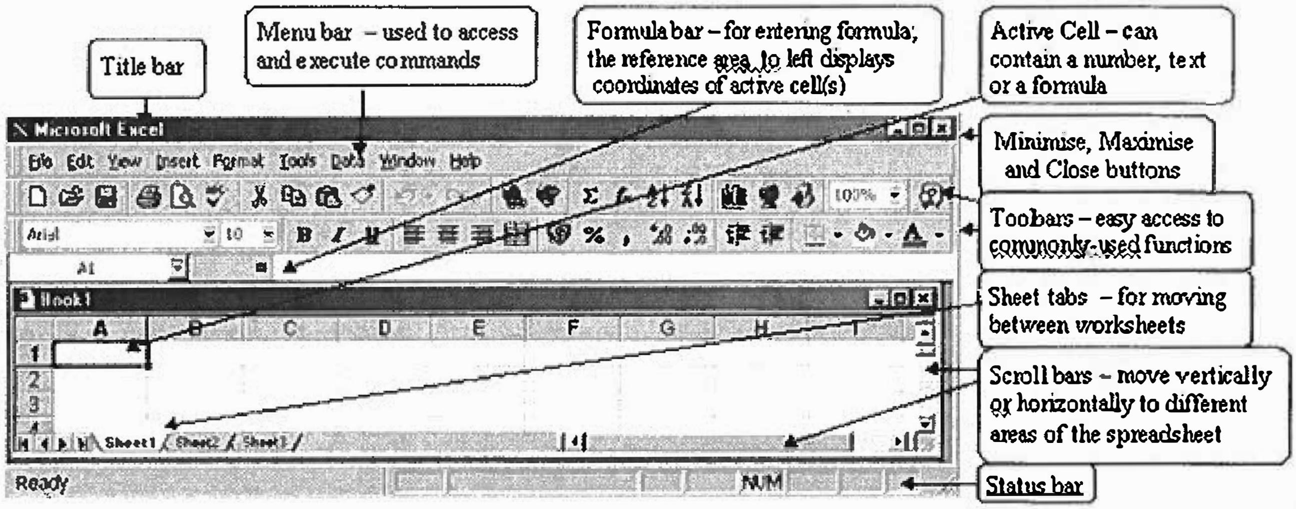
- Click on File-Open (Ctrl+O) to open/retrieve an existing workbook; change the directory area or drive to look for files in other locations
- To create a new workbook, click on File-New-Blank Document.

Saving and Closing a Document:

To save your document with its current filename, location and file format either click on File – Save. If you are saving for the first time, click File-Save; choose/type a name for your document; then click OK. Also use File-Save if you want to save to a different filename/location.

When you have finished working on a document you should close it. Go to the File menu and click on Close. If you have made any changes since the file was last saved, you will be asked if you wish to save them.

The Excel screen


Title bar
Menu bar – used to access and execute commands
Formula bar – for entering formula; the reference area to left displays coordinates of active cell(s)
Active Cell – can contain a number, text or a formula
Minimise, Maximise and Close buttons
Toolbars – easy access to commonly-used functions
Sheet tabs – for moving between worksheets
Scroll bars – move vertically or horizontally to different areas of the spreadsheet
Status bar
Microsoft Excel
File Edit View Insert Format Tools Data Window Help
Arial
10
A1
Book1
A B C D E F G H
1 2 3
Sheet1 Sheet2 Sheet3
Ready
NUM

Workbooks and worksheets:

When you start Excel, a blank worksheet is displayed which consists of a multiple grid of cells with numbered rows down the page and alphabetically-titled columns across the page. Each cell is referenced by its coordinates (e.g., A3 is used to refer to the cell in column A and row 3; B10:B20 is used to refer to the range of cells in column B and rows 10 through 20).

Your work is stored in an Excel file called a workbook. Each workbook may contain several worksheets and/or charts – the current worksheet is called the active sheet. To view a different worksheet in a workbook click the appropriate Sheet Tab.

You can access and execute commands directly from the main menu or you can point to one of the toolbar buttons (the display box that appears below the button, when you place the cursor over it, indicates the name/action of the button) and click once.

Moving Around the Worksheet:

It is important to be able to move around the worksheet effectively because you can only enter or change data at the position of the cursor. You can move the cursor by using the arrow keys or by moving the mouse to the required cell and clicking. Once selected the cell becomes the active cell and is identified by a thick border; only one cell can be active at a time.

To move from one worksheet to another click the sheet tabs. (If your workbook contains many sheets, right-click the tab scrolling buttons then click the sheet you want.) The name of the active sheet is shown in bold.

Moving Between Cells:

Here is a keyboard shortcuts to move the active cell:

- Home – moves to the first column in the current row
- Ctrl+Home – moves to the top left corner of the document
- End then Home – moves to the last cell in the document

To move between cells on a worksheet, click any cell or use the arrow keys. To see a different area of the sheet, use the scroll bars and click on the arrows or the area above/below the scroll box in either the vertical or horizontal scroll bars.

Note that the size of a scroll box indicates the proportional amount of the used area of the sheet that is visible in the window. The position of a scroll box indicates the relative location of the visible area within the worksheet.

Entering Data

A new worksheet is a grid of **rows** and **columns**. The rows are labeled with numbers, and the columns are labeled with letters. Each intersection of a row and a column is a **cell.** Each cell has an **address**, which is the column letter and the row number. The arrow on the worksheet to the right points to cell A1, which is currently **highlighted**, indicating that it is an **active cell**. A cell must be active to enter information into it. To highlight (select) a cell, click on it.

To select more than one cell:

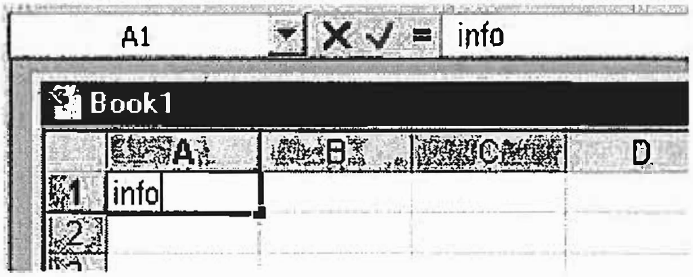
- Click on a cell (e.g. A1), then hold the shift key while you click on another (e.g. D4) to select all cells between and including A1 and D4.
- Click on a cell (e.g. A1) and drag the mouse across the desired range, unclicking on another cell (e.g. D4) to select all cells between and including A1 and D4.
- To select several cells which are not adjacent, press "control" and click on the cells you want to select. Click a number or letter labeling a row or column to select that entire row or column.

One worksheet can have up to 256 columns and 65,536 rows, so it'll be a while before you run out of space.

Each cell can contain a **label, value**, **logical value**, or **formula.**

- Labels can contain any combination of letters, numbers, or symbols.
- Values are numbers. Only values (numbers) can be used in calculations. A value can also be a date or a time
- Logical values are "true" or "false."
- Formulas automatically do calculations on the values in other specified cells and display the result in the cell in which the formula is entered (for example, you can specify that cell D3 is to contain the sum of the numbers in B3 and C3; the

number displayed in D3 will then be a function of the numbers entered into B3 and C3).



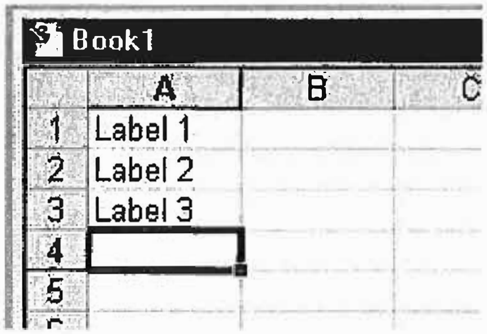
To enter information into a cell, select the cell and begin typing.

Note that as you type information into the cell, the information you enter also displays in the formula bar. You can also enter information into the formula bar, and the information will appear in the selected cell.

When you have finished entering the label or value:

- Press "Enter" to move to the next cell below (in this case, A2)
- Press "Tab" to move to the next cell to the right (in this case, B1)
- Click in any cell to select it

Entering Labels



Unless the information you enter is formatted as a value or a formula, Excel will interpret it as a label, and defaults to align the text on the left side of the cell.

If you are creating a long worksheet and you will be repeating the same label information in many different cells, you can use the **AutoComplete** function. This function will look at other entries in the same column and attempt to match a previous entry with your current entry. For example, if you have already typed "Wesleyan" in another cell and you type "W" in a new cell, Excel will automatically enter "Wesleyan." If you intended to type "Wesleyan" into the cell, your task is done, and you can move on to the next cell. If you intended to type something else, e.g. "Williams," into the cell, just continue typing to enter the term.

To turn on the AutoComplete function, click on "Tools" in the menu bar, then select "Options," then select "Edit," and click to put a check in the box beside "Enable AutoComplete for cell values."

Another way to quickly enter repeated labels is to use the **Pick List** feature. Right click on a cell, then select "Pick From List." This will give you a menu of all other entries in cells in that column. Click on an item in the menu to enter it into the currently selected cell.

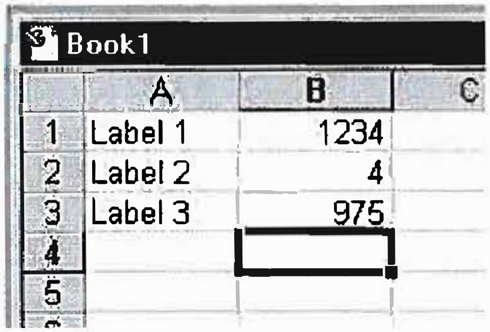
Entering Values

A value is a number, date, or time, plus a few symbols if necessary to further define the numbers [such as: . , + - ( ) % $ /].

Numbers are assumed to be positive; to enter a negative number, use a minus sign "-" or enclose the number in parentheses "()".

Dates are stored as MM/DD/YYYY, but you do not have to enter it precisely in that format. If you enter "jan 9" or "jan-9", Excel will recognize it at January 9 of the current year, and store it as 1/9/2002. Enter the four-digit year for a year other than the current year (e.g. "jan 9, 1999"). To enter the current day's date, press "control" and ";" at the same time.

Times default to a 24 hour clock. Use "a" or "p" to indicate "am" or "pm" if you use a 12 hour clock (e.g. "8:30 p" is interpreted as 8:30 PM). To enter the current time, press "control" and ":" (shift-semicolon) at the same time.



	A	B	C
1	Label 1	1234	
2	Label 2	4	
3	Label 3	975	
4			
5			

An entry interpreted as a value (number, date, or time) is aligned to the right side of the cell, to reformat a value.

To Copy and Paste All Cells in a Sheet

1. Select the cells in the sheet by pressing Ctrl+A (in Excel 2003, select a cell in a blank area before pressing Ctrl+A, or from a selected cell in a Current Region/List range, press Ctrl+A+A).
   OR
   Click Select All at the top-left intersection of rows and columns.

2. Press Ctrl+C.
3. Press Ctrl+Page Down to select another sheet, then select cell A1.
4. Press Enter.

To Copy the Entire Sheet

Copying the entire sheet means copying the cells, the page setup parameters, and the defined range Names.

Option 1:

1. Move the mouse pointer to a sheet tab.
2. Press Ctrl, and hold the mouse to drag the sheet to a different location.
3. Release the mouse button and the Ctrl key.

Option 2:

1. Right-click the appropriate sheet tab.
2. From the shortcut menu, select Move or Copy. The Move or Copy dialog box enables one to copy the sheet either to a different location in the current workbook or to a different workbook. Be sure to mark the Create a copy checkbox.

Option 3:

1. From the Window menu, select Arrange.
2. Select Tiled to tile all open workbooks in the window.
3. Use Option 1 (dragging the sheet while pressing Ctrl) to copy or move a sheet.

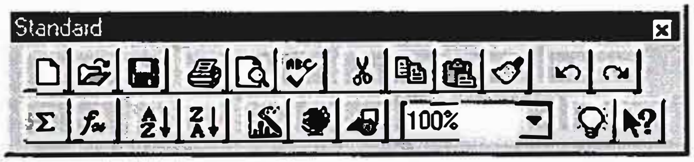
Sorting by Columns
The default setting for sorting in Ascending or Descending order is by row. To sort by columns:

1. From the Data menu, select Sort, and then Options.
2. Select the Sort left to right option button and click OK.
3. In the Sort by option of the Sort dialog box, select the row number by which the columns will be sorted and click OK.

Don't want the value to change when doing calculations:
Assume we have the value 0.105 in cell d1 and we want to multiply every value in column E by that value and place in column F. In cell F1 type
=E1*d$1$ <enter>
The $ tell Excel to always use the value in D1 and not go to the next cell. Then we may drag the formula cell down.

Descriptive Statistics & Displays

The Data Analysis ToolPak has a Descriptive Statistics tool that provides you with an easy way to calculate summary statistics for a set of sample data. Summary statistics includes Mean, Standard Error, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count. This tool eliminates the need to type individual functions to find each of these results. Excel includes elaborate and customizable toolbars, for example the "standard" toolbar shown here:

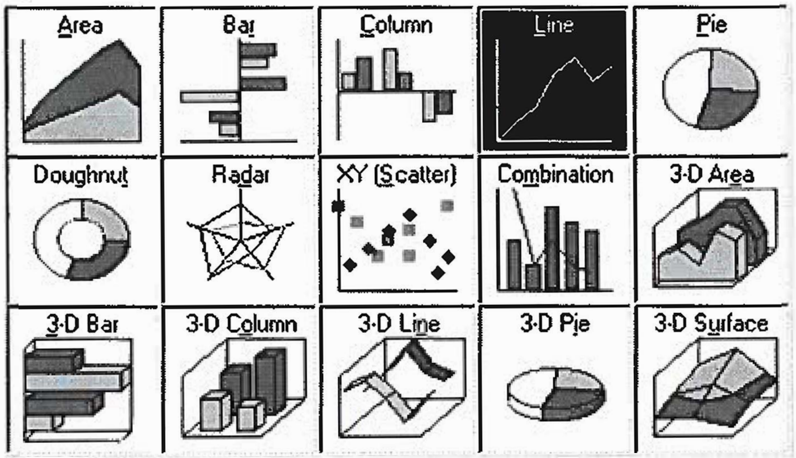


Some of the icons are useful mathematical computation:

Σ is the "Autosum" icon, which enters the formula "=sum()" to add up a range of cells.

fx is the "FunctionWizard" icon, which gives you access to all the functions available.

📊 is the "GraphWizard" icon, giving access to all graph types available, as shown in this display:



Excel can be used to generate measures of location and variability for a variable. Suppose we wish to find descriptive statistics for a sample data: 2, 4, 6, and 8.

Step 1. Select the Tools *pull-down menu, if you see **data analysis, click on this option, otherwise, click on add-in.. option to install** analysis TOOLPAK.
Step 2. Click on the data analysis option.
Step 3. Choose **Descriptive Statistics** from Analysis Tools list.
Step 4. When the dialog box appears:
Enter A1:A4 in the **input range** box, **A1** is a value in **column A** and **row 1**, in this case this value is 2. Using the same technique enter other VALUES until you reach the last one. If a sample consists of 20 numbers, you can select for example A1, A2, A3, etc. as the input range.

Step 5. Select an **output range**, in this case B1. Click on **summary statistics** to see the results.
Select **OK**.
When you click **OK**, you will see the result in the selected range.

As you will see, the mean of the sample is 5, the median is 5, the standard deviation is 2.581989, the sample variance is 6.666667, the range is 6 and so on. Each of these factors might be important in your calculation of different statistical procedures.

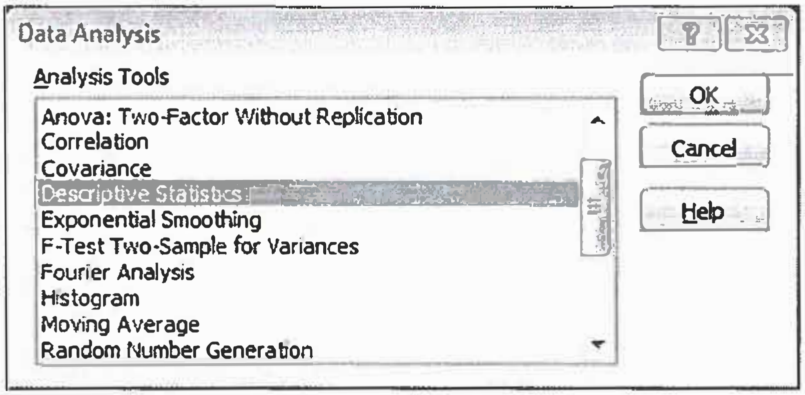
Assume we have the following data:

DA Students-Survey Information

Height	Weight	Travel time	Mode	Family	Rank
70	186	4	1	3	3
66	180	4	1	2	3
69	200	5	4	4	4
72	225	17	4	1	4
66	170	6	4	1	4
72	185	14	2	5	4
70	188	1	1	4	4
75	215	13	1	5	4
72	185	1	2	2	3
71	220	5	3	4	4
69	185	11	4	5	5
71	185	11	3	1	4
71	175	9	3	4	6
69	170	1	2	4	4
68	180	15	2	6	3
70	170	6	1	2	3
70	210	11	3	2	4
72	165	9	3	3	4
70	202	17	2	2	2
74	225	10	3	6	4
64	156	9	2	3	4
71	185	1	2	6	4
69	168	18	3	1	4
72	165	20	2	4	4
71	185	1	3	3	4
72	185	17	3	2	4
75	230	11	3	2	4

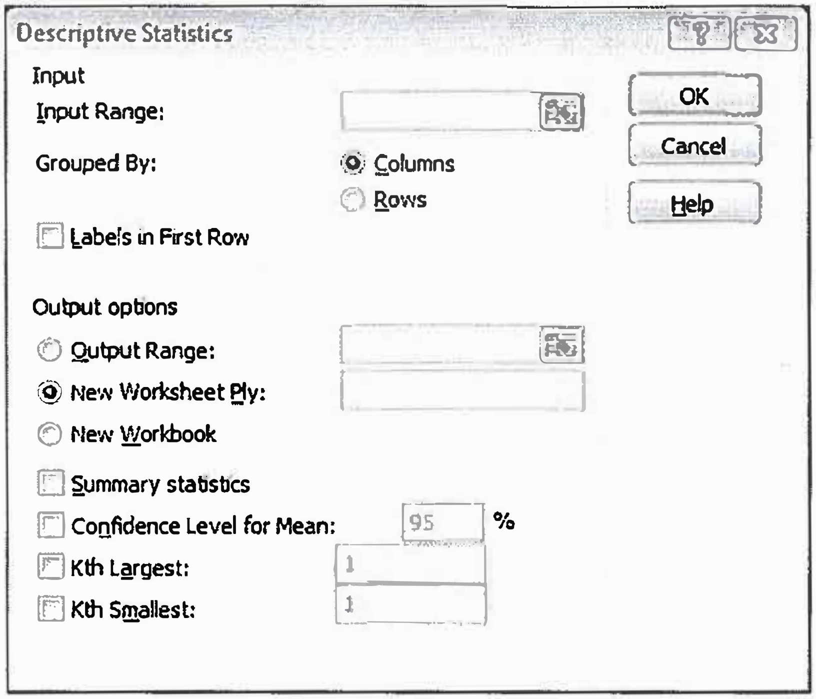
The process:

Go to DATA and Open DATA Analysis ToolPak
Highlightt Descriptive Statistics


Data Analysis

Analysis Tools

- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation


Descriptive Statistics

Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

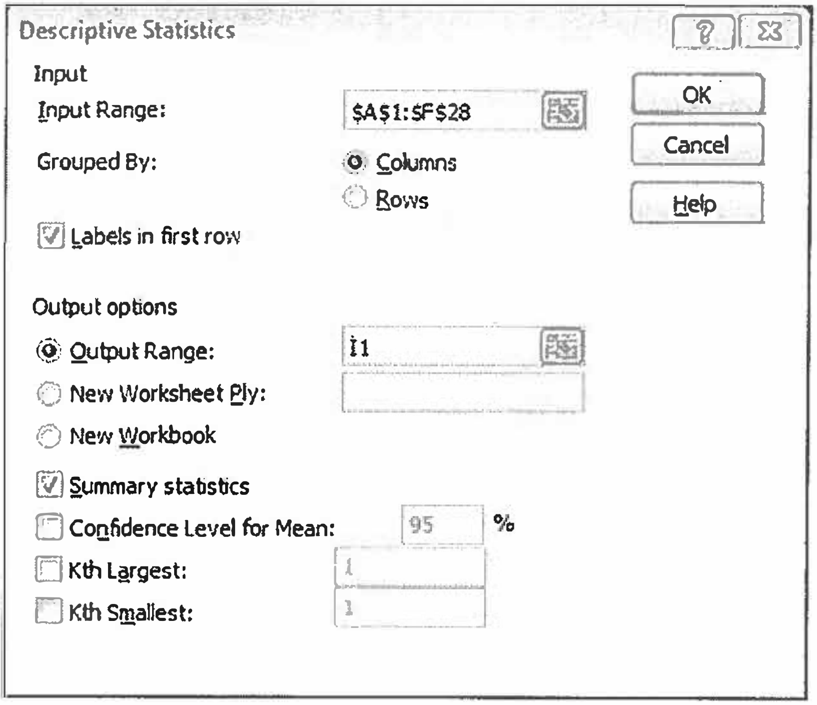
□ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

Enter the columns in the INPUT. Insure to tell Excel whether or not the labels are highlighted. Tell EXCEL where to put the output, either in the current page (cell location) or a different worksheet.
Click on Summary Statistics



Descriptive Statistics

Input

Input Range: $A$1:$F$28

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: I1

○ New Worksheet Ply:

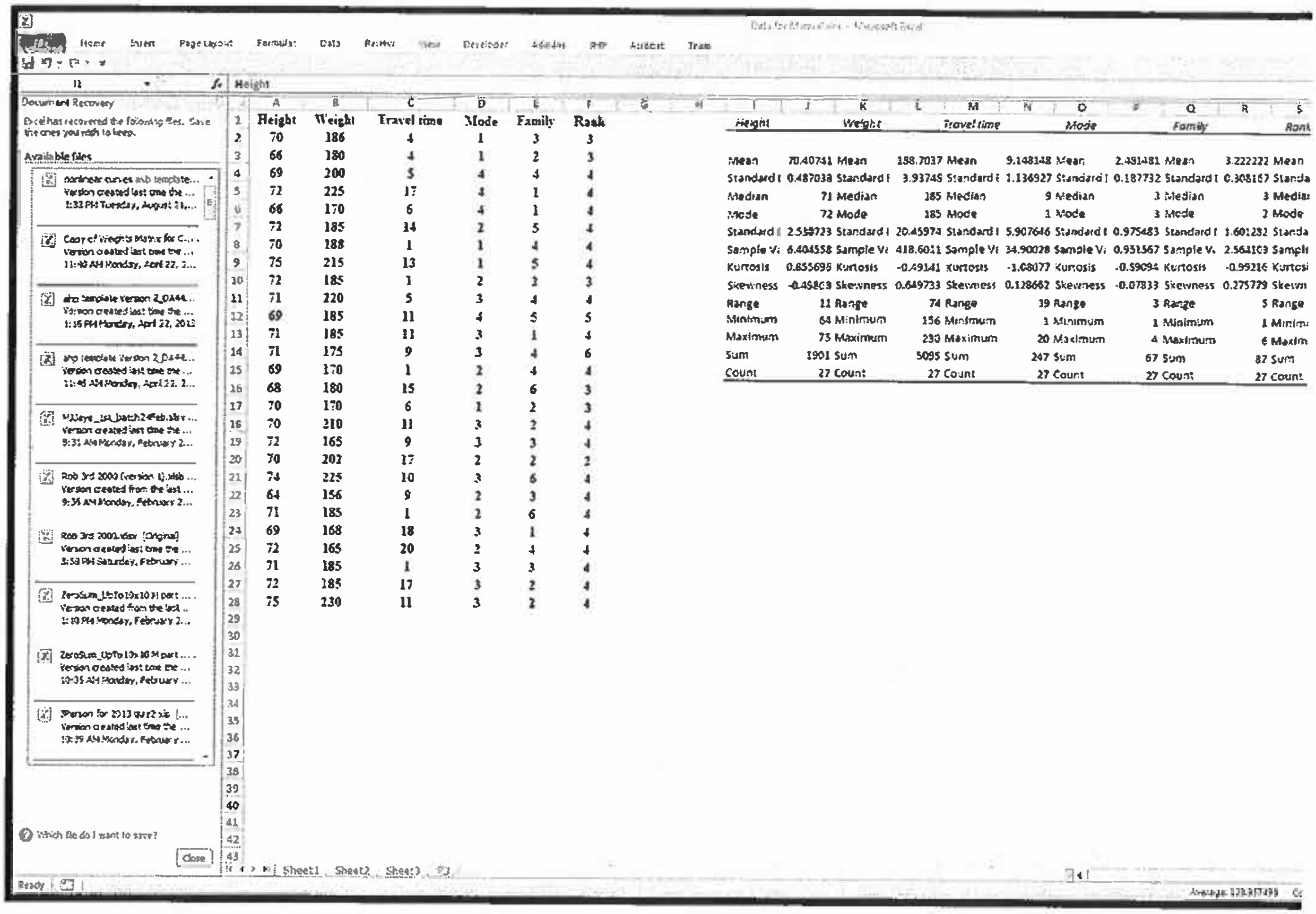
○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1



Document Recovery

Excel has recovered the following files. Save the ones you wish to keep.

Available files

- nonlinear curves lab template... Version created last time the ... 1:33 PM Tuesday, August 21,...
- Copy of Weights Matrix for C... Version created last time the ... 11:40 AM Monday, April 22, 2...
- ahp template version 2_DA44... Version created last time the ... 1:16 PM Monday, April 22, 2013
- ahp template Version 2_DA4... Version created last time the ... 11:46 AM Monday, April 22, 2...
- MJJaye_1st_batch24Feb.xlsx ... Version created last time the ... 9:31 AM Monday, February 2...
- Rob 3rd 2000 (version 1).xlsb ... Version created from the last ... 9:35 AM Monday, February 2...
- Rob 3rd 2002.xlsx [Original] Version created last time the ... 3:53 PM Saturday, February ...
- ZeroSum_UpTo10x10 M part ... Version created from the last ... 1:18 PM Monday, February 2...
- ZeroSum_UpTo10x10 M part ... Version created last time the ... 10:35 AM Monday, February ...
- Pierson for 2013 quiz2.xls [... Version created last time the ... 10:39 AM Monday, February ...

Which file do I want to save?

Close

Height	Weight	Travel time	Mode	Family	Rank
70	186	4	1	3	3
66	180	4	1	2	3
69	200	5	4	4	4
72	225	17	4	1	4
66	170	6	4	1	4
72	185	14	2	5	4
70	188	1	1	4	4
75	215	13	1	5	4
72	185	1	2	2	3
71	220	5	3	4	4
69	185	11	4	5	5
71	185	11	3	1	4
71	175	9	3	4	6
69	170	1	2	4	4
68	180	15	2	6	3
70	170	6	1	2	3
70	210	11	3	2	4
72	165	9	3	3	4
70	202	17	2	2	2
74	225	10	3	6	4
64	156	9	2	3	4
71	185	1	2	6	4
69	168	18	3	1	4
72	165	20	2	4	4
71	185	1	3	3	4
72	185	17	3	2	4
75	230	11	3	2	4

Height		*Weight*		*Travel time*		*Mode*		*Family*		*Rank*
Mean	70.40741	Mean	188.7037	Mean	9.148148	Mean	2.481481	Mean	3.222222	Mean
Standard E	0.487038	Standard E	3.93745	Standard E	1.136927	Standard E	0.187732	Standard E	0.308167	Standa
Median	71	Median	185	Median	9	Median	3	Median	3	Media
Mode	72	Mode	185	Mode	1	Mode	3	Mode	2	Mode
Standard D	2.530723	Standard D	20.45974	Standard D	5.907646	Standard D	0.975483	Standard D	1.601232	Standa
Sample Va	6.404558	Sample Va	418.6011	Sample Va	34.90028	Sample Va	0.951567	Sample Va	2.564103	Sampl
Kurtosis	0.655695	Kurtosis	-0.49141	Kurtosis	-1.08077	Kurtosis	-0.59094	Kurtosis	-0.99216	Kurtosi
Skewness	-0.45803	Skewness	0.649733	Skewness	0.128662	Skewness	-0.07833	Skewness	0.275779	Skewn
Range	11	Range	74	Range	19	Range	3	Range	5	Range
Minimum	64	Minimum	156	Minimum	1	Minimum	1	Minimum	1	Minim
Maximum	75	Maximum	230	Maximum	20	Maximum	4	Maximum	6	Maxim
Sum	1901	Sum	5095	Sum	247	Sum	67	Sum	87	Sum
Count	27	Count	27	Count	27	Count	27	Count	27	Count

Average: 178.9173498

Descriptive statistics output in Excel on data:

Height		*Weight*		*Travel time*	
Mean	70.40741	Mean	188.7037	Mean	9.148148
Standard Error	0.487038	Standard Error	3.93748	Standard Error	1.136927
Median	71	Median	185	Median	9
Mode	72	Mode	185	Mode	1
Standard Deviatic	2.530723	Standard Deviation	20.45974	Standard Deviation	5.907646
Sample Variance	6.404558	Sample Variance	418.6011	Sample Variance	34.90028
Kurtosis	0.855696	Kurtosis	-0.49141	Kurtosis	-1.08677
Skewness	-0.49808	Skewness	0.649733	Skewness	0.128662
Range	11	Range	74	Range	19
Minimum	64	Minimum	156	Minimum	1
Maximum	75	Maximum	230	Maximum	20
Sum	1901	Sum	5095	Sum	247
Count	27	Count	27	Count	27

Mode		*Family*		*Rank*	
Mean	2.481481	Mean	3.222222	Mean	3.851852
Standard Error	0.187732	Standard Error	0.308167	Standard Error	0.138199
Median	3	Median	3	Median	4
Mode	3	Mode	2	Mode	4
Standard Deviation	0.975483	Standard Deviatior	1.601282	Standard Deviation	0.718101
Sample Variance	0.951567	Sample Variance	2.564103	Sample Variance	0.51567
Kurtosis	-0.89094	Kurtosis	-0.99216	Kurtosis	3.442181
Skewness	-0.07833	Skewness	0.275779	Skewness	0.22958
Range	3	Range	5	Range	4
Minimum	1	Minimum	1	Minimum	2
Maximum	4	Maximum	6	Maximum	6
Sum	67	Sum	87	Sum	104
Count	27	Count	27	Count	27

Data Displays

Univariate data:

For height, weight, and travel time we most likely will use ***histograms*** or ***stem & leaf*** plots. Occasionally to compare spread we will use Boxplots. Boxplots ad stem & leaf plots are supplemental programs .For mode of travel, family size, and rank we will use either **pie** or **bar charts**.

Displaying Categorical Data

PIE Chart

The *pie chart* is useful to show the division of a total quantity into component parts. A pie chart, if done correctly, is usually safe from misinterpretation. The total quantity, or 100%, is shown as the entire circle. Each wedge of the circle represents a component part of the total. These parts are usually labeled with *percentages* of the total. Thus, a pie chart helps us see what part of the whole each group forms.

Let's review percentages. Let a represent the partial amount and b represent the total amount. Then P represents a percentage calculated by $P=a/b\ (100)$.

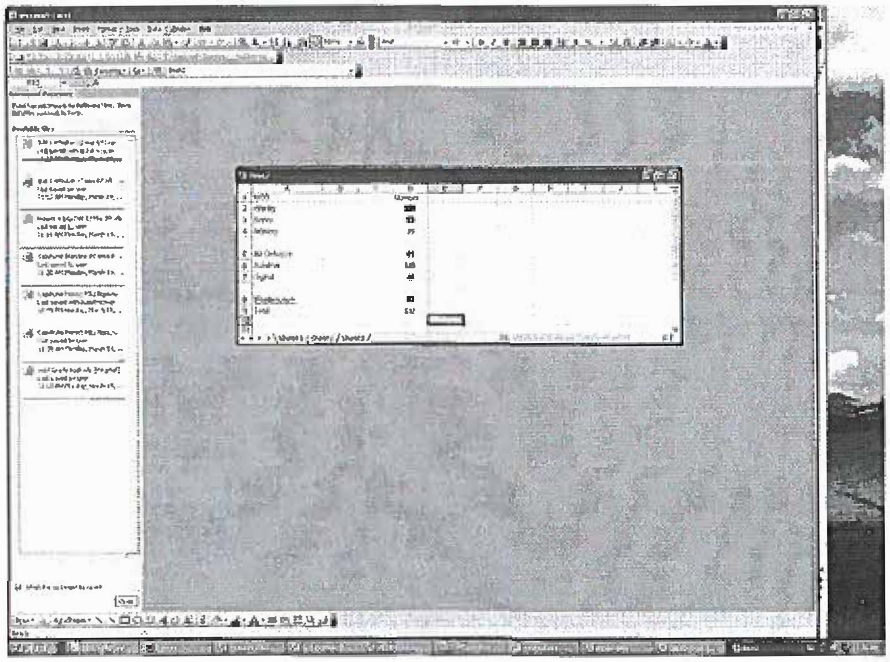
A percentage is thus a part of a whole. For example, $0.25 is what part of $1.00? We let $a=25$ and $b=100$. Then, $P= 25/100\ (100)= 25\%$.

Now, let's see how Excel would create a pie chart for us in the following scenario.

Consider soldiers choosing their MOS. Out of the 632 new soldiers recruited in SC that actually choose a MOS, the breakdown of selection is as follows.

Infantry	250
Armor	53
Artillery	35
Air Defense	41
Aviation	125
Signal	45
Maintenance	83
Total	632

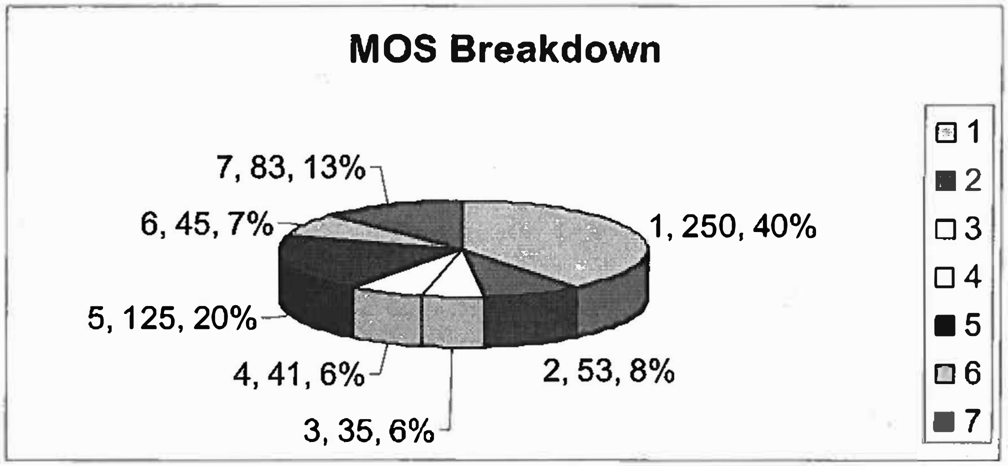
We begin by entering the data into labeled columns.



To select a pie chart, we highlight the labels and the numbers for the seven majors. We then click on Chart Wizard with the mouse, click on Pie, and follow editing directions.

In the Pie chart section, we select the type of chart to display. Select *No Legend*, show label and percent, and choose to display as a new sheet.

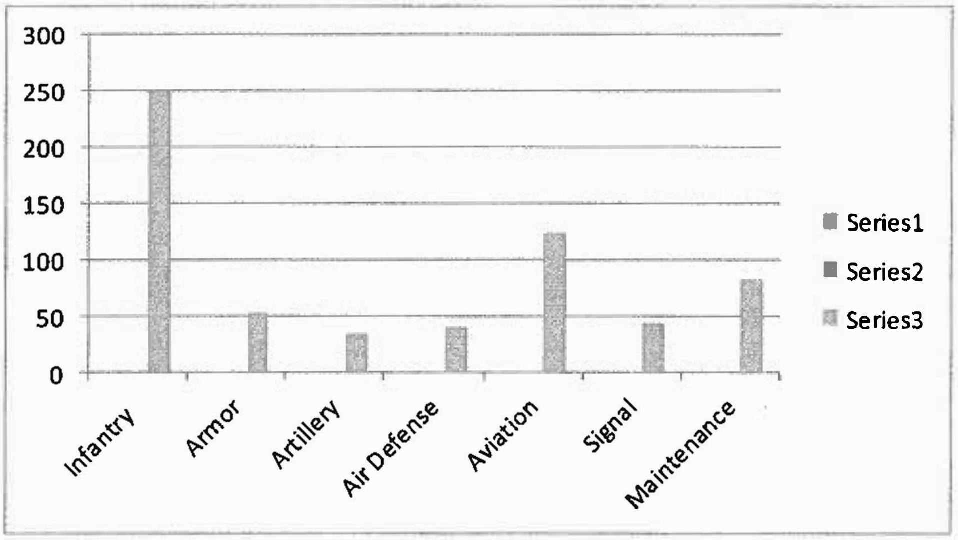
The output for the Pie Chart looks as follows:



Each of the shaded regions displays the percentage (%) of soldiers out of 632 that chose that MOS. Clearly Infantry has the largest percent of recruits. Which MOS appears to have the least?

What advantages and disadvantages can you see with using Pie Charts?

Let's view a bar chart:



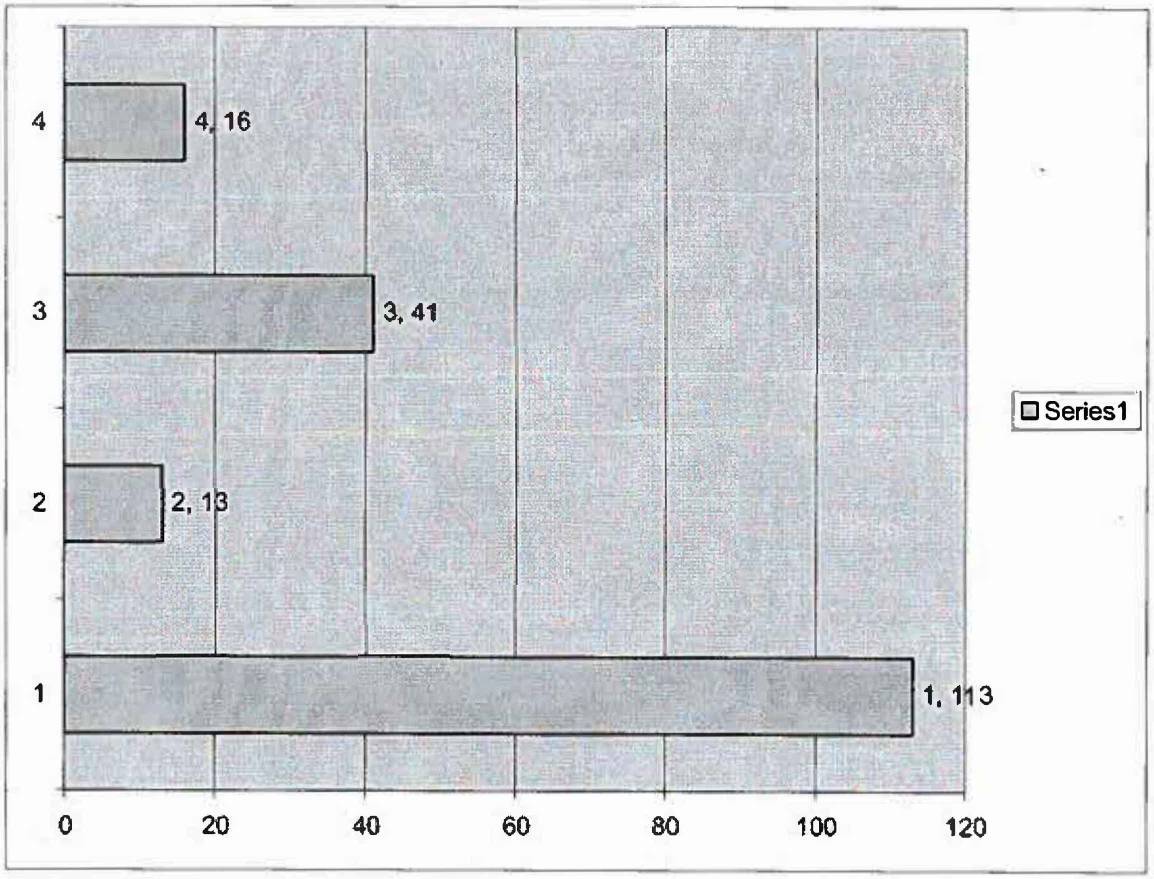
Is this clearer to make your point that the pie chart?

Bar Chart

Bar charts are useful when comparing relative sizes of data groups especially when they come from **categorical** variables. For example, consider the eye color from patients visiting the local eye clinic last year.

Eye Color	***Count***	***Percent***
Blue	***113***	***61.7486***
Green	***13***	***7.10383***
brown	***41***	***22.4044***
mixed	***16***	***8.74317***
Total	***183***	***100***

Enter the data then
Click on Chart Wizard and obtain a Bar Chart.
Or Insert CHart



Displaying Quantitative Data

In quantitative data we are concerned with the shape of the data. Shape refers to symmetry of data.” Is it symmetric?”” Is it skewed?” are questions we ask and answer.

Stem and Leaf

A stem-and Leaf plot uses the real data points in making a plot. The plot will appear strange because your plot is sideways. The rules are as follows:

Step 1: Order the data

Step 2: Separate according to the one or more leading digits. List stems in a vertical column.

Step 3: Leading digit is the stem and trailing digit is the leaf. For example 32, 3 is the stem and 2 is the leaf. Separate the stem from the leafs by a vertical line.

Step 4. Indicate the units for stems and leafs in the display.

You will probably create these plots by hand (***Excel*** will not produce a stem and leaf plot).

Example: Grades for 20 students in a course
53, 55, 66, 69, 71, 78, 75, 79, 77, 75, 76, 73, 82, 83, 85, 74, 90, 92, 95, 99

Stems are the leading digit:
5
6
7
8
9
Standing for 50-s, 60's, 70's, 80's, and 90's.

If there had been a score of 100, then the leading digit is in 100's. So we would need:
05
06
07
08
09
10
for 50's, 60's, 70's, 80's, 90's, and 100's

Draw a vertical line after each stem.

5|
6|
7|
8|
9|

Now add the leafs, which are the trailing digits,

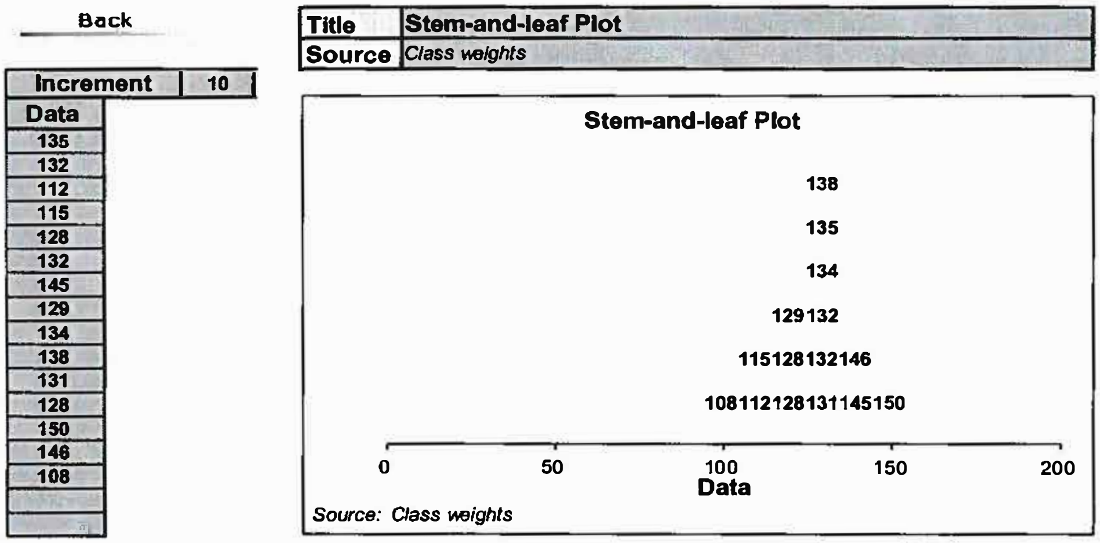
53, 55, 66, 69, 71, 73, 74, 75, 75, 76, 77, 78,79, 82, 83, 85, 90, 92, 95, 99

5| 3, 5
6| 6, 9
7| 1, 3, 4, 5, 5, 6, 7, 8, 9
8| 2, 3, 5
9| 0, 2, 5, 9

We can characterize this shape as almost *symmetric*. Note how we read the values from the stem-and-leaf.

For example, we read
5| 3, 5
as data elements 53 and 55.

We have a program in Excel that gives us a stem & leaf plot. It is still up to us to determine the shape. Here is an example with class weights.



Back

Increment	10
Data	
135	
132	
112	
115	
128	
132	
145	
129	
134	
138	
131	
128	
150	
146	
108	

Title	Stem-and-leaf Plot
Source	*Class weights*

1.4 Displaying quantitative data with Histograms

We begin by stating that there is a difference between bar charts and histograms. Bar charts have discrete values as their horizontal axis. Thus, bars are centered at discrete values. A histogram has continuous values as its horizontal axis. Thus, there are no spaces between the bars unless no data in in that range. Since most of the data that you will use are large, we will go quickly to displays with technology.

Steps:

(1) Obtain descriptive statistics for the data or order the data smaller to larger
(2) Determine the Interval [smallest, largest]
(3) Calculate the class intervals (largest-smallest)/n where n is the number of intervals desired. The value of n must be between 5 and 20. Start with 5 and go up until a good view of the histogram is obtained.
(4) List the endpoints as Bin values
(5) Go to Data Analysis, Histogram and bring up dialog box. Put data in data input and endpoints in bins.
(6) The output is a table.
(7) Highlight the frequencies of the table and go to insert Bar chart
(8) Right click in bar char (on a bar) and close GAP size to 0.
(9) Comment on shape in regard to symmetry and skewness.

Histograms of data series can be created using the Analysis ToolPak's Histogram tool. Data is grouped into intervals (known as bins) and the number of observations that fall into each are displayed both in a table and, also graphically, as a bar chart. We must edit the bar chart so that the gap width is 0 to be a histogram.

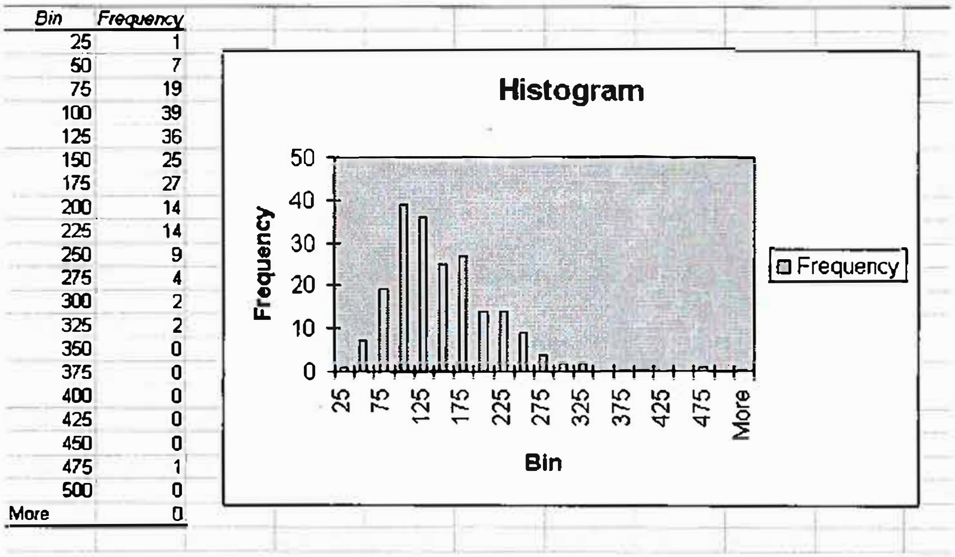
Example Histogram with corrected Excel problem

Using the potato data select **Tools > Data Analysis... > Histogram...** from the menu bar and a dialog box will appear. Insert the input range by either entering the reference by keyboard or highlighting the input range on the worksheet. Tick the **label** box if labels are included in the input range.

Set up the ranges for the histogram divisions (e.g. a column of numbers starting at 50 and going up in steps of 50) and enter the reference for the bin range. If cells E2=50,E3=100,...E11=500, then the reference given is $E$2:$E$11. If this box is left empty Excel will generate a default range. Define where you want the output to appear by entering the cell reference of the top left corner of where you want it to go.

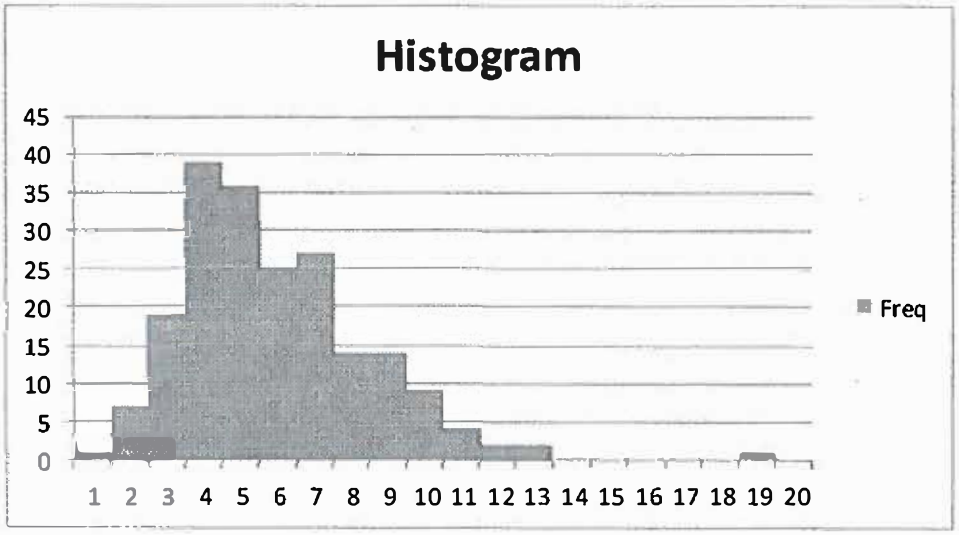
Tick the chart output box to obtain the histogram and click on the OK button when you have finished. The histogram for the variable **weight** could look like the one on the right, after you have stretched it vertically.

Below is the output generated by the Histogram tool for the **weight** data using a step-size of 25 instead of 50 as in the previous graph.



Bin	Frequency
25	1
50	7
75	19
100	39
125	36
150	25
175	27
200	14
225	14
250	9
275	4
300	2
325	2
350	0
375	0
400	0
425	0
450	0
475	1
500	0
More	0

We must close the gap width to 0 for a histogram:



Our examination shows the data is skewed right.

Histogram

Using the Histogram tool will allow us to make a histogram and create a frequency distribution chart at the same time.

1. Click Tools > Data Analysis.... If you don't have a Data Analysis option under Tools, see step #3 of "How do I get started with Excel?".

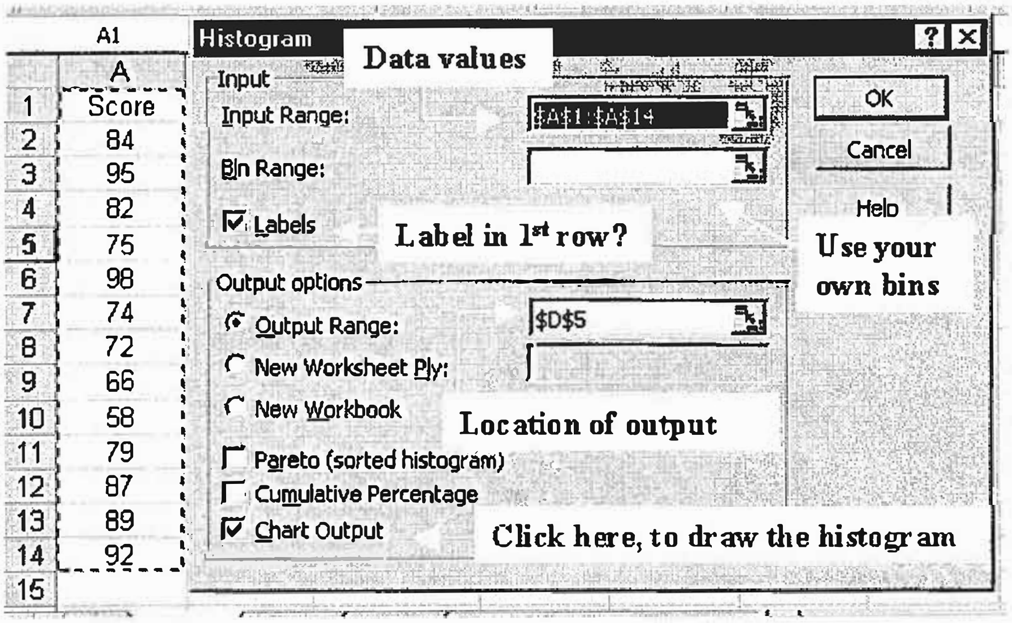
2. In the Data Analysis window, select Histogram and click OK.

3. A new window titled Histogram should appear. This window has many options. Below is a brief explanation of each:

    - Input Range is where the data being used to create the histogram goes. Simply put your cursor back into the spreadsheet and highlight the variable name and all the data in that column.
    - More information about Bin Range.
    - Click Labels. If a variable name was highlighted in the Input Range, then this needs to be checked.
    - You must select one of the following Output options:
        - Click Output Range if you want the histogram to be placed on the current sheet. Next, simply input the cell where you want the output to be placed.
        - Click New Worksheet Ply if you want the histogram to be placed on a new sheet. Next, type the name of the new sheet where you want the output to be placed.
    - Clicking Cumulative Percentage will list the cumulative percentage for each class and include a cumulative percentage line on your histogram.
    - Click Chart Output under Output options. This step is ***necessary*** to obtain the histogram. If this is not highlighted you will only receive a frequency distribution chart.

4. Click OK. The histogram and frequency distribution chart should be placed onto your spreadsheet.

An example:



	A
1	Score
2	84
3	95
4	82
5	75
6	98
7	74
8	72
9	66
10	58
11	79
12	87
13	89
14	92
15	

Input Range: $A$1:$A$14

Bin Range:

☑ Labels

● Output Range: $D$5

○ New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

☑ Chart Output

Data values

Label in 1st row?

Use your own bins

Location of output

Click here, to draw the histogram

We will present the information on how to construct a histogram using EXCEL.

Histogram:

Step 1. Determine and select the classes, 5-15 classes. Find the range (lowest to highest value). Classes should be evenly spaced if possible.

Step 2. Tally the data in the classes.

Step 3. Find the numerical (relative) frequencies from the tallies.

Step 4. Find the cumulative frequencies.

Histogram: connects class interval as a base and tallies (or relative frequencies) as the height of a rectangle. Rectangle is centered at the mid-point of class interval.

53, 55, 66, 69, 71, 73, 74, 75, 75, 76, 77, 78,79, 82, 83, 85, 90, 92, 95, 99

Possible class intervals:

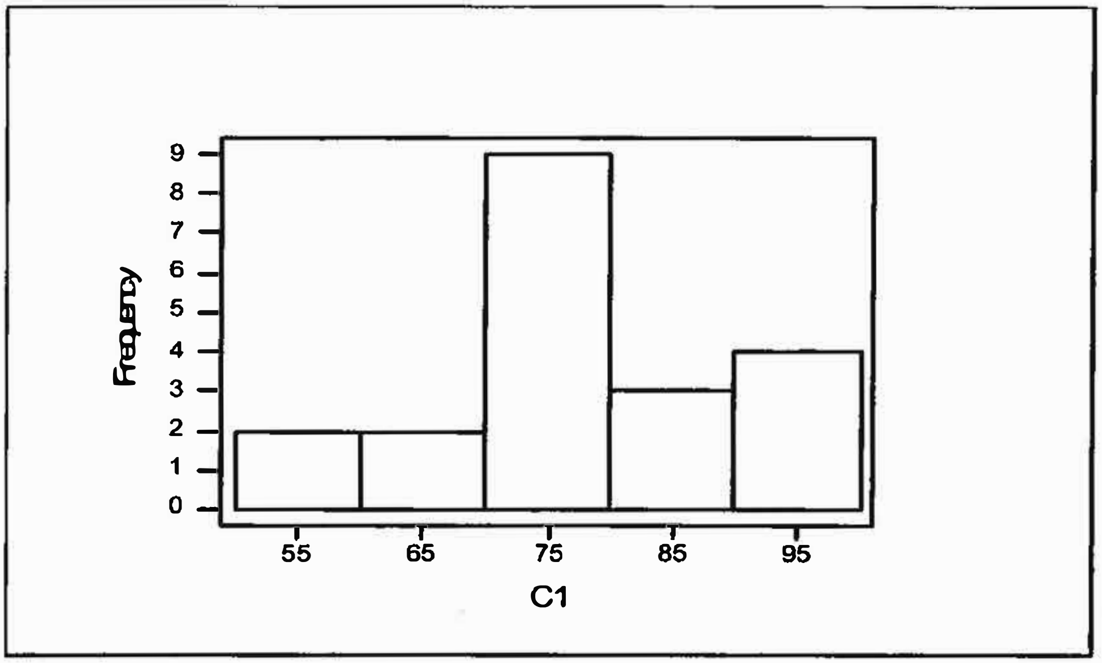
(a) Classes 51-60,61-70,71-80,81-90, 91- 100
(5 classes intervals)

(b) Classes 50-59, 60-69,70-79, 80-89, 90-99, 100-109
(6 classes intervals)

(c) Classes 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, 81-85, 86-90, 91-95, 96-100
(10 class intervals)

Let's use selection (a)

Interval	Tally	Decimal
51 - 60	2	2/20 = .10
61 - 70	2	2/20 = .10
71 - 80	9	9/20 = .45
81 - 90	4	4/20 = .2
91 - 100	3	3/20 = .15
Total	20	20/20 = 1.00



We note the data is somewhat symmetric.

We will present the information on how to construct and use a boxplot. I believe a have an Excel program to do boxplots. Boxplots are a good way to compare data sets from multiple sources. For example, let's look at violence in a 10 regions in Afghanistan. Putting the 10 boxplots together allows us to compare many aspects such as medians, ranges, and dispersions.

Boxplot

Step 1. Draw a horizontal measurement scale that includes all data within the range of data.

Step 2. Construct a rectangle (the box) whose left edge is the lower quartile value and whose right edge is the upper quartile value.

Step 3. Draw a vertical line segment in the box for the median value.

Step 4. Extend line segments from rectangle to the smallest and largest data values (these are called whiskers).

53, 55, 66, 69, 71, 73, 74, 75, 75, 76, 77, 78, 79, 82, 83, 85, 90, 92, 95, 99

The values are in numerical order. What is needed are the range, the quartiles, and the median.

Range is the smallest and largest values from the data: 53 and 99.

The median is the middle value. It is the average of the 10th and 11th values as we will see later: *(76 + 77)/ 2 = 76.5*

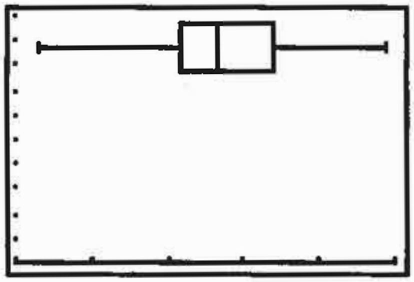
The quartiles values are the median of the lower and upper half of the data.

Lower quartile values: *53, 55, 66, 69, 71, 73, 74, 75, 75, 76.* Its median is 72.

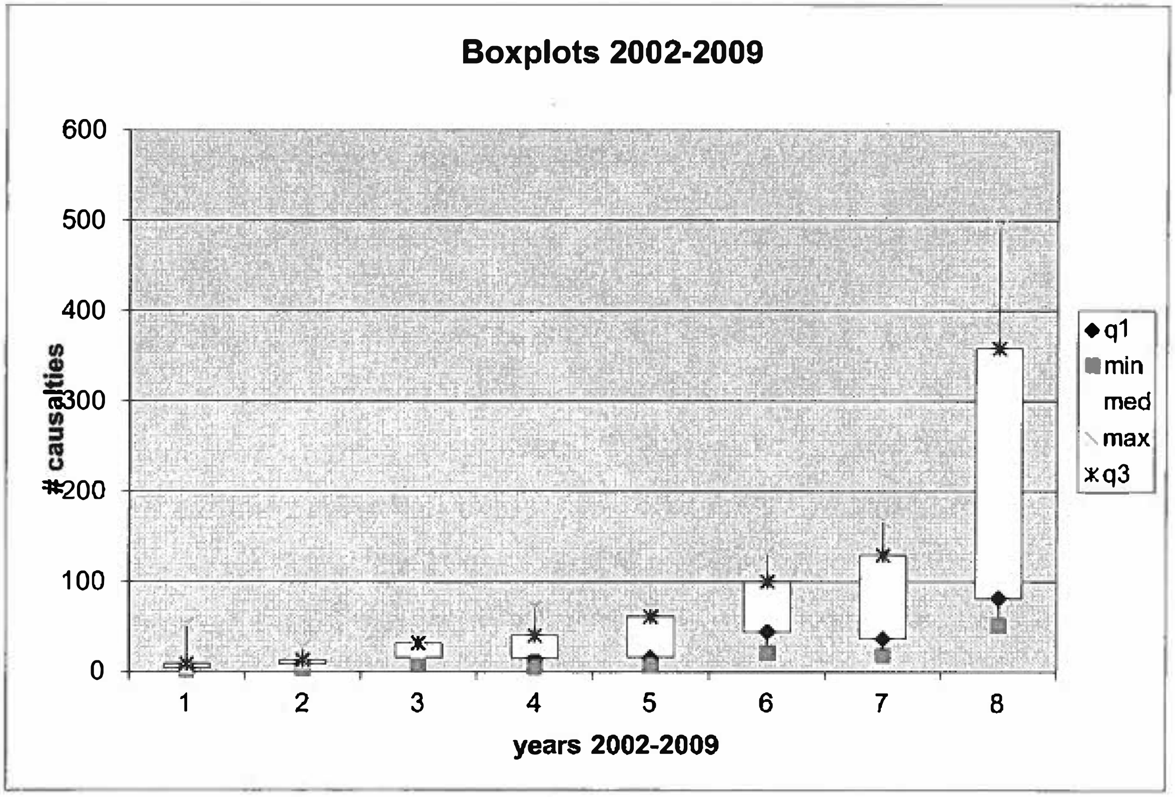
Upper quartile values: *77, 78,79, 82, 83, 85, 90, 92, 95, 99.* Its median is 84.

You draw a rectangle from 72 to 84 with a vertical line at 76.5

Then draw a whisker to the left to 53 and to the right to 99.
It would look something like this:



Comparisons;
Consider our data for casualties in Afghanistan through the years 2002-2009. This is presented to you as a commander. What information is this telling you?



Bivariate data:

We will look at this later.

To display the data, we use the chart wizard or INSERT. We have pie charts, bar charts, line, XY-scatter, and histograms that we might obtain. Just look at the options in the available charts. Presenting the most appropriate chart is the key.

Distributions and Probabilities from Distributions

Discrete distributions: Binomial, & Poisson.

Continuous distribution: Normal, exponential

Binomial Distribution

Command is =BINOMDIST(number_s,trials,probability_s,Cumulative)

The BINOMDIST function syntax has the following arguments:

- **Number_s Required. The number of successes in trials.**
- **Trials Required. The number of independent trials.**
- **Probability_s Required. The probability of success on each trial.**
- **Cumulative Required. A logical value that determines the form of the function. If cumulative is TRUE, then BINOMDIST returns the cumulative distribution function, which is the probability that there are at most number_s successes; if FALSE, it returns the probability mass function, which is the probability that there are number_s successes.**

Example: Assume we have a binomial distribution problem

For our problem, n=15, p(s)=.672, we want to find the following probabilities

a) P(X=4)

b) $P(X\leq4)$

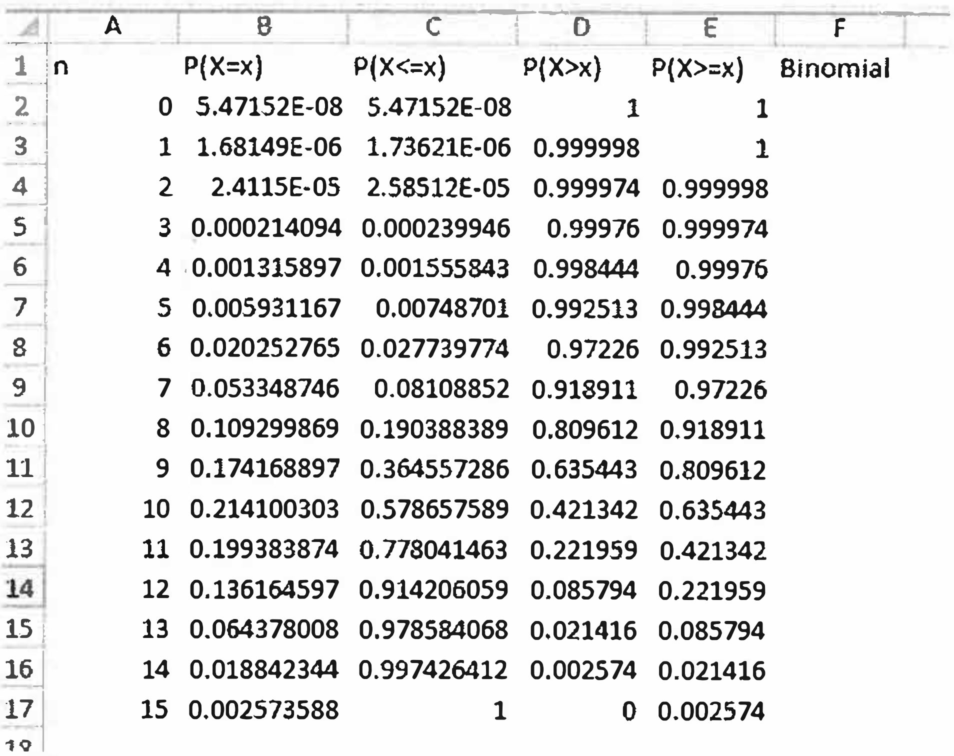
c) P(X<4)

d) P(X>4)

e) $P(X\geq4)$

To start, open Excel. Label your column in row 1. In column A list the numbers from 0 to 15. In column B2 type

BINOMDIST(a2,15,.672,false) to get P(X=x). In column C2 type BINOMDIST(a2,15,.672,true) to get P(X≤x). To get P(X>x), in cell D2 type 1-BINOMDIST(a2,15,.673,true). To get P(X>=x) in Cell E2 type =b2+d2



	A	B	C	D	E	F
1	n	P(X=x)	P(X<=x)	P(X>x)	P(X>=x)	Binomial
2	0	5.47152E-08	5.47152E-08	1	1	
3	1	1.68149E-06	1.73621E-06	0.999998	1	
4	2	2.4115E-05	2.58512E-05	0.999974	0.999998	
5	3	0.000214094	0.000239946	0.99976	0.999974	
6	4	0.001315897	0.001555843	0.998444	0.99976	
7	5	0.005931167	0.00748701	0.992513	0.998444	
8	6	0.020252765	0.027739774	0.97226	0.992513	
9	7	0.053348746	0.08108852	0.918911	0.97226	
10	8	0.109299869	0.190388389	0.809612	0.918911	
11	9	0.174168897	0.364557286	0.635443	0.809612	
12	10	0.214100303	0.578657589	0.421342	0.635443	
13	11	0.199383874	0.778041463	0.221959	0.421342	
14	12	0.136164597	0.914206059	0.085794	0.221959	
15	13	0.064378008	0.978584068	0.021416	0.085794	
16	14	0.018842344	0.997426412	0.002574	0.021416	
17	15	0.002573588	1	0	0.002574	

a) P(X=4)=0.001315897

b) P(X≤4)=0.00155843

c) P(X<4)=P(X<=3)=0.000239946

d) P(X>4)=P(X>=5)=0.998444

e) P(X≥4)=0.99976

Poisson Distribution

Command:

=Poisson(x, mean, cumulative)

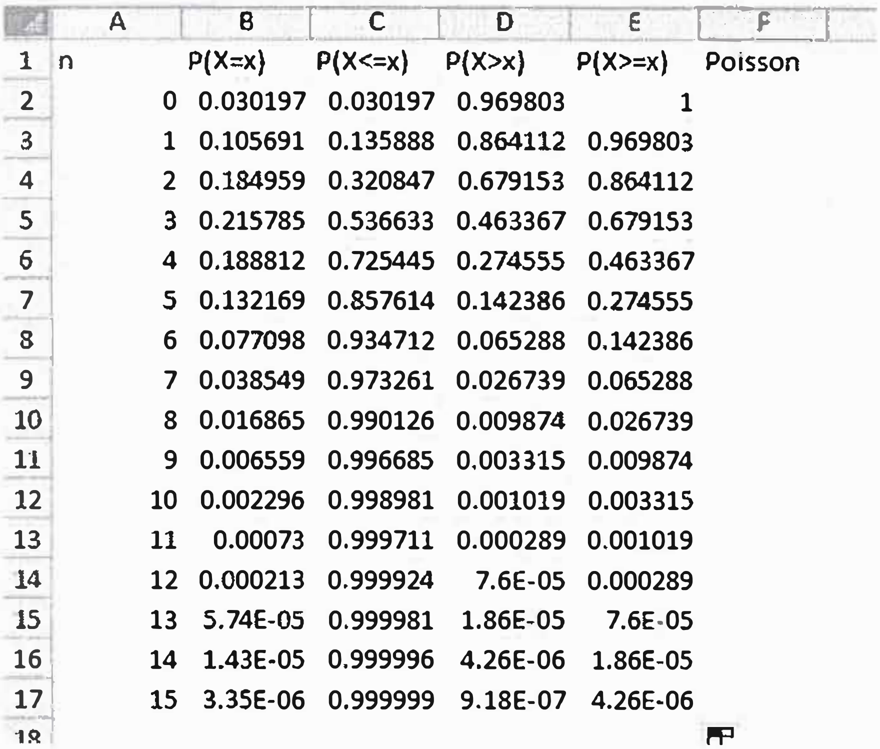
For p(X=x) use =Poisson(x, mean, False). For P(X≤x) use =Poisson(x, mean, true).

Example, assume the number of IEDS follow a Poisson distribution with mean 3.5.

In Row 1 place your labels

N in A1, P(X=x) in B1, P(X<=x) in C1, P(X>x) in D1 and P(X>=x) in E1.

Issue: we do not know in advance how far to go for n. Start with n from 0 to some reasonable value like 15. Expand if necessary.



	A	B	C	D	E	F
1	n	P(X=x)	P(X<=x)	P(X>x)	P(X>=x)	Poisson
2	0	0.030197	0.030197	0.969803	1	
3	1	0.105691	0.135888	0.864112	0.969803	
4	2	0.184959	0.320847	0.679153	0.864112	
5	3	0.215785	0.536633	0.463367	0.679153	
6	4	0.188812	0.725445	0.274555	0.463367	
7	5	0.132169	0.857614	0.142386	0.274555	
8	6	0.077098	0.934712	0.065288	0.142386	
9	7	0.038549	0.973261	0.026739	0.065288	
10	8	0.016865	0.990126	0.009874	0.026739	
11	9	0.006559	0.996685	0.003315	0.009874	
12	10	0.002296	0.998981	0.001019	0.003315	
13	11	0.00073	0.999711	0.000289	0.001019	
14	12	0.000213	0.999924	7.6E-05	0.000289	
15	13	5.74E-05	0.999981	1.86E-05	7.6E-05	
16	14	1.43E-05	0.999996	4.26E-06	1.86E-05	
17	15	3.35E-06	0.999999	9.18E-07	4.26E-06	
18						

N=15 is far enough.

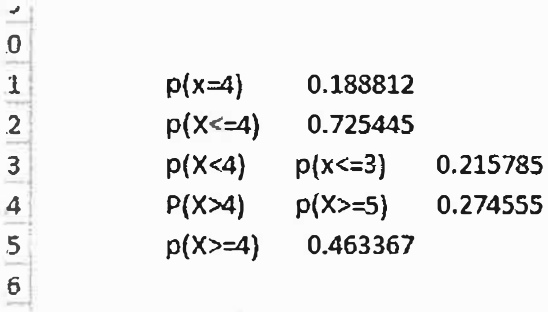
a) **$P(X=4)$**

b) **$P(X\le4)$**

c) **$P(X<4)$**

d) **$P(X>4)$**

e) **$P(X\ge4)$**



0			
1	p(x=4)	0.188812	
2	p(X<=4)	0.725445	
3	p(X<4)	p(x<=3)	0.215785
4	P(X>4)	p(X>=5)	0.274555
5	p(X>=4)	0.463367	
6			

Continuous Distributions

Exponential distribution

$f(x)= 1/a \exp(-x/a)$ for $x >0$

where a = mean.

The probability for P(X=x) does not exist for continuous distributions. Also P(X<x) and P(X<=x) are equivalent.

$P(X<x)=P(X<=x)=1-exp(x/a)$

For example, a radar sensor operates and finds speeders at a rate of 0.119 per hour. Find the probability that we our next contact is within 10 minutes. We convert 10 minutes to 1/6 of an hour to keep units straight.

X=1/6

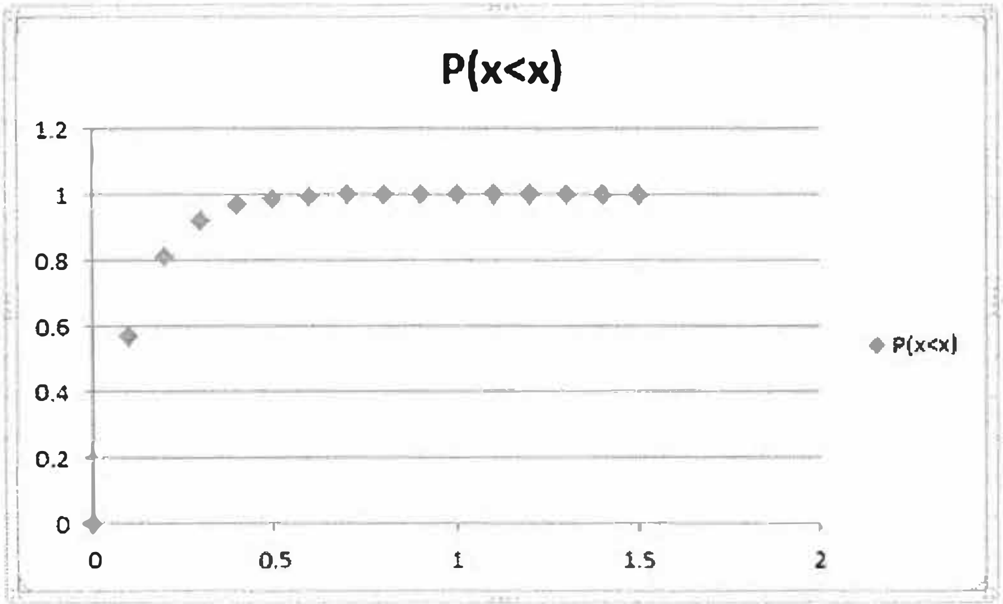
Mu=8.4

1-exp(-8.4/6)=0.7535

P(X<1/6) = 0.7535

P(X>1.6)=1-.7535=0.2465

The cumulative probability looks like



Normal Distribution

Consider the problem of finding the probability of getting less than a certain value under any normal probability distribution. As an illustrative example, let us suppose the SAT scores nationwide are normally distributed with a mean and standard deviation of 500 and 100, respectively. Answer the following questions based on the given information:

A: What is the probability that a randomly selected student score will be less than 600 points?
B: What is the probability that a randomly selected student score will exceed 600 points?
C: What is the probability that a randomly selected student score will be between 400 and 600?

Hint: Using Excel you can find the probability of getting a value approximately less than or equal to a given value. In a problem, when the mean and the standard deviation of the population are given, you have to use common sense to find different probabilities based on the question since you know the area under a normal curve is 1.

Solution:

In the work sheet, select the cell where you want the answer to appear. Suppose, you chose cell number one, A1. From the menus, select "insert pull-down".

Steps 2-3 From the menus, select insert, then click on the Function option.

Step 4. After clicking on the Function option, the Paste Function dialog appears from Function Category. Choose **Statistical** then ***NORMDIST*** from the ***Function Name*** box; Click ***OK***

Step 5. After clicking on OK, the NORMDIST distribution box appears:
i. Enter 600 in X (the value box);
ii. Enter 500 in the Mean box;
iii. Enter 100 in the Standard deviation box;
iv. Type "true" in the cumulative box, then click OK.

As you see the value 0.84134474 appears in A1, indicating the probability that a randomly selected student's score is below 600 points. Using common sense we can answer part "b" by subtracting 0.84134474 from 1. So the part "b" answer is 1-0.8413474 or 0.158653. This is the probability that a randomly selected student's score is greater than 600 points. To answer part "c", use the same techniques to find the probabilities or area in the left sides of values 600 and 400. Since these areas or probabilities overlap each other to answer the question you should subtract the smaller probability from the larger probability. The answer equals
0.84134474 - 0.15865526 = 0.68269.

Inverse Case

Calculating the value of a random variable often called the "x" value

You can use **NORMINV** from the function box to calculate a value for the random variable - if the probability to the left side of this variable is given. Actually, you should use this function to calculate different percentiles. In this problem one could ask what is the score of a student whose percentile is 90? This means approximately 90% of students scores are less than this number. On the other hand if we were asked to do this problem by hand, we would have had to calculate the x value using the normal distribution formula x = m + zd. Now let's use Excel to calculate P90. In the Paste function, dialog click on statistical, then click on **NORMINV**. The screen shot would look like the following:

When you see **NORMINV** the dialog box appears.
i. Enter 0.90 for the probability (this means that approximately 90% of students' score is less than the value we are looking for)
ii. Enter 500 for the mean (this is the mean of the normal distribution in our case)
iii. Enter 100 for the standard deviation (this is the standard deviation of the normal distribution in our case)

At the end of this screen you will see the formula result which is approximately 628 points. This means the top 10% of the students scored better than 628.

Confidence Interval for the Mean.

Suppose we wish for estimating a confidence interval for the mean of a population. Depending on the size of your sample size you may use one of the following cases:

Large Sample Size (n is larger than, say 30):

The general formula for developing a confidence interval for a population means is:

$$\bar{x} \pm Z \bullet (S / \sqrt{n})$$

In this formula $\bar{x}$ is the mean of the sample; Z is the interval coefficient, which can be found from the normal distribution table (for example the interval coefficient for a 95% confidence level is 1.96). S is the standard deviation of the sample and n is the sample size.

Now we would like to show how Excel is used to develop a certain confidence interval of a population mean based on a sample information. As you see in order to evaluate this formula you need $\bar{x}$ "the mean of the sample" and the margin of error "$Z \bullet (S / \sqrt{n})$." Excel will automatically calculate these quantities for you.

The only things you have to do are:

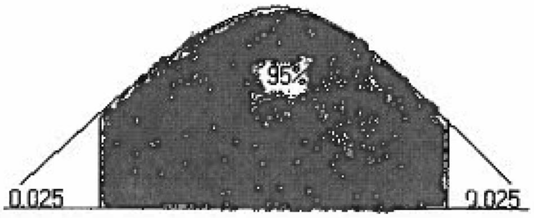
add the margin of error "$Z \bullet (S / \sqrt{n})$." to the mean of the sample, $\bar{x}$; Find the upper limit of the interval and subtract the margin of error from the mean to the lower limit of the interval. To demonstrate how Excel finds these quantities we will use the data set, which contains the hourly income of 36 work-study students here, at the University of Baltimore. These numbers appear in cells A1 to A36 on an Excel work sheet.

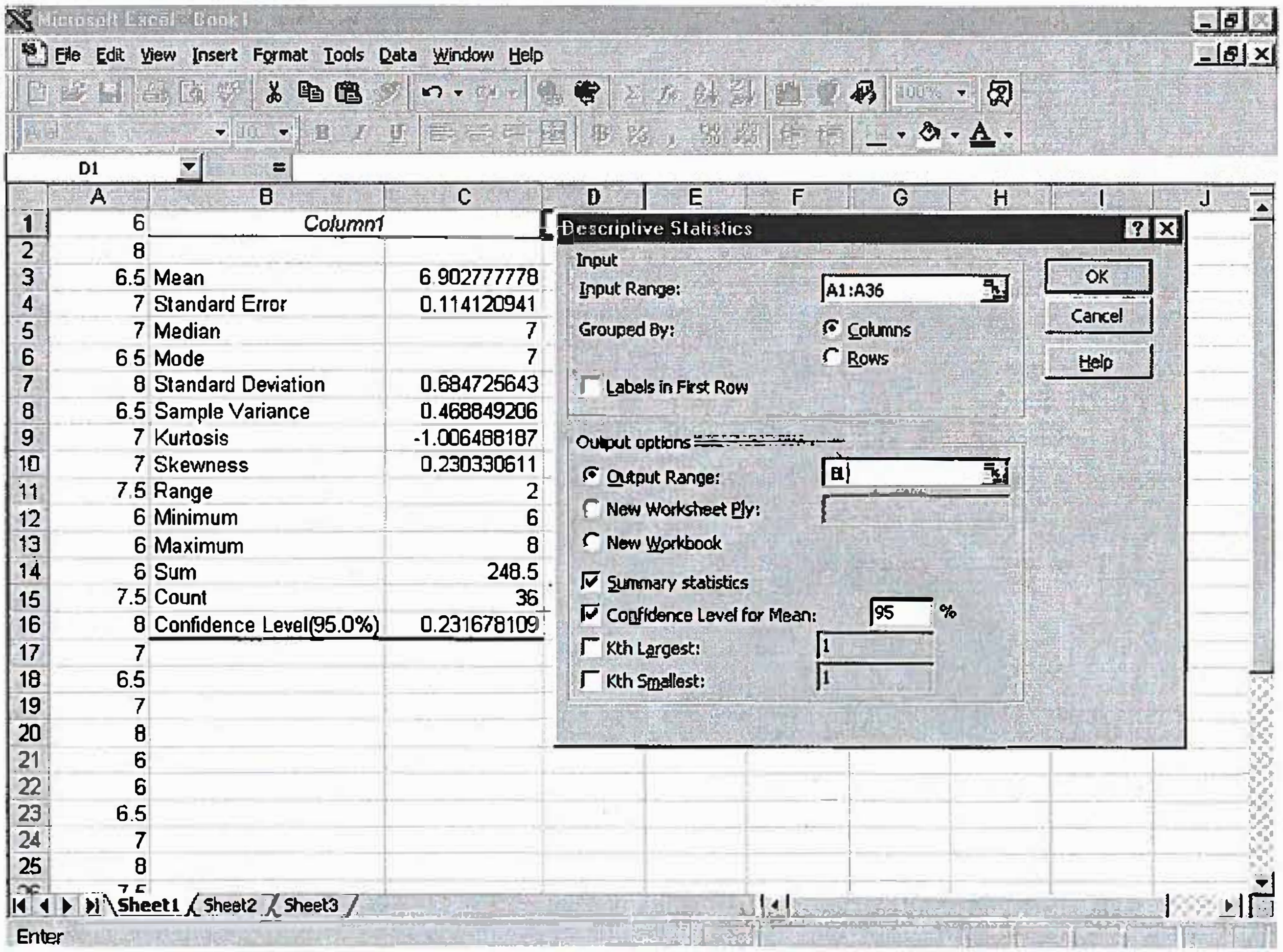
After entering the data, we followed the descriptive statistic procedure to calculate the unknown quantities. The only additional step is to click on the confidence interval in the descriptive statistics dialog box and enter the given confidence level, in this case 95%.

Here is, the above procedures in step-by-step:

Step 1. Enter data in cells A1 to A36 (on the spreadsheet)
Step 2. From the menus select **Tools**
Step 3. Click on **Data Analysis** then choose the **Descriptive Statistics** option then click **OK**.

On the descriptive statistics dialog, click on Summary Statistic. After you have done that, click on the confidence interval level and type 95% - or in other problems whatever confidence interval you desire. In the Output Range box enter B1 or what ever location you desire.
Now click on **OK**. The screen shot would look like the following minus the graph below (graph not producible by EXCEL):





	A	B	C
1	6	*Column1*	
2	8		
3	6.5	Mean	6.902777778
4	7	Standard Error	0.114120941
5	7	Median	7
6	6.5	Mode	7
7	8	Standard Deviation	0.684725643
8	6.5	Sample Variance	0.468849206
9	7	Kurtosis	-1.006488187
10	7	Skewness	0.230330611
11	7.5	Range	2
12	6	Minimum	6
13	6	Maximum	8
14	6	Sum	248.5
15	7.5	Count	36
16	8	Confidence Level(95.0%)	0.231678109
17	7		
18	6.5		
19	7		
20	8		
21	6		
22	6		
23	6.5		
24	7		
25	8		

Descriptive Statistics

Input

Input Range: A1:A36

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

● Output Range: B1

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

☑ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

OK Cancel Help

As you see, the spreadsheet shows that the mean of the sample is $\bar{x}$ = 6.902777778 and the absolute value of the margin of error $|\pm Z\bullet(S/\sqrt{n})|$ = 0.231678109. This mean is based on this sample information. A 95% confidence interval for the hourly income of the UB work-study students has an upper limit of 6.902777778 + 0.231678109 and a lower limit of 6.902777778 - 0.231678109.

On the other hand, we can say that of all the intervals formed this way 95% contains the mean of the population. Or, for practical purposes, we can be 95% confident that the mean of the population is between 6.902777778 - 0.231678109 and 6.902777778 + 0.231678109. We can be at least 95% confident that interval [$6.68 and $7.13] contains the average hourly income of a work-study student.

Small Sample Size (say less than 30) If the sample n is less than 30 or we must use the small sample procedure to develop a confidence interval for the mean of a population. The general formula for developing confidence intervals for the population mean based on small a sample is:

$$\bar{x} \pm t_{\alpha/2} \bullet (S/\sqrt{n})$$

In this formula $\bar{x}$ is the mean of the sample. $t_{\alpha/2}$ is the interval coefficient providing an area of $\alpha/2$ in the upper tail of a t distribution with n-1 degrees of freedom which can be found from a t distribution table (for example the interval coefficient for a 90% confidence level is 1.833 if the sample is 10). S is the standard deviation of the sample and n is the sample size.

Now you would like to see how Excel is used to develop a certain confidence interval of a population mean based on this small sample information.

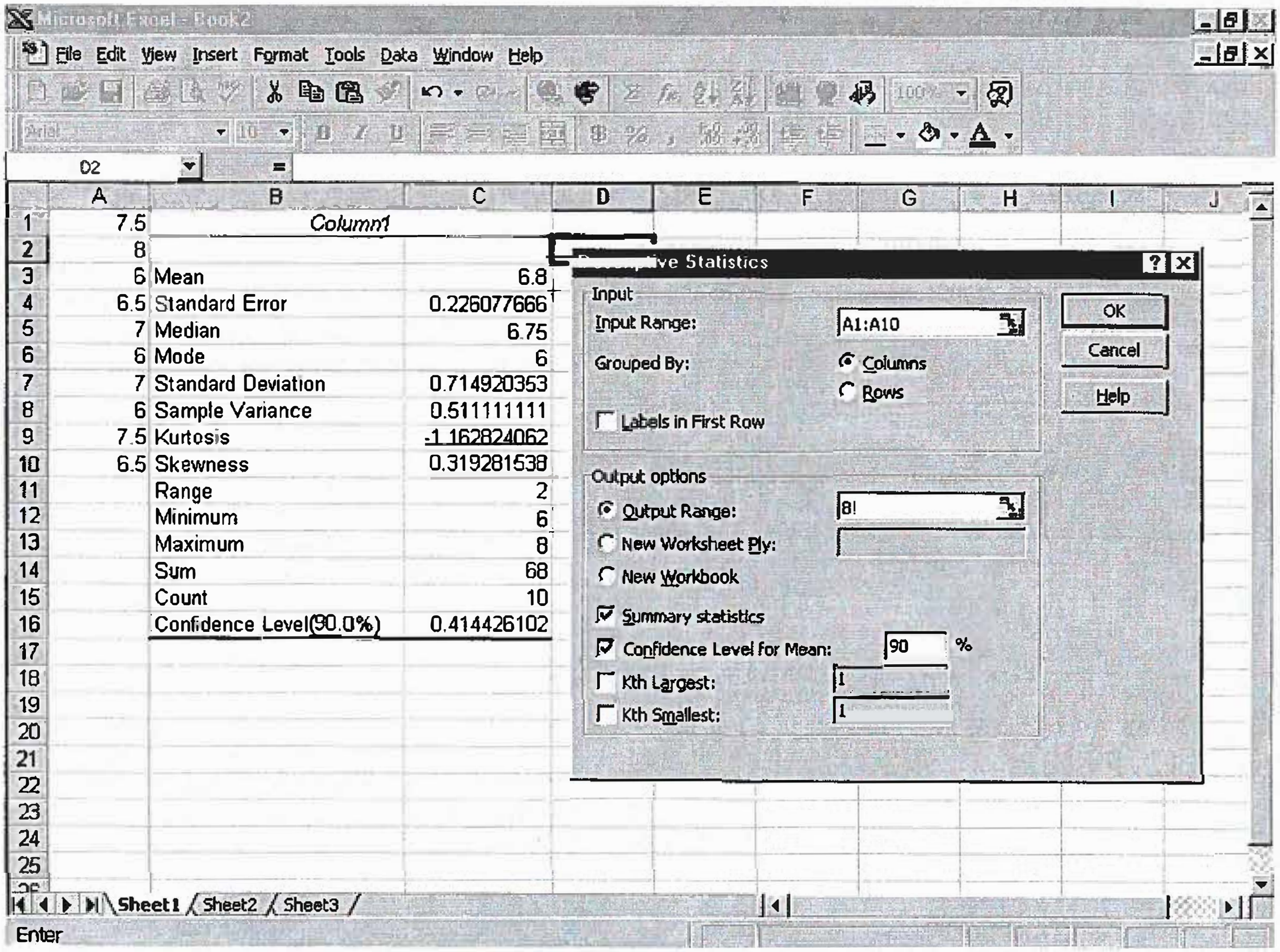
As you see, to evaluate this formula you need $\bar{x}$ "the mean of the sample" and the margin of error "$t_{\alpha/2} \bullet (S/\sqrt{n})$" Excel will automatically calculate these quantities the way it did for large samples.

Again, the only things you have to do are: add the margin of error "$t_{\alpha/2} \bullet (S/\sqrt{n})$" to the mean of the sample, $\bar{x}$, find the upper limit of the interval and to subtract the margin of error from the mean to find the lower limit of the interval.

To demonstrate how Excel finds these quantities we will use the data set, which contains the hourly incomes of 10 work-study students here, at the University of Baltimore. These numbers appear in cells A1 to A10 on an Excel work sheet.

After entering the data we follow the descriptive statistic procedure to calculate the unknown quantities (exactly the way we found quantities for large sample). Here you are with the procedures in step-by-step form:

Step 1. Enter data in cells A1 to A10 on the spreadsheet
Step 2. From the menus select **Tools**
Step 3. Click on **Data Analysis** then choose the **Descriptive Statistics** option. Click **OK** on the descriptive statistics dialog, click on Summary Statistic, click on the confidence interval level and type in 90% or in other problems whichever confidence interval you desire. In the Output Range box, enter B1 or whatever location you desire. Now click on **OK**. The screen shot will look like the following:



	A	B	C
1	7.5	*Column1*	
2	8		
3	6	Mean	6.8
4	6.5	Standard Error	0.226077666
5	7	Median	6.75
6	6	Mode	6
7	7	Standard Deviation	0.714920353
8	6	Sample Variance	0.511111111
9	7.5	Kurtosis	-1.162824062
10	6.5	Skewness	0.319281538
11		Range	2
12		Minimum	6
13		Maximum	8
14		Sum	68
15		Count	10
16		Confidence Level(90.0%)	0.414426102

ive Statistics

Input

Input Range: A1:A10

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

● Output Range: B1

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

☑ Confidence Level for Mean: 90 %

□ Kth Largest: 1

□ Kth Smallest: 1

Now, like the calculation of the confidence interval for the large sample, calculate the confidence interval of the population based on this small sample information. The confidence interval is:

$$6.8 \pm 0.414426102$$

or

$6.39<===>$7.21.

We can be at least 90% confidant that the interval [$6.39 , $7.21] contains the true mean of the population.

Test of Hypothesis Concerning the Population Mean

We have templates that we may use for hypothesis testing. However, if you lose the templates you can still use Excel as follows.

Again, we must distinguish two cases with respect to the size of your sample

Large Sample Size (say, over 30): In this section you wish to know how Excel can be used to conduct a hypothesis test about a population mean. We will use the hourly incomes of different work-study students than those introduced earlier in the confidence interval section. Data are entered in cells A1 to A36. The objective is to test the following **Null** and **Alternative** hypothesis:

$$H_0: \mu = 7$$
$$H_a: \mu \neq 7$$

The null hypothesis indicates that the average hourly income of a work-study student is equal to $7 per hour; however, the alternative hypothesis indicates that the average hourly income is not equal to $7 per hour.

I will repeat the steps taken in descriptive statistics and at the very end will show how to find the value of the test statistics in this case, z, using a cell formula.

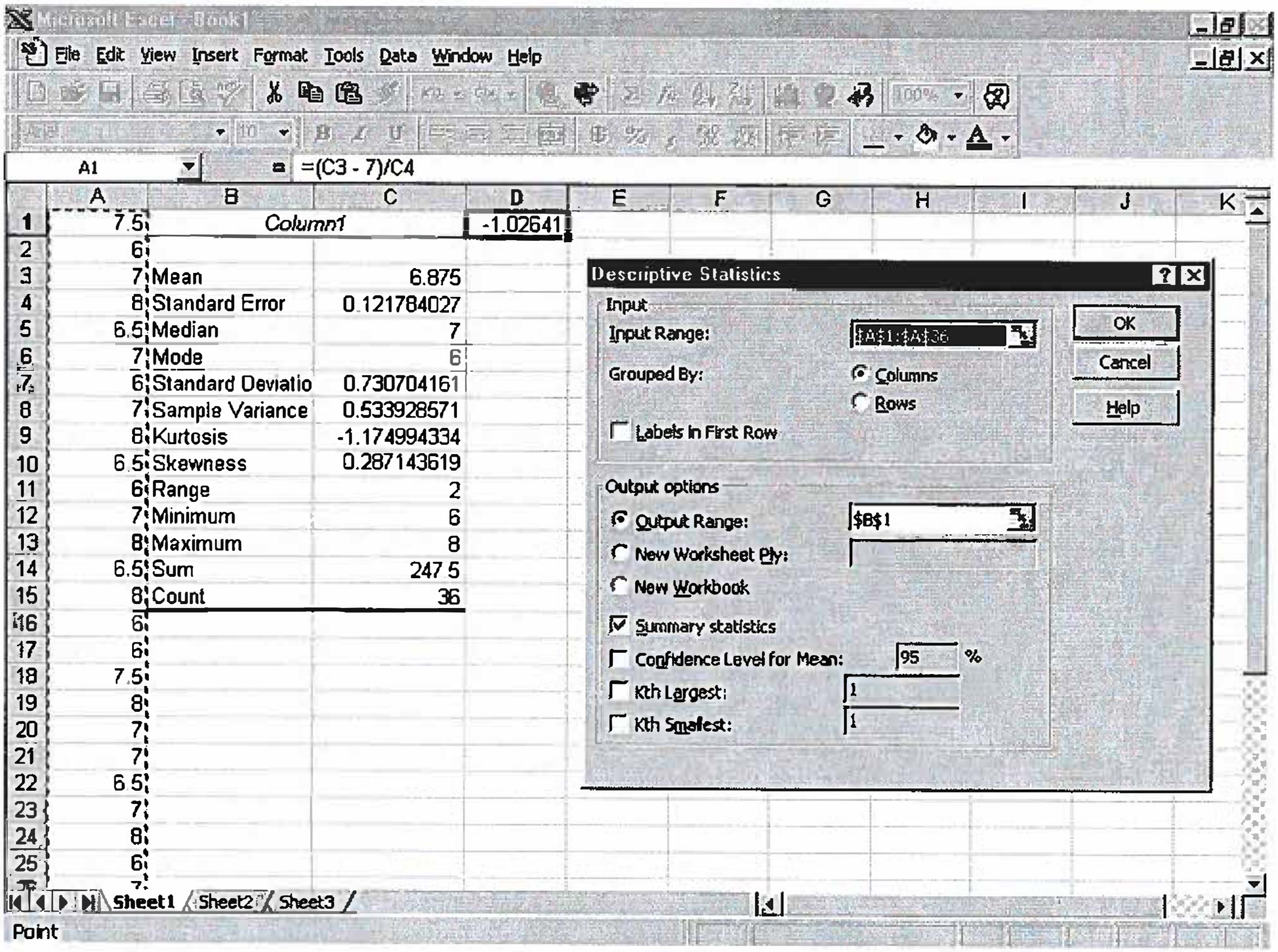
Step 1. Enter data in cells A1 to A36 (on the spreadsheet)

Step 2. From the menus select **Tools**

Step 3. Click on **Data Analysis** then choose the **Descriptive Statistics** option, click **OK**. On the descriptive statistics dialog, click on Summary Statistic. Select the **Output Range** box, enter B1 or whichever location you desire. Now click **OK**.

(To calculate the value of the test statistics search for the mean of the sample then the standard error. In this output, these values are in cells C3 and C4.)

Step 4. Select cell D1 and enter the cell formula = (C3 - 7)/C4. The screen shot should look like the following:



A1 = =(C3 - 7)/C4

	A	B	C	D
1	7.5	Column1		-1.02641
2	6			
3	7	Mean	6.875	
4	8	Standard Error	0.121784027	
5	6.5	Median	7	
6	7	Mode	6	
7	6	Standard Deviatio	0.730704161	
8	7	Sample Variance	0.533928571	
9	8	Kurtosis	-1.174994334	
10	6.5	Skewness	0.287143619	
11	6	Range	2	
12	7	Minimum	6	
13	8	Maximum	8	
14	6.5	Sum	247.5	
15	8	Count	36	
16	6			
17	6			
18	7.5			
19	8			
20	7			
21	7			
22	6.5			
23	7			
24	8			
25	6			
26	7			

Descriptive Statistics

Input
Input Range: $A$1:$A$36
Grouped By: ● Columns ○ Rows
□ Labels in First Row

Output options
● Output Range: $B$1
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
□ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

The value in cell D1 is the value of the test statistics. Since this value falls in acceptance range of -1.96 to 1.96 (from the normal distribution table), we fail to reject the null hypothesis.

Small Sample Size (say, less than 30):

Using steps taken the large sample size case, Excel can be used to conduct a hypothesis for small-sample case. Let's use the hourly income of 10 work-study students at UB to conduct the following hypothesis.

$$H_0: \mu = 7$$
$$H_a: \mu \neq 7$$

The null hypothesis indicates that average hourly income of a work-study student is equal to $7 per hour .The alternative hypothesis indicates that average hourly income is not equal to $7 per hour.

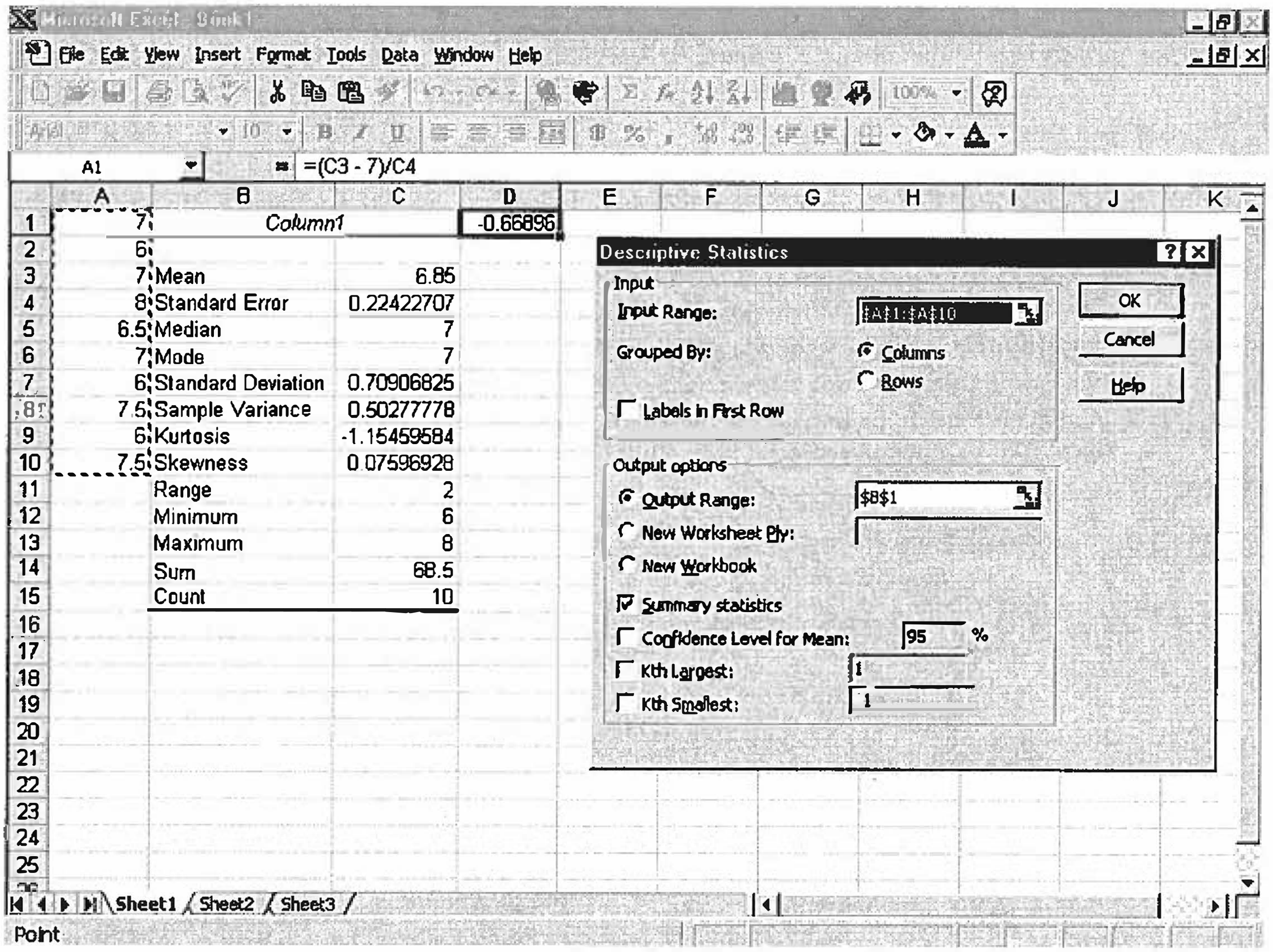
I will repeat the steps taken in descriptive statistics and at the very end will show how to find the value of the test statistics in this case "t" using a cell formula.

Step 1. Enter data in cells A1 to A10 (on the spreadsheet)

Step 2. From the menus select **Tools**

Step 3. Click on **Data Analysis** then choose the **Descriptive Statistics** option. Click **OK**. On the descriptive statistics dialog, click on Summary Statistic. Select the Output Range boxes, enter B1 or whatever location you chose. Again, click on **OK**.
(To calculate the value of the test statistics search for the mean of the sample then the standard
error, in this output these values are in cells C3 and C4.)

Step 4. Select cell D1 and enter the cell formula = (C3 - 7)/C4. The screen shot would look like the following:



A1 = =(C3 - 7)/C4

	A	B	C	D
1	7	*Column1*		-0.66896
2	6			
3	7	Mean	6.85	
4	8	Standard Error	0.22422707	
5	6.5	Median	7	
6	7	Mode	7	
7	6	Standard Deviation	0.70906825	
8	7.5	Sample Variance	0.50277778	
9	6	Kurtosis	-1.15459584	
10	7.5	Skewness	0.07596928	
11		Range	2	
12		Minimum	6	
13		Maximum	8	
14		Sum	68.5	
15		Count	10	

Descriptive Statistics

Input
Input Range: $A$1:$A$10
Grouped By: ● Columns ○ Rows
□ Labels in First Row

Output options
● Output Range: $B$1
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
□ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

Since the value of test statistic t = -0.66896 falls in acceptance range -2.262 to +2.262 (from t table, where $\alpha/2$ = 0.025 and the degrees of freedom is 9), we fail to reject the null hypothesis.

Difference Between Mean of Two Populations

In this section we will show how Excel is used to conduct a hypothesis test about the difference between two population means assuming that populations have equal variances. The data in this case are taken from various offices here at the University of Baltimore. I collected the hourly income data of 36 randomly selected work-study students and 36 student assistants. The hourly income range for work-study students was $6 - $8 while the hourly income range for student assistants was $6-$9. The main objective in this hypothesis testing is to see whether there is a significant difference between the means of the two populations. The **NULL** and the **ALTERNATIVE** hypothesis is that the means are equal and the means are not equal, respectively.

Referring to the spreadsheet, I chose **A1** and **A2** as label centers. The work-study students' hourly income for a sample size 36 are shown in cells **A2:A37**, and the student assistants' hourly income for a sample size 36 is shown in cells **B2:B37**

Data for Work Study Student: 6, 6, 6, 6, 6, 6, 6, 6.5, 6.5, 6.5, 6.5, 6.5, 6.5, 7, 7, 7, 7, 7, 7, 7, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 8, 8, 8, 8, 8, 8, 8, 8, 8.

Data for Student Assistant: 6, 6, 6, 6, 6, 6.5, 6.5, 6.5, 6.5, 6.5, 7, 7, 7, 7, 7, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 8, 8, 8, 8, 8, 8, 8, 8.5, 8.5, 8.5, 8.5, 8.5, 9, 9, 9, 9.

Use the *Descriptive Statistics* procedure to calculate the variances of the two samples. The Excel procedure for testing the difference between the two population means will require information on the variances of the two populations. Since the variances of the two populations are unknowns they should be replaced with sample variances. The descriptive for both samples show that the variance of first sample is s_1^2 = **0.55546218**, while the variance of the second sample s_2^2 =**0.969748**.

work-study student		**student assistant**	
Mean	7.05714286	Mean	7.471429
Standard Error	0.12597757	Standard Error	0.166454
Median	7	Median	7.5
Mode	8	Mode	8
Standard Deviation	0.74529335	Standard Deviation	0.984758

Sample Variance	0.55546218	Sample Variance	0.969748
Kurtosis	-1.38870558	Kurtosis	-1.192825
Skewness	-0.09374375	Skewness	-0.013819
Range	2	Range	3
Minimum	6	Minimum	6
Maximum	8	Maximum	9
Sum	247	Sum	261.5
Count	35	Count	35

To conduct the desired test hypothesis with Excel the following steps can be taken:

Step 1. From the menus select *Tools* then click on the *Data Analysis* option.

Step 2. When the *Data Analysis* dialog box appears:
Choose z-Test: Two Sample for means then click OK

Step 3. When the **z-Test: Two Sample for means** dialog box appears:

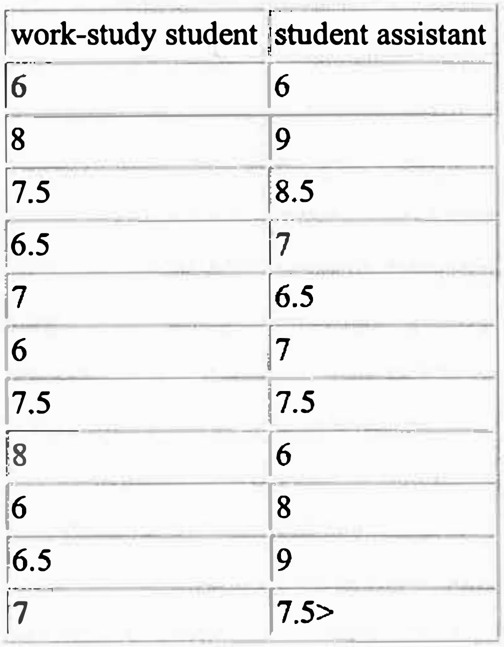
Enter **A1:A36** in the **variable 1 range box (work-study students' hourly income)**
Enter **B1:B36** in the **variable 2 range box (student assistants' hourly income)**
Enter 0 in the ***Hypothesis Mean Difference*** box (if you desire to test a mean difference other than 0, enter that value)
Enter the variance of the first sample in the **Variable 1 Variance box**
Enter the variance of the second sample in the **Variable 2 Variance box** and select Labels
Enter 0.05 or, whatever **level of significance** you desire, in the **Alpha box**
Select a suitable **Output Range** for the results, I chose **C19**, then click OK.

The value of test statistic **z=-1.9845824** appears in our case in cell D24. The rejection rule for this test is z < -1.96 or z > 1.96 from the normal distribution table. In the Excel output these values for a two-tail test are z<-1.959961082 and z>+1.959961082. Since the value of the test statistic z=-1.9845824 is less than -1.959961082 we reject the null hypothesis. We can also draw this conclusion by comparing the p-value for a two tail - test and the alpha value.

Since p-value **0.047190813** is less than a=0.05 we reject the null hypothesis. Overall we can say, based on the sample results, the two populations' means are different.

Small Samples: n_1 OR n_2 are less than 30

In this section we will show how Excel is used to conduct a hypothesis test about the difference between two population means. - Given that the populations have equal variances when two small independent samples are taken from both populations. Similar to the above case, the data in this case are taken from various offices here at the University of Baltimore. I collected hourly income data of 11 randomly selected work-study students and 11 randomly selected student assistants. The hourly income range for both groups was similar range, $6 - $8 and $6-$9. The main objective in this hypothesis testing is similar too, to see whether there is a significant difference between the means of the two populations. The **NULL** and the ALTERNATIVE hypothesis are that the means are equal and they are not equal, respectively.



work-study student	student assistant
6	6
8	9
7.5	8.5
6.5	7
7	6.5
6	7
7.5	7.5
8	6
6	8
6.5	9
7	7.5>

Referring to the spreadsheet, we chose **A1** and **A2** as label centers. The work-study students' hourly income for a sample size 11 are shown in cells **A2:A12**, and the student assistants' hourly income for a sample size 11 is shown in cells **B2:B12**. Unlike previous case, you do not have to calculate the variances of the two samples, Excel will automatically calculate these quantities and use them in the calculation of the value of the test statistic.

Similar to the previous case, but a bit different in step # 2, to conduct the desired test hypothesis with Excel the following steps can be taken:

Step 1. From the menus select *Tools* then click on the *Data Analysis* option.

Step 2. When the *Data Analysis* dialog box appears:
Choose **t-Test: Two Sample Assuming Equal Variances** then click OK

Step 3 When the **t-Test: Two Sample Assuming Equal Variances dialog box appears**:

Enter A1:A12 in the **variable 1 range box** (work-study student hourly income)
Enter B1:B12 in the **variable 2 range box** (student assistant hourly income)
Enter 0 in the **Hypothesis Mean Difference** box(if you desire to test a mean difference other than zero, enter that value) then select Labels

Enter 0.05 or, whatever *level of significance* you desire, in the *Alpha* box

Select a suitable *Output Range* for the results, I chose C1, then click OK.

The value of the **test statistic t=-1.362229828** appears, in our case, in cell D10. The rejection rule for this test is t<-2.086 or t>+2.086 from the **t distribution table** where the t value is based on a t distribution with n_1-n_2-2 degrees of freedom and where the area of the upper one tail is 0.025 (that is equal to alpha/2).

In the Excel output the values for a two-tail test are t<-2.085962478 and t>+2.085962478. Since the value of the test statistic t=-1.362229828, is in an acceptance range of t<-2.085962478 and t>+2.085962478, we fail to reject the null hypothesis.

We can also draw this conclusion by comparing the p-value for a two-tail test and the alpha value.

Since the **p-value 0.188271278 is greater than a=0.05 again,** we fail to reject the null hypothesis.

Overall we can say, based on sample results, the two populations' means are equal.

	work-study student	student assistant
Mean	6.909090909	7.454545455
Variance	0.590909091	1.172727273
Observations	11	11
Pooled Variance	0.881818182	
Hypothesized Mean Difference	0	
Df	20	
t Stat	-1.362229828	
P(T<=t) one tail	0.094135639	
t Critical one tail	1.724718004	
P(T<=t)two tail	0.188271278	
t Critical two tail	2.085962478	

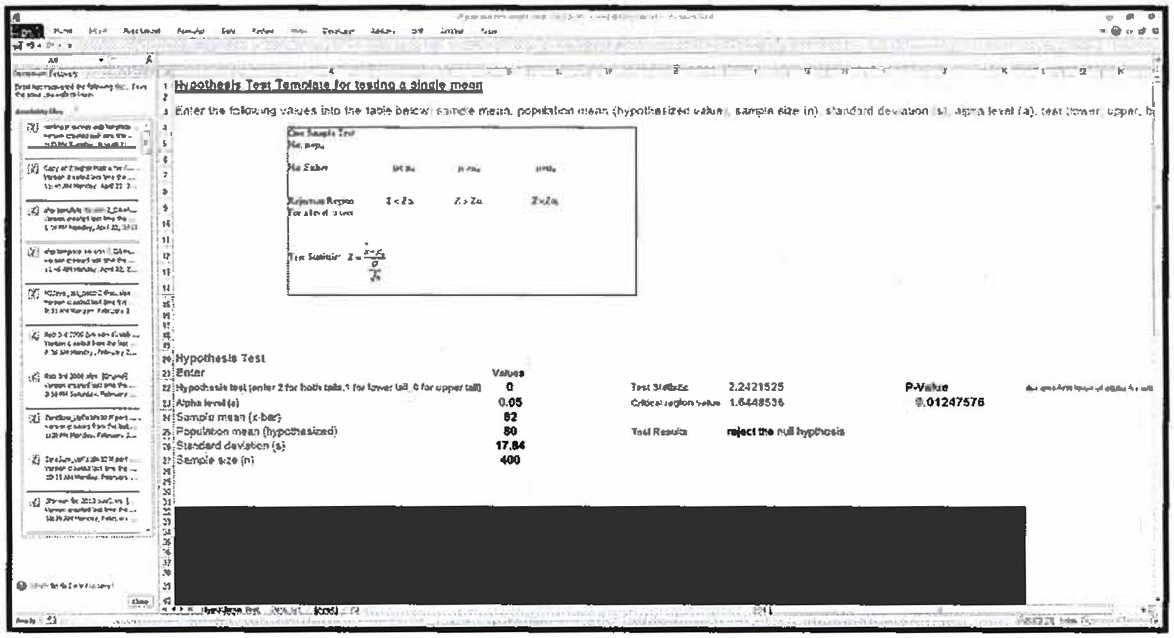
Hypothesis Testing and Templates

The templates requires you to have the following information;

Sample mean or sample proportion for each sample.
Sample standard deviation for each sample.
Size of each sample

These can be attended from the output of descriptive statistics in the Analysis Toolpak.

The user enters the information into the template and informs the template which test is desired: left tail, right tail, or two tails. Enter the values in the green area.



Hypothesis Test Template for testing a single mean

Enter the following values into the table below: sample mean, population mean (hypothesized value), sample size (n), standard deviation (s), alpha level (a), test (lower, upper, b

One Sample Test
Ho: μ=μ0

Ha: Either	μ<μ0	μ>μ0	μ≠μ0
Rejection Region For a level α test	Z < Zα	Z > Zα	Z>Zα/2

Test Statistic $Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$

Hypothesis Test

Enter	Values				
Hypothesis test (enter 2 for both tails,1 for lower tail, 0 for upper tail)	0	Test Statistic	2.2421525	P-Value	
Alpha level (a)	0.05	Critical region value	1.6448536	0.01247576	
Sample mean (x-bar)	82				
Population mean (hypothesized)	80	Test Results	reject the null hypothesis		
Standard deviation (s)	17.84				
Sample size (n)	400				

[redacted]

This type interface is available for each template:

Hypothesis test for a single mean
Hypothesis test to compare means
Hypothesis test for a single proportion
Hypothesis test to compare proportions

Linear Correlation and Regression Analysis

In this section the objective is to see whether there is a correlation between two variables and to find a model that predicts one variable in terms of the other variable. There are so many examples that we could mention but we will mention the popular ones in the world of the military, industry, or government. Usually independent variable is presented by the letter *x* and the dependent variable is presented by the letter *y*. A business man would like to see whether there is a relationship between the number of cases of sold and the temperature in a hot summer day based on information taken from the past. He also would like to estimate the number cases of soda which will be sold in a particular hot summer day in a ball game. He clearly recorded temperatures and number of cases of soda sold on those particular days. The following table shows the recorded data from June 1 through June 13. The weatherman predicts a 94F degree temperature for June 14. The businessman would like to meet all demands for the cases of sodas ordered by customers on June 14.

DAY	Cases of Soda	Temperature
1-Jun	57	56
2-Jun	59	58
3-Jun	65	63
4-Jun	67	66
5-Jun	75	73
6-Jun	81	78
7-Jun	86	85
8-Jun	88	85
9-Jun	88	87
10-Jun	84	84
11-Jun	82	88
12-Jun	80	84
13-Jun	83	89

Now let's use Excel to find the ***linear correlation coefficient*** and then the regression line equation. The linear correlation coefficient is a quantity between -1 and +1. This quantity is denoted by **R**. The closer **R to +1** the stronger positive (direct) correlation and similarly the closer **R** to -1 the stronger negative (inverse) correlation exists between the two variables. The general form of the regression line is $y = mx + b$. In this formula, m is the slope of the line and b is the y-intercept. You can find these quantities from the Excel output. In this situation the variable y (the dependent variable) is the number of cases of soda and the x (independent variable) is the temperature. To find the Excel output the following steps can be taken:

Step 1. From the menus choose Tools and click on Data Analysis.

Step 2. When Data Analysis dialog box appears, click on correlation.

Step 3. When correlation dialog box appears, enter B1:C14 in the input range box. Click on Labels in first row and enter a16 in the output range box. Click on OK.

	Cases of Soda	Temperature
Cases of Soda	1	
Temperature	0.96659877	1

As you see the correlation between the number of cases of soda demanded and the temperature is a very strong positive correlation where this indicates a strong positive linear relationship. This means as the temperature increases the demand for cases of soda is also increasing. The linear correlation coefficient is 0.966598577 which is very close to +1.

Note: you may obtain the correlation coefficient between two sets of data (*x,y*) by typing

=Correl(series1, series 2)

Regression

Now let's follow same steps but a bit different to find the regression equation.

Step 1. From the menus choose *Tools* and click on *Data Analysis*

Step 2. When *Data Analysis* dialog box appears, click on *regression.*

Step 3. When *Regression* dialog box appears, enter b1:b14 in the y-range box and c1:c14 in the x-range box. Click on *labels.*

Step 4. Enter a19 in the *output range box.*

Note: The regression equation in general should look like $Y=m\,X+b$. In this equation m is the slope of the regression line and b is its y-intercept.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.966598577
R Square	0.934312809
Adjusted R Square	0.928341246
Standard Error	2.919383191
Observations	13

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1333.479989	1333.479989	156.4603497	7.58511E-08
Residual	11	93.75078034	8522798213		
Total	12	1427.230769			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	9.17800767	5.445742836	1.685354587	0.120044801	-2.80799756	21.16401
Temperature	0.879202711	0.07028892	12.50841116	7.58511E-08	0.724497763	1.033908

The relationship between the number of cans of soda and the temperature is:

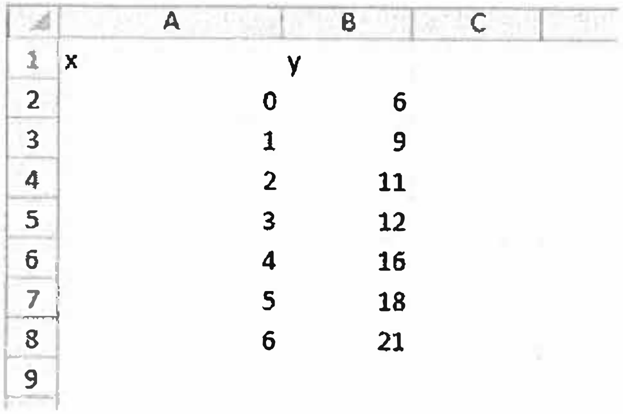
Y = 0.879202711 X + 9.17800767

This is our model.

Predicting: we use our model to predict using substitution. However, a single prediction has less meaning than an interval answer as we will show.

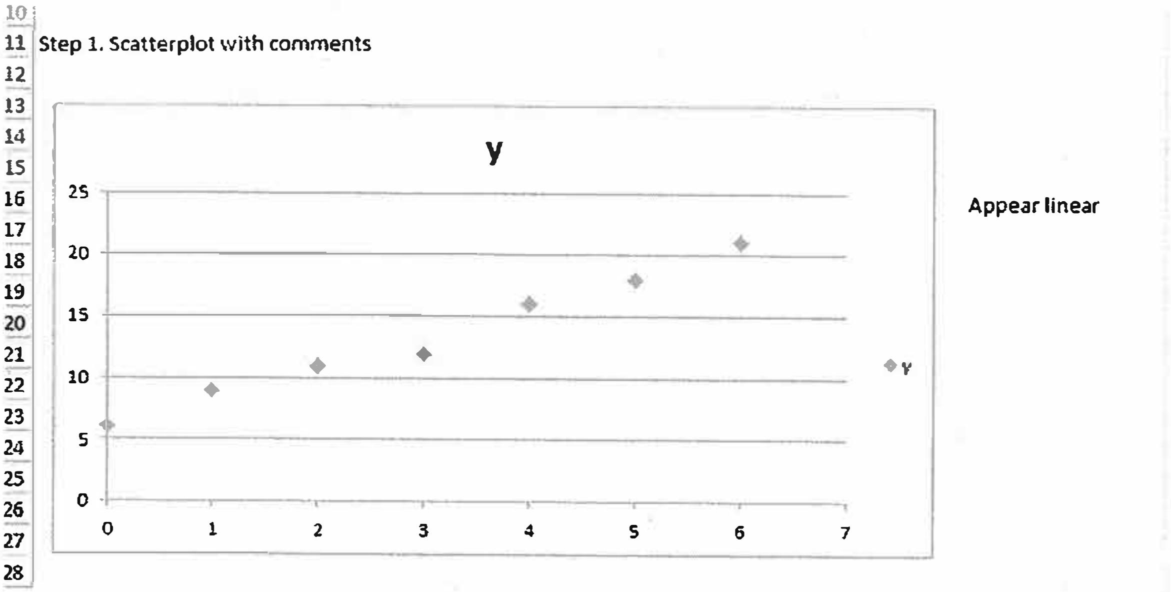
Another example of linear regression in Excel.

The data



	A	B	C
1	x	y	
2	0	6	
3	1	9	
4	2	11	
5	3	12	
6	4	16	
7	5	18	
8	6	21	
9			

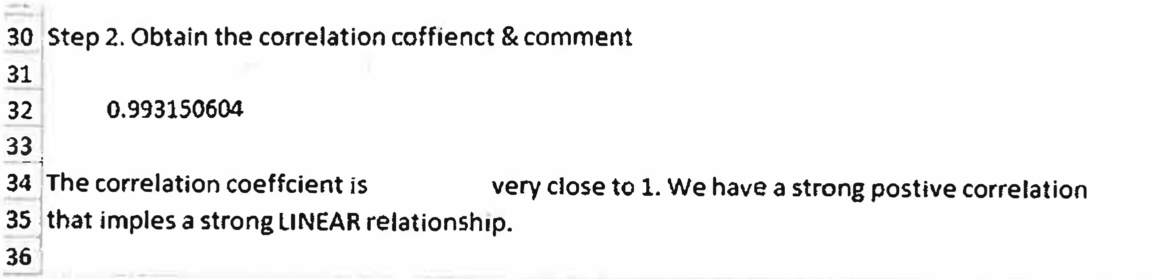
The scatterplot with comments



Step 1. Scatterplot with comments

Appear linear

Obtain the correlation coefficient & comment

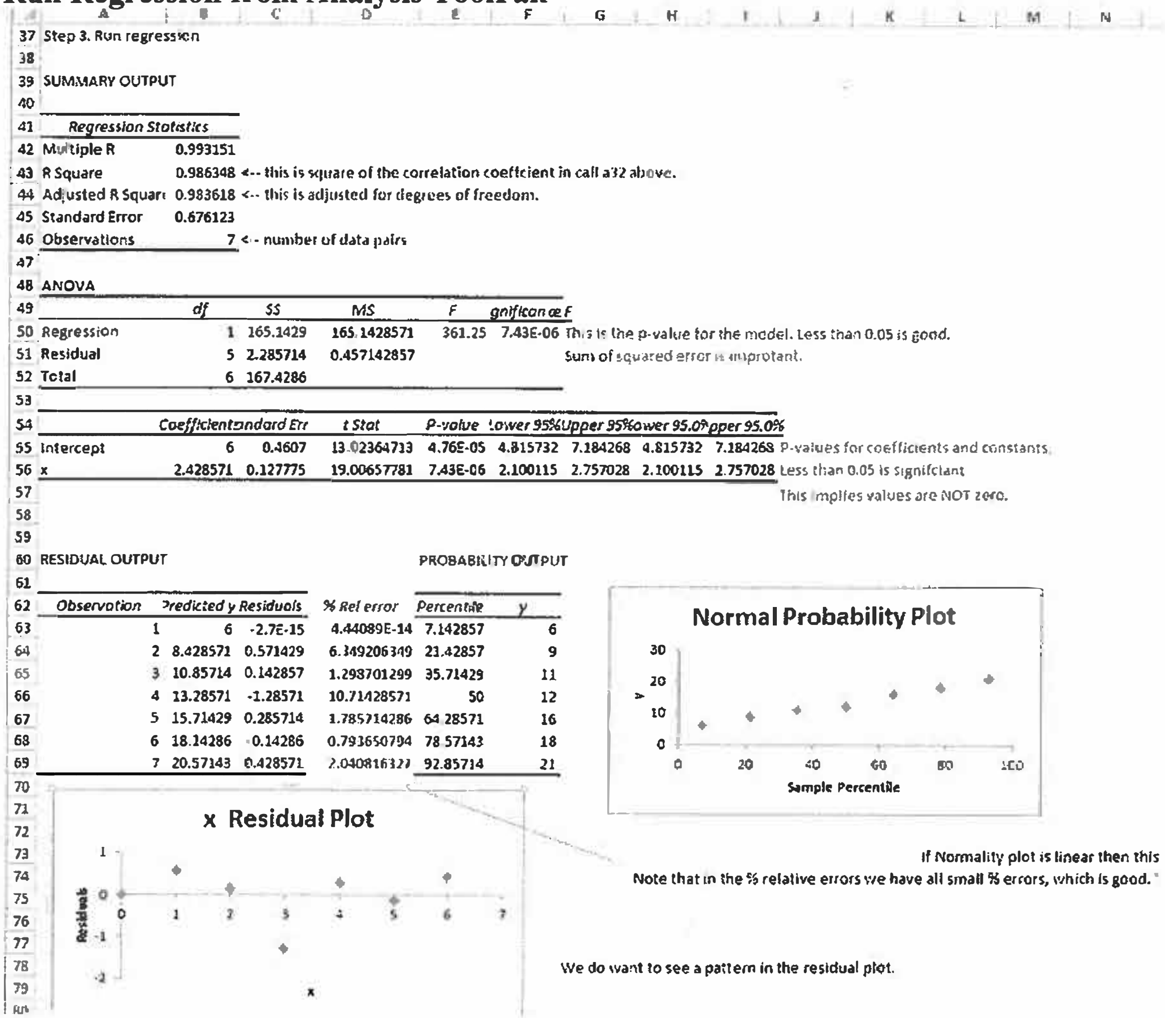


Step 2. Obtain the correlation coffienct & comment

0.993150604

The correlation coeffcient is very close to 1. We have a strong postive correlation that imples a strong LINEAR relationship.

Run Regression from Analysis ToolPak



37 Step 3. Run regression

39 SUMMARY OUTPUT

		Regression Statistics	
42	Multiple R	0.993151	
43	R Square	0.986348	<-- this is square of the correlation coefficient in call a32 above.
44	Adjusted R Square	0.983618	<-- this is adjusted for degrees of freedom.
45	Standard Error	0.676123	
46	Observations	7	<- - number of data pairs

48 ANOVA

49		df	SS	MS	F	gnificance F	
50	Regression	1	165.1429	165.1428571	361.25	7.43E-06	This is the p-value for the model. Less than 0.05 is good.
51	Residual	5	2.285714	0.457142857			Sum of squared error is improtant.
52	Total	6	167.4286				

54		Coefficient	tandard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%	
55	Intercept	6	0.4607	13.02364733	4.76E-05	4.815732	7.184268	4.815732	7.184268	P-values for coefficients and constants.
56	x	2.428571	0.127775	19.00657781	7.43E-06	2.100115	2.757028	2.100115	2.757028	Less than 0.05 is significant
57										This implies values are NOT zero.

60 RESIDUAL OUTPUT PROBABILITY OUTPUT

62	Observation	Predicted y	Residuals	% Rel error	Percentile	y
63	1	6	-2.7E-15	4.44089E-14	7.142857	6
64	2	8.428571	0.571429	6.349206349	21.42857	9
65	3	10.85714	0.142857	1.298701299	35.71429	11
66	4	13.28571	-1.28571	10.71428571	50	12
67	5	15.71429	0.285714	1.785714286	64.28571	16
68	6	18.14286	-0.14286	0.793650794	78.57143	18
69	7	20.57143	0.428571	2.040816327	92.85714	21

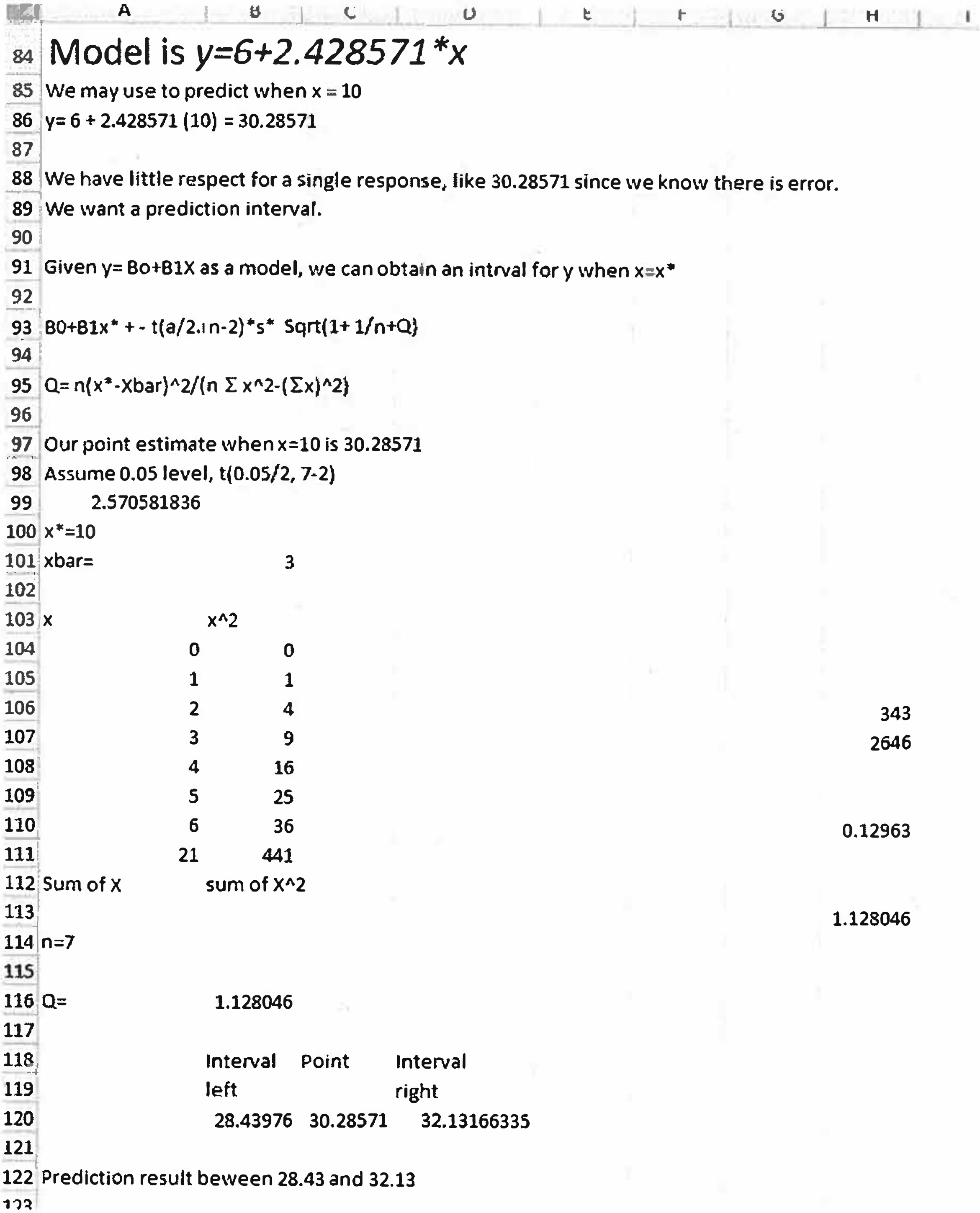
If Normality plot is linear then this

Note that in the % relative errors we have all small % errors, which is good.

We do want to see a pattern in the residual plot.

Note comments in Red and additions in Blue.

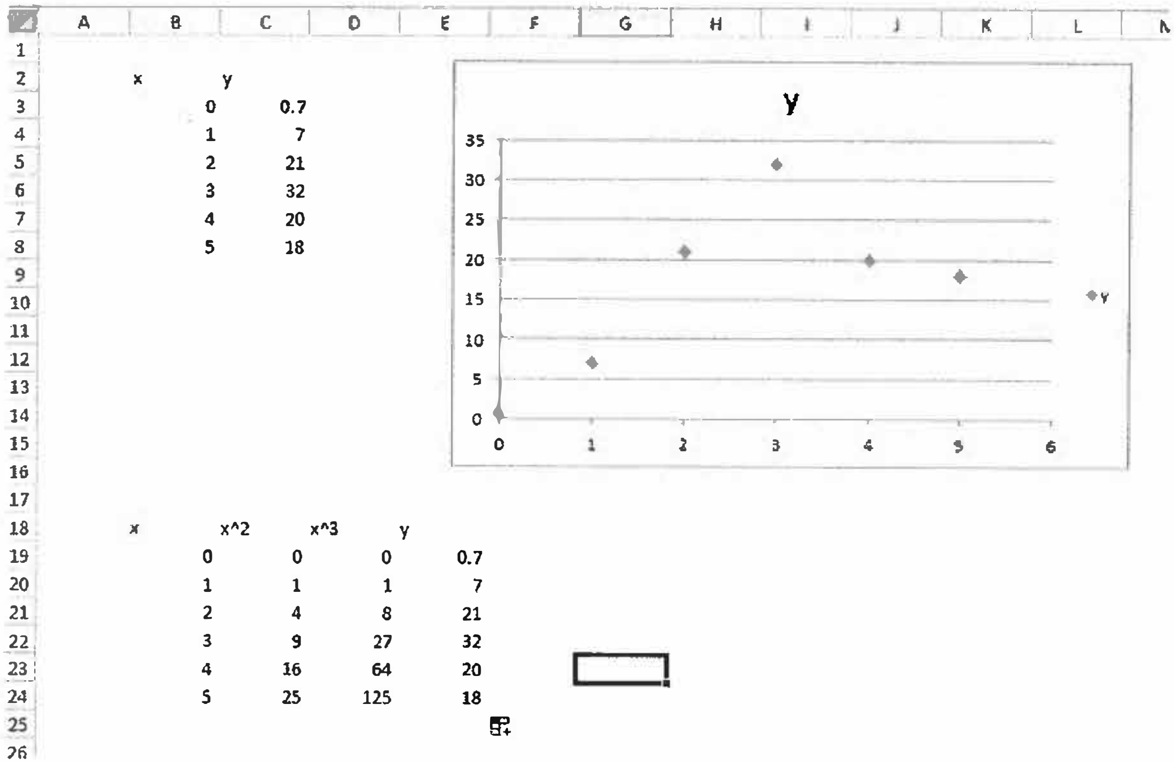
Predictions and Prediction Intervals



	A	B	C	D	E	F	G	H
84	Model is *y=6+2.428571*x*							
85	We may use to predict when x = 10							
86	y= 6 + 2.428571 (10) = 30.28571							
87								
88	We have little respect for a single response, like 30.28571 since we know there is error.							
89	We want a prediction interval.							
90								
91	Given y= Bo+B1X as a model, we can obtain an intrval for y when x=x*							
92								
93	B0+B1x* + - t(a/2.,n-2)*s* Sqrt(1+ 1/n+Q)							
94								
95	Q= n(x*-Xbar)^2/(n Σ x^2-(Σx)^2)							
96								
97	Our point estimate when x=10 is 30.28571							
98	Assume 0.05 level, t(0.05/2, 7-2)							
99	2.570581836							
100	x*=10							
101	xbar=	3						
102								
103	x	x^2						
104	0	0						
105	1	1						
106	2	4						343
107	3	9						2646
108	4	16						
109	5	25						
110	6	36						0.12963
111	21	441						
112	Sum of X	sum of X^2						
113								1.128046
114	n=7							
115								
116	Q=	1.128046						
117								
118		Interval	Point	Interval				
119		left		right				
120		28.43976	30.28571	32.13166335				
121								
122	Prediction result beween 28.43 and 32.13							
123								

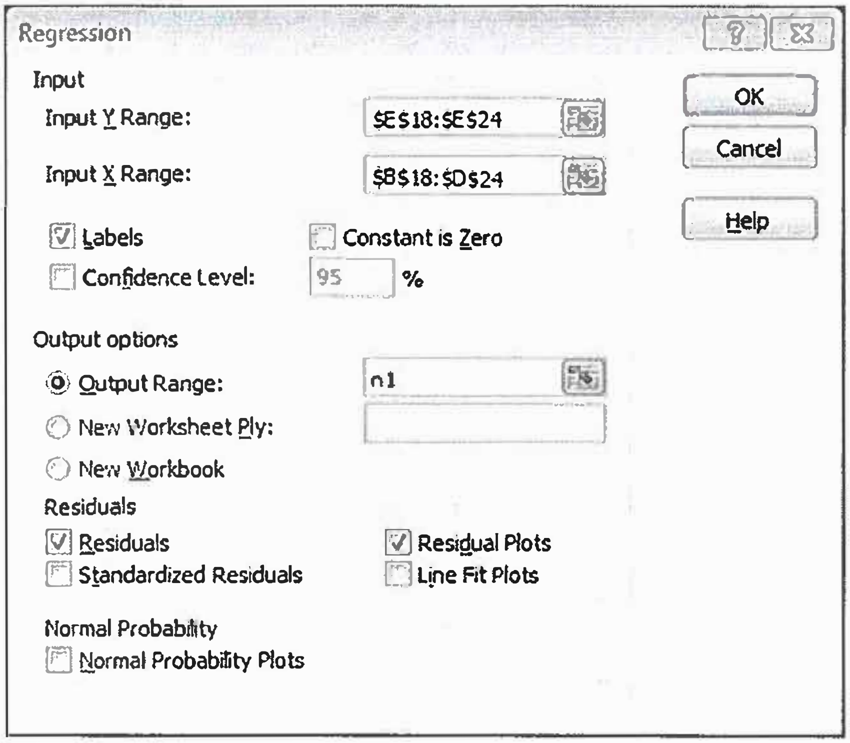
Advanced Regression Techniques

Multivariable—This can be done using the Analysis->Regression command in Excel. First you need a column of each independent variable in the function you want to build by regression.
If you want $y= a+bx+cx^2+dx^3$, then we need three columns in order of $\{x,x^2,x^3\}$. All of these will be entered as the X-variable in the Regression dialog box.



x	y
0	0.7
1	7
2	21
3	32
4	20
5	18

x	x^2	x^3	y
0	0	0	0.7
1	1	1	7
2	4	8	21
3	9	27	32
4	16	64	20
5	25	125	18



Regression

Input

Input Y Range: $E$18:$E$24

Input X Range: $B$18:$D$24

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: n1

○ New Worksheet Ply:

○ New Workbook

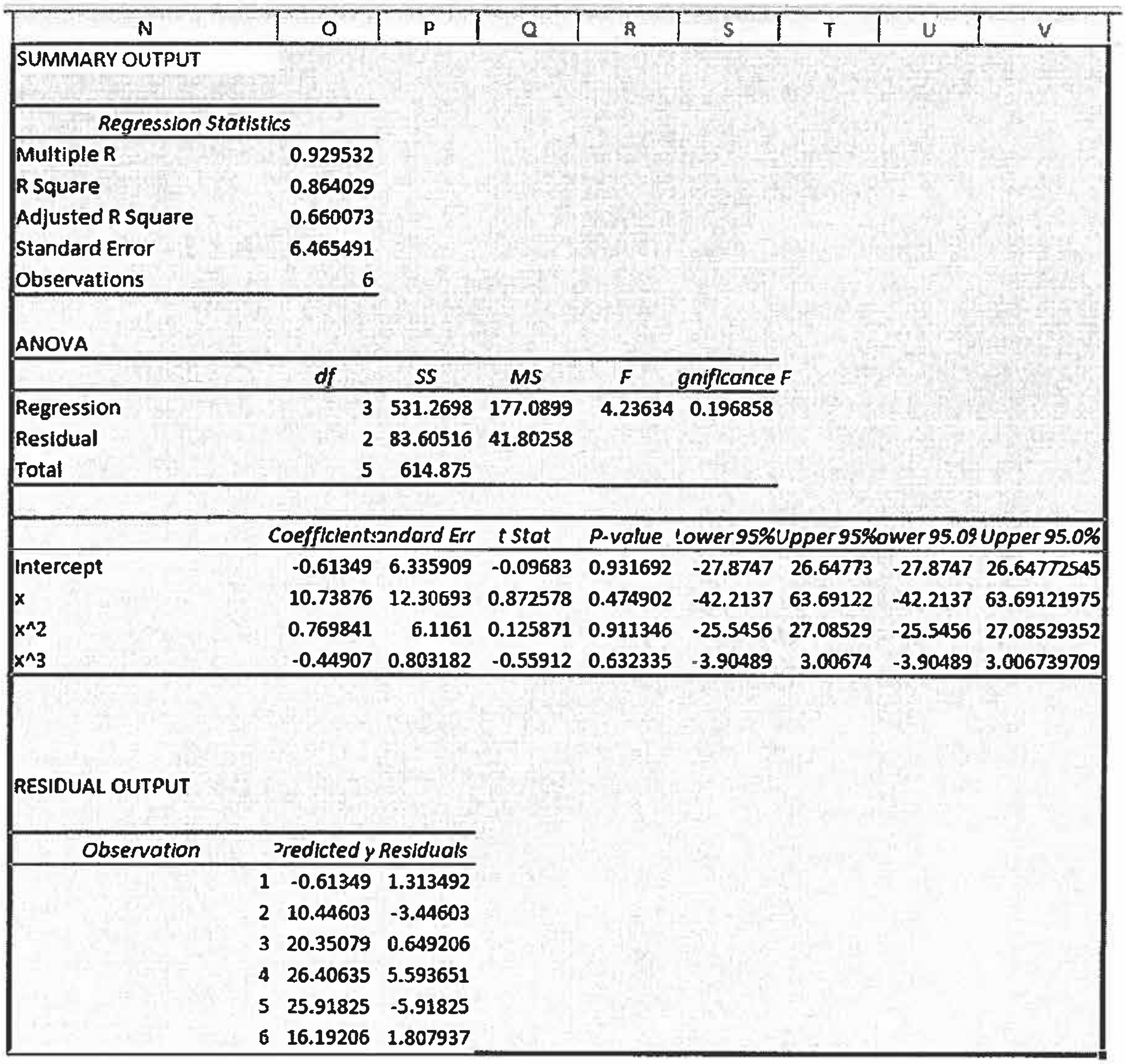
Residuals

☑ Residuals ☑ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots



N	O	P	Q	R	S	T	U	V

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.929532
R Square	0.864029
Adjusted R Square	0.660073
Standard Error	6.465491
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	531.2698	177.0899	4.23634	0.196858
Residual	2	83.60516	41.80258		
Total	5	614.875			

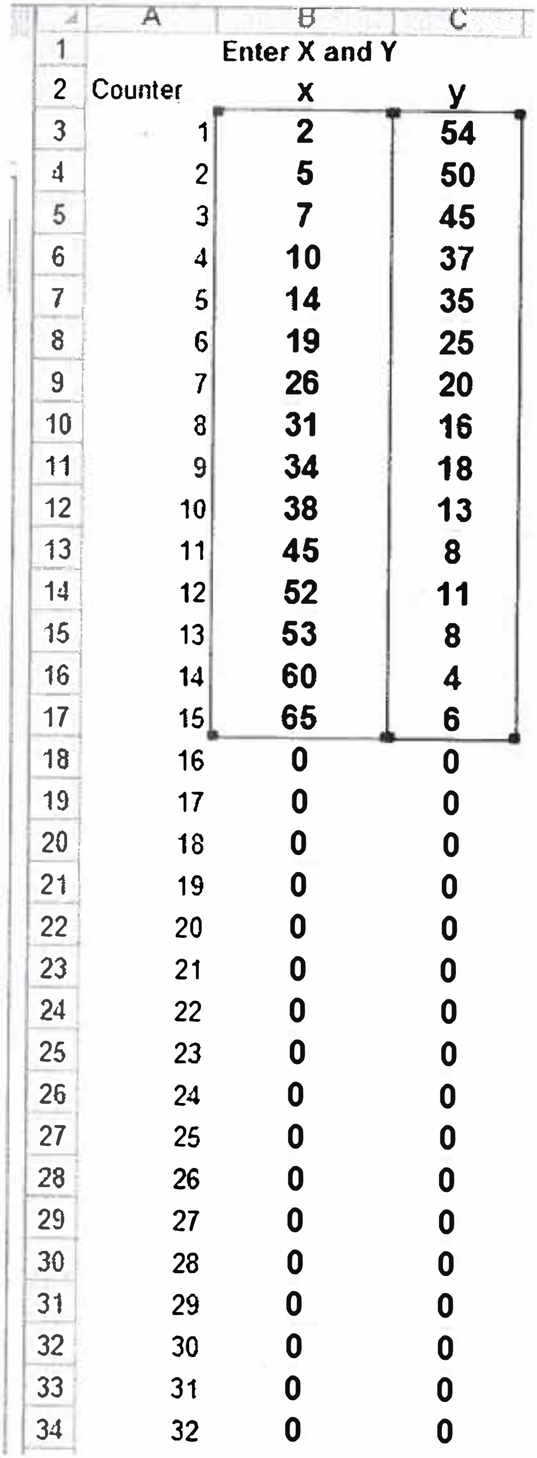
	Coefficient:	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	-0.61349	6.335909	-0.09683	0.931692	-27.8747	26.64773	-27.8747	26.64772545
x	10.73876	12.30693	0.872578	0.474902	-42.2137	63.69122	-42.2137	63.69121975
x^2	0.769841	6.1161	0.125871	0.911346	-25.5456	27.08529	-25.5456	27.08529352
x^3	-0.44907	0.803182	-0.55912	0.632335	-3.90489	3.00674	-3.90489	3.006739709

RESIDUAL OUTPUT

Observation	*Predicted y*	*Residuals*
1	-0.61349	1.313492
2	10.44603	-3.44603
3	20.35079	0.649206
4	26.40635	5.593651
5	25.91825	-5.91825
6	16.19206	1.807937

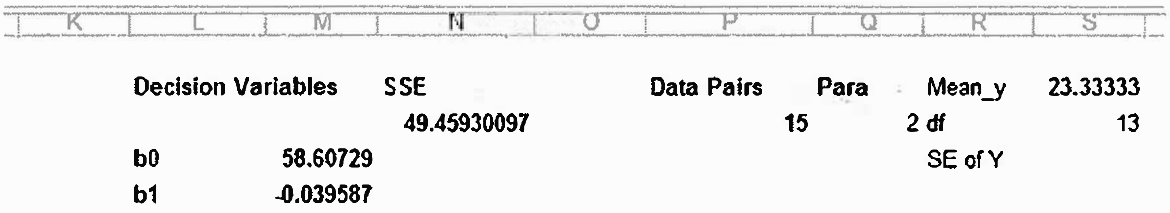
Nonlinear-Use template that should be self-explanatory

1. **Enter the data**



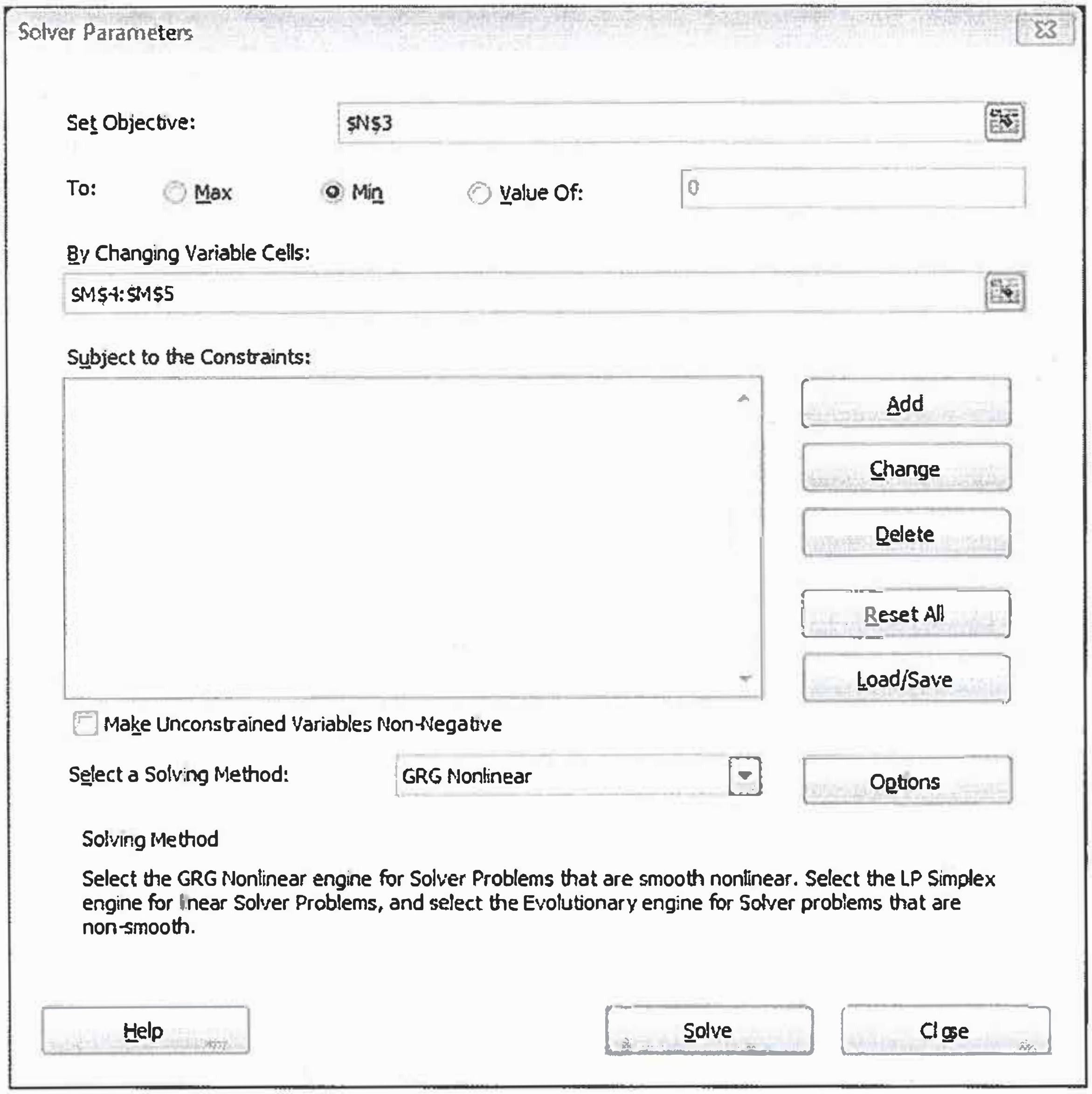
	A	B	C
1		Enter X and Y	
2	Counter	x	y
3	1	2	54
4	2	5	50
5	3	7	45
6	4	10	37
7	5	14	35
8	6	19	25
9	7	26	20
10	8	31	16
11	9	34	18
12	10	38	13
13	11	45	8
14	12	52	11
15	13	53	8
16	14	60	4
17	15	65	6
18	16	0	0
19	17	0	0
20	18	0	0
21	19	0	0
22	20	0	0
23	21	0	0
24	22	0	0
25	23	0	0
26	24	0	0
27	25	0	0
28	26	0	0
29	27	0	0
30	28	0	0
31	29	0	0
32	30	0	0
33	31	0	0
34	32	0	0

2. **Fill in yellow highlight cells with initial Solver guess , # data pairs, # of unknown parameters 9B0 and B1 here). Bo and b1 values were set an zero initially**



K	L	M	N	O	P	Q	R	S
	Decision Variables		SSE		Data Pairs	Para	Mean_y	23.33333
			49.45930097		15	2	df	13
	b0	58.60729					SE of Y	
	b1	-0.039587						

3. **Use Solver to Minimize cell N3 by changing cells M4:M5**



Solver Parameters

Set Objective: $N$3

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:

$M$4:$M$5

Subject to the Constraints:

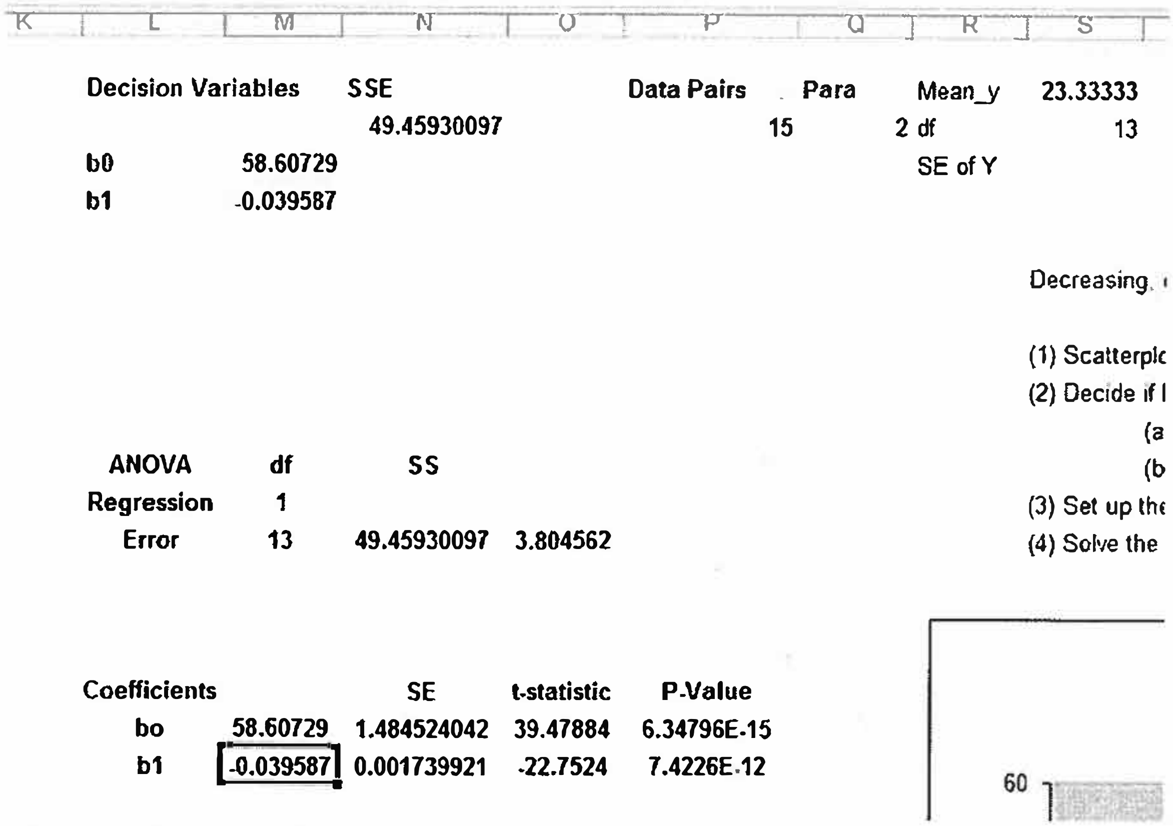
□ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

4. View output values and statistics.



Decision Variables		SSE		Data Pairs	Para	Mean_y	23.33333
		49.45930097		15	2	df	13
b0	58.60729					SE of Y	
b1	-0.039587						

Decreasing,

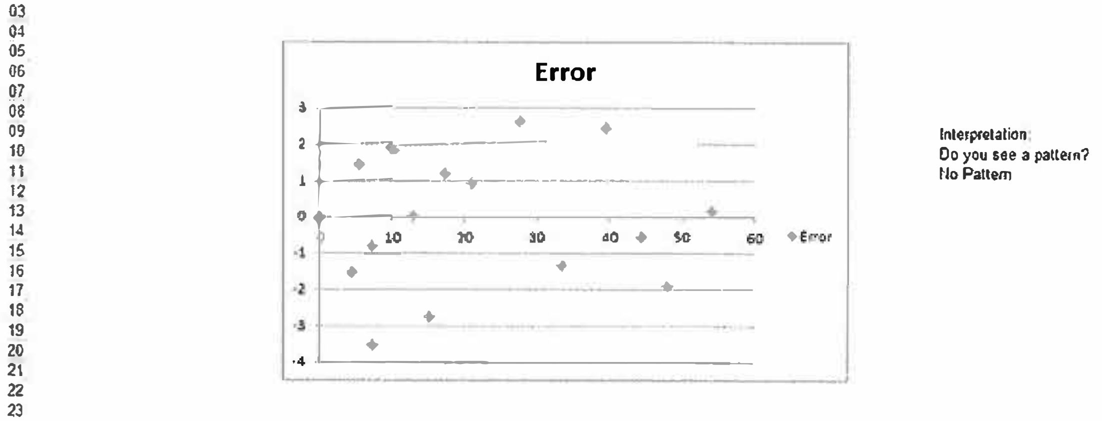
(1) Scatterplc
(2) Decide if l
(a
(b
(3) Set up the
(4) Solve the

ANOVA	df	SS	
Regression	1		
Error	13	49.45930097	3.804562

Coefficients		SE	t-statistic	P-Value
bo	58.60729	1.484524042	39.47884	6.34796E-15
b1	-0.039587	0.001739921	-22.7524	7.4226E-12

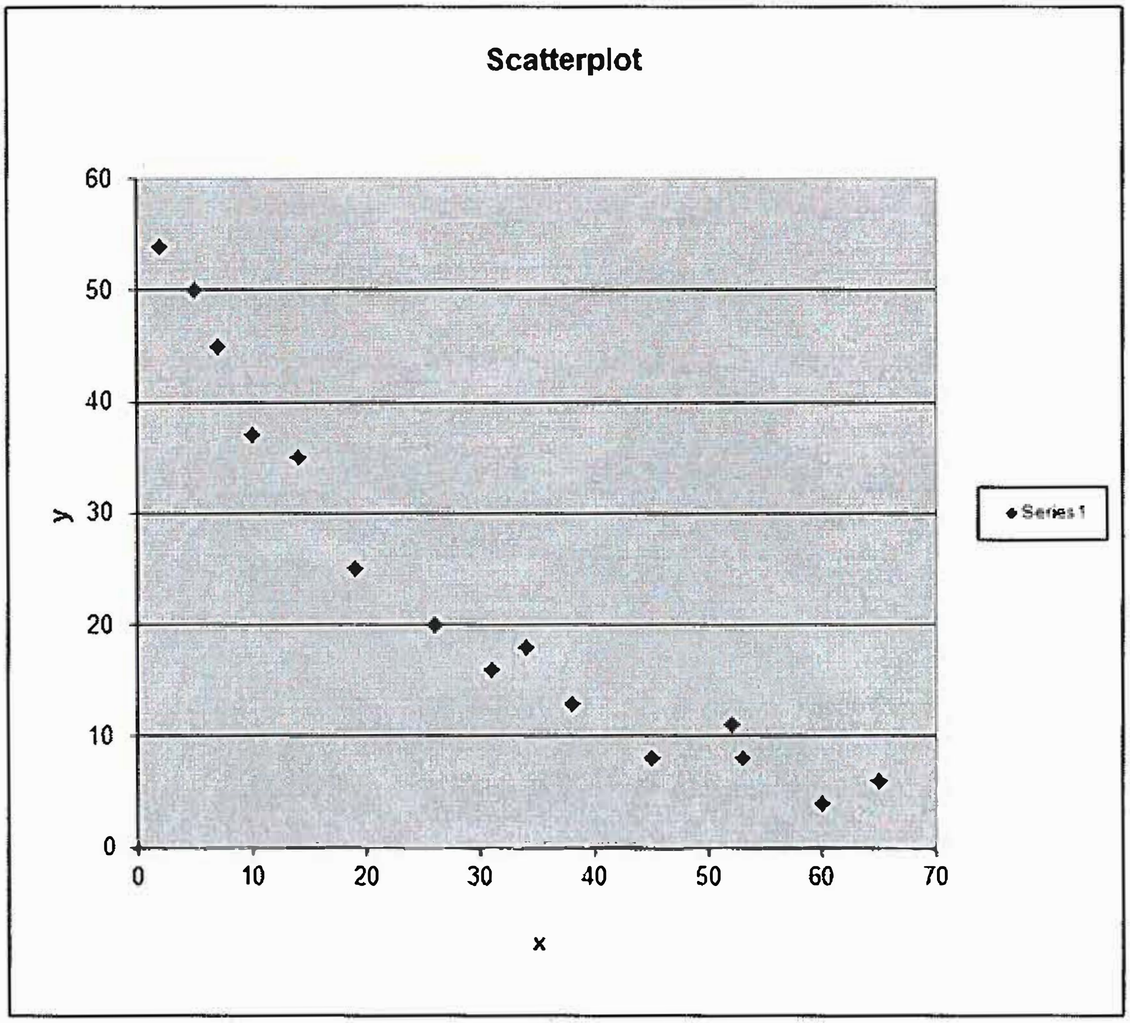
5. Obtain plots of data and residuals on your own.

Residuals



Interpretation:
Do you see a pattern?
No Pattern

Scatterplot



The other regression templates follow the same type format. Occasionally, you must engage the Solver more than once to obtain all the necessary calculations

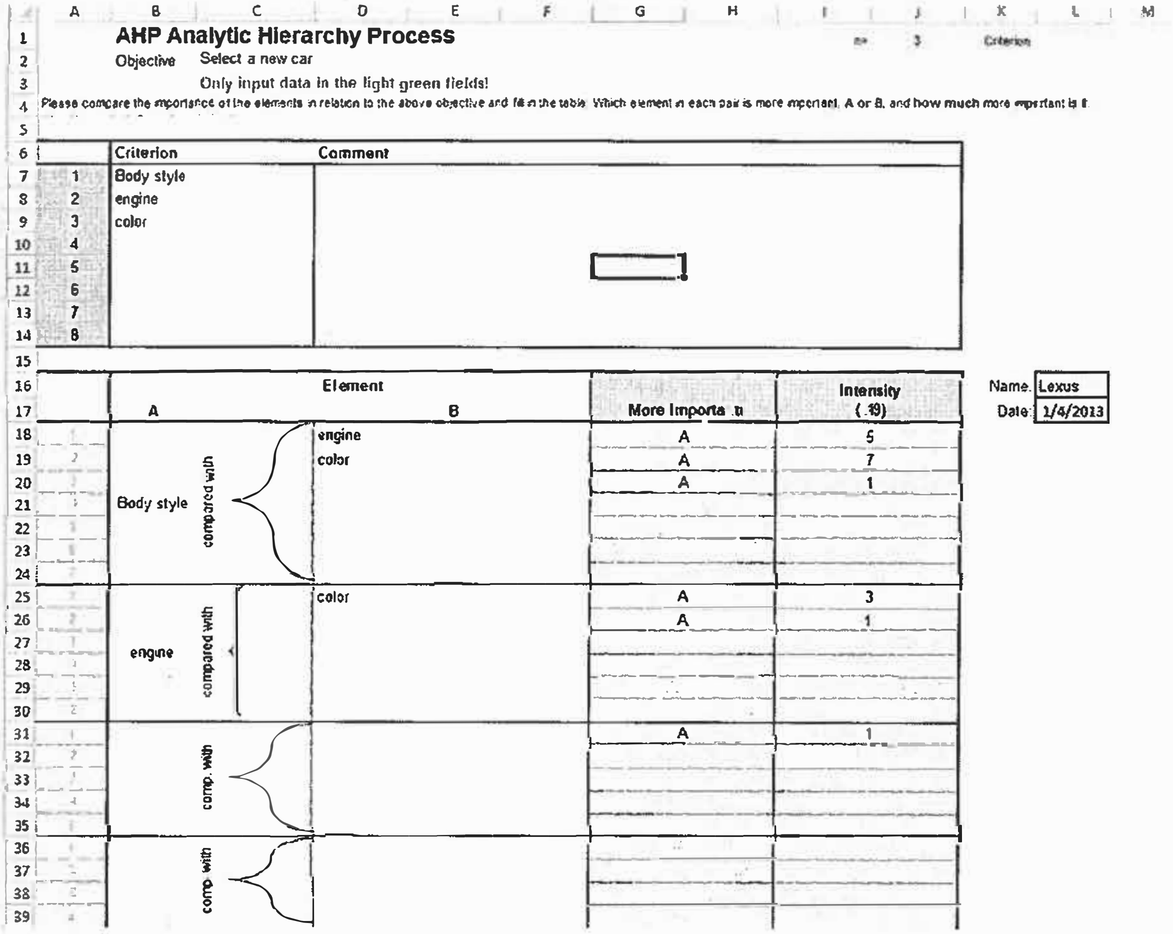
Logistics-Use template that should be self-explanatory

Poisson Regression-Use template that should be self-explanatory

DEA
Concept is to use linear programming to optimize efficiencies of units to compare. Since you have used the Solver for Linear programming in several modeling courses, I think the study guide will be sufficient.

AHP-use template for up to *8 x 8* of alternatives and criterion.

Fill out template information and interpret results.



AHP Analytic Hierarchy Process n= 3 Criterion

Objective Select a new car

Only input data in the light green fields!

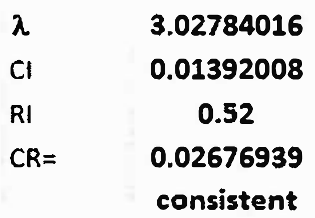
Please compare the importance of the elements in relation to the above objective and fill in the table: Which element in each pair is more important, A or B, and how much more important is it

	Criterion	Comment
1	Body style	
2	engine	
3	color	
4		
5		
6		
7		
8		

Element A		Element B	More Importa .t	Intensity (1-9)
Body style	compared with	engine	A	5
		color	A	7
			A	1
engine	compared with	color	A	3
			A	1
	comp. with		A	1
	comp. with			

Name: Lexus
Date: 1/4/2013

Insure your matrix inputs above provide a CR ratio < 0.1.

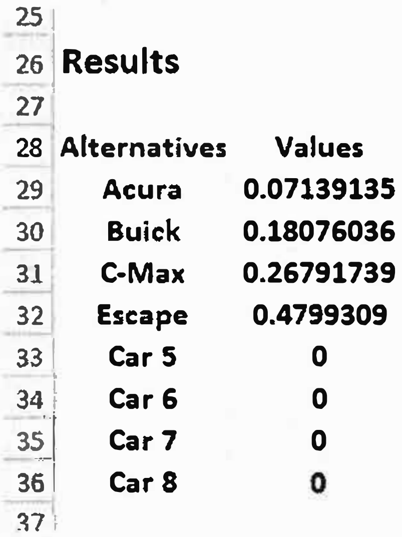


λ	3.02784016
CI	0.01392008
RI	0.52
CR=	0.02676939
	consistent

In this case the CR is 0.0267 < 0.1 so we are OK.

Repeat for all alternatives versus criterion

Go to Summary for results



25		
26	Results	
27		
28	Alternatives	Values
29	Acura	0.07139135
30	Buick	0.18076036
31	C-Max	0.26791739
32	Escape	0.4799309
33	Car 5	0
34	Car 6	0
35	Car 7	0
36	Car 8	0
37		

Put these in numerical value order: Escape #1, C-max #2, Buick #3, Acura #4.

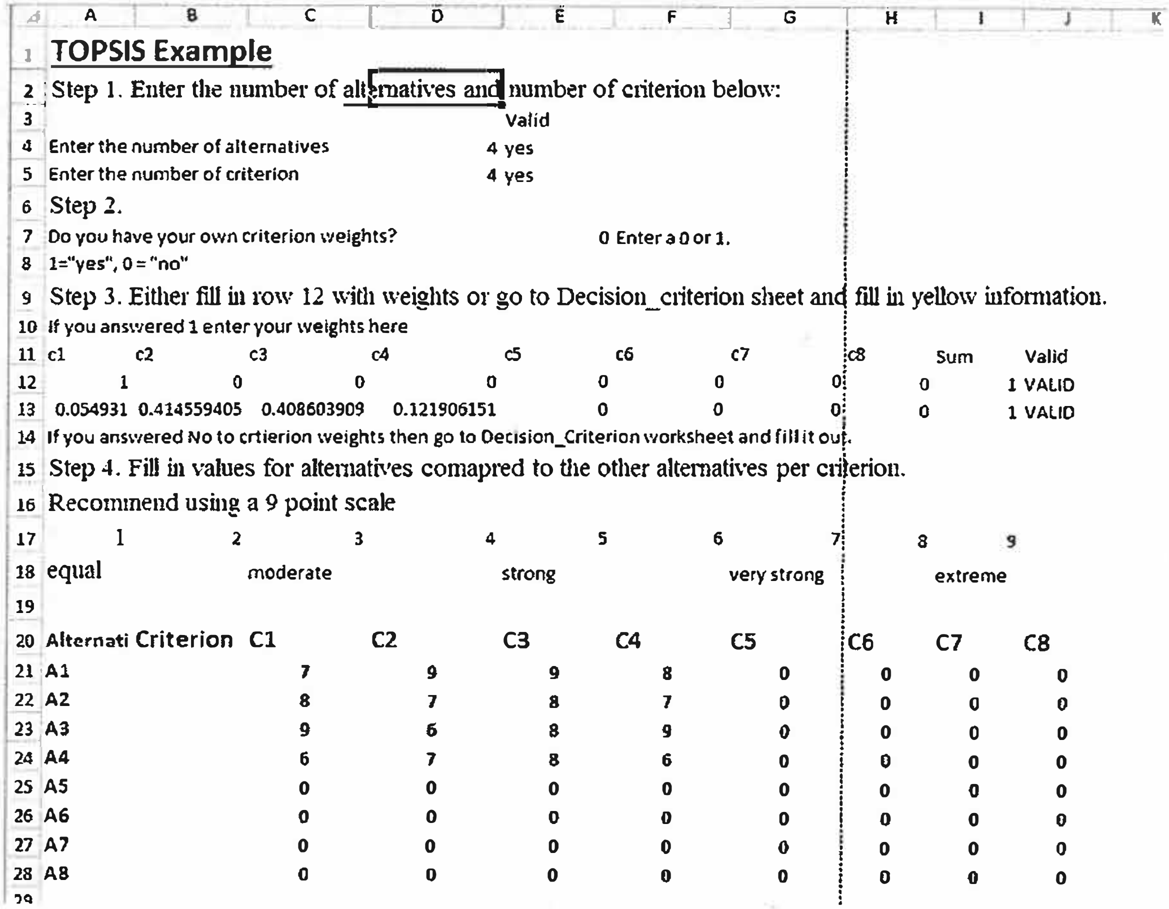
MADM procedure that ranks alternatives based upon real and subjective criterion weights. You fill in the pairwise comparison. YOU MUST Have a CR less than or equal to 0.1 or GO back and change your pair-wise values.

The result is a ranking of alternatives based upon 100% total.

TOPSIS-template

If you need Decision weights, then use the AHP template to obtain the criterion weights. Transfer the values to TOPSIS.

Again we will allow up to 8 alternatives and up to 8 criterion for our template. TOPSIS ranks order alternatives based upon distances from an ideal solution. You may use either inputs weights or use the criterion weight methods from AHP. The values in TOPSIS are either real values (where bigger is better) or subjectively entered values where bigger is better. I think the template will be self-explanatory.



TOPSIS Example

Step 1. Enter the number of alternatives and number of criterion below:

		Valid
Enter the number of alternatives	4	yes
Enter the number of criterion	4	yes

Step 2.

Do you have your own criterion weights? 0 Enter a 0 or 1.

1="yes", 0 = "no"

Step 3. Either fill in row 12 with weights or go to Decision_criterion sheet and fill in yellow information.

If you answered 1 enter your weights here

c1	c2	c3	c4	c5	c6	c7	c8	Sum	Valid
1	0	0	0	0	0	0	0	1	VALID
0.054931	0.414559405	0.408603909	0.121906151	0	0	0	0	1	VALID

If you answered No to crtierion weights then go to Decision_Criterion worksheet and fill it out.

Step 4. Fill in values for alternatives comapred to the other alternatives per criterion.

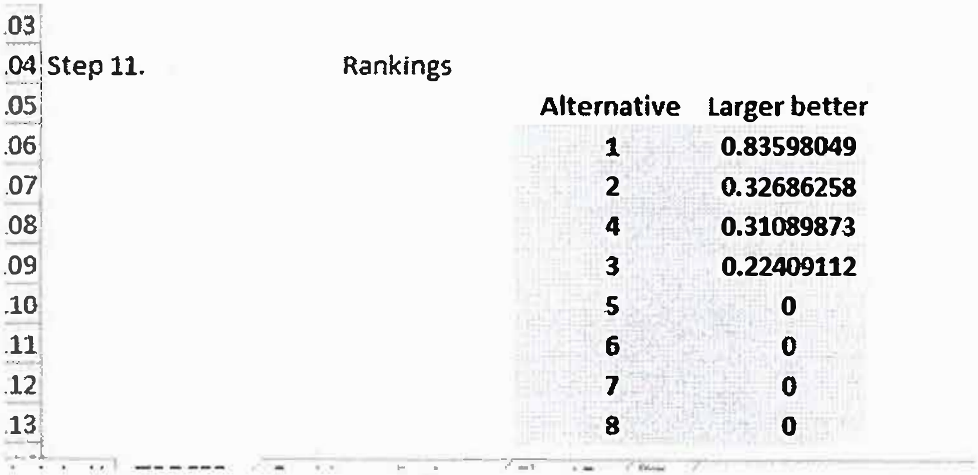
Recommend using a 9 point scale

1	2	3	4	5	6	7	8	9
equal		moderate		strong		very strong		extreme

Alternati	Criterion	C1	C2	C3	C4	C5	C6	C7	C8
A1		7	9	9	8	0	0	0	0
A2		8	7	8	7	0	0	0	0
A3		9	6	8	9	0	0	0	0
A4		6	7	8	6	0	0	0	0
A5		0	0	0	0	0	0	0	0
A6		0	0	0	0	0	0	0	0
A7		0	0	0	0	0	0	0	0
A8		0	0	0	0	0	0	0	0

Enter into yellow spaces only.

Read your output in green.



.03				
.04	Step 11.	Rankings		
.05			Alternative	Larger better
.06			1	0.83598049
.07			2	0.32686258
.08			4	0.31089873
.09			3	0.22409112
.10			5	0
.11			6	0
.12			7	0
.13			8	0

Alternative #1 is best followed by alt #2, alt #4, and alt #3.

Simulation Modeling in Excel

We stress that you need a good algorithm before you write a simulation even in Excel. Walk through one iteration of your algorithm to see if it is doing what you expect or want.

Monte Carlo simulation involves the use of random numbers, so we begin with random numbers.

Random Numbers in Excel

EXCEL has several choices to generate random numbers.

RAND

Returns an evenly distributed random real number greater than or equal to 0 and less than 1. A new random real number is returned every time the worksheet is calculated.

Syntax

RAND( )

Remarks

To generate a random real number between a and b, use:

RAND()*(b-a)+a

If you want to use RAND to generate a random number but don't want the numbers to change every time the cell is calculated, you can enter =RAND() in the formula bar, and then press F9 to change the formula to a random number.

Example

The example may be easier to understand if you copy it to a blank worksheet.

⊞ How to copy an example

	A	**B**
	Formula	**Description (Result)**
1	=RAND()	A random number between 0 and 1 (varies)
2 **3**	=RAND()*100	A random number greater than or equal to 0 but less than 100 (varies)

Examples

=Rand()   =rand()*100

0.79505492.04806

RANDBETWEEN

Returns a random integer number between the numbers you specify. A new random integer number is returned every time the worksheet is calculated.

If this function is not available, and returns the #NAME? error, install and load the Analysis ToolPak add-in.

⊞ How?

1. On the **Tools** menu, click **Add-Ins.**
2. In the **Add-Ins available** list, select the **Analysis ToolPak** box, and then click **OK.**
3. If necessary, follow the instructions in the setup program.

Syntax

RANDBETWEEN(bottom,top)

Bottom is the smallest integer RANDBETWEEN will return.

Top is the largest integer RANDBETWEEN will return.

Example

The example may be easier to understand if you copy it to a blank worksheet.

⊞ How to copy an example

1. Create a blank workbook or worksheet.
2. Select the example in the Help topic.

   Note Do not select the row or column headers.

   Selecting an example from Help

3. Press CTRL+C.
4. In the worksheet, select cell A1, and press CTRL+V.
5. To switch between viewing the results and viewing the formulas that

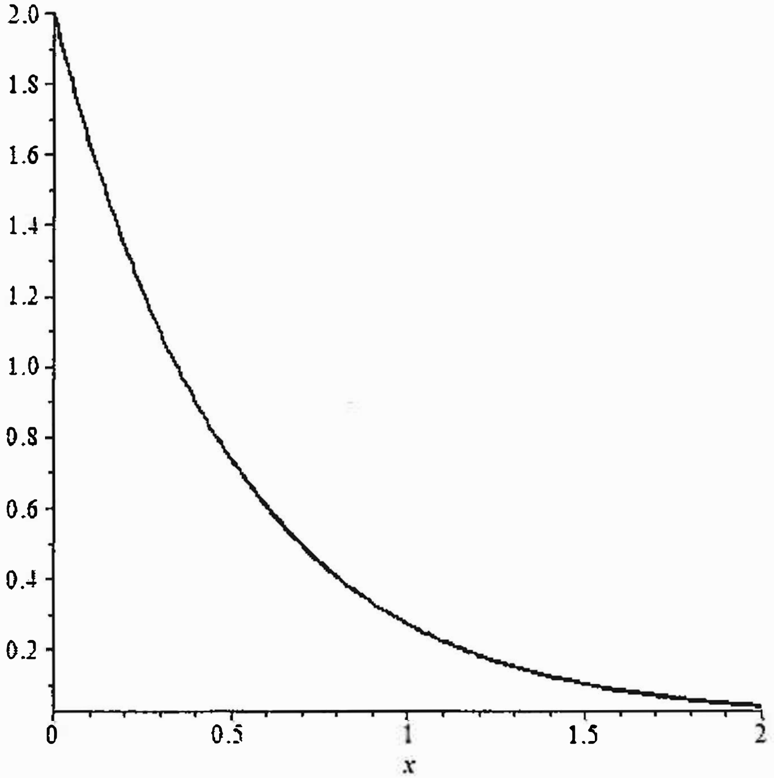
return the results, press CTRL+` (grave accent), or on the **Tools** menu, point to **Formula Auditing**, and then click **Formula Auditing Mode**.

	A	B
1	**Formula**	**Description (Result)**
2	=RANDBETWEEN(1,100)	Random number between 1 and 100 (varies)
3	=RANDBETWEEN(-1,1)	Random number between -1 and 1 (varies)

Simulations are independently built in Excel. For that reason you need an algorithm to know how to go from inputs to outputs.

Simple Simulation in Excel

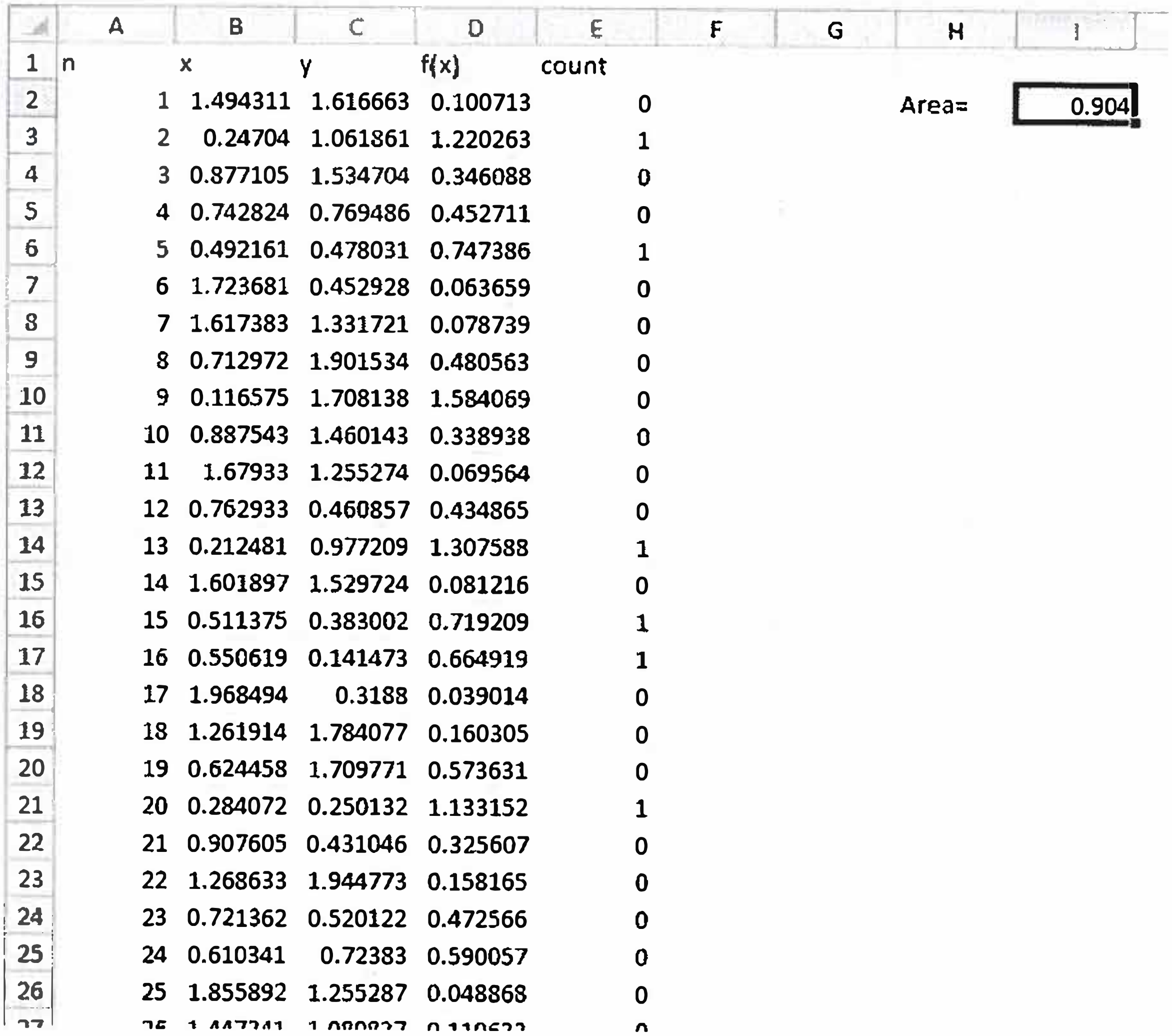
Area under the curve y=2exp(-2x) from x= [0,2].



We will use the algorithm from class.
Inputs: Number of trails, domain for x, range for y, function
Outputs: Area under the curve between $0 \leq x \leq 2$.

Let's decide we want N= 1000 trials in our simulation.
Steps.
Labels in row 1
Column A will be N starting from 1 to 1000, A2 to A1001.
Column B will be the generated x coordinate values for the random x between [0,2] using =0+(2-0)*rand().
Column C will be the generated y coordinate values for the random y which we see from the graph should also be between 0 and 2 so we use =0+(2-0)*rand().
Column D will be the value of our function evaluated at the random *x*. V=2*exp(-2*b2).
Column E Compare the y random value to the value of the function using an IF statement. We only want to count the result if *y≤f(x). Use =If(C2<=D2,1,0)*
Copy column B-E down through A1001.
Sum column E.

*Area = (2-0)(2-0)*sum of column E/ 1000*



	A	B	C	D	E	F	G	H	I
1	n	x	y	f(x)	count				
2	1	1.494311	1.616663	0.100713	0			Area=	0.904
3	2	0.24704	1.061861	1.220263	1				
4	3	0.877105	1.534704	0.346088	0				
5	4	0.742824	0.769486	0.452711	0				
6	5	0.492161	0.478031	0.747386	1				
7	6	1.723681	0.452928	0.063659	0				
8	7	1.617383	1.331721	0.078739	0				
9	8	0.712972	1.901534	0.480563	0				
10	9	0.116575	1.708138	1.584069	0				
11	10	0.887543	1.460143	0.338938	0				
12	11	1.67933	1.255274	0.069564	0				
13	12	0.762933	0.460857	0.434865	0				
14	13	0.212481	0.977209	1.307588	1				
15	14	1.601897	1.529724	0.081216	0				
16	15	0.511375	0.383002	0.719209	1				
17	16	0.550619	0.141473	0.664919	1				
18	17	1.968494	0.3188	0.039014	0				
19	18	1.261914	1.784077	0.160305	0				
20	19	0.624458	1.709771	0.573631	0				
21	20	0.284072	0.250132	1.133152	1				
22	21	0.907605	0.431046	0.325607	0				
23	22	1.268633	1.944773	0.158165	0				
24	23	0.721362	0.520122	0.472566	0				
25	24	0.610341	0.72383	0.590057	0				
26	25	1.855892	1.255287	0.048868	0				
27	26	1.447241	1.080827	0.110622	0				

Our approximate answer is 0.904 and the exact answer is 0.98168.

Simple Queue Model in Excel

For example, a bank manager wants to improve customer satisfaction by offering service such that (1) the average waiting time does not exceed 2 minutes and (2) the average queue length is 2 or fewer customers. The bank's historical records indicate that the banks gets on average 150 customers each day, Data has been collected for customer arrivals and service times as show in the tables below.

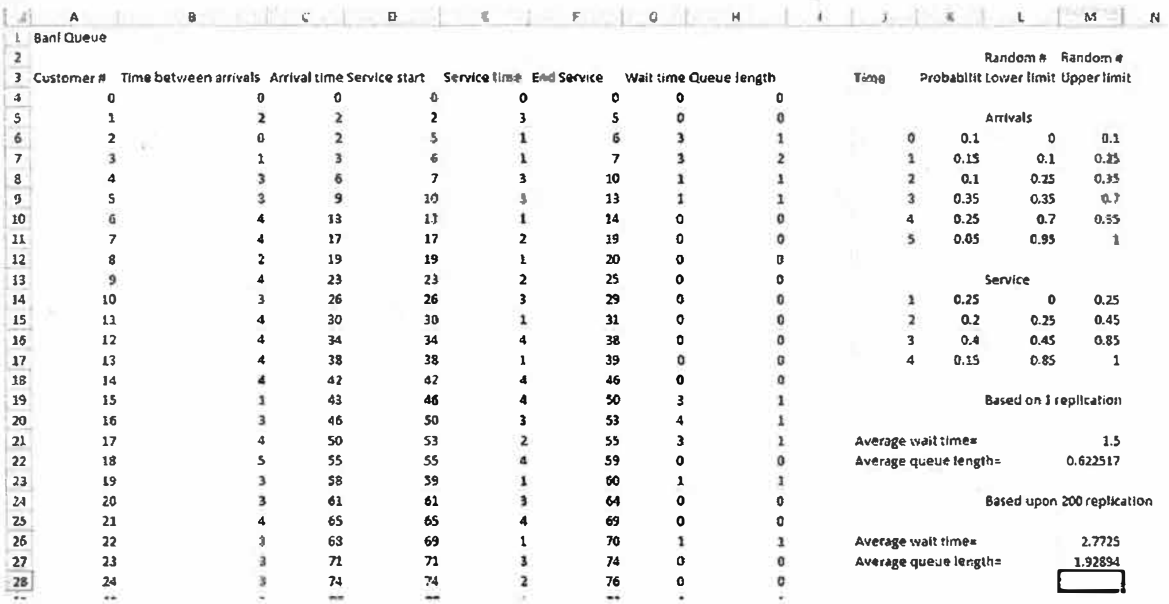
Service Time (minutes)	Probability		Time between Arrivals (minutes)	Probability
1	0.25		0	0.10
2	0.20		1	0.15
3	0.40		2	0.10
4	0.15		3	0.35
			4	0.25
			5	0.05

Each replication of the simulation corresponds to a day's operation of the bank; arrival and service of 150 customers. Note we start with customer 0 which allows us to start the simulation clock and initialize at 0 the other column values (this is a good practice in discrete event simulations).

Column A: Customers 0 – 150
Column B: the arrival time of each customer, this can be done several ways using random numbers. Here we use the LOOKUP command.
Column C calculates the actual clock time of the arrival of each customer.
Column D computes the actual start time for service.
Column E calculates the service time of each customer. Again we choose to use the LOOKUP command.
Column F calculates the clock time the customer ends service.
Column G calculates the wait time in the queue
Column H calculates the queue length using the MATCH function.

We will calculate descriptive statistics on Columns G and H for our analysis.

Let's illustrate:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Banf Queue												
2												Random #	Random #
3	Customer #	Time between arrivals	Arrival time	Service start	Service time	End Service	Wait time	Queue length		Time	Probabilit	Lower limit	Upper limit
4	0	0	0	0	0	0	0	0					
5	1	2	2	2	3	5	0	0				Arrivals	
6	2	0	2	5	1	6	3	1		0	0.1	0	0.1
7	3	1	3	6	1	7	3	2		1	0.15	0.1	0.25
8	4	3	6	7	3	10	1	1		2	0.1	0.25	0.35
9	5	3	9	10	3	13	1	1		3	0.35	0.35	0.7
10	6	4	13	13	1	14	0	0		4	0.25	0.7	0.95
11	7	4	17	17	2	19	0	0		5	0.05	0.95	1
12	8	2	19	19	1	20	0	0					
13	9	4	23	23	2	25	0	0				Service	
14	10	3	26	26	3	29	0	0		1	0.25	0	0.25
15	11	4	30	30	1	31	0	0		2	0.2	0.25	0.45
16	12	4	34	34	4	38	0	0		3	0.4	0.45	0.85
17	13	4	38	38	1	39	0	0		4	0.15	0.85	1
18	14	4	42	42	4	46	0	0					
19	15	1	43	46	4	50	3	1				Based on 1 replication	
20	16	3	46	50	3	53	4	1					
21	17	4	50	53	2	55	3	1		Average wait time=			1.5
22	18	5	55	55	4	59	0	0		Average queue length=			0.622517
23	19	3	58	59	1	60	1	1					
24	20	3	61	61	3	64	0	0				Based upon 200 replication	
25	21	4	65	65	4	69	0	0					
26	22	3	68	69	1	70	1	1		Average wait time=			2.7725
27	23	3	71	71	3	74	0	0		Average queue length=			1.92894
28	24	3	74	74	2	76	0	0					

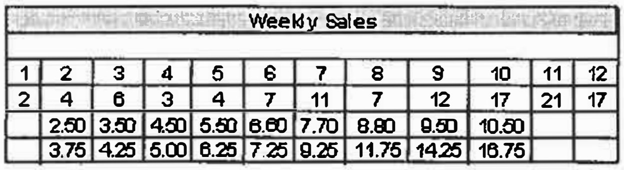
The bank manager has shown, via simulation, that unless they improve the service times that will not meet the goal.

(optional) Moving Average and Exponential Smoothing

Moving Average Models: Use the Add Trendline option to analyze a moving average forecasting model in Excel. You must first create a graph of the time series you want to analyze. Select the range that contains your data and make a scatter plot of the data. Once the chart is created, follow these steps:

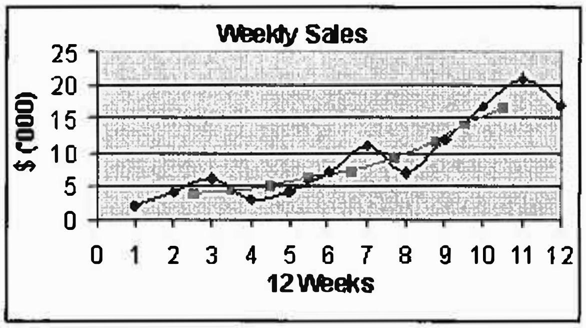
1. Click on the chart to select it, and click on any point on the line to select the data series. When you click on the chart to select it, a new option, Chart, s added to the menu bar.
2. From the Chart menu, select Add Trendline.

The following is the moving average of order 4 for weekly sales:



Weekly Sales

1	2	3	4	5	6	7	8	9	10	11	12
2	4	6	3	4	7	11	7	12	17	21	17
	2.50	3.50	4.50	5.50	6.60	7.70	8.80	9.50	10.50		
	3.75	4.25	5.00	6.25	7.25	9.25	11.75	14.25	16.75		



Exponential Smoothing Models: The simplest way to analyze a timer series using an Exponential Smoothing model in Excel is to use the data analysis tool. This tool works almost exactly like the one for Moving Average, except that you will need to input the value of a instead of the number of periods, k. Once you have entered the data range and the damping factor, 1- α , and indicated what output you want and a location, the analysis is the same as the one for the Moving Average model.

1. **Time Series data analysis:**

Time series data consists of numerical values recorded at intervals of time. Time series data are often used in conjunction with regression techniques, which are covered here.

In time series analysis the independent variable (x) is given as a period of time. A linear regression equation is used to calculate the trend that the dependent variable (y) adheres to as time passes.

But when time is used as the independent variable there are a number of complications that are introduced to the regression method. These originate from the fact that the dependent variable will usually be subject to a number of influences that, in themselves, are affected by the units that are used to measure time.

For example, if annual data are used, it will be impossible to identify the seasonal factors that may well influence the data. So, if we are looking at data about the consumption of ice cream products, we would probably want to view quarterly figures rather than merely annual data, as we would expect there to be an increase in purchases of these products in the Summer quarter.

So, the objective of time series analysis must be to develop techniques to divide the raw time series data into its component parts. These are: a trend value (t), a seasonal element (s), and a residual element (r).

Firstly, then we want to find the trend value of the time series data that we are analyzing. There are a range of techniques designed to discover the trend line. Some of these are based on trends to produce a straight line of best fit. For most purposes a line is drawn by hand or by using averages over periods of time, to smooth out fluctuations and show the general trend.

The most commonly used trend is the moving average, which is a process of repeatedly calculating a series of different average values along a time series in order to produce a trend line.

There is more detail available on moving averages in the 'Digging' section of TimeWeb.

Regression analysis is used as a more advanced method of trend identification. If we assume that the form of model used to identify the component parts of time series data is as follows:

$y = t + s +$ r then we can discover the trend on the basis of a least squares regression equation.

Regression analysis of time series data using Excel.

You should now open a new Excel workbook and enter the following data:

In cell A1: time (x)
In cells A2:A21

```
1   2   3   4   5   6   7   8   9   10
11  12  13  14  15  16  17  18  19  20
```

In cell B1: y
In cells B2:B21

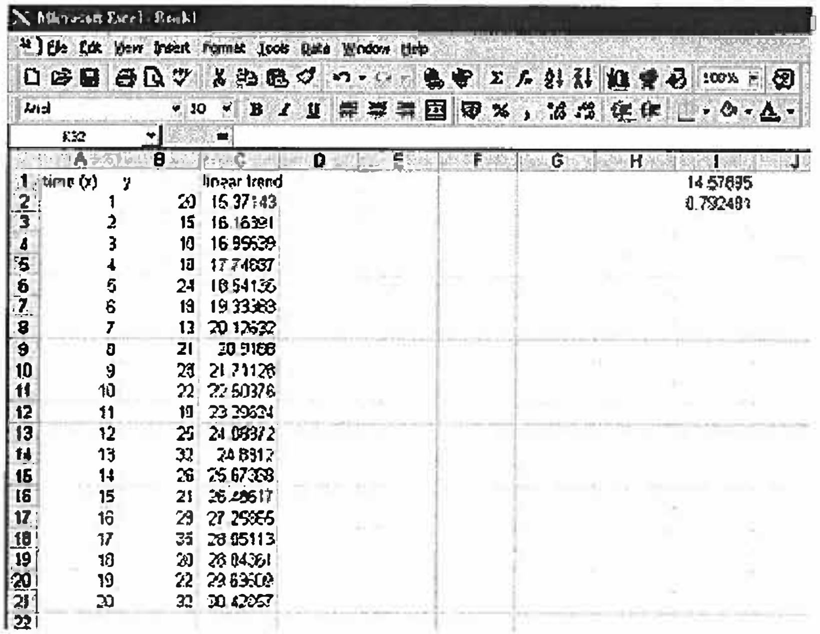
```
20  15  10  18  24  18  13  21  28  22
19  25  32  26  21  29  35  28  22  32
```

This data represents a series of quarterly observations over a five year period, where x is the number of the quarter (1 to 20).

Now compute the intercept and the gradient of the trend line and use these to calculate the estimated trend in column C, using the following formulae:

In I1: =INTERCEPT(B2:B21,A2:A21)
In I2: =SLOPE(B2:B21,A2:A21)
In C2: =I$1+I$2*A2 copied into C3:C21

The resulting sheet should look like the following:



	A	B	C	I
1	time (x)	y	linear trend	14.57895
2	1	20	15.37143	0.792481
3	2	15	16.16391	
4	3	10	16.95639	
5	4	18	17.74887	
6	5	24	18.54135	
7	6	18	19.33383	
8	7	13	20.12632	
9	8	21	20.9188	
10	9	28	21.71128	
11	10	22	22.50376	
12	11	19	23.29624	
13	12	25	24.08872	
14	13	32	24.8812	
15	14	26	25.67368	
16	15	21	26.46617	
17	16	29	27.25865	
18	17	35	28.05113	
19	18	28	28.84361	
20	19	22	29.63609	
21	20	32	30.42857	
22				

Column C gives us the linear trend values predicted by the regression equation of y on x, as follows:
y = 14.57895 + 0.792481x

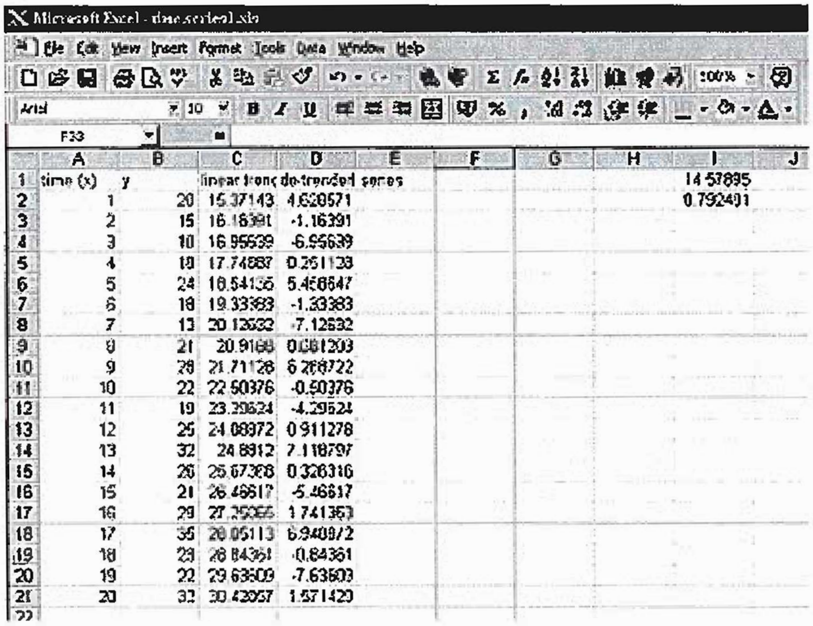
The next step is to identify the de-trended series. In the form of the equation set up earlier, this will be shown as:

De-trended series = y - t

This means that the entries in Column D should be arrived at by subtracting the values in Column C from those in Column B.

We can do this by entering the following formula into cell D2 (and copying it into D3:D21).
=B2-C2

Now we can label the column 'De-Trended Series' in cell D1.
The sheet should now appear as below:


Microsoft Excel - time series1.xls
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
F33

	A	B	C	D	E	F	G	H	I	J
1	time (x)	y	linear trend	de-trended series					14.57895	
2	1	20	15.37143	4.628571					0.792481	
3	2	15	16.16391	-1.16391						
4	3	10	16.95639	-6.95639						
5	4	19	17.74887	1.251128						
6	5	24	18.54135	5.458647						
7	6	18	19.33383	-1.33383						
8	7	13	20.12632	-7.12632						
9	8	21	20.9188	0.081203						
10	9	28	21.71128	6.288722						
11	10	22	22.50376	-0.50376						
12	11	19	23.29624	-4.29624						
13	12	25	24.08872	0.911278						
14	13	32	24.8812	7.118797						
15	14	26	25.67368	0.326316						
16	15	21	26.46617	-5.46617						
17	16	29	27.25865	1.741353						
18	17	35	28.05113	6.948872						
19	18	28	28.84361	-0.84361						
20	19	22	29.63609	-7.63609						
21	20	32	30.42857	1.571429						
22										

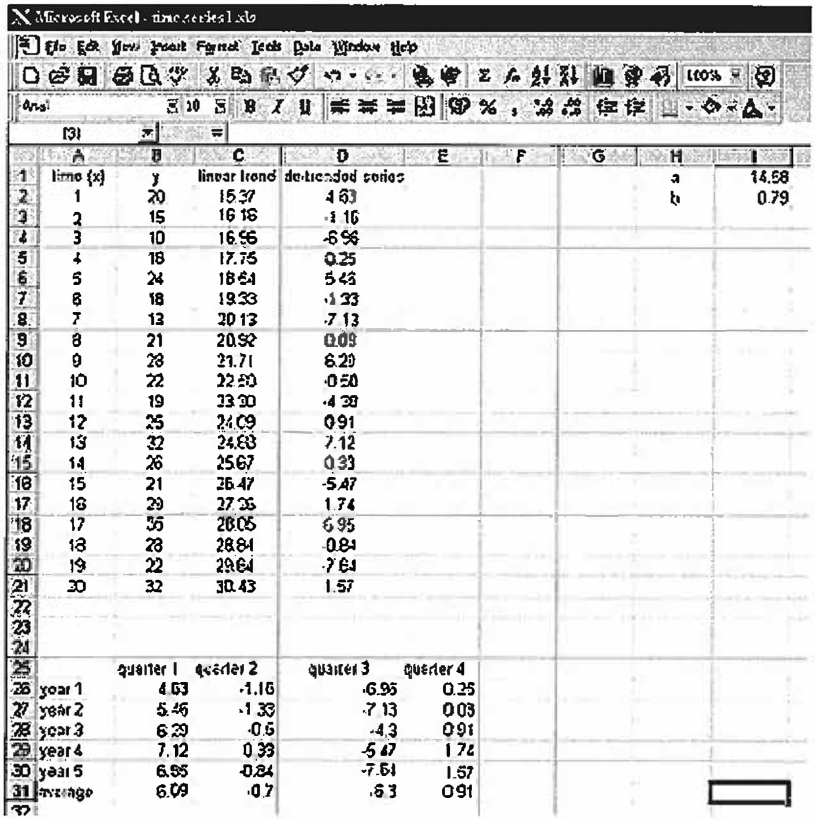
OK. Now we want to see if we can identify any seasonal component contained in the de-trended series. To do this we have to group all the values for each quarter of the year. This means that over the 5 year period, the grouped quarters will be as follows:

1, 5, 9, 13, 17.
2, 6, 10, 14, 18.

3, 7, 11, 15, 19.
4, 8, 12, 16, 20.

With this operation we want to find the difference between the trend value and the actual value for y. So for the first quarter of year 1 the difference between the actual value (20) and the trend value (15.37143) is 4.63.

These equivalent quarter values are placed in the worksheet and are averaged to produce an average value for each season that they represent. The worksheet should now resemble the following:

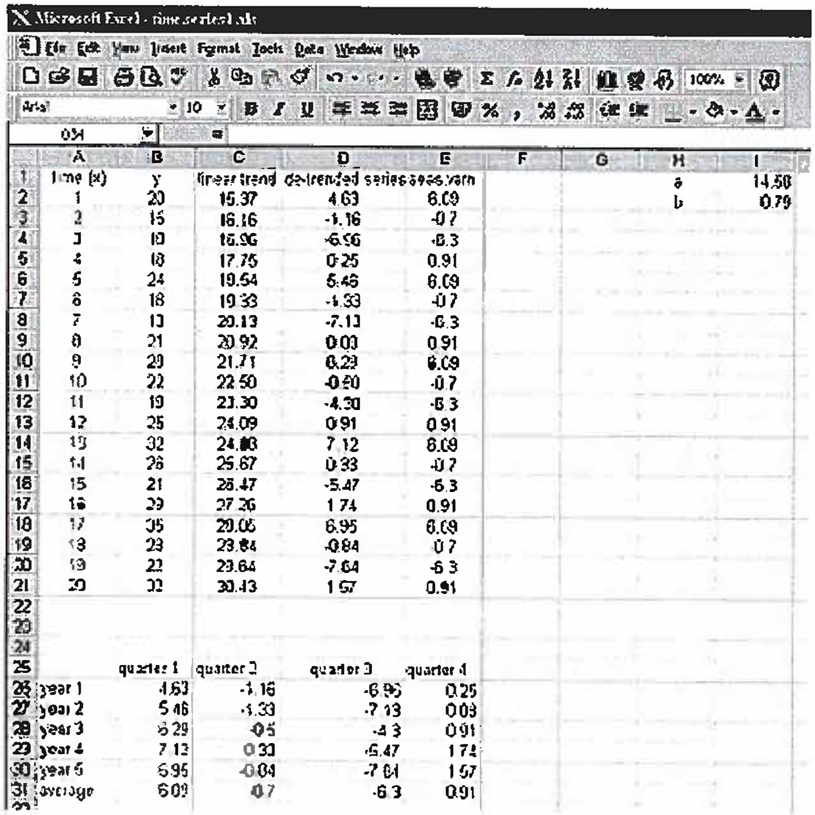

Microsoft Excel - time series1.xls
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
I31
=

	A	B	C	D	E	F	G	H	I
1	time (x)	y	linear trend	de-trended series				a	14.58
2	1	20	15.37	4.63				b	0.79
3	2	15	16.16	-1.16					
4	3	10	16.96	-6.96					
5	4	18	17.75	0.25					
6	5	24	18.64	5.43					
7	6	18	19.33	-1.33					
8	7	13	20.13	-7.13					
9	8	21	20.92	0.09					
10	9	28	21.71	6.29					
11	10	22	22.50	-0.50					
12	11	19	23.30	-4.30					
13	12	25	24.09	0.91					
14	13	32	24.88	7.12					
15	14	26	25.67	0.33					
16	15	21	26.47	-5.47					
17	16	29	27.26	1.74					
18	17	35	28.05	6.95					
19	18	28	28.84	-0.84					
20	19	22	29.64	-7.64					
21	20	32	30.43	1.57					
22									
23									
24									
25		quarter 1	quarter 2	quarter 3	quarter 4				
26	year 1	4.63	-1.16	-6.96	0.25				
27	year 2	5.46	-1.33	-7.13	0.03				
28	year 3	6.29	-0.5	-4.3	0.91				
29	year 4	7.12	0.33	-5.47	1.74				
30	year 5	6.95	-0.84	-7.64	1.57				
31	average	6.09	-0.7	-6.3	0.91				

Notice that the four entries in the Row labeled 'Average' (A31 to E31) is an estimate of the seasonal variation of the series. The next step is to place these values in Column E, alongside their respective quarters. These are then subtracted from the actual time series values. This produces the seasonally adjusted series in Column F.

Do this by following these steps:
In E2, E3, E4 and E5 enter
=B$31 =C$31 =D$31 =E$31
then copy E2:E5 into E6:E21

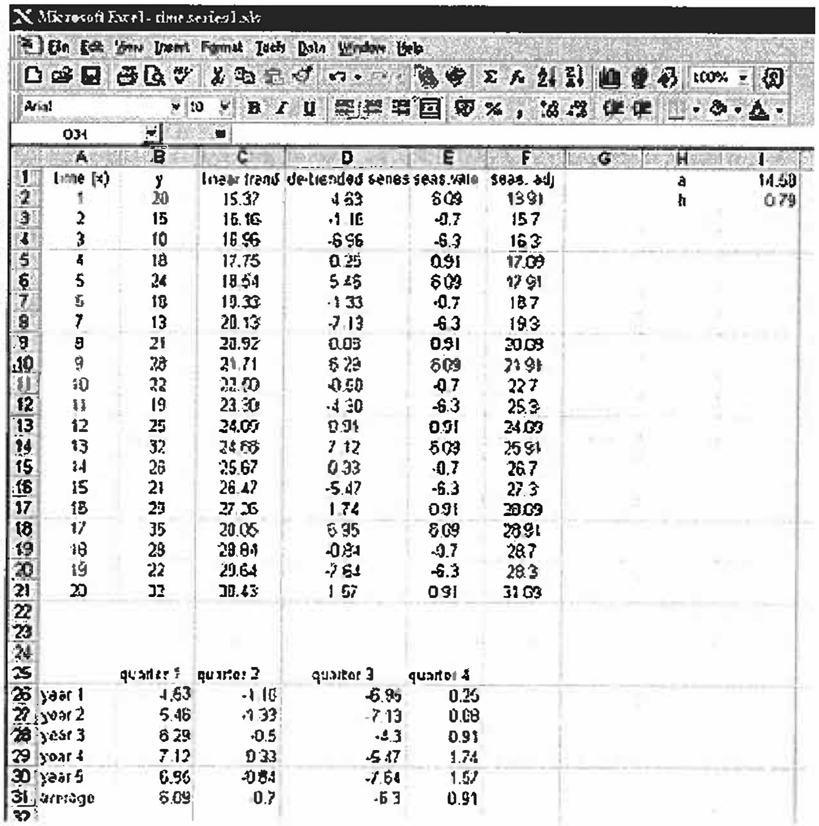
This transfers the equivalent quarter seasonal variations into Column E and should produce the following result:


Microsoft Excel - time series1.xls
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
O34
=

	A	B	C	D	E	F	G	H	I
1	time (x)	y	linear trend	de-trended series	seas.varn			a	14.58
2	1	20	15.37	4.63	6.09			b	0.79
3	2	15	16.16	-1.16	-0.7				
4	3	10	16.96	-6.96	-6.3				
5	4	18	17.75	0.25	0.91				
6	5	24	18.54	5.46	6.09				
7	6	18	19.33	-1.33	-0.7				
8	7	13	20.13	-7.13	-6.3				
9	8	21	20.92	0.08	0.91				
10	9	28	21.71	6.29	6.09				
11	10	22	22.50	-0.50	-0.7				
12	11	19	23.30	-4.30	-6.3				
13	12	25	24.09	0.91	0.91				
14	13	32	24.88	7.12	6.09				
15	14	26	25.67	0.33	-0.7				
16	15	21	26.47	-5.47	-6.3				
17	16	29	27.26	1.74	0.91				
18	17	35	28.05	6.95	6.09				
19	18	28	28.84	-0.84	-0.7				
20	19	22	29.64	-7.64	-6.3				
21	20	32	30.43	1.57	0.91				
22									
23									
24									
25		quarter 1	quarter 2	quarter 3	quarter 4				
26	year 1	4.63	-1.16	-6.96	0.25				
27	year 2	5.46	-1.33	-7.13	0.08				
28	year 3	6.29	-0.5	-4.3	0.91				
29	year 4	7.12	0.33	-5.47	1.74				
30	year 5	6.95	-0.84	-7.64	1.57				
31	average	6.09	-0.7	-6.3	0.91				

Now in F2 enter =B2-E2 and copy this into F3:F21 This will subtract the seasonal variations from the actual series values and will produce the seasonally adjusted series in Column F.

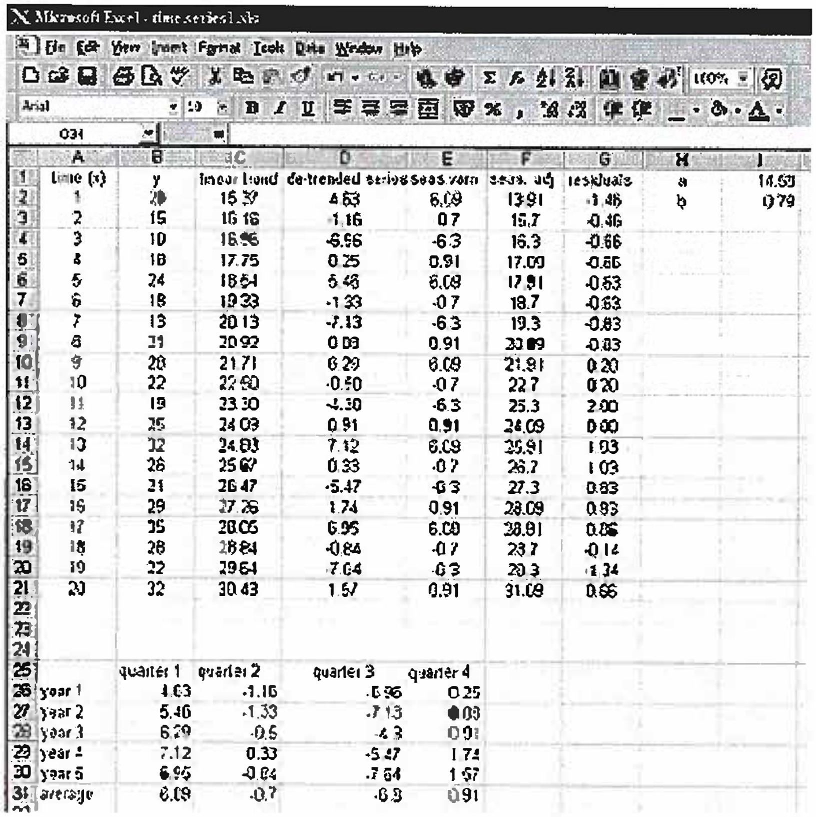
The result should resemble the following:


Microsoft Excel - time series1.xls
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
O34

	A	B	C	D	E	F	G	H	I
1	time (x)	y	linear trend	de-trended series	seas.vale	seas. adj		a	14.58
2	1	20	15.37	4.63	6.09	13.91		h	0.79
3	2	15	16.16	-1.16	-0.7	15.7			
4	3	10	16.96	-6.96	-6.3	16.3			
5	4	18	17.75	0.25	0.91	17.09			
6	5	24	18.54	5.46	6.09	17.91			
7	6	18	19.33	-1.33	-0.7	18.7			
8	7	13	20.13	-7.13	-6.3	19.3			
9	8	21	20.92	0.08	0.91	20.09			
10	9	28	21.71	6.29	6.09	21.91			
11	10	22	22.50	-0.50	-0.7	22.7			
12	11	19	23.30	-4.30	-6.3	25.3			
13	12	25	24.09	0.91	0.91	24.09			
14	13	32	24.88	7.12	6.09	25.91			
15	14	26	25.67	0.33	-0.7	26.7			
16	15	21	26.47	-5.47	-6.3	27.3			
17	16	29	27.26	1.74	0.91	28.09			
18	17	35	28.05	6.95	6.09	28.91			
19	18	28	28.84	-0.84	-0.7	28.7			
20	19	22	29.64	-7.64	-6.3	28.3			
21	20	32	30.43	1.57	0.91	31.09			
22									
23									
24									
25		quarter 1	quarter 2	quarter 3	quarter 4				
26	year 1	4.63	-1.16	-6.96	0.25				
27	year 2	5.46	-1.33	-7.13	0.08				
28	year 3	6.29	-0.5	-4.3	0.91				
29	year 4	7.12	0.33	-5.47	1.74				
30	year 5	6.96	-0.84	-7.64	1.57				
31	average	6.09	0.7	-6.3	0.91				
32									

The final step is to identify the **residual elements**.
These are found by the following expression: y - s - t.
To produce these enter the following formula into Cell G2:
=B2-E2-C2 and copy it into G3:G21.

Having carried this out and added labels, the worksheet should look like the following:

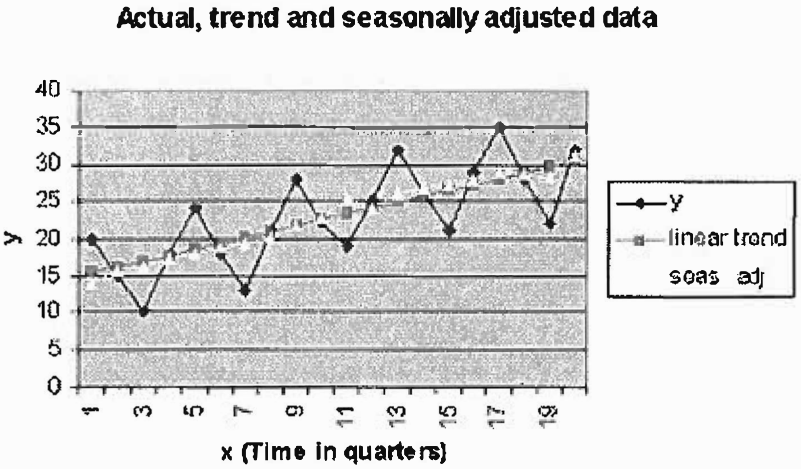

Microsoft Excel - timeseries1.xls
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
O31

	A	B	C	D	E	F	G	H	I
1	time (x)	y	linear trend	de-trended series	seas.varn	seas. adj	residuals	a	14.59
2	1	20	15.37	4.63	6.09	13.91	-1.46	b	0.79
3	2	15	16.16	-1.16	-0.7	15.7	-0.46		
4	3	10	16.96	-6.96	-6.3	16.3	-0.66		
5	4	18	17.75	0.25	0.91	17.09	-0.66		
6	5	24	18.54	5.46	6.09	17.91	-0.63		
7	6	18	19.33	-1.33	-0.7	18.7	-0.63		
8	7	13	20.13	-7.13	-6.3	19.3	-0.83		
9	8	21	20.92	0.08	0.91	20.09	-0.83		
10	9	28	21.71	6.29	6.09	21.91	0.20		
11	10	22	22.50	-0.50	-0.7	22.7	0.20		
12	11	19	23.30	-4.30	-6.3	25.3	2.00		
13	12	25	24.09	0.91	0.91	24.09	0.00		
14	13	32	24.88	7.12	6.09	25.91	1.03		
15	14	26	25.67	0.33	-0.7	26.7	1.03		
16	15	21	26.47	-5.47	-6.3	27.3	0.83		
17	16	29	27.26	1.74	0.91	28.09	0.83		
18	17	35	28.05	6.95	6.09	28.91	0.86		
19	18	28	28.84	-0.84	-0.7	28.7	-0.14		
20	19	22	29.64	-7.64	-6.3	28.3	-1.34		
21	20	32	30.43	1.57	0.91	31.09	0.66		
22									
23									
24									
25		quarter 1	quarter 2	quarter 3	quarter 4				
26	year 1	4.63	-1.16	-6.96	0.25				
27	year 2	5.46	-1.33	-7.13	0.08				
28	year 3	6.29	-0.5	-4.3	0.91				
29	year 4	7.12	0.33	-5.47	1.74				
30	year 5	6.95	-0.84	-7.64	1.57				
31	average	6.09	-0.7	-6.3	0.91				

Summary of Time Series analysis:

You should note that the seasonally adjusted series is one of the most important parts of this analysis. Almost all statistics that you find here in TimeWeb and generally will be seasonally adjusted. What this does is to indicate how the dependent variable would have behaved if it had not been affected by seasonal variation.

In order to see this more clearly, produce a graph from your spreadsheet of the actual series, the trend and the seasonally adjusted series, all plotted on the same axes. You should produce the following type of graph:



As you can see quite clearly, the seasonally adjusted data closely follows the trend line, rather than the actual series. This is the whole reason for carrying out seasonal adjustment.

Examples

1. Applications and Numerical Examples

Descriptive Statistics: Suppose you have the following, n = 10, data:

1.2, 1.5, 2.6, 3.8, 2.4, 1.9, 3.5, 2.5, 2.4, 3.0

1. Type your n data points into the cells A1 through An.
2. Click on the "Tools" menu. (At the bottom of the "Tools" menu will be a submenu "Data Analysis...", if the Analysis Tool Pack has been properly installed.)
3. Clicking on "Data Analysis..." will lead to a menu from which "Descriptive Statistics" is to be selected.
4. Select "Descriptive Statistics" by pointing at it and clicking twice, or by highlighting it and clicking on the "Okay" button.
5. Within the Descriptive Statistics submenu,

a. for the "input range" enter "A1:Dn", assuming you typed the data into cells A1 to An.

b. click on the "output range" button and enter the output range "C1:C16".

c. click on the Summary Statistics box

d. finally, click on "Okay."

The Central Tendency: The data can be sorted in ascending order:

1.2, 1.5, 1.9, 2.4, 2.4, 2.5, 2.6, 3.0, 3.5, 3.8

The mean, median and mode are computed as follows:

(1.2 1.5 2.6 3.8 2.4 1.9 3.5 2.5 2.4 3.0) / 10 = 2.48

(2.4 + 2.5) / 2 = 2.45

The mode is 2.4, since it is the only value that occurs twice.

The midrange is (1.2+ 3.8) / 2 = 2.5.

Note that the mean, median and mode of this set of data are very close to each other. This suggests that the data is very symmetrically distributed.

Variance: The variance of a set of data is the average of the cumulative measure of the squares of the difference of all the data values from the mean.

The sample variance-based estimation for the population variance are computed differently. The sample variance is simply the arithmetic mean of the squares of the difference between each data value in the sample and the mean of the sample. On the other hand, the formula for an estimate for the variance in the population is similar to the formula for the sample variance, except that the denominator in the fraction is (n-1) instead of n. However, you should not worry about this difference if the sample size is large, say over 30. ***Compute an estimate for the variance of the population***, given the following sorted data:

1.2, 1.5, 1.9, 2.4, 2.4, 2.5, 2.6, 3.0, 3.5, 3.8 mean = 2.48 as computed earlier. An estimate for the population variance is: $s^2 = 1 / (10-1)$ [$(1.2 - 2.48)^2 + (1.5 - 2.48)^2 + (1.9 - 2.48)^2 + (2.4 -2.48)^2 + (2.4 - 2.48)^2 + (2.5 - 2.48)^2 + (2.6 - 2.48)^2 + (3.0 - 2.48)^2 + (3.5 -2.48)^2 + (3.8 - 2.48)^2$]
= (1 / 9) (1.6384 + 0.9604 + 0.3364 + 0.0064 + 0.0064 + 0.0004 + 0.0144 + 0.2704 + 1.0404 + 1.7424) = 0.6684

Therefore, **the standard deviation** is $s = (0.6684)^{1/2} = 0.8176$

2. **Probability and Expected Values:** Newsweek reported that "average take" for bank robberies was $3,244 but 85 percent of the robbers were caught. Assuming 60 percent of those caught lose their entire take and 40 percent lose half, graph the probability mass function using EXCEL. Calculate the expected take from a bank robbery. Does it pay to be a bank robber?

To construct the probability function for bank robberies, first define the random variable x, bank robbery take. If the robber is not caught, x = $3,244. If the robber is caught and manages to keep half, x = $1,622. If the robber is caught and loses it all, then x = 0. The associated probabilities for these x values are 0.15 = (1 - 0.85), 0.34 = (0.85)(0.4), and 0.51 = (0.85)(0.6). After entering the x values in cells A1, A2 and A3 and after entering the associated probabilities in B1, B2, and B3, the following steps lead to the probability mass function:

1. Click on ChartWizard. The "ChartWizard Step 1 of 4" screen will appear.
2. Highlight "Column" at "ChartWizard Step 1 of 4" and click "Next."
3. At "ChartWizard Step 2 of 4 Chart Source Data," enter "=B1:B3" for "Data range," and click "column" button for "Series in." A graph will appear. Click on "series" toward the top of the screen to get a new page.
4. At the bottom of the "Series" page, is a rectangle for "Category (X) axis labels:" Click on this rectangle and then highlight A1:A3.
5. At "Step 3 of 4"; move on by clicking on "Next," and at "Step 4 of 4", click on "Finish."

The expected value of a robbery is $1,038.08.

$$E(X) = (0)(0.51)+(1622)(0.34) + (3244)(0.15) = 0 + 551.48 + 486.60 = 1038.08$$

The expected return on a bank robbery is positive. On average, bank robbers get $1,038.08 per heist. If criminals make their decisions strictly on this expected value, then it pays to rob banks. A decision rule based only on an expected value, however, ignores the risks or variability in the returns. In addition, our expected value calculations do not include the cost of jail time, which could be viewed by criminals as substantial.

3. **Discrete & Continuous Random Variables:**

Binomial Distribution Application: A multiple choice test has four unrelated questions. Each question has five possible choices but only one is correct. Thus, a person who guesses randomly has a probability of 0.2 of guessing correctly. Draw a tree diagram showing the different ways in which a test taker could get 0, 1, 2, 3 and 4 correct answers. Sketch the probability mass function for this test. What is the probability a person who guesses will get two or more correct?

Solution: Letting Y stand for a correct answer and N a wrong answer, where the probability of Y is 0.2 and the probability of N is 0.8 for each of the four questions, the probability tree diagram is shown in the textbook on page 182. This probability tree diagram shows the "branches" that must be followed to show the calculations captured in the binomial mass function for n = 4 and = 0.2. For example, the tree diagram shows the six different branch systems that yield two correct and two wrong answers (which corresponds to 4!/(2!2!) = 6. The binomial mass function shows the probability of two correct answers as

P(x = 2 | n = 4, p = 0.2) = 6(.2)2(.8)2 = 6(0.0256) = 0.1536 = P(2)

Which is obtained from excel by using the "BINOMDIST" Command, where the first entry is x, the second is n, and the third is mass (0) or cumulative (1); that is, entering

=BINOMDIST(2,4,0.2,0) IN ANY EXCEL CELL YIELDS 0.1536 AND
=BINOMDIST(3,4,0.2,0) YIELDS P(x=3|n=4, p = 0.2) = 0.0256
=BINOMDIST(4,4,0.2,0) YIELDS P(x=4|n=4, p = 0.2) = 0.0016
=1-BINOMDIST(1,4,0.2,1) YIELDS P(x ≥ 2 | n = 4, p = 0.2) = 0.1808

Normal Example: If the time required to complete an examination by those with a certain learning disability is believed to be distributed normally, with mean of 65 minutes and a standard deviation of 15 minutes, then when can the exam be terminated so that 99 percent of those with the disability can finish?

Solution: Because the average and standard deviation are known, what needs to be established is the amount of time, above the mean time, such that 99 percent of the distribution is lower. This is a distance that is measured in standard deviations as given by the Z value corresponding to the 0.99 probability found in the body of Appendix B, Table 5,as shown in the textbook OR the commands entered into any cell of Excel to find this Z value is =NORMINV(0.99,0,1) for 2.326342.

The closest cumulative probability that can be found is 0.9901, in the row labeled 2.3 and column headed by .03, Z = 2.33, which is only an approximation for the more exact 2.326342 found in Excel. Using this more exact value the calculation with mean μ and standard deviation σ in the following formula would be

$Z = (X - \mu) / \sigma$
That is, Z = (x - 65)/15
Thus, x = 65 + 15(2.32634) = 99.9 minutes.

Alternatively, instead of standardizing with the Z distribution using Excel we can simply work directly with the normal distribution with a mean of 65 and standard deviation of 15 and enter "=NORMINV(0.99,65,15)". In general to obtain the x value for which alpha percent of a normal random variable's values are lower, the following "NORMINV" command may be used, where the first entry is α, the second is μ , and the third is σ.

Another Example: In the early 1980s, the Toro Company of Minneapolis, Minnesota, advertised that it would refund the purchase price of a snow blower if the following winter's snowfall was less than 21 percent of the local average. If the average snowfall is 45.25 inches, with a standard deviation of 12.2 inches, what is the likelihood that Toro will have to make refunds?

Solution: Within limits, snowfall is a continuous random variable that can be expected to vary symmetrically around its mean, with values closer to the mean occurring most often. Thus, it seems reasonable to assume that snowfall (x) is approximately normally distributed with a mean of 45.25 inches and standard deviation of 12.2 inches. Nine and one half inches is 21 percent of the mean snowfall of 45.25 inches and, with a standard deviation of 12.2 inches, the number of standard deviations between 45.25 inches and 9.5 inches is Z:

$Z = (x - \mu) / s = (9.50 - 45.25)/12.2 = -2.93$

Using Appendix B, Table 5, the textbook demonstrates the determination of $P(x \leq 9.50) = P(z \leq -2.93) = 0.17$, the probability of snowfall less than 9.5 inches. Using Excel, this normal probability is obtained with the "NORMDIST" command, where the first entry is x, the second is mean μ , the third is standard deviation s, and the fourth is CUMULATIVE (1). Entering

=NORMDIST(9.5,45.25,12.2,1), Gives $P(x \leq 9.50) = 0.001693$.

4. **Sampling Distribution and the Central Limit Theorem :** A bakery sells an average of 24 loaves of bread per day. Sales (x) are normally distributed with a standard deviation of 4.

If a random sample of size n = 1 (day) is selected, what is the probability this x value will exceed 28?

If a random sample of size n = 4 (days) is selected, what is the probability that xbar ≥ 28?

Why does the answer in part 1 differ from that in part 2?

Solutions:

1. The sampling distribution of the sample mean xbar is normal with a mean of 24 and a standard error of the mean of 4. Thus, using Excel, 0.15866 =1-NORMDIST(28,24,4,1).

2. The sampling distribution of the sample mean xbar is normal with a mean of 24 and a standard error of the mean of 2 using Excel, 0.02275 =1-NORMDIST(28,24,2,1).

5. **Regression Analysis:** The highway deaths per 100 million vehicle miles and highway speed limits for 10 countries, are given below:

(Death, Speed) = (3.0, 55), (3.3, 55), (3.4, 55), (3.5, 70), (4.1, 55), (4.3, 60), (4.7, 55), (4.9, 60), (5.1, 60), and (6.1, 75).

From this we can see that five countries with the same speed limit have very different positions on the safety list. For example, Britain ... with a speed limit of 70 is demonstrably safer than Japan, at 55. Can we argue that, speed has little to do with safety. Use regression analysis to answer this question.

Solution: Enter the ten paired y and x data into cells A2 to A11 and B2 to B11, with the "death" rate label in A1 and "speed" limits label in B1, the following steps produce the regression output.

Choose "Regression" from "Data Analysis" in the "Tools" menu. The Regression dialog box will appear.

Note: Use the mouse to move between the boxes and buttons. Click on the desired box or button. The large rectangular boxes require a range from the worksheet. A range may be typed in or selected by highlighting the cells with the mouse after clicking on the box. If the dialog box blocks the data, it can be moved on the screen by clicking on the title bar and dragging.

For the "Input Y Range," enter A1 to A11, and for the "Input X Range" enter B1 to B11.

Because the Y and X ranges include the "Death" and "Speed" labels in A1 and B1, select the "Labels" box with a click.

Click the "Output Range" button and type reference cell, which in this demonstration is A13.

To get the predicted values of Y (Death rates) and residuals select the "Residuals" box with a click.

Your screen display should show a Table, clicking "OK" will give the "SUMMARY OUTPUT," "ANOVA" AND RESIDUAL OUTPUT"

The first section of the EXCEL printout gives "SUMMARY OUTPUT." The "Multiple R" is the square root of the "R Square;" the computation and interpretation of which we have already discussed. The "Standard Error" of estimate (which will be discussed in the next chapter) is s = 0.86423, which is the square root of "Residual SS" = 5.97511 divided by its degrees of freedom, df = 8, as given in the "ANOVA" section. We will also discuss the adjusted R-square of 0.21325 in the following chapters.

Under the "ANOVA" section are the estimated regression coefficients and related statistics that will be discussed in detail in the next chapter. For now it is sufficient to recognize that the calculated coefficient values for the slope and y intercept are provided (b = 0.07556 and a = -0.29333). Next to these coefficient estimates is information on the variability in the distribution of the least-squares estimators from which these specific estimates were drawn: the column titled "Std. Error" contains the standard deviations (standard errors) of the intercept and slope distributions; the "t-ratio" and "p" columns give the calculated values of the t statistics and associated p-values. As shown in Chapter 13, the t statistic of 1.85458 and p-value of 0.10077, for example, indicates that the sample slope (0.07556) is sufficiently different from zero, at even the 0.10 two-tail Type I error level, to conclude that there is a significant relationship between deaths and speed limits in the population. This conclusion is contrary to assertion that "speed has little to do with safety."

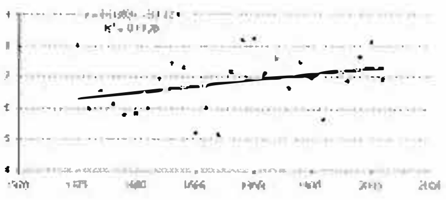
SUMMARY OUTPUT: Multiple R = 0.54833, R Square = 0.30067, Adjusted R Square = 0.21325, Standard Error = 0.86423, Observations = 10

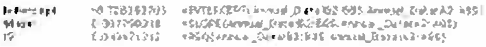
ANOVA	df	SS	MS	F	P-value
Regression	1	2.56889	2.56889	3.43945	0.10077
Residual	8	5.97511	0.74689		
Total	9	8.54400			

Coeffs.	Estimate	Std. Error	T Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.29333	2.45963	-0.11926	0.90801	-5.96526	5.37860
Speed	0.07556	0.04074	1.85458	0.10077	-0.01839	0.16950

Residual Output:

Predicted	Residuals
3.86222	-0.86222
3.86222	-0.56222
3.86222	-0.46222
4.99556	-1.49556
3.86222	0.23778
4.24000	0.06000
3.86222	0.83778
4.24000	0.66000
4.24000	0.86000
5.37333	0.72667

.Excel Regression Approaches		
Approach	**Description**	**Example** - *(Click to enlarge)*
Chart with Trend Line	The simplest way to get the regression formula for your data is to create a simple XY chart and to add the Trendline formula and r2 values from the Options dialogue. This simple technique only provides the equation and r^2 values as text on the chart. If you want to use the equation or conduct significance test, you need to use one of the	

	other techniques.	
Manual Calculations	Since Excel can reproduce just about any calculation that you want, you can add the necessary statistical formulas from a statistical text and produce any statistical parameter you want. While this technique gives you maximum flexibility, you may want to evaluate Excel's built-in statistical functions before you re-invent the wheel	$Y = a + bx$ $b = \frac{n\sum(xy) - \sum x \sum y}{n\sum(x^2) - (\sum x)^2}$ $a = \frac{\sum y - b\sum x}{n}$ $r = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{[n\sum(x^2) - (\sum x)^2][n\sum(y^2) - (\sum y)^2]}}$
Excel's SLOPE, INTERCEPT and RSQ Functions	Excel's built-in SLOPE, INTERCEPT and RSQ functions allow the user to calculate these values directly. The functions return the same results as adding the trendline to the chart technique. In this case, however, the statistical parameters are in a user specified cell so they can be used for further analysis.	
Excel's LINEST Function	The LINEST function returns ten statistical parameters for a simple linear regression: o Regression coefficients for a_1 and a_0 o Standard error values for coefficients o r^2 and Standard error of the Y estimate o F observed value o Degrees of freedom o Regression sum of squares o Residual sum of squares The LINEST function returns an array of values so that it must be entered as an array formula. Rather than use the entire array, the user can return individual LINEST parameters to a cell by using the Index function, as shown in this formula to return the F value: =INDEX(LINEST(y,x,1,1),r,c) *Where*: r = row of LINEST array (4 for F	

	value) c = column of LINEST array (1 for F value)	
Analysis Toolpak Regression Procedure	Microsoft's Analysis Toolpak add-in includes a number of advanced analysis data analysis tools, including Regression. The Regression procedure provides regression statistics, ANOVA, regression coefficients, their standard errors, t stat's, p values and upper and lower confidence values. The Regression procedure results are comparable to output from statistical packages.	
Combination of Excel Statistical Functions	In addition to the 10 regression measures provided by LINEST, Excel provides T and F value test functions so that users can reproduce the Analysis Toolpak Regression procedure results in a series of cell formulas that use a combination of LINEST, T and F test functions. The picture to the right shows a screen print of an worksheet that allows the user to interactively select the regression period, then, prepares the regression analysis using a combination of Excel LINEST, T and F tests functions as well as conditional formatting to evaluate the regression.