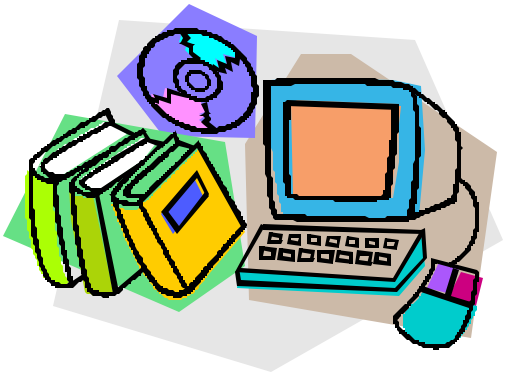


Getting Started with MegaStat ®

J. B. Orris

Butler University

Copyright 2002 by J. B. Orris

Table of Contents

1. Basic Procedures *2*

Buttons **3**

Data Selection **3**

Data Labels **5**

Output **5**

Repeat Last Option **5**

Generate Random Numbers **6**

Utilities **6**

Insert descriptive information 6

ChartDataSheet utilities 6

Start new output sheet 7

Delete output sheet 7

Deactivate MegaStat 7

Uninstall MegaStat 7

Help/Information **8**

Help System 8

About MegaStat 8

2. Tutorial Examples *9*

Example 1: Frequency distribution **9**

Example 2: Normal Distribution **13**

Example 3: Entering Proportions **14**

3. Reference *15*

Descriptive Statistics **15**

Frequency Distributions **17**

Quantitative 17

Qualitative 18

Counting Rules 20

Discrete Probability Distributions 21

Continuous Probability Distributions 21

Confidence Intervals / Sample Size **22**

Confidence interval – mean 22

Confidence interval – p 22

Sample size –mean, mean with specified α and β, and p 23
Hypothesis Tests 24
Mean vs. Hypothesized Values 24
Compare Two Independent Groups 25
Paired Observations 25
Proportion vs. Hypothesized Value 26
Compare Two Independent Proportions 26
Chi-square Variance Test 27
Analysis of Variance 28
One-Factor ANOVA 28
Randomized Blocks ANOVA 29
Two Factor ANOVA 30
Correlation / Regression 31
Scatterplot 31
Correlation Matrix 32
Regression Analysis 33
Predictor values from spreadsheet cells 33
Type in predictor values 34
Select Options: 34
Select Residuals options 35
Time Series / Forecasting 36
Trendline Curve Fit 36
Deseasonalization 37
Moving Average 38
Exponential Smoothing 38
Simple Exponential Smoothing 38
Two-factor Exponential Smoothing 39
Chi-Square / Crosstab 40
Contingency Table 40
Crosstabulation 41
Goodness of Fit Test 42
Nonparametric Tests 43
Sign Test 43
Runs Test for Random Sequence 43
Wilcoxon – Mann/Whitney Test 44
Wilcoxon – Signed Rank Test 44
Kruskal – Wallis Test 45
Friedman Test 46

Kendall Coefficient of Concordance 46
Spearman Coefficient of Rank Correlation 47
Fisher Exact Test 47
Quality Control Process Charts 48
Control chart for variables (Xbar and R chart) 48
Control chart for proportion nonconforming (p chart) 48
Control chart for number of defects per sample (c chart) 48
Generate Random Numbers 49

Getting Started with MegaStat®

J. B. Orris, Ph.D.

Butler University

MegaStat[1][2] is an Excel add-in that performs statistical analyses within an Excel workbook. After it is installed it appears on the Excel menu and works like any other Excel option. The purpose of this Getting Started document is to introduce you to how MegaStat works. The first chapter will describe the general operating procedures and conventions that are common throughout MegaStat. The second chapter will work through a few tutorials. The Reference section shows the dialog boxes for all of the options and notes briefly what data/input is expected and any unique aspects of each option.

While MegaStat is an excellent tool for learning statistics, this document focuses on using MegaStat and is not intended to teach statistics. Indeed, it assumes that you know what the various procedures do and are familiar with the terminology. It also assumes you have a basic working knowledge of Excel.

[1] MegaStat is copyrighted and is a registered trademark of J. B. Orris

[2] This document was written for version 9.0 of MegaStat; however, most of it will be relevant for other versions.

1. Basic Procedures

When you click on MegaStat the main Excel menu appears (Figure 1). Most of the menu options display sub-menus. If a menu item is followed by an ellipsis (…) clicking it will display the dialog box for that option.

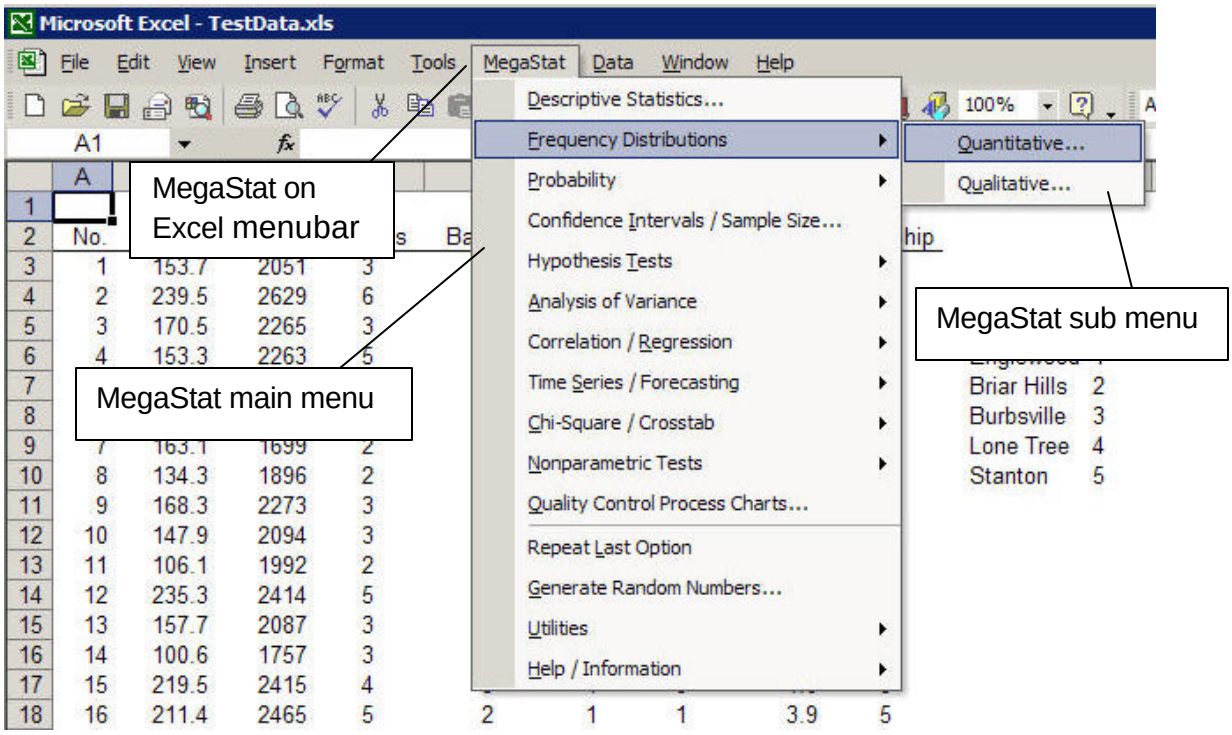


Figure 1. MegaStat menu

A dialog box allows you to specify the data to be used and other inputs and options. Figure 2 shows a typical dialog box. After you have selected the data and options you click OK, the dialog box disappears and MegaStat performs the analysis.

Before we look at specific dialog boxes let's take a minute to look at some issues that are common to all of the options. MegaStat use is intuitive and very much like other Excel operations; however, there are some features unique to MegaStat and some ways to make using it more efficient so it will be worth your time to look at the following material.

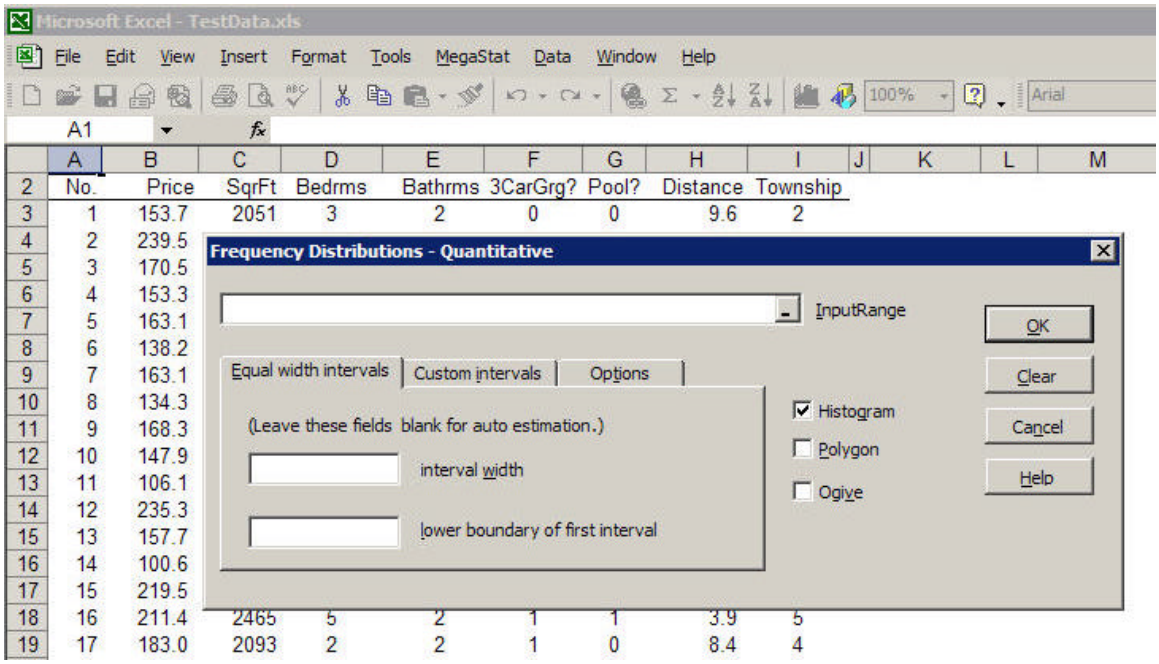


Figure 2. MegaStat dialog box

Buttons

Every dialog box has the four buttons shown on Figure 2.

OK This button could also be labeled "Calculate", "Go", "Execute" or "Do it". It tells MegaStat that you are done specifying inputs and you are turning control over to it to do its thing. First your input values are validated and then the dialog box disappears and the output worksheet is displayed. When the dialog box disappears, it is still in memory and will contain the same inputs if recalled later.

Clear This button removes all input values and resets any default options on the form.

Cancel This button could be labeled "Never mind". It simply hides the dialog box. The dialog box is not cleared or removed from memory. Userforms do not take much memory and there is no problem with having several of them in memory. However, if you really want to unload the form, click the "X" in the upper right corner of the form.

Help As you have guessed, this button displays context sensitive help for the active userform. If you want to see the full Help System, use the Help selection on the main menu.

Data Selection

Most MegaStat dialog boxes have fields where you select input ranges that contain the data to be used. Input ranges can be selected four ways:

1. **Pointing and dragging with the mouse (the most common method).**

Since the dialog box "pops-up" on the screen it may block some of your data. You can move dialog boxes around on the screen by placing the mouse pointer over the title bar (colored area at the top), clicking and holding the left mouse button while dragging the dialog box to a new location. You can even drag it partially off the screen.

You will also notice that when you start selecting data by dragging the mouse pointer, the dialog box will collapse to a smaller size to help you see the underlying data. It will automatically return to full size when you release the mouse button. You can also collapse and uncollapse the dialog box manually by clicking the Collapse button at the right end of the field. Clicking the button again will uncollapse the form. (Never use the 'X' button to uncollapse a form.)

2. Using MegaStat's AutoExpand feature

Pointing and dragging to select data can be tedious if you have a lot of data. When you drag the mouse down it is easy to 'over-shoot' the selection and then you have to drag the mouse back until you get the area correctly selected.

AutoExpand allows rapid data selection without having to drag through the entire column of data. Here is how it works:

- Make sure the input box has the focus. (Click in it or tab to it.) An input box has the focus when the insertion pointer is blinking in it.
- Select one row of data by clicking in one cell of the column you want. If more than one column is being selected, drag the mouse across the columns.
- Right-click over the input field ***or*** left-click the label next to input box. The data range will expand to include all of the rows in the region where you selected one row.

With a little practice you will find this is a very efficient way to select data. The only time you cannot use it is when you want to use a partial column of data. You should also be aware that the AutoExpand stops when it finds a blank cell; thus any summations or other calculations at the bottom of a column would be selected. It is good practice to leave a blank cell at the bottom of each column before inserting formulas.

Note: When using the above methods of data selection you may select variables in an alternate sequence by holding the CTRL key while making multiple selections.

3. Typing the name of a named range.

If you have previously identified a range of cells using Excel's name box, you may use that name to specify a data range in a MegaStat userform. This method can be very useful if you are using the same data for several different statistical procedures.

4. Typing a range address

You may type in any valid Excel range address, e.g. B5:B43. This is the least efficient way to specify data ranges but it certainly works.

Data Labels

For most procedures the first cell in each input range can be a label**. If the first cell in a range is text it is considered a label**; **if the first cell is a numeric value it is considered data**. If you want to use numbers as variable labels you must enter the numbers as text by preceding them with a single quote mark *e.g.* '2. Even though Excel stores times and dates as numbers, MegaStat will recognize them as labels if they are formatted as time/date values.

If data labels are not part of the input range, the program automatically uses the cell immediately above the data range as a label if it contains a text value.

If an option can consider the entire first row (or column) of an input range as labels, any numeric value in the row will cause the entire row to be treated as data.

If the program detects sequential integers (1,2,3…) in a location where you might want labels it will display a warning message otherwise the rule is: **text cells are labels, numeric cells are data**.

Output

When you click OK on a MegaStat dialog box it performs some statistical analysis and needs a place to put its output. It looks for a worksheet named Output. If it finds one it goes to the end of it and appends its output; if it doesn't find an Output worksheet it creates one. MegaStat will never make any changes to the user's worksheets, it only sends output to its Output sheet.

MegaStat makes a good attempt at formatting the output but it is important to remember that the Output sheet is just a standard Excel worksheet and can be modified in any way by the user. You can adjust column widths and change any formatting that you think needs improvement. You can insert, delete and modify cells. You can copy all or part of the output to another worksheet or to another application such as a word processor.

MegaStat charts get their values from cells on the Output sheet (or one of your worksheets in the case of the Scatterplot). You can click a chart and select 'Source Data' to see what values are being displayed.

When you click a chart the MegaStat menu item will disappear from the main menubar since the Chart menu becomes active. Click outside the chart to bring back the main menu that contains the MegaStat menu item.

When the program generates output it adjusts column widths for the current output. If you have previous output from a different option already in the Output sheet, the column widths for the previous output may get messed up. You can attempt to fix this by manually adjusting the columns widths or by always starting a new output sheet.

The Utilities menu has options for deleting the Output sheet or making a copy of it and starting a new one.

Repeat Last Option

Once you have performed a MegaStat option, this menu selection will allow you to re-display the last dialog box without having to go through the menu selections. This can be handy if you need to make a change or when you need to repeat the same operation with the different data sets.

Generate Random Numbers

This option allows you to create random numbers. It is described on page 49.

Utilities

The Utilities menu contains some items that perform useful functions.

Insert descriptive information

This option is used for identifying output. It will insert rows with labels for Description, Name, Data source, Time and Version. You then use the adjacent cells to type in the appropriate information. Figure 7 in the next chapter shows an Output sheet after clicking this option.

This option is not limited to MegaStat output sheets – it can be used on any Excel worksheet.

ChartDataSheet utilities

In order to display a graphical output (*e.g.,* a chart) Excel must reference values in a worksheet. If the values are not available as a part of the output sheet, MegaStat stores them in a hidden worksheet called "ChartDataSheet_". If MegaStat needs a ChartDataSheet it creates one unless one already exists, in which case it appends its values to the end.

The following options use ChartDataSheets:

Descripitive Statistics – BoxPlot
Descriptive Statistics – DotPlot
Regression Analysis – Plot residuals by X values
Regression Analysis – Normal Probability Plot
Quality Control Process Charts

In general you do not need to be concerned with ChartDataSheets; however, the following utilities exist for advanced users.

View

ChartDataSheets exist only to provide values to charts and thus the output is not labeled. However, at the top of each output section is a label telling what type of chart it is used for and a time/date stamp. Under the corresponding chart there is also a time/date stamp so you can associate the data with a chart.

If you change or delete any of the values on ChartDataSheet the corresponding chart will be changed.

Hide

This will hide the ChartDataSheet after viewing it. You can also use Excel's Format | Sheet command to view/hide ChartDataSheets.

Delete

You would use this option if you wanted to delete a ChartDataSheet that no longer has any associated charts. If you delete a ChartDataSheet that has

an existing chart, the chart will still exist but will not have any values plotted. There is no UnDo so make sure before you click OK.

Start new output sheet

If there is an existing Output sheet it will be renamed Output(2) so that your next output will be on a fresh Output sheet. You can rename Output(2) to whatever you wish by double-clicking the name tab.

Delete output sheet

This option deletes the current Output sheet. It will present a warning message because there is no way to recover a sheet once it is deleted.

Deactivate MegaStat

This option is used to remove the 'MegaStat' item from the main menubar. It does not delete any files or uninstall MegaStat. To restore the 'MegaStat' menu item click the tools Tools | Add-Ins item on the Excel's main menubar and then check the MegaStat option that you will see in the list of available add-ins.

Uninstall MegaStat

This menu item does not actually uninstall MegaStat. It displays a dialog box prompting you on how to start the uninstallation process described below.

Uninstalling is the process of removing the installed MegaStat files from your system. It does not remove any data files nor does it remove the file you used to install MegaStat. You may delete the installation file (MegaStat_Setup.exe) if it is still on your system.

Uninstall steps:

1. Deactivate MegaStat using the Utilities menu

2. Exit Excel

3. Click: Start | Settings | Control Panel | Add/Remove Programs

4. Find MegaStat in the list of programs, click it and then click the Add/Remove button

[If you installed MegaStat manually without using a setup program you uninstall it by finding and deleting: Megastat.xla and Megastat.hlp.]

Help/Information

Help System

This option displays the full MegaStat help program.

The 'How it works (General Operating Procedures)" section contains all of the information in this tutorial. You can click specific topics or search for a particular item by clicking on Index.

Figure 3 MegaStat Help System

About MegaStat

This option displays current version information. There are also links for the MegaStat website and e-mail for tech support. These links will only work if you have an active Internet connection and your system is setup to properly respond to Internet and e-mail links.

The form also contains a **System Information** button, which causes the form to expand and display some technical information regarding system parameters and file locations. If you click the Insert button, the information will be placed in the current Output sheet. If you click Hide System Information, the form will go back to its original size.

2. Tutorial Examples

Although MegaStat performs many different statistical options the various dialog boxes all work the same way and have standard Excel objects (input boxes, buttons, checkboxes, etc.). Thus it is not necessary to show graphical examples of every MegaStat option. This chapter will work through a few detailed examples and will point out a few things that are unique to MegaStat. The next chapter will provide a reference source for the various options.

Example 1: Frequency distribution

The first tutorial example will perform a quantitative frequency distribution on the Price variable of the Sheet1 worksheet of the Testdata.xls workbook. If you want to work through this example, start Excel and open Testdata.xls and click on the Sheet1 worksheet tab.

a. Open the Quantitative Frequency Distributions dialog box by clicking: MegaStat | Frequency Distributions | Quantitative. (As shown on pages 2 and 3)

b. Click the Help button. Review the Help screen and then exit Help.

Since we want to do the frequency distribution on the Price variable we need to select all the data in column B. This example will illustrate the use of the AutoExpand feature although you could get the same result with any of the other methods of data selection.

c. Make sure the insertion cursor is blinking in the input box. If necessary, click in the box with the left mouse button.

d. Click anywhere in column B, e.g. cell B6

e. Place the mouse pointer over the input box label and click the left mouse button.

Figure 4 shows the dialog box at this point.

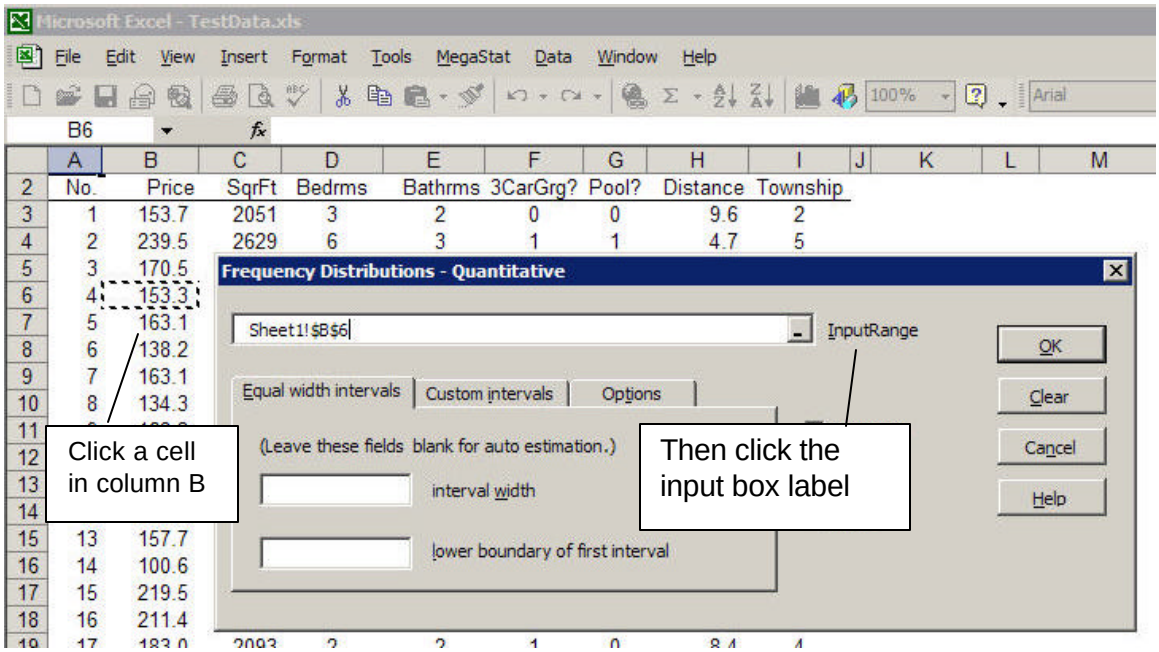


Figure 4. Illustration of AutoExpand

After clicking the label, the dialog box appears as follows with all of the data in column B being selected. The selection expands up/down until it encounters a blank cell or the top/bottom of the column.

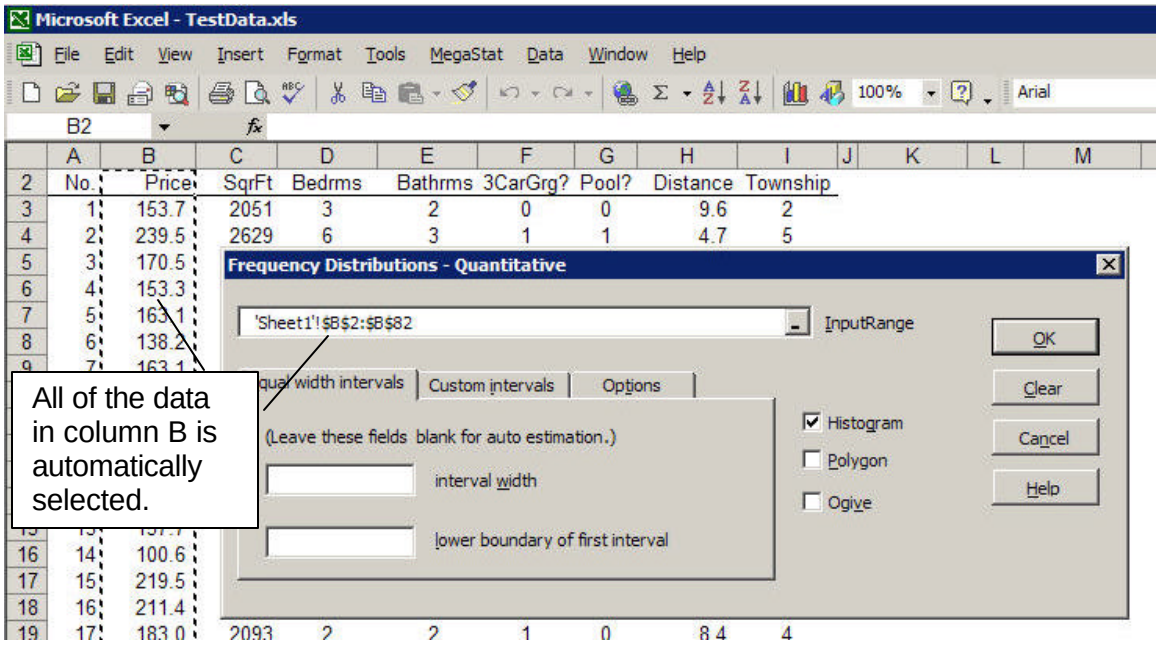


Figure 5. All of the data in column B is selected.

f. Type 50 in the Interval width box and click each of the checkboxes. Click OK.

Determining the proper number to enter is something you will learn in your statistics course. If you leave either input box empty the MegaStat will calculate an appropriate value. Prior to clicking OK the dialog box will look like Figure 6.

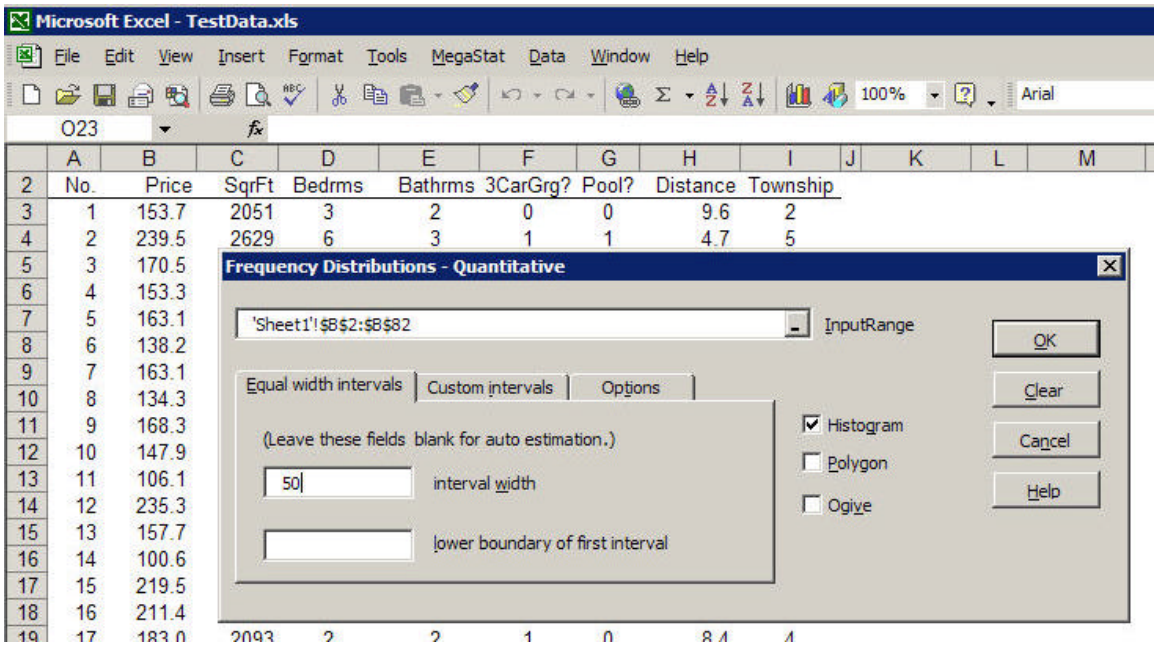


Figure 6. Completed dialog box waiting for OK click.

After you click OK the dialog box disappears and MegaStat does the requested calculations. A new Output (shown in Figure 7) is created and displayed. If there were already an Output sheet MegaStat would have appended the new output to the end.

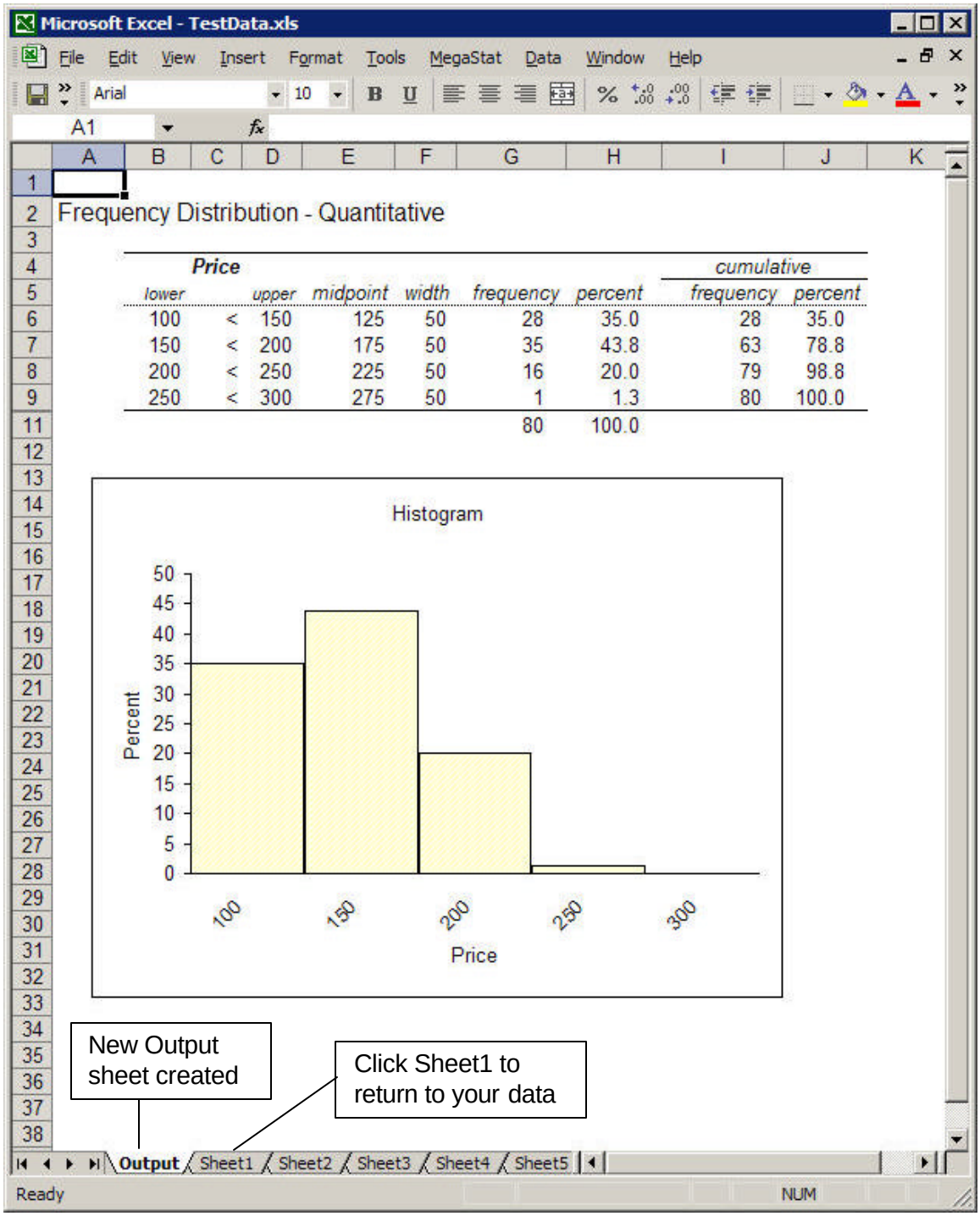


Figure 7. MegaStat Output sheet

The amount displayed on your screen will be different depending on the size of your screen. You may need to use the vertical and/or horizontal scrollbars to view the entire output.

The Output sheet is just an Excel worksheet and it can be formatted and manipulated just like any other worksheet. You may change the font style / size / color; increase / decrease decimal places; add annotations, etc.. However, you do need to be careful if you delete part of the output. For example, if you delete the midpoint cells, the Frequency Polygon cannot be plotted.

Also note that the charts are Excel charts and may be edited also to change colors, axis scaling and anything else. One thing you should routinely do is change the generic title (e.g. 'Histogram') to something more meaningful and relevant to the data.

Remember that when you are editing charts the Excel Chart menu takes over the Excel menubar and the MegaStat menu item is not visible. Click outside the chart to bring back the standard menubar.

Insert descriptive information

If you are submitting MegaStat output as part of a report or homework project you can have MegaStat insert descriptive cells.

g. Click in cell A2 and then click MegaStat | Utilities | Insert descriptive information.

This will insert rows with descriptive labels (as shown in Figure 8) and you can type in the appropriate information.

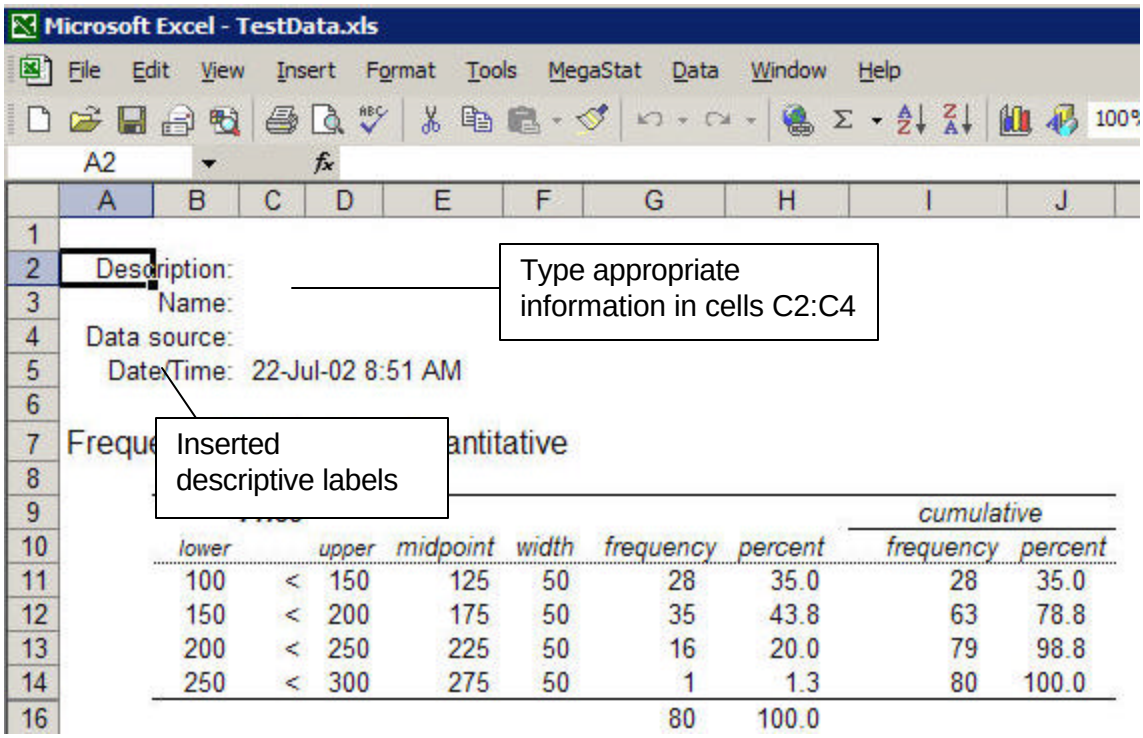


Figure 8. Inserted descriptive information.

Example 2: Normal Distribution

Most MegaStat options work with data; however, some options use other types of input. For example, when using probability distributions you only need to specify the appropriate input values. The tutorial example will use the normal distribution.

a. Open the Normal Distribution dialog box by clicking: MegaStat | Probability | Continuous Probability Distributions.

The Normal Distribution should be active, if not, click it.

b. Type 1.73 in the input box labeled 'z' and then click the Preview button.

The dialog box will now look like Figure 9.

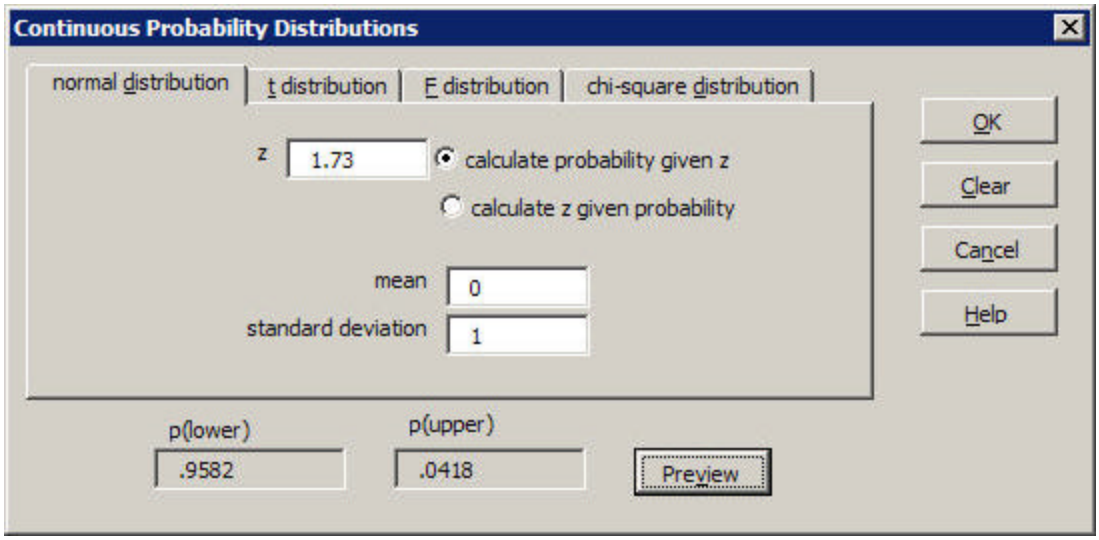


Figure 9. Normal distribution userform.

The Preview button is on several forms and is used where you might just want a quick look at an answer without the full output. You may select either of the preview values and copy them using CTRL-C and then paste them in any application. Of course, you may also click OK to get the full output on an Output sheet.

Example 3: Entering Proportions

Several MegaStat dialog boxes require that you enter proportions. This particular example calculates the confidence interval if the sample proportion is 11/38. You could enter the decimal value of 0.28947368421; however, this would be time consuming and error prone. It would be better to just enter the fraction as shown in Figure 10.

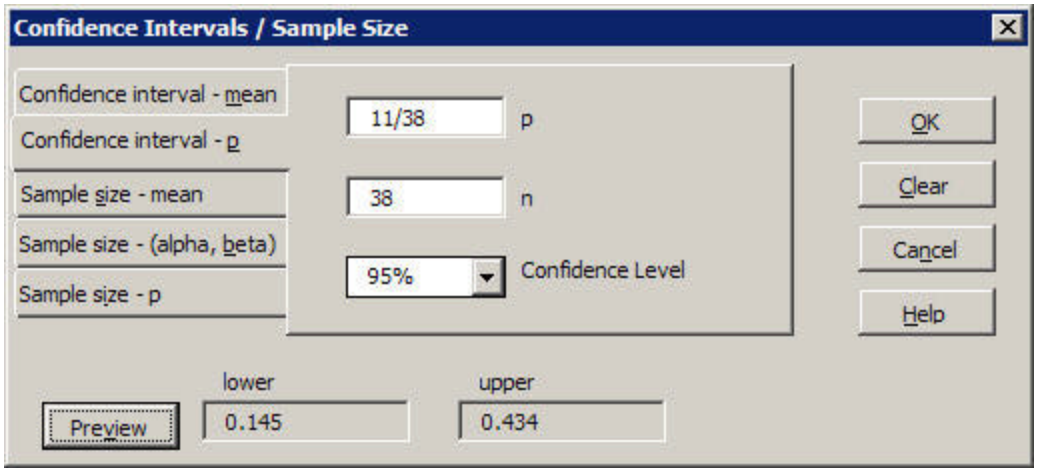


Figure 10. Proportion entered as a fraction.

An even better method is to just enter the numerator portion of the proportion and the program will divide the n value entered in the n box. When you enter a value in the p box that is greater than or equal to 1, the label automatically changes to x and the program knows that is supposed to divide by n (see Figure 11).

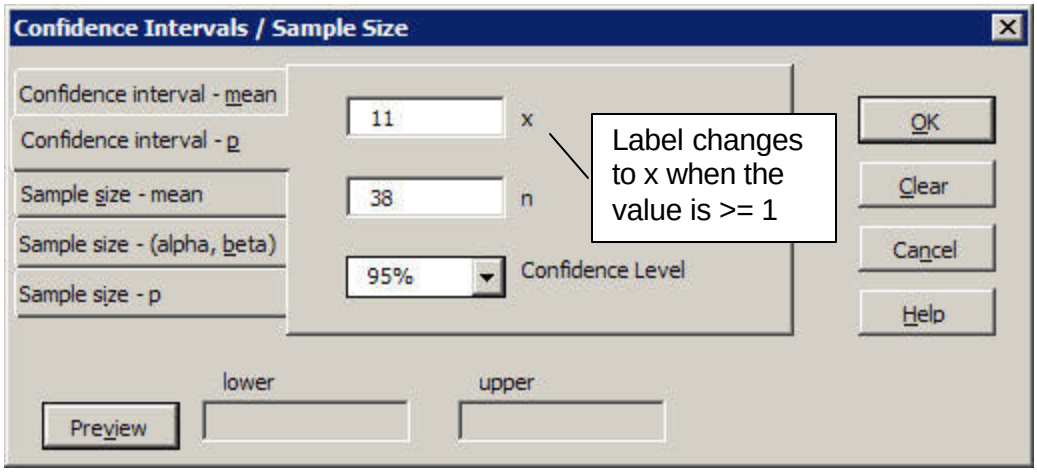


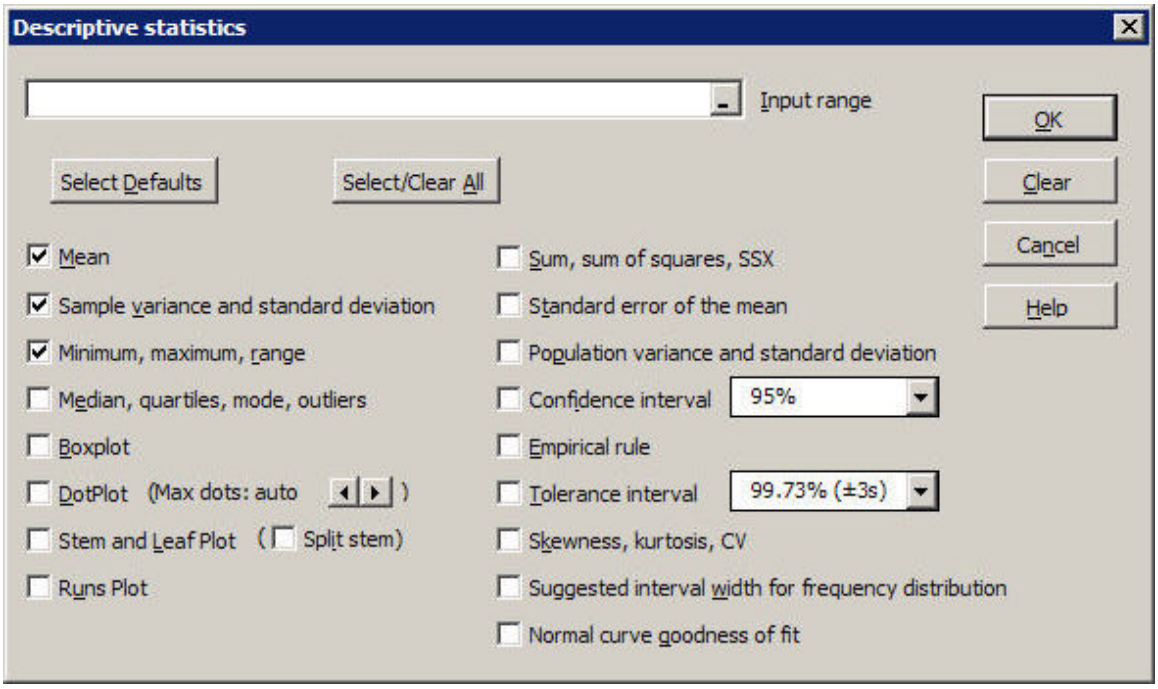
Figure 11. Proportion entered by typing only numerator.

This method works whenever MegaStat requires a proportion.

3. Reference

This section lists MegaStat options and briefly discusses any issues relevant to the option. Each dialog box is displayed; however, only features and procedures that are not self-apparent are discussed.

Descriptive Statistics



The Select Defaults button selects the options shown. Select/Clear All toggles 'Select All' and 'Clear All' options.

Note 1: The option that calculates **outliers** defines outliers and extremes as follows:

Let: Q1 = 25th percentile
Q3 = 75th percentile
H = Q3 – Q1 (the interquartile range)

An outlier is defined as any value less than Q1 - 1.5*H or greater than Q3 + 1.5*H

An extreme is defined as any value less than Q1 – 3.0*H or greater than Q3 + 3.0*H

Note 2: Boxplot fences

If the data contains outliers or extremes (see Note 1) they are plotted individually on the boxplots and the fences are displayed as dashed vertical lines. Fences are the values that define outliers and extremes.

Note 3: Boxplot and DotPlot charts

The data values needed to display the chart(s) are stored in ChartDataSheets as discussed in the Utilities section. If you paste the chart into another application

you might want to do a Paste Special | Picture that does not require a link to these values.

Note 4: DotPlot option

You can select the number of stacked dots; however, this is done by rounding the data so you may not get exactly the number of dots you specify.

Note 5: Normal curve goodness of fit

This option uses the mean and standard deviation of the data to determine the intervals boundaries that would divide a normal distribution into equal-probability intervals. The number of intervals is determined by Sturge's rule: $1 + \log_2(n)$. The row header for the observed values shows the z-value corresponding the upper end of each interval.

The program then calculates the frequency distribution. The expected frequencies are n/(number of intervals). If the data are normally distributed each interval should have the same number of observations.

Interpretation: A larger p-value indicates that the observed distribution closely matches a normal distribution. As the p-value gets smaller the fit is not as good. If the p-value is under .05 you would reject the hypothesis that the data could have come from a normal distribution.

Note: The sample size for the first variable is used to determine the number of intervals for all of the variables. If some of the other variables contain substantial amounts of missing data you should run the test on them individually.

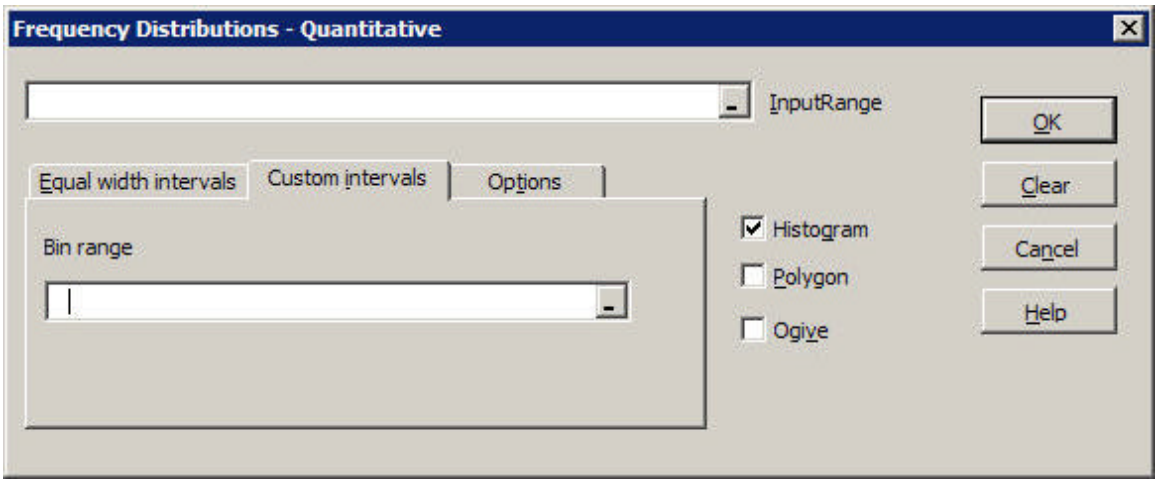
Frequency Distributions

Quantitative

The basic operation of this option is described in the Tutorial Example section above.

If the interval width or lower boundary boxes are left empty, MegaStat will attempt to calculate appropriate values: To estimate the interval width it first calculates the range of the data excluding any outliers or extremes. It then determines the interval width by the method discussed in David P. Doane "Aesthetic Frequency Classifications", The American Statistician, November 1976, Vol 30, No. 4.

If your data includes outliers the resulting frequency distribution will probably have some empty intervals in order to reach the outliers. You might exclude some data or consider custom intervals. The dialog box for frequency distributions with equal width intervals is shown and discussed in the Tutorial Example section above. However, when you have a wide range of data, perhaps with some outliers you would click the Custom Intervals tab and the dialog box would appear as follows:



Use this option to specify unequal width or open-ended intervals. Select the worksheet cells (called the 'bin range') containing the interval boundaries. Each cell will be the lower boundary of an interval except for the last cell, which will be the upper boundary of the last interval. The values must be selected in a single column.

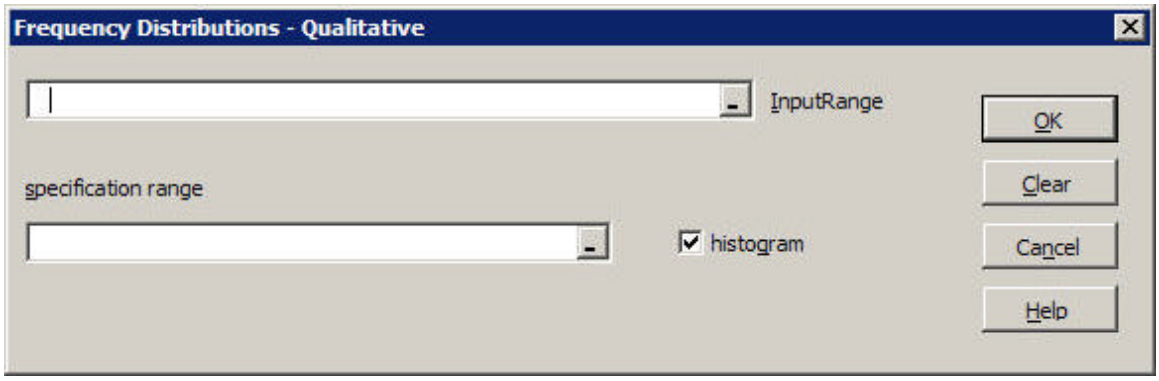
Example: The following bin range

0
100
200
1000
4000
10000

Would create the following intervals:

lower	upper
0	99.99
100	199.99
200	999.99
1000	3999.99
4000	10000.00

Qualitative



A qualitative frequency distribution is used to count the number of occurrences of specified data values. A **specification range** is used to specify what values are to be counted.

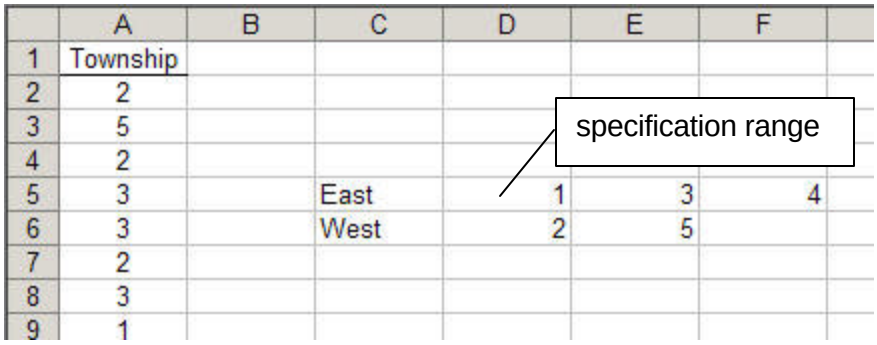
Suppose you had a variable representing five townships coded 1-5. If you wanted to include the name of the townships on the output you would select cells C6:D10 as the specification range as shown below. (Note: the cell borders shown are not required as a part of the specification range.)

	A	B	C	D	E	F	G
1	Twnship						
2	5						
3	4						
4	3						
5	2						
6	1		Englewood	1			
7	1		Briar Hills	2			
8	3		Burbsville	3			
9	2		Stanton	4			
10	1		Lone Tree	5			
11	4						
12	3						

specification range

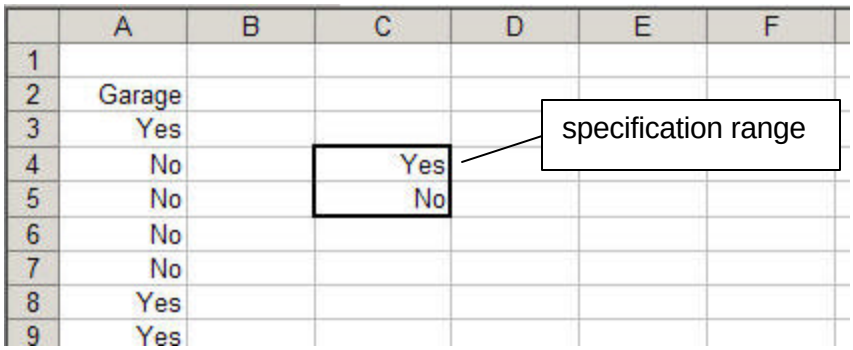
Suppose townships 1, 3 and 4 were in the eastern part of the city and townships 2 and 5 were in the western part. If you wanted to count the number of homes in the East and West region you would set the specification range as shown below. Values listed on the same row are counted in the same category. If no label is specified, the left hand data value would be used as the label.

Any data values not listed in the specification range are ignored.



	A	B	C	D	E	F
1	Township					
2	2					
3	5					
4	2					
5	3		East	1	3	4
6	3		West	2	5	
7	2					
8	3					
9	1					

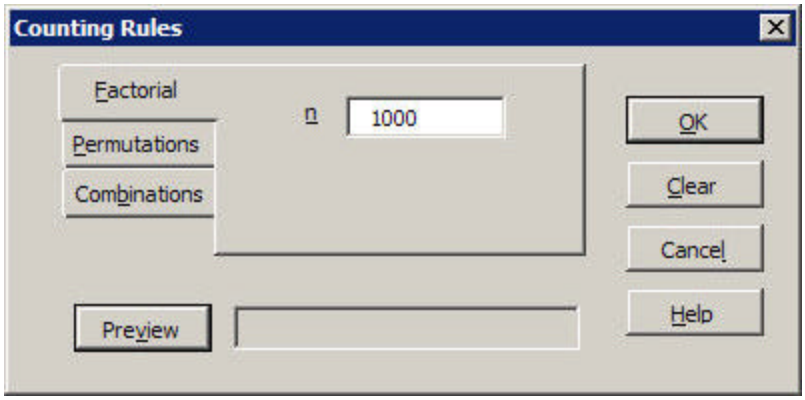
MegaStat also allows qualitative data in the data range. For example, if the Garage variable was coded Yes and No the specification range would be as follows:



	A	B	C	D	E	F
1						
2	Garage					
3	Yes					
4	No		Yes			
5	No		No			
6	No					
7	No					
8	Yes					
9	Yes					

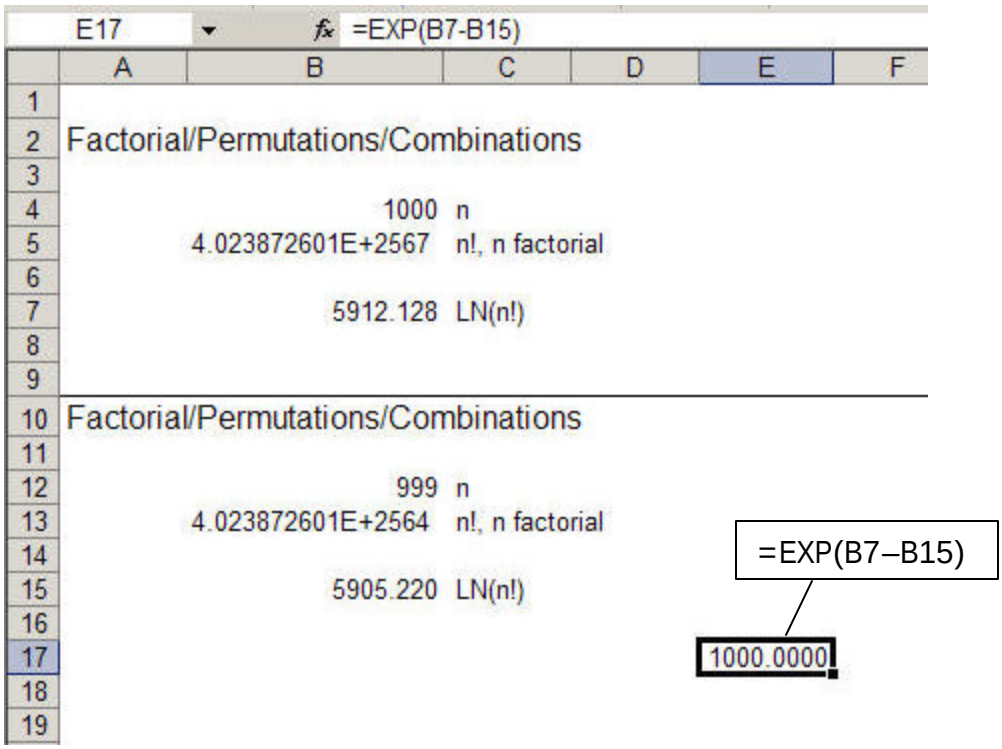
If you enter qualitative data as text if must be typed exactly the same way every time and it is case sensitive: Yes and yes would be two different responses. You could get around this by typing Yes and yes on the same line in the specification range but generally it would be better to enter numeric values such as 0 and 1. These could also be used as indicator variables in regression.

Counting Rules

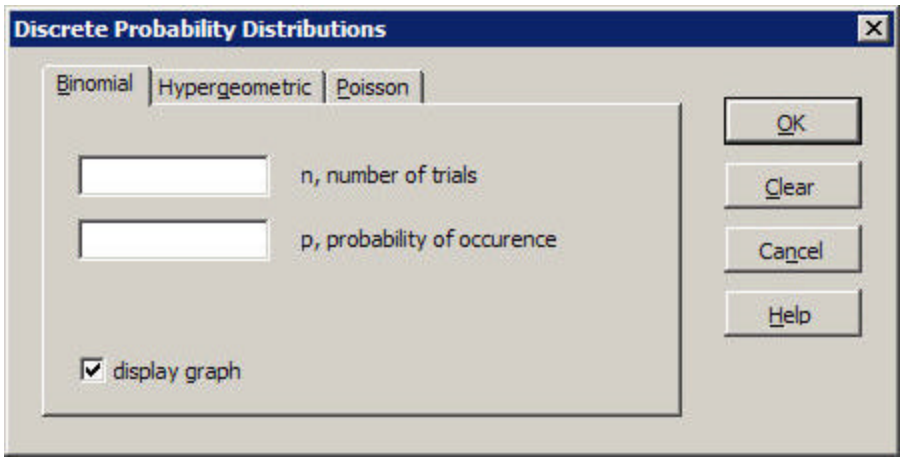


The largest number Excel can handle is approximately 1.0E306, i.e., a number with 306 places to the right of the decimal. That is certainly a large number but factorial/permutation/combination calculations often generate larger values. The largest factorial that can be displayed as a number is 170!. Values larger than that are calculated but the results are shown in text format. The Preview example above shows 1000!

Technical note: The output also includes the natural log of the answer. You may use this value for further computations and testing. For example, to verify that the factorials are correct, recall that n!/(n-1)! = n, therefore exp(ln(n!)-ln((n-1)!) should equal n, e.g. exp(ln(1000!)-ln(999!) = 1000. Values larger than 10,000,000! start to show a little error but the percent error is very small. These computations are shown below in cell E17:



Discrete Probability Distributions

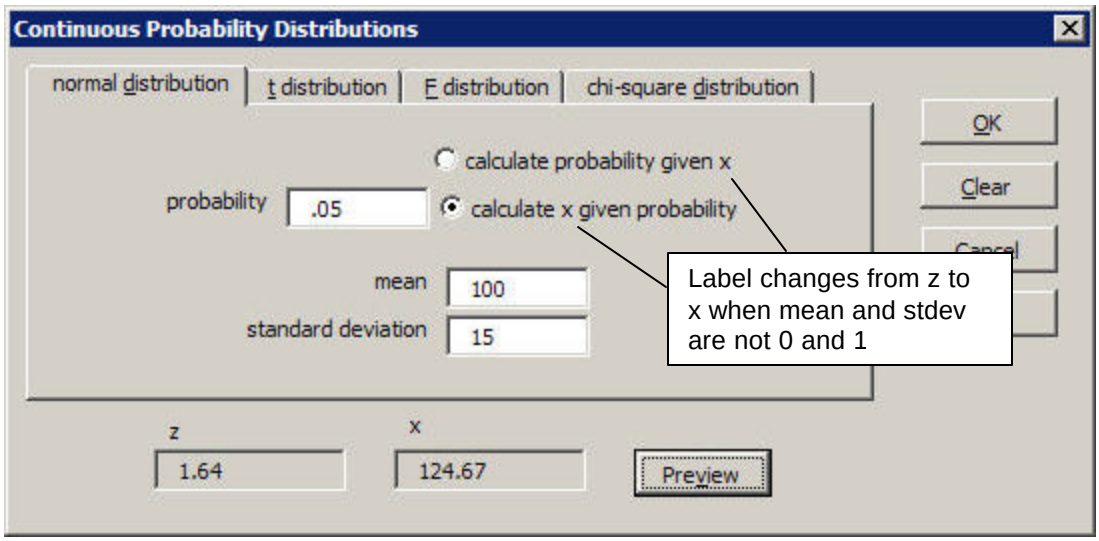


Enter the values required by each distribution. The Help file gives details regarding size limitations.

The output is formatted to five decimal places so a value that shows as .00000 is not truly zero, it is just 0 out to five decimal places. Like any MegaStat output, the formats can be changed. You will need to use scientific notation to display very small probabilities.

Continuous Probability Distributions

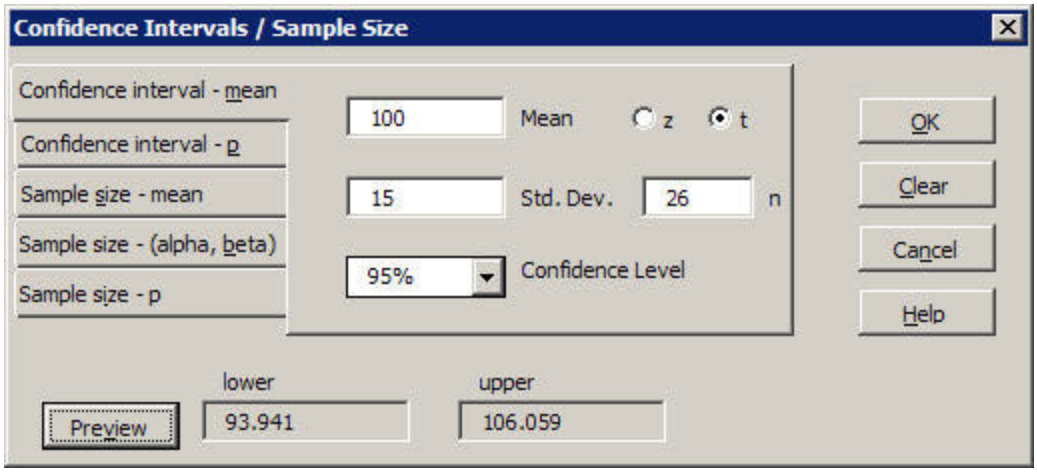
An example of continuous probability distributions is shown in the Tutorial Example chapter. The Normal Distribution dialog box below shows some other aspects. First of all it shows the calculation of values corresponding to a specified upper tail probability. It also shows that when the mean and standard deviation are anything other than 0 and 1 respectively the labels automatically change from 'z' to 'x'. Compare this dialog box with Figure 9.



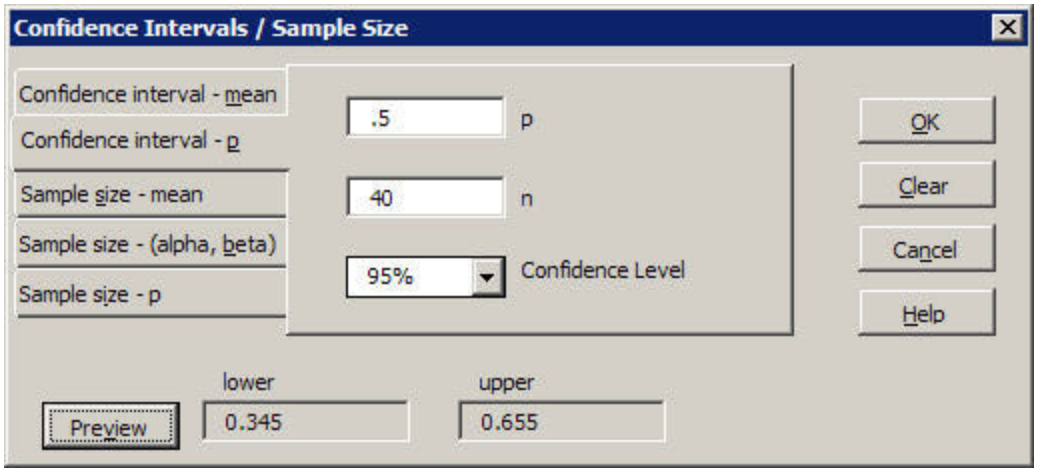
Confidence Intervals / Sample Size

For each of the dialog boxes confidence levels of .99, .95, or .90 may be selected by the dropdown arrow or you may type in any other value.

Confidence interval – mean

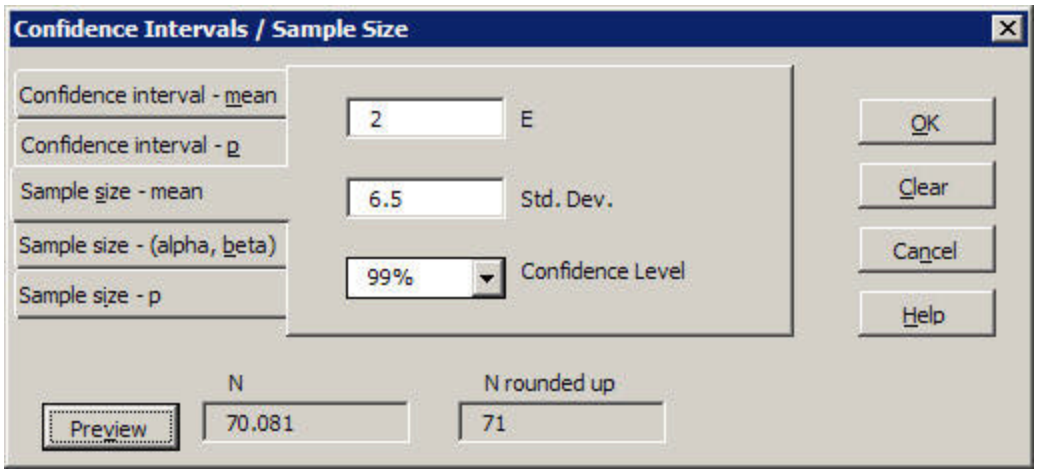


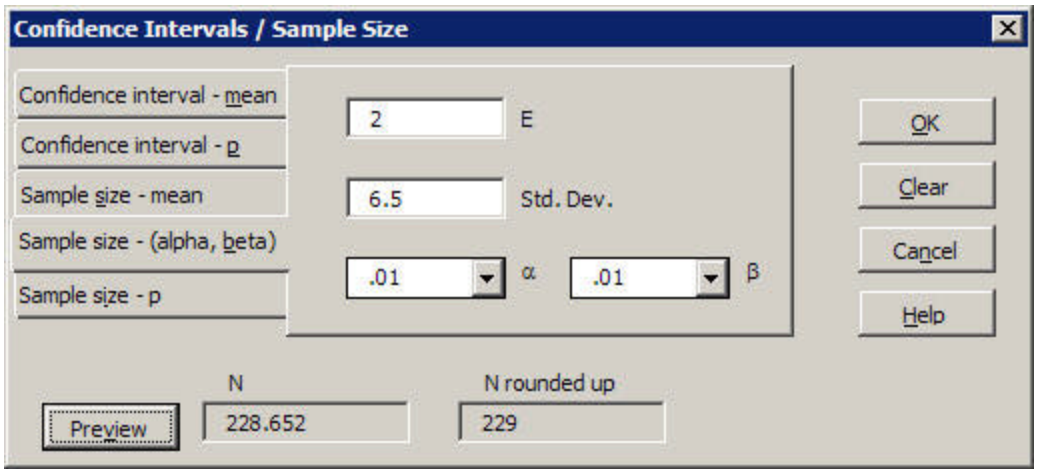
Confidence interval – p

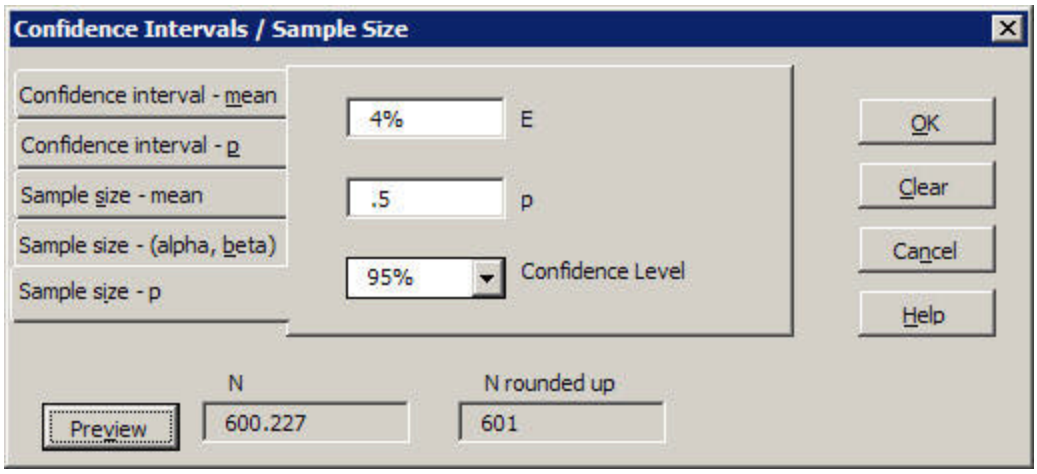


Note: Instead of typing .5 you could have typed 20 and the label would have automatically changed to ‘x’

Sample size –mean, mean with specified a and b, and p



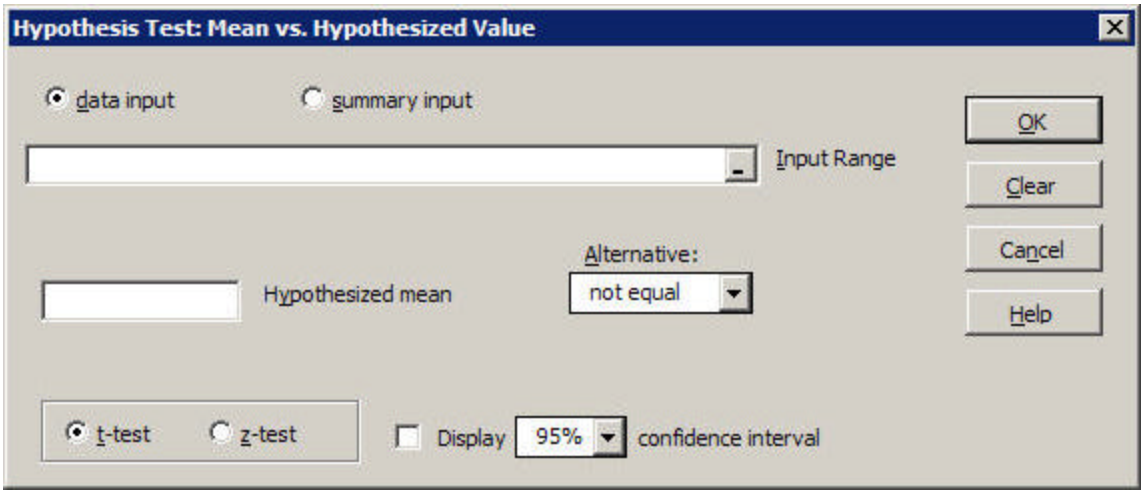




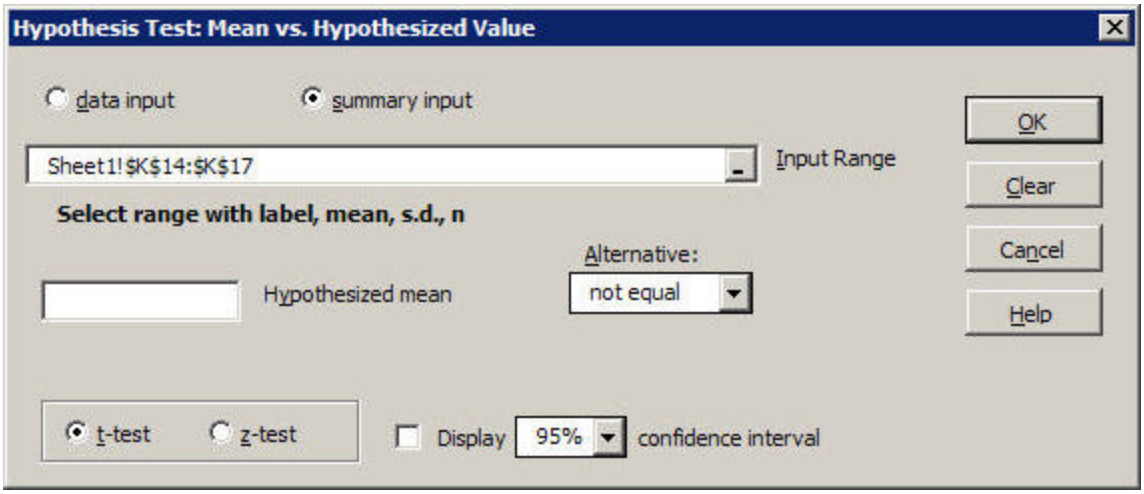
Note that the sample size calculation is rounded up to the next higher integer value.

Hypothesis Tests

Mean vs. Hypothesized Values

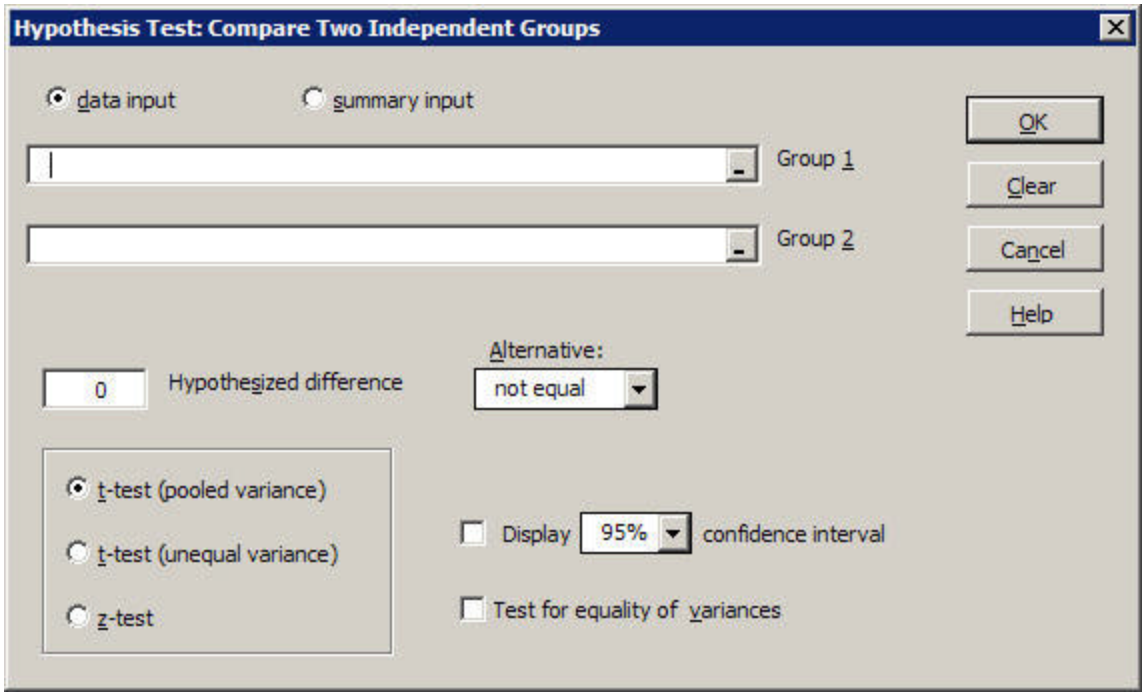


This dialog box is usually used with actual data as shown above; however, there may be situations where you already have calculated values and want to perform the hypothesis test. For example, if cells K14 through K17 contain a label, mean, standard deviation and n respectively, you would click summary input and select the cells as shown below.



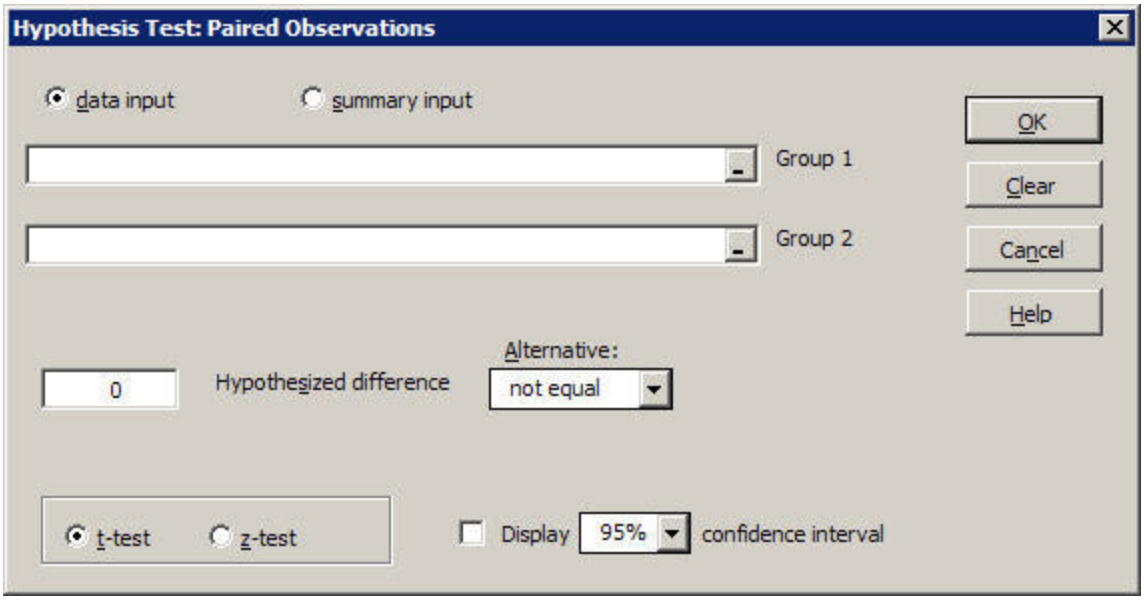
If the summary values are not in contiguous cells you may select any four cells in the proper sequence by holding the CTRL key.

Compare Two Independent Groups



All of the data selected for each group will be treated as a single group even if you select multiple columns.

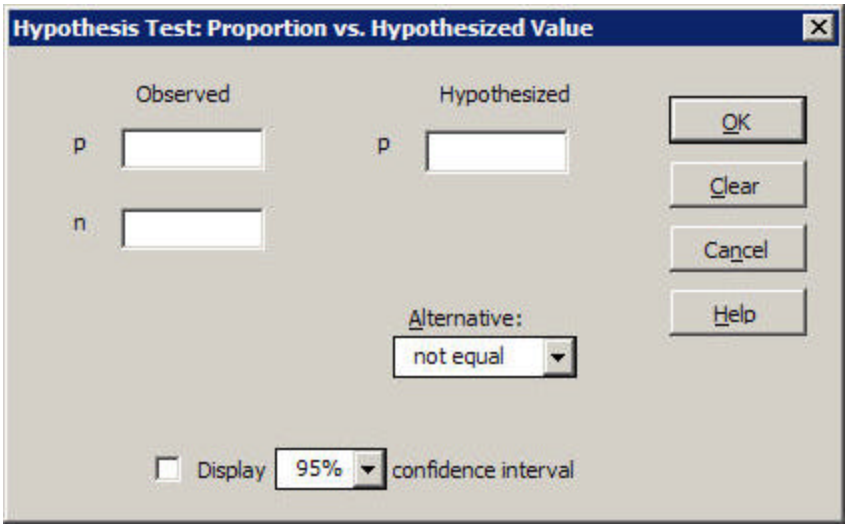
Paired Observations



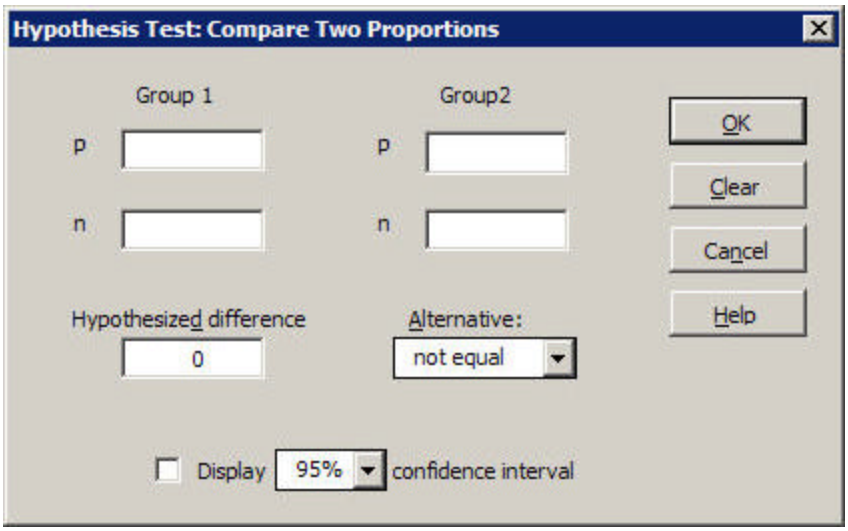
Both groups must have the same number of observations. If your data is already in the form of differences you would use the Mean vs. Hypothesized Value test.

Proportion vs. Hypothesized Value

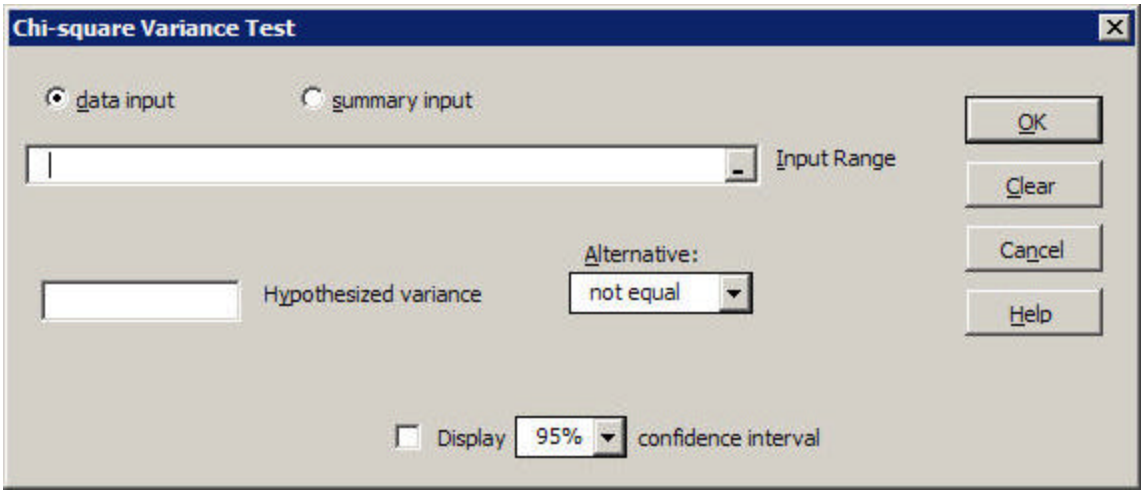
For this test and comparing two proportions, review the Entering Proportions section of the Tutorial Examples chapter.



Compare Two Independent Proportions



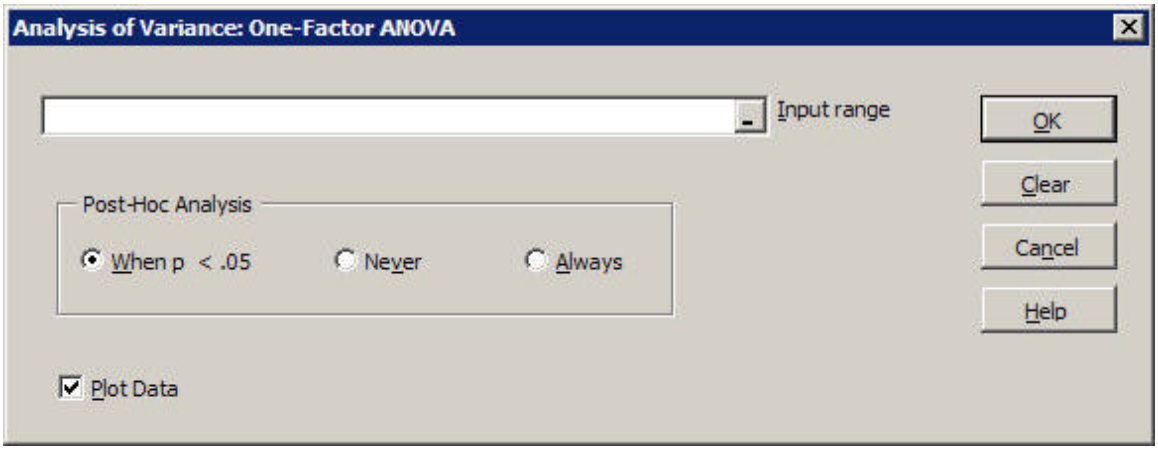
Chi-square Variance Test



To compare the variances from two groups use the hypothesis test for comparing two means and check the variance test option

Analysis of Variance

One-Factor ANOVA



Within the input range each column will be considered a group. The groups do not have to be the same size so there may be some empty cells – just make the selection block large enough to include the largest group.

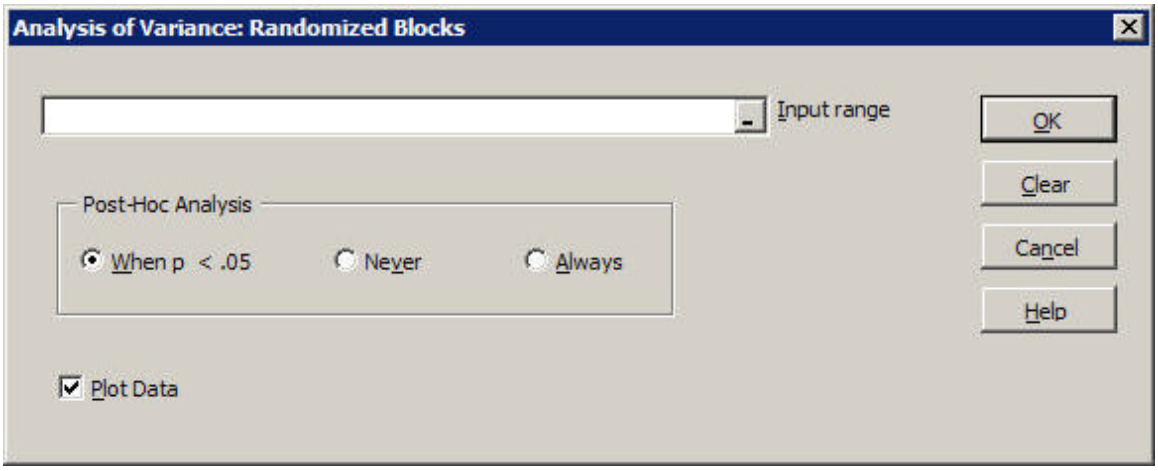
The data groups for this procedure must be side by side. For example, if you were doing an ANOVA with three groups the groups might look like this:

	A	B	C	D	E	F
1						
2		Group 1	Group 2	Group 3		
3		23	32	37		
4		25	29	31		
5		24	26	29		
6			33	31		
7			27			
8						

The post-hoc analysis is a table showing the Tukey simultaneous comparison t-values and p-values for the pairwise independent groups t-test.

Checking the Plot Data box gives a plot of the data with the group means and grand mean.

Randomized Blocks ANOVA



Within the input range each column will be considered a treatment and each row will be a block. No missing or invalid data are allowed.

Each row (block) may have a label as well as the columns (treatments). Just remember that labels must be text. If you want to use AutoFill to generate block labels, enter the first two values as '1 and '2 to force them to be text.

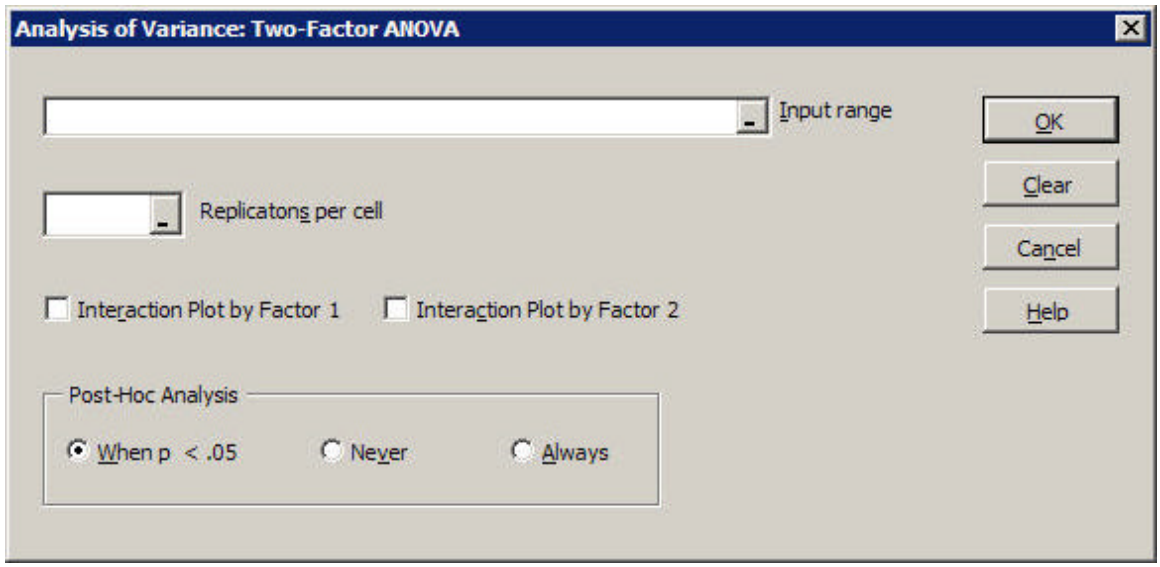
The data groups for this procedure must be side by side. For example, if you were doing an ANOVA with three treatments and five blocks the data cells must look like this:

	A	B	C	D	E	F
1						
2			Treatment 1	Treatment 2	Treatment 3	Treatment 4
3		Block 1	23	23	32	37
4		Block 2	25	25	29	31
5		Block 3	24	24	26	29
6		Block 4	27	27	31	34
7		Block 5	28	28	24	33
8						

The post-hoc analysis is a table showing the Tukey simultaneous comparison t-values and p-values for the pairwise independent groups t-test.

Checking the Plot Data box gives a plot of the data with the group means and grand mean. The data points for each block are connected.

Two Factor ANOVA



The data must be in the form shown below. Assume you have two column treatments, three row treatments and three replications per cell:

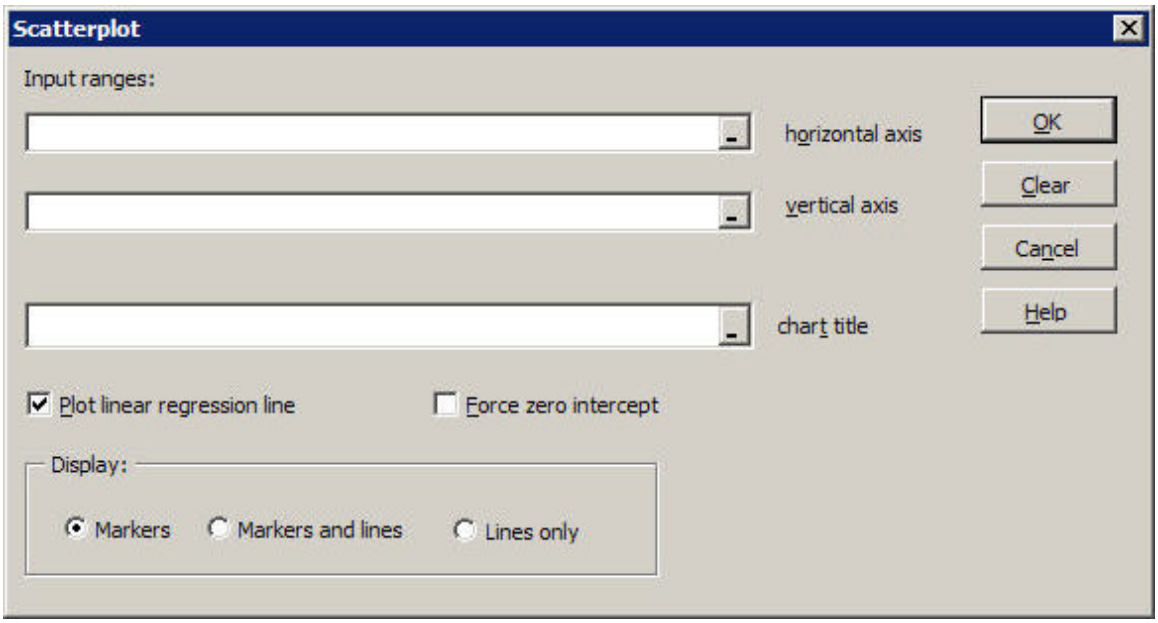
	A	B	C	D	E
1					
2			Column A	Column B	
3			48.2	55.7	
4		Row 1	53.7	52.5	
5			55.8	58.9	
6			73.0	76.2	
7		Row 2	78.1	78.4	
8			75.4	82.1	
9			52.4	54.0	
10		Row 3	49.7	52.1	
11			50.9	49.9	
12					

All cells must have the same number of replications and no missing data are allowed.

The row treatment labels may be placed in any of the cells to the left of the treatment.

Correlation / Regression

Scatterplot



Both input ranges must be the same size and be in a single column.

A title is not required but it is recommended. The title can have more than one line. Use the scroll bar on the edit box to view multiple lines. You may type in the title or click on cell(s) that contain title text.

Specify if you want the linear regression line forced through the intercept (The r^2 value is not shown for zero intercept since Excel calculates it incorrectly.)

Post output editing: MegaStat will do a good first approximation of the scatterplot, but remember that it is an Excel chart and if you right-click a chart object (*e.g.,* an axis) you go into editing mode where you can make changes. Some common things you might want to fine-tune are:

- Move the text box that contains the regression equation and r^2 if it is covering data points.

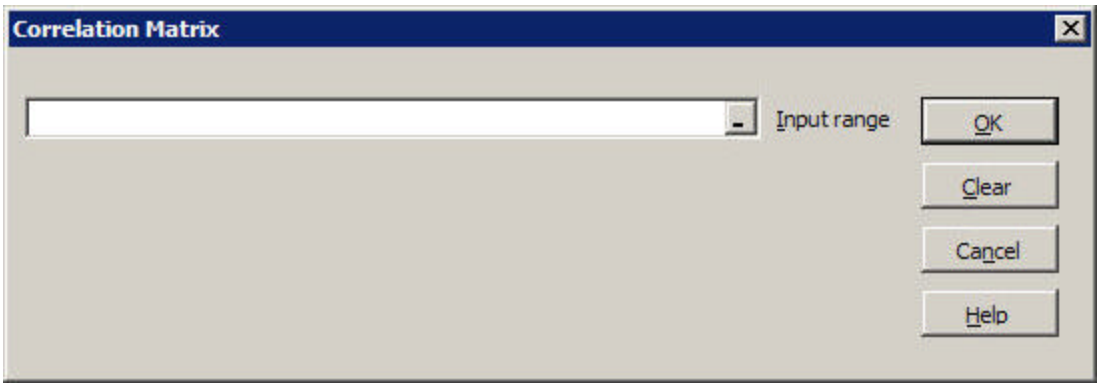
- Right-click the regression line (Trendline) and select Format Trendline | Options to extend the line forward or backward or try a non-linear curve fit.

- Rescale the axes. The program attempts to scale the axes properly; however, you can change the scaling to your preference.

- Resize the Scatterplot.

Note: The Scatterplot is linked to the data you selected in the input ranges. This means that if you subsequently change the data, the chart will be updated. Also if you drag any of the data points on the chart, the data sheet will change. If you want to lock it or paste it to another document without linking, do a Copy and Paste Special|Picture into Microsoft Word. It will look just the same, but you cannot do any further editing.

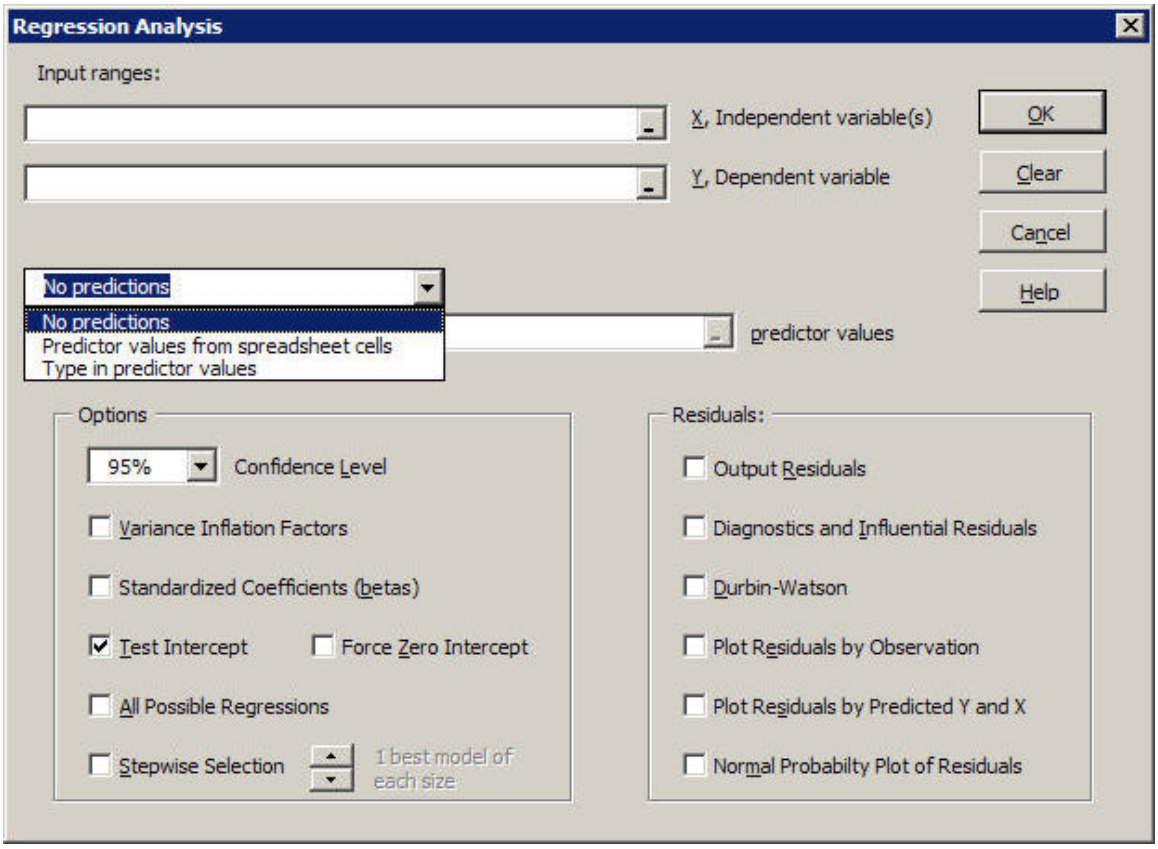
Correlation Matrix



If you want to change the order of the variables you may select multiple data ranges by holding the Ctrl key . Each selection area must have the same number of rows. No missing or invalid data are allowed.

Only the lower half of the correlation table is displayed; however, the upper portion output is in hidden format and may be displayed by formatting the cells.

Regression Analysis



Within the input range each column is considered a variable. If there is more than one independent variable MegaStat will automatically do multiple regression. You may select multiple input ranges by holding the Ctrl key. This is handy if you are doing multiple regression and the independent variables are not contiguous. The independent and dependent variable ranges must have the same number of rows.

If you want to calculate predicted values, click the dropdown arrow and select one of the options (as shown in the dialog box above):

Predictor values from spreadsheet cells

If you select this option you will need to select an input range for the values of the independent variable(s) to be used for prediction. Within this range each column will correspond to the independent variable(s) and each row will specify a different prediction. For example, if you are doing one prediction for simple regression the predictor range would be a single cell; or if you are doing three predictions for a multiple regression with four independent variables, the predictor range would be three rows by four columns.

The predictor range can be any cells, even in a different sheet; however, the obvious place to put the predictor variables would be right below the independent variables of the input data.

Type in predictor values

The input box allows you to type predictor values for the independent variable. The entries must be in the same order as the independent variables in the input data. The values must be separated by spaces or commas. You may make multiple predictions by typing a semi-colon between each prediction. If you have several predictions and/or several independent variables it will be easier to enter them into spreadsheet cells and use the method described above.

For example: if you wanted to make three predictions using X1 = 20, X2 =9; X1 = 25, X2 = 12 and X1 = 30, X2 = 15 you would type in:

20,9;25,12;30,15

You can also enter spaces to make it easier to read

20,9; 25,12; 30,15

(Secret tip: You really don't need the semi-colons and commas if you put one or more spaces between each number. You could enter: 20 9 25 12 30 15)

Select Options:

-Confidence level for confidence intervals.

-Variance Inflation Factors to measure multicollinearity in multiple regression

-Standardized regression coefficients (sometimes called betas)

-Test for zero intercept

-Force zero intercept

-All Possible Regressions

This option prints a summary line for each possible combination of variables in a multiple regression. Since each summary line includes the p-values for the variables in the model, this output is useful for model building since you can easily see which combinations of variables work best.

The number of summary lines is $2^k - 1$ where k is the number of independent variables. Thus 12 variables will generate 4095 regression summaries. Above 10 variables the calculations can take quite a while and there may be memory issues. Stepwise Selection is usually a better option for more than 8 variables. The program still has to calculate the regression summaries but there is a lot less output since only the best regressions are displayed.

-Stepwise Selection

This option is similar All Possible Regression but it displays only the best models for any given number of variables, which gives more compact output.

All Possible Regressions and/or Stepwise Selection are more powerful and flexible than traditional Stepwise Regression. Stepwise Regression determines the one best model where MegaStat not only does this but also shows if there are other models nearly as good.

Selecting All Possible Regressions or Stepwise Selection causes all other options to be deactivated (grayed out). To restore the other options, uncheck All Possible Regressions or Stepwise Selection.

The Force zero intercept option will not work if the slope is negative or the intercept is not close to zero.

You cannot select Test intercept = 0 and Force zero intercept simultaneously. Selecting one will uncheck the other one.

Select Residuals options

- Output of dependent variable, predicted value, residual

  This option will automatically be selected if you choose any other residuals option.

- Durbin-Watson
- Plot Residuals by Observation
- Plot Residuals by Predicted Values and by each Independent Variable
- Diagnostic and Influential Residuals

  These are advanced options that would be important for upper level courses and research use.

  - --Leverage

    Shaded blue if the value is > 2 * (Nvar + 1) / n
  - --Studentized Residual
  - --Studentized Deleted Residual

    These are t values and they are shaded light blue if the p-value is <=.05 and blue if the p-value is <= .01
  - Cooks D

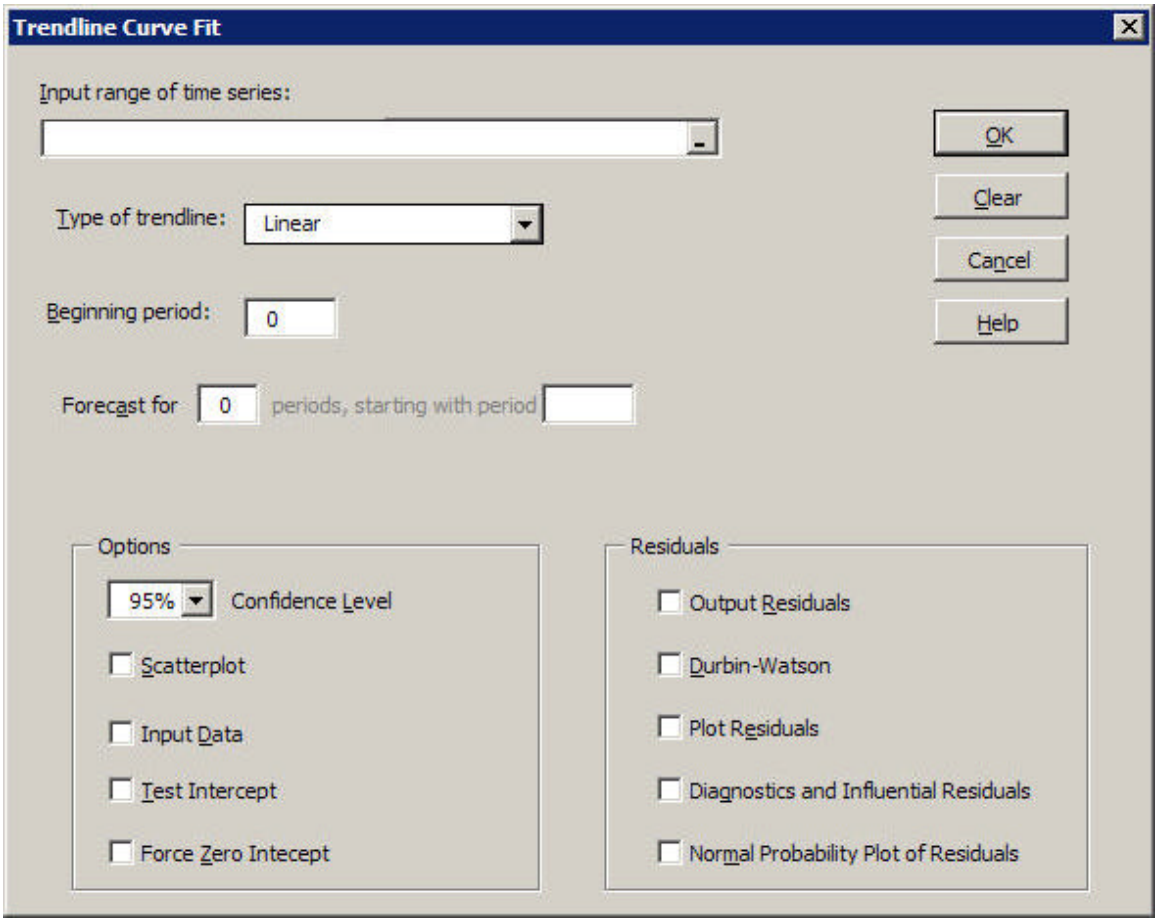
    This value is an F ratio and it is shaded light blue if the p-value is <= .80 and blue if the p-value is <=.50.

- Normal Probability Plot of Residuals

  Note: The data values needed to display the Normal Probability Plot are stored in a ChartDataSheet as discussed in the Utilities section.

Time Series / Forecasting

Trendline Curve Fit



Select a single column of data as the input range of the data to be fit with a trendline (*i.e.*, the dependent variable, Y). The program will create the time series independent variable (X). Then select the type of trendline you wish to fit: Linear, Exponential or Polynomial using the dropdown selection menu.

Select the beginning period for the time series variable (X). This is usually 0 or 1 but it can be any value, for example, a year. If you enter a non-integer value it will be rounded to the nearest integer.

If you wish to make one or more forecasts, specify how many and the starting period for the forecasts.

Specify if you want any of the output options:

Scatterplot of the data with the trendline. (In order to get the scatterplot you must also have the input data displayed so that box will automatically be checked also.

[In rare instances when you specify an exponential curve fit, the scatterplot will not be able to display the trendline. This is caused

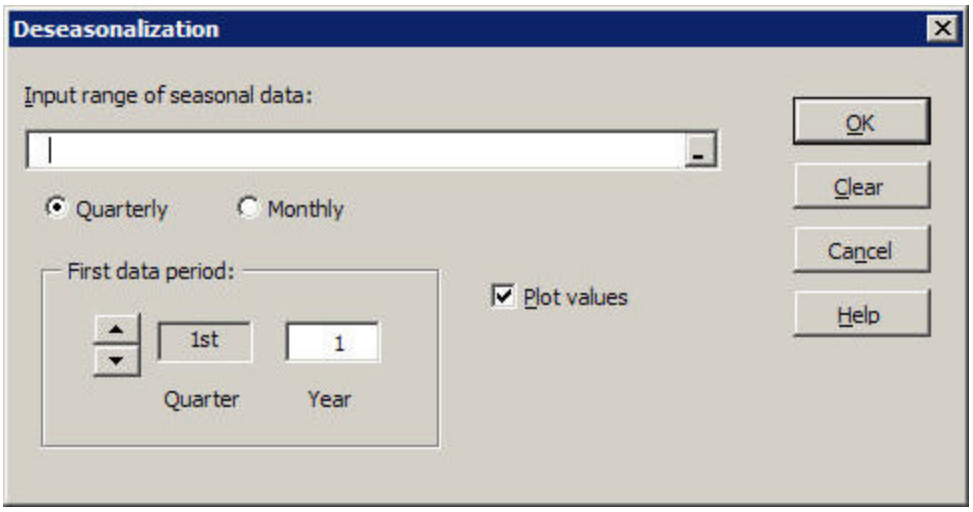
when you use large values (*e.g.*, years) for your time series. If you start the time series at t=0 the trendline will probably be displayed.]

Display the input data, *i.e.* the time series variable and the input data.

Output of residuals, Durbin-Watson and plot of residuals. Note that you cannot get the latter two unless you output the residuals.

Options for Adjusted R^2, Test for intercept and Force intercept through zero. (See regression section for details on these items.)

Deseasonalization

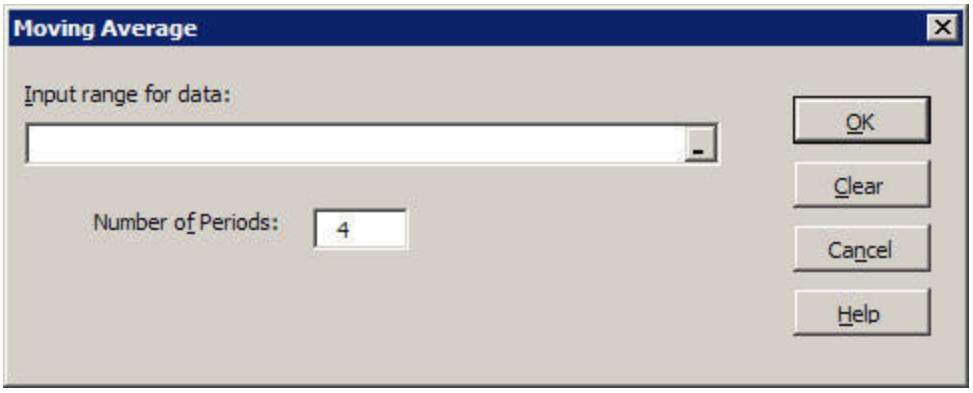


In order to do deseasonalization you must have at least two years of data (*i.e.*, 8 quarters or 24 months). You cannot have missing data; however, you can have a partial year at the beginning and end of the data.

Indicate if the data is quarterly or monthly. Then specify the month and year of the first data value. The example shows the first data value is the 3rd quarter of 1994.

After you have run the analysis you can use the Trendline or Regression options to make predictions using the deseasonalized data. (Make sure to remember to reseasonalize the prediction by multiplying by the appropriate seasonal index.)

Moving Average

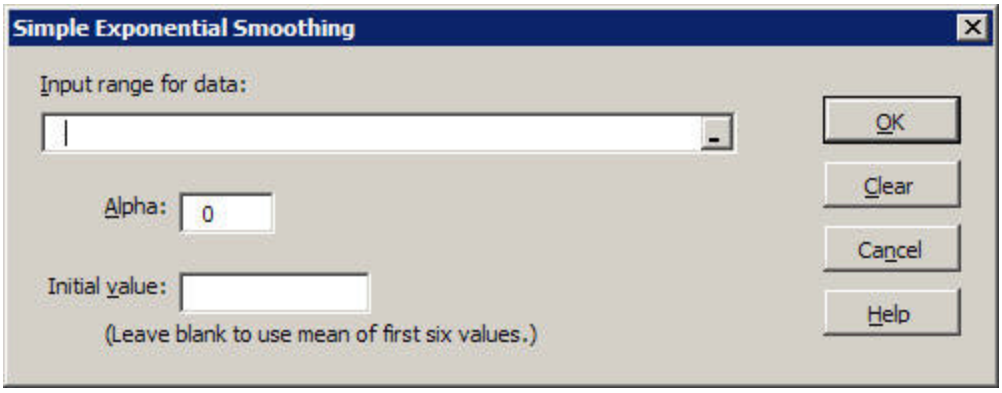


Select a single column of data as the input range. Then specify the number of periods in the moving average (usually 3 to 5 but can be any value between 1 and n). The example above shows a 5 term moving average.

The moving average is displayed adjacent to the last period averaged; however, you can cut/paste or drag/drop the averaged values to position them anywhere.

Exponential Smoothing

Simple Exponential Smoothing

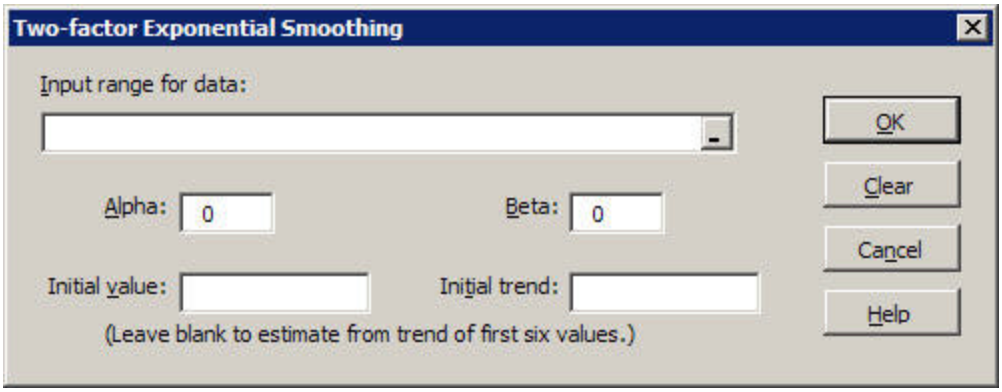


Select a single column of data as the input range of the data to be smoothed.

Specify Alpha, the weight to be given to each new value of the smoothed series. Alpha must be between 0 and 1.

Specify an initial data value for the smoothed series. If you leave the initial value blank, it will use the mean of the first six data values.

Two-factor Exponential Smoothing

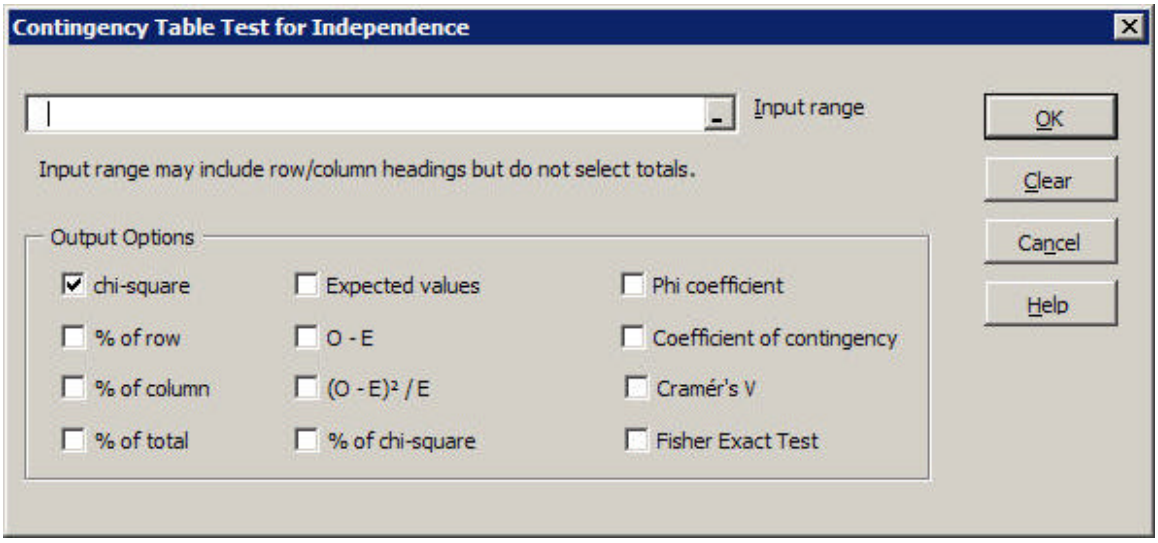


Select a single column of data as the input range of the data to be smoothed.

Specify Alpha, the weight to be given to each new value of the smoothed series. Alpha must be between 0 and 1. If you leave this blank it will use the intercept of the linear trend of the first six values. Then Specify Beta, the weight to be given to the trend. Beta must be between 0 and 1. If you leave this blank it will use the slope of the linear trend of the first six values.

Chi-Square / Crosstab

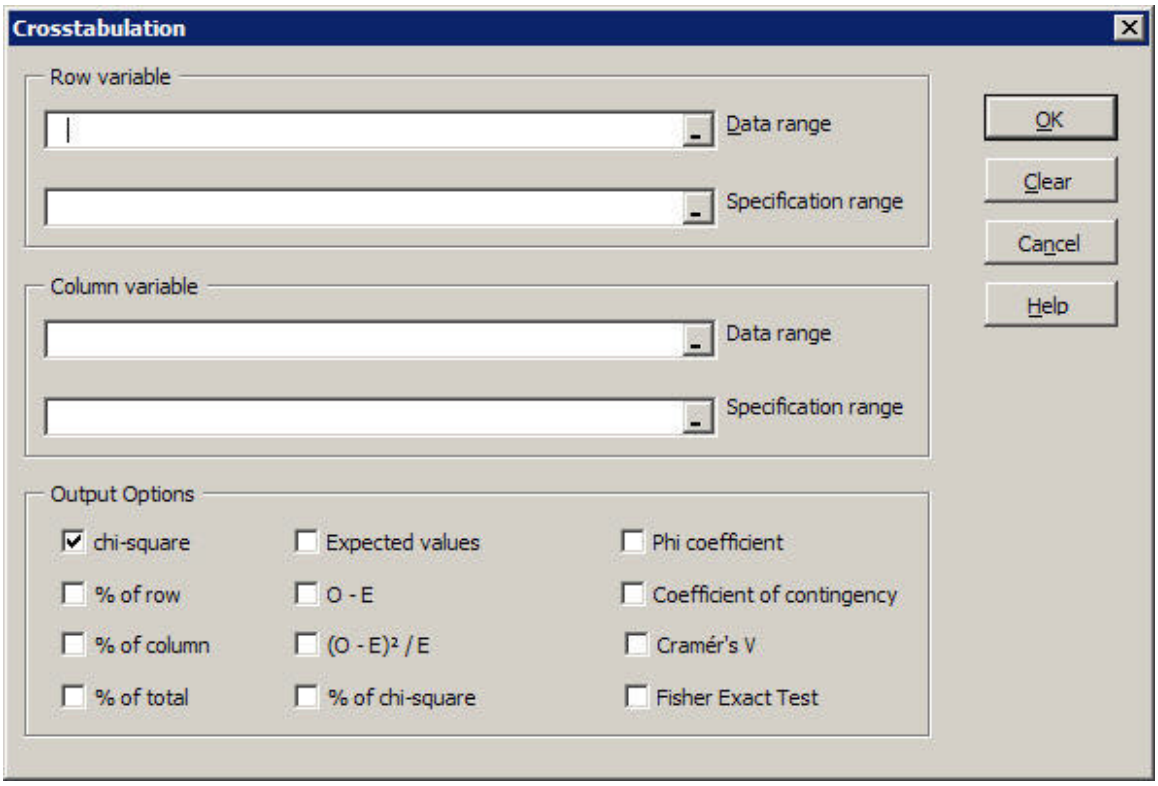
Contingency Table



You may include row and column headings in the input range; however, you should not include row or columns totals. The table cannot contain any empty or non-numeric cells

Click on the Output Options checkboxes to select the values to be output. In general, it is not a good idea to select all of the options for one output since this will give a cluttered output.

Crosstabulation

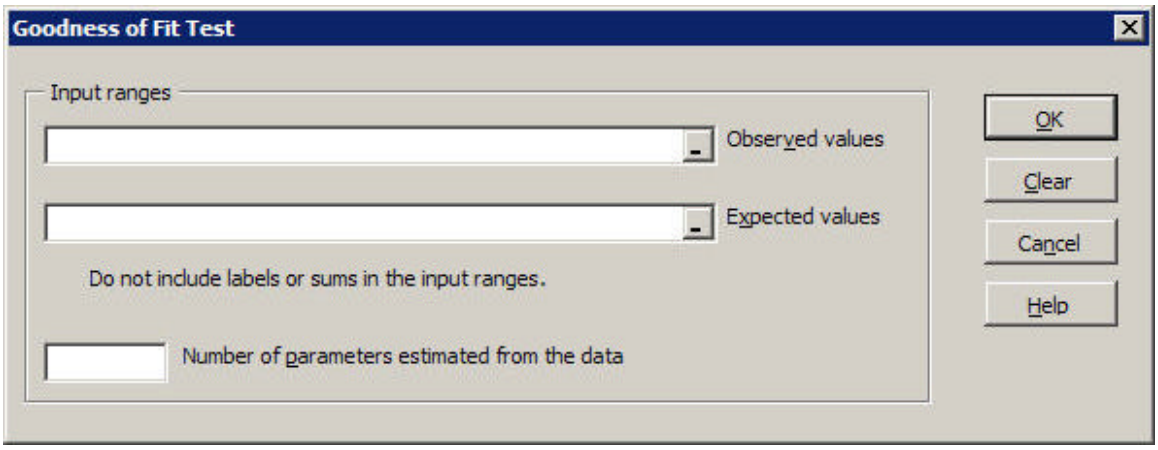


This option calculates a two-factor crosstabulation table from qualitative data.

Select the input data ranges for the row and column variables. Then select the specification range for the row and column variables. (See the discussion of specification ranges under Qualitative Frequency Distributions.)

Click on the Output Options checkboxes to select the values to be output. If all you want are the counts, leave all options unchecked.

Goodness of Fit Test

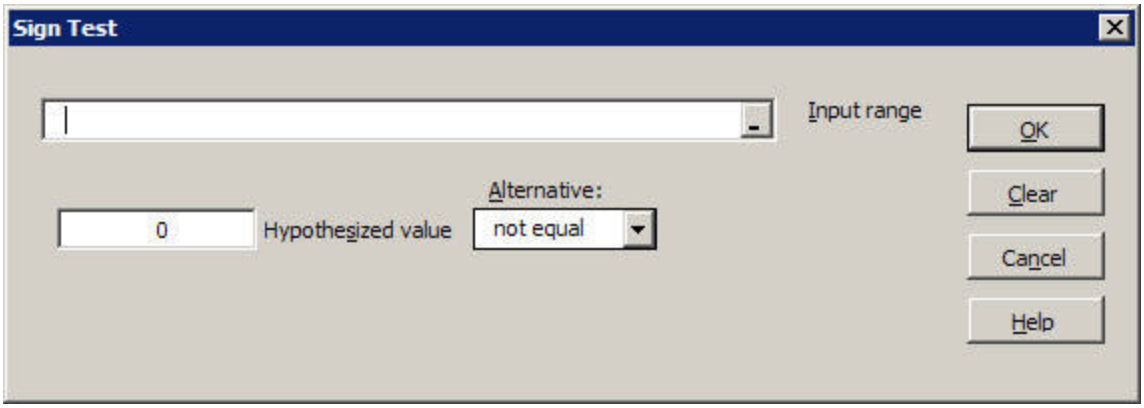


Select the input data ranges for the observed and expected frequencies. You must not include labels in the range. Both ranges must include the same number of cells although they don't have to be the same shape; however, if the ranges are not the same shape, carefully check the output to make sure observed and expected values were paired properly.

Specify the number of input parameters estimated from the data. The default value is none. For example, if you were doing a Goodness of Fit for a normal distribution and you used the mean and standard deviation from the actual data to find the expected values, then this value would be 2.

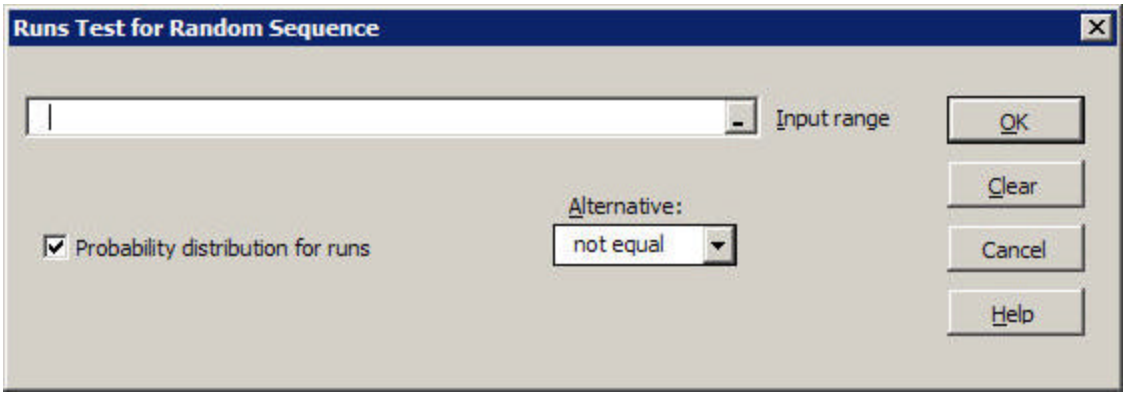
Nonparametric Tests

Sign Test



This test can also be performed with the binomial distribution with n = sample size and p = .5.

Runs Test for Random Sequence



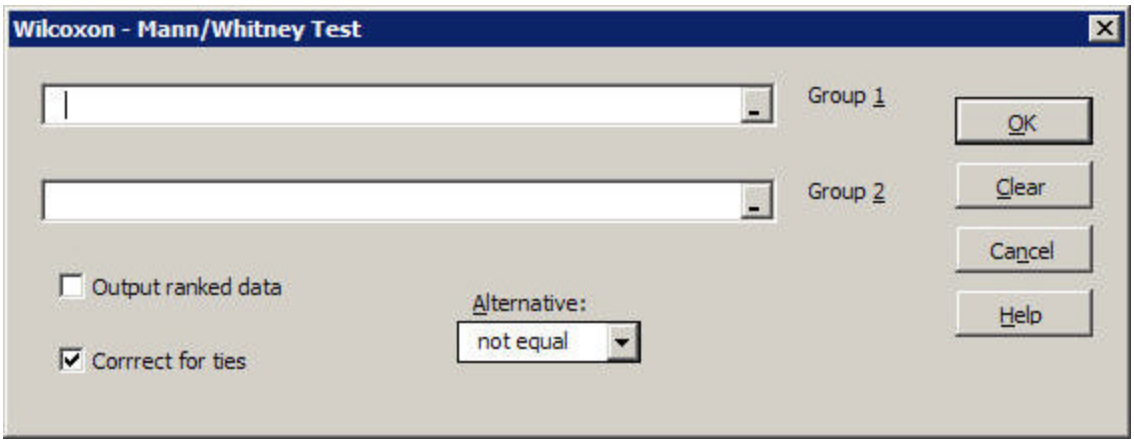
The program expects each cell in the input range to contain one of two unique values and it tests whether there are too few or too many runs than would be expected by chance. The two values can be either alpha or numeric but if more than two different values are detected in the range an error message will be displayed.

If you want to use this test to check runs above and below the median, use the Excel IF function to code your data into two values representing above and below the median.

Note: The z-value is calculated using a correction for continuity.

Check if you want to display the complete probability distribution for the total number of runs. Hint: If you want to see the probability distribution and do not have a set of data, create a dummy dataset with n1 A's and n2 B's (or any two values).

Wilcoxon – Mann/Whitney Test

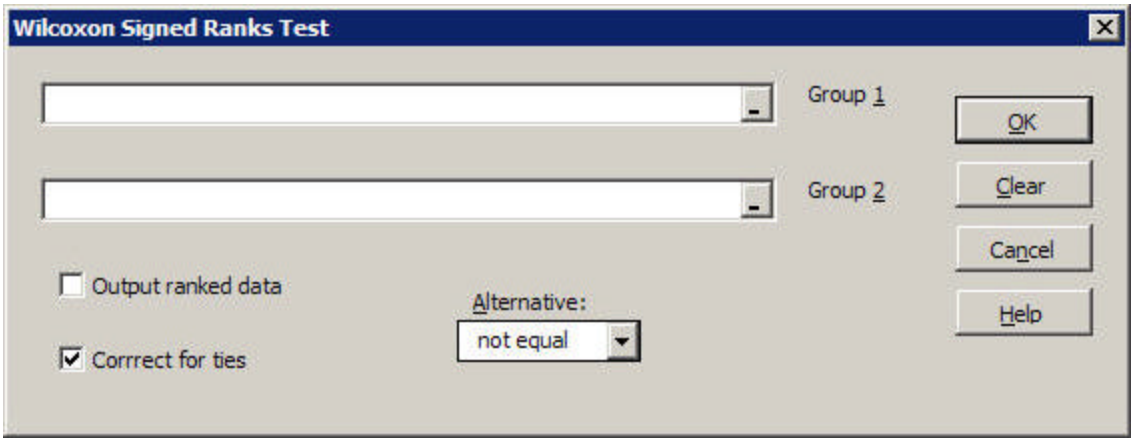


This test works with ranked data; however, the data does not have to be ranked. The program will convert the data to ranks as required by the test.

If you check that you want to output ranked data, the ranked data will be on the output sheet; your original data will not be changed.

The default is to correct for ties and generally should be selected.

Wilcoxon – Signed Rank Test

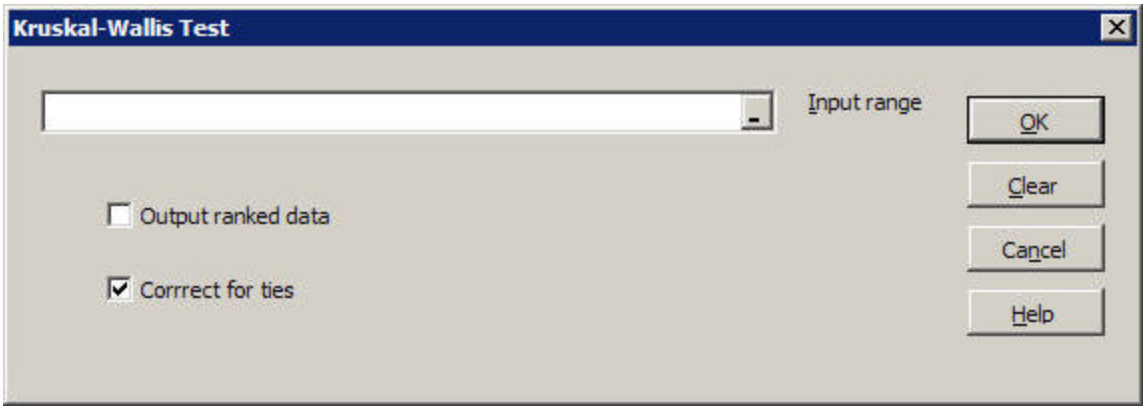


This test works with ranked data; however, the data does not have to be ranks. The program will convert the data to ranks as required by the test. Since this test works with paired observations, both groups must be the same size.

If you check that you want to output ranked data, the ranked data will be on the output sheet; your original data will not be changed.

The default is to correct for ties and generally should be selected.

Kruskal – Wallis Test



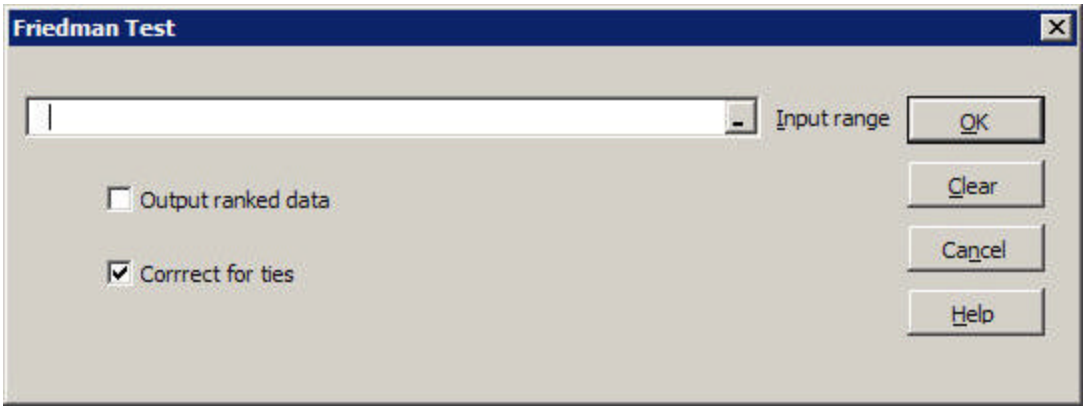
Within the input range each column will be considered a group. The groups do not have to be the same size so there may be some empty cells – just make the selection block large enough to include the largest group. See One Factor ANOVA for details on data layout.

This test works with ranked data; however, the data does not have to be ranks. The program will convert the data to ranks as required by the test.

If you check that you want to output ranked data, the ranked data will be on the output sheet; your original data will not be changed.

The default is to correct for ties and generally should be selected.

Friedman Test



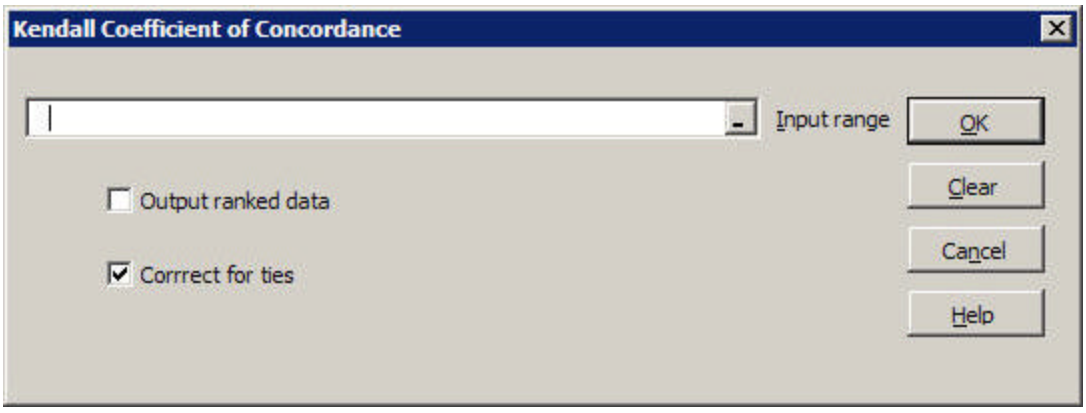
Within the input range each column will be considered a treatment and each row will be a block. No missing or invalid data are allowed. See Randomized Blocks ANOVA for details on data layout.

This test works with ranked data; however, the data does not have to be ranks. The program will convert the data to ranks as required by the test.

If you check that you want to output ranked data, the ranked data will be on the output sheet; your original data will not be changed.

The default is to correct for ties and generally should be selected.

Kendall Coefficient of Concordance



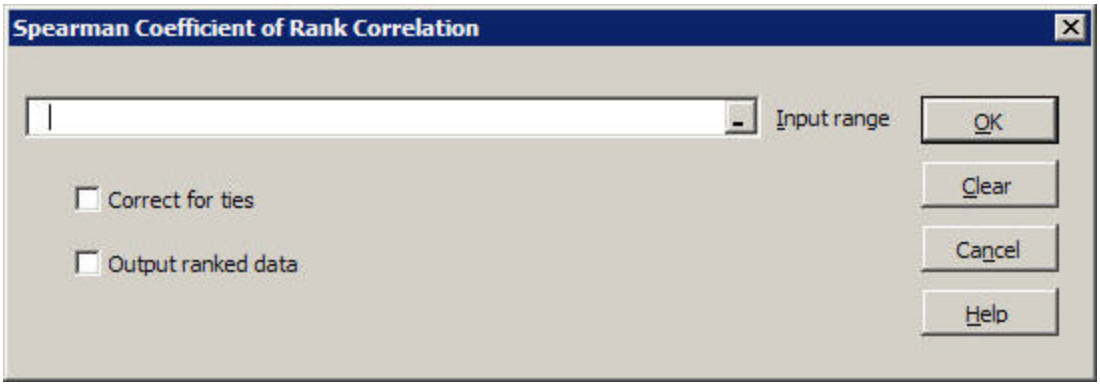
Within the input range each column represents an item being judged and each row represents a judge. No missing or invalid data are allowed.

This test works with ranked data; however, the data does not have to be ranks. The program will convert the data to ranks as required by the test.

If you check that you want to output ranked data, the ranked data will be on the output sheet; your original data will not be changed.

The default is to correct for ties and generally should be selected

Spearman Coefficient of Rank Correlation

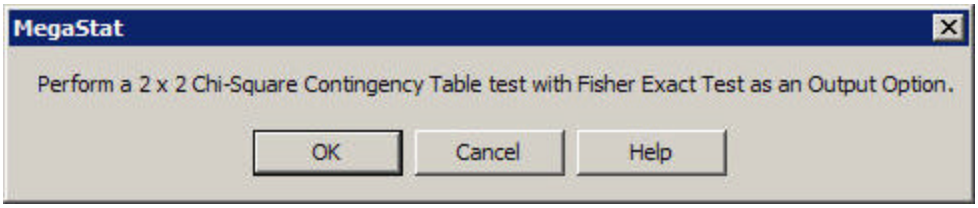


Within the input range each column will be considered a variable. No missing or invalid data are allowed.

This test works with ranked data; however, the data does not have to be ranks. The program will convert the data to ranks as required by the test. The program corrects for tied ranks.

Check whether you want to correct for ties and check whether you want to output the ranked data. The ranked data will be on the output sheet; your original data will not be changed.

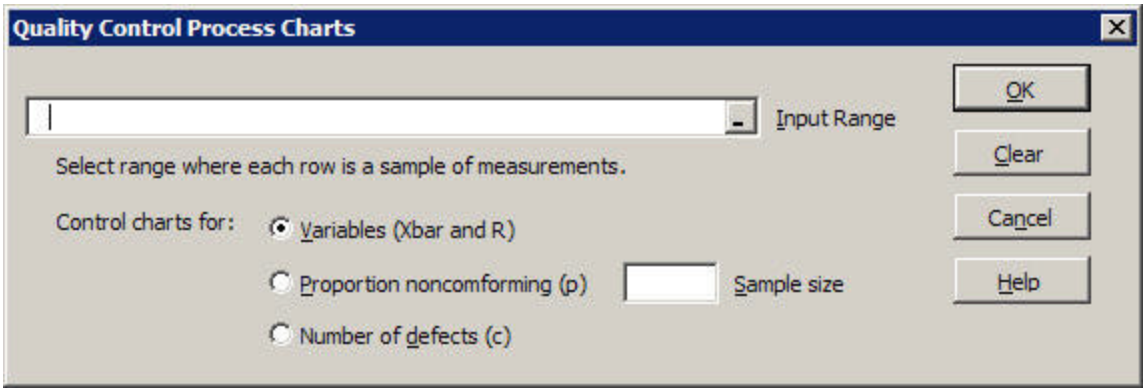
Fisher Exact Test



This test is an output option of the chi-square Contingency Table Test and assumes you have data in the form of a contingency table. When you click OK the Contingency Table Test will be loaded with the Fisher Exact Test checked as an output option. You may also select other output options.

Quality Control Process Charts

This option allows you to plot the most common quality control process charts.



Control chart for variables (Xbar and R chart)

Select a range of data where each row is a sample of measurements. You may have missing data in any sample; however, the sample size for the chart will be determined by the number of columns in the selected range.

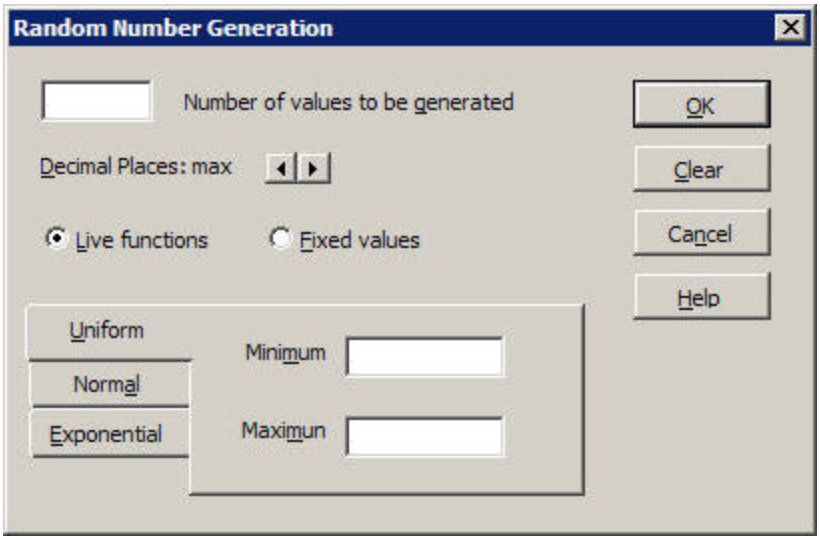
Control chart for proportion nonconforming (p chart)

Specify a sample size and select a range where the values are the proportion or number of defective items in each sample. If the number is less than one it will be considered a proportion; if the number is greater than or equal to one it will be divided by the specified sample size to calculate the proportion.

Control chart for number of defects per sample (c chart)

Select a range where each value is the number of defects in a sample.

Generate Random Numbers



If you specify live functions MegaStat places Excel functions in the specified number of cells. You may recalculate the values by pressing the F9 function key. If you find a set of values you want to keep, you can 'freeze' them by selecting the cells and doing a Copy followed by a Paste Special | Values.

If you select uniform random numbers you need to specify the minimum and maximum values; the normal distribution will need a mean and standard deviation; and the exponentially distributed numbers require, mu, the mean rate of occurrence.

The exponential distribution creates a distribution skewed to the right. If you want a distribution skewed to the left, subtract the values from a constant larger than the largest value.