Statistical Analysis Tools

Statistical Analysis Tools

Introduction

There are numerous statistical functions provided by Excel. Excel also provides a set of special analysis tools called the Analysis ToolPak. These tools include statistical analyses.

Note: Statistical packages, such as SPSS for Windows and Stata are more suitable for serious data analysis, particularly when analysing large data sets.

Analysis ToolPak

The Analysis ToolPak includes the following tools:

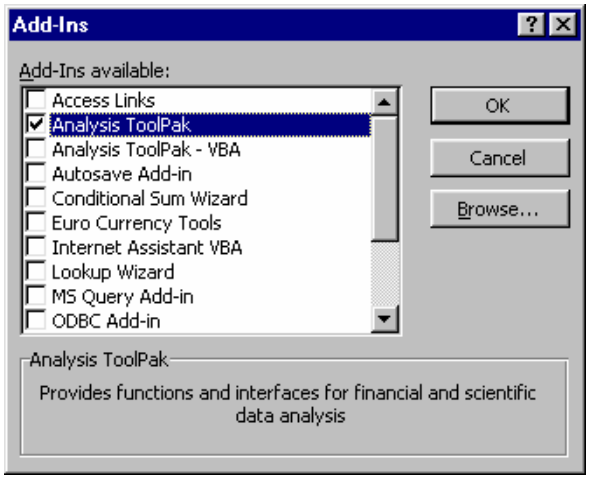
Type	Tool name
Statistical	Anova: single factor (one way)
	Anova: two factor with replication
	Anova: two factor without replication
	Covariance
	Correlation
	Descriptive Statistics
	Exponential Smoothing
	F-Test: two-sample for variances
	Histogram
	Moving average
	Random number generation
	Rank and percentile
	Regression
	t-test: paired two-sample for means
	t-test: two-sample assuming equal variances
	t-test: two-sample assuming unequal variances
	z-test: two-sample for means
Engineering	Fourier analysis
	Sampling

Where to find Data Analysis tools

To use an analysis tool, choose **Tools | Data Analysis**. In the Analysis Tools box, select the name of the Analysis Tool you want to use. Then specify the input and output ranges and any other options you want.

If the Data Analysis command does not appear on the Tools menu, click **Tools | Add-Ins.** Check the ***Analysis ToolPak*** box and click OK.

You might be requested to insert your Microsoft Office CD-ROM at this point.



Using a Data Analysis tool

The Data Analysis tools assume a certain amount of knowledge. Before you use one, you must enter and organise the data you want to analyse into columns or rows on your worksheet. This is your input range. Using Named ranges of cells can make this easier. You can also include a text label in the first cell of a row or column to identify your variables.

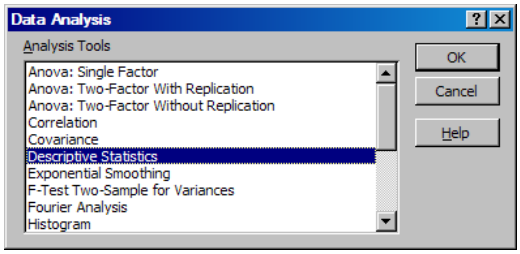
You enter cell ranges in the Input Range and Output Range boxes by typing a cell reference in the box or by selecting the contents of each box and then selecting the cell range on the worksheet. You can also enter references to other sheets or to other workbooks in the Input Range and Output Range boxes.

When you use an analysis tool to analyse data in an input range, Excel, creates an output table of the results. The contents of the output table depend on the analysis tool you are using. If you included labels in the input range, Excel uses them to label data in the output table. If you did not include labels in the input range, Excel automatically generates data labels for the results in the output table.

You can choose to save the output table on the same sheet as the input range, on a separate sheet in the same workbook (this is created in on a new worksheet inserted into your workbook), or in a new workbook. If you attempt to save your output table in location where data already exists, Excel warns you and gives you the opportunity to specify a new location.

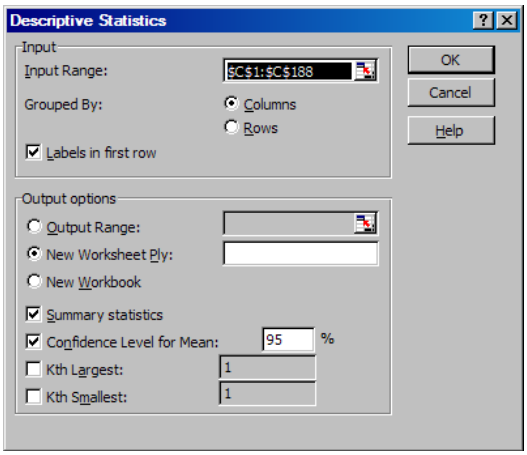
Descriptive Statistics

The Descriptive Statistics tool creates summary statistics for one or more columns of continuous data.



The Input Range for a single column was selected including the first row where the cell contains a label.
The Labels in first row box must be ticked therefore.
Output can be on an existing worksheet as specified by Output Range, a new worksheet, or a new workbook.

To obtain output the Summary Statistics and Confidence Level for Mean boxes have been checked.



Output for a single column with Descriptive Statistics

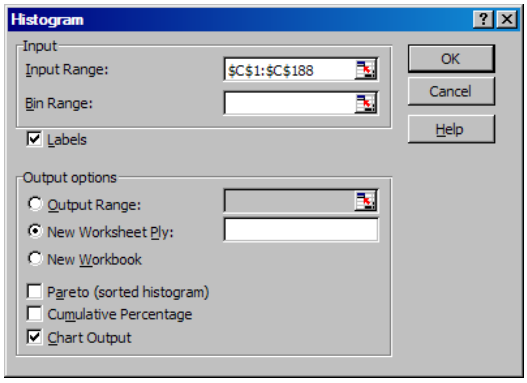
	A	B
1	*Birthwt*	
2		
3	Mean	3.466527
4	Standard Error	0.042152
5	Median	3.544
6	Mode	3.629
7	Standard Deviation	0.571778
8	Sample Variance	0.32693
9	Kurtosis	0.244394
10	Skewness	-0.29414
11	Range	3.475
12	Minimum	1.6
13	Maximum	5.075
14	Sum	637.841
15	Count	184
16	Confidence Level(95.0%)	0.083166

Note *The Confidence Level is is the value which must be subtracted and added to the mean to obtain the Confidence Interval.*

Note *You cannot use the Undo command on the Edit menu to reverse the creation of an output table if you choose to overwrite existing data.*
Updating data does not change the calculated values. The analysis tool must be rerun.

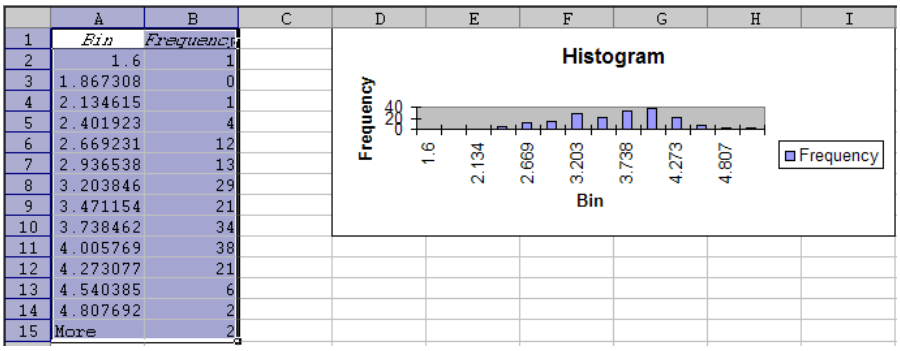
Histograms

The Histogram tool produces a bar chart and not a true histogram



The Input Range identifies the data to be plotted. In this instance a label is in the first cell C1 and the data in cells C2 to C188.

Output will be placed in a new worksheet.
The Chart Output box must be ticked to obtain the histogram



	A	B
1	*Bin*	*Frequency*
2	1.6	1
3	1.867308	0
4	2.134615	1
5	2.401923	4
6	2.669231	12
7	2.936538	13
8	3.203846	29
9	3.471154	21
10	3.738462	34
11	4.005769	38
12	4.273077	21
13	4.540385	6
14	4.807692	2
15	More	2

The Bin Range selected by Excel does not provide convenient breakpoint.
For this data a range of 1.5 to 5.0 in steps of 0.25 might be more appropriate.
Set up a range of values using Auto Fill.
Rerun the Histogram tool and enter the Bin Range: e.g., L3:L13

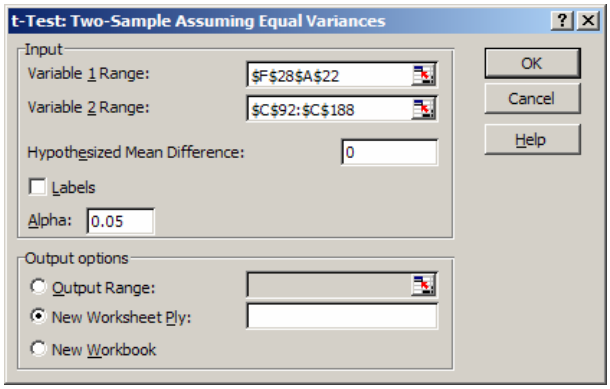
1.50
1.75
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
4.25
4.50
4.75
5.00

t-Test Two-sample assuming equal variances
We wish to test whether there is a difference between the means values of two unrelated groups. The assumption is that the data is taken from a population of values that follow a Normal distribution. Our example is the birth weights of two groups of babies.

The data must organised into two ranges within the spreadsheet. For example data from one column (col C in this example) was sorted using values in another column (col B) that defined the two groups. Alternatively the data for two groups might be found in two different columns.

Select **Tools | Data Analysis**
Select **t-Test Two-Sample Assuming Equal Variances**

Variable 1 Range and ***Variable 2 Range*** define the data for the two groups we wish to compare.
The ***Hypothesized Mean Difference:*** is entered as 0. This is the Null hypothesis of no difference between the two groups.



Alpha is the significance level (0.05 = 95%; (0.01 = 99%), entered here as 0.05.
Output options control where the output is placed.

The output has to be interpreted carefully. (It is not as easy to read as SPSS output).

	A	B	C
2			
3		*Variable*	*Variable*
4	Mean	3.397943	3.529396
5	Variance	0.347907	0.302809
6	Observations	88	96
7	Pooled Variance	0.324367	
8	Hypothesized Mean Difference	0	
9	df	182	
10	t Stat	-1.56394	
11	P(T<=t) one-tail	0.059785	
12	t Critical one-tail	1.65327	
13	P(T<=t) two-tail	0.11957	
14	t Critical two-tail	1.973085	

df are the degrees of freedom.
t Stat is the calculated t value.
Normally we use only two-tail tests, one-tail information can be ignored.
P(T<=t) two-tail is the P value; P= 0.119
t Critical two-tail is the value that t Stat must exceed for significance at the 95% level. This is calculated from the t distribution.

Note *Excel is inaccurate, convention statistical packages give t-Stat as -1.559 and P=0.121. So Beware the results are approximate!*