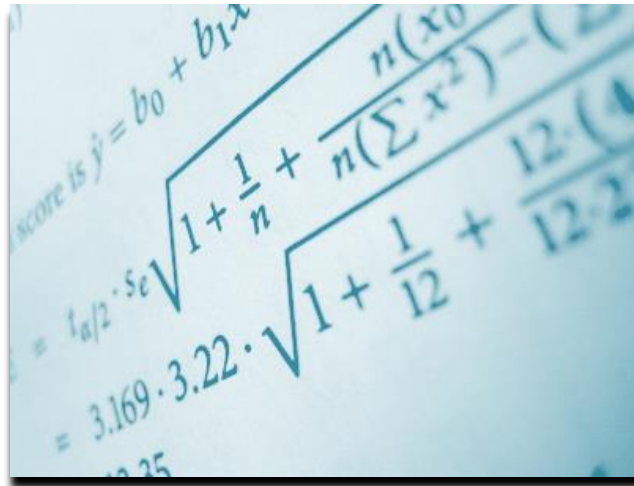




Statistician

Powerful, Easy to Use Statistics Add-in for Excel



User Manual for

Statistician (Lite)

Statistician (Standard)

from xlQA (www.StatisticianAddin.com)

Preface

The *Statistician* Excel 2007/2010/2013/365 addin performs a range of statistical analysis within the Microsoft Excel environment. Excel is a powerful and widely used data analysis tool that was not originally designed for statistical analysis. Although Excel contains a number of statistical functions (see Appendix A), it is often difficult to implement in Excel a number of the statistical tests to the level required by many researchers and students. *Statistician* overcomes this shortcoming of Excel. *Statistician* is totally integrated within Excel making the extensive graphical, data analysis and presentation features of Microsoft Excel available to the user without having to cut and paste between various software packages.

Statistician is unique amongst Excel statistical addins because of its extensive use of meaningful input forms that make the use of the software easy and intuitive. Because statistical analysis is performed upon random variables, *Statistician* forces the user to define *Data Variables* which are consequently used for statistical analysis. The data observations which comprise a *Data Variable* are stored in a hidden worksheet and are saved when the user exits Excel. This feature gives *Statistician* the look and feel of a professional statistical package. All output from *Statistician* is sent to a spreadsheet. This output is second to none in terms of clarity for the end user. Rather than simply reporting a few cryptic numbers which are the result of a statistical test, *Statistician* outputs a meaningful report of a statistical test clearly stating the null and alternative hypotheses, alphas, test statistics, critical values, p-values, the decision rule, conclusion and any other relevant factors.

The engine behind *Statistician* is Microsoft Visual Studio, a fast and powerful development environment that integrates seamlessly into the suite of Microsoft Office products. The use Microsoft Visual Studio as the development tool offers the end user processing speeds that Excel VBA could not achieve. This guarantees that the software underpinning *Statistician* will have a very long shelf life.

Dr Bernard Bollen (PhD in Econometrics, Monash University),
(March 2012).

Table of Contents

Preface.....	2
Section (1) - Manage Data.....	6
(a) Data Variables.....	7
(b) Importing Data Variables.....	7
(c) Selecting Data Variables.....	9
(d) Exporting Data Variables.....	9
(e) Removing Data Variables.....	9
(f) Renaming a Data Variable.....	10
Section (2) - Controls Common to all Forms.....	11
Section (3) – Tools.....	13
(a) Summary Statistics.....	13
(b) Covariance and Correlation.....	17
(c) Autocorrelation.....	21
(d) Statistical Tables.....	23
(e) Sort and Rank.....	26
(f) Sampling.....	27
(g) Optimizer.....	28
(h) Make Lagged Data Variables.....	29
Section (4) – Standard Tests.....	30
(a) Test Population Mean.....	30
(b) Test Difference in Population Means: Independent Samples.....	34
(c) Test Difference in Matched Pairs: Dependent Samples.....	40
(d) Test Population Proportion.....	43
(e) Test Difference in Population Proportions.....	47
Section (5) – Variance Tests.....	51
(a) One Sample.....	52
(b) Two Samples.....	53
(c) Many Samples.....	54
Section (6) Normality Tests.....	57
(a) Jacque-Bera Test.....	58
(b) Anderson-Darling Test.....	58
(c) Shapiro-Wilk Test.....	59

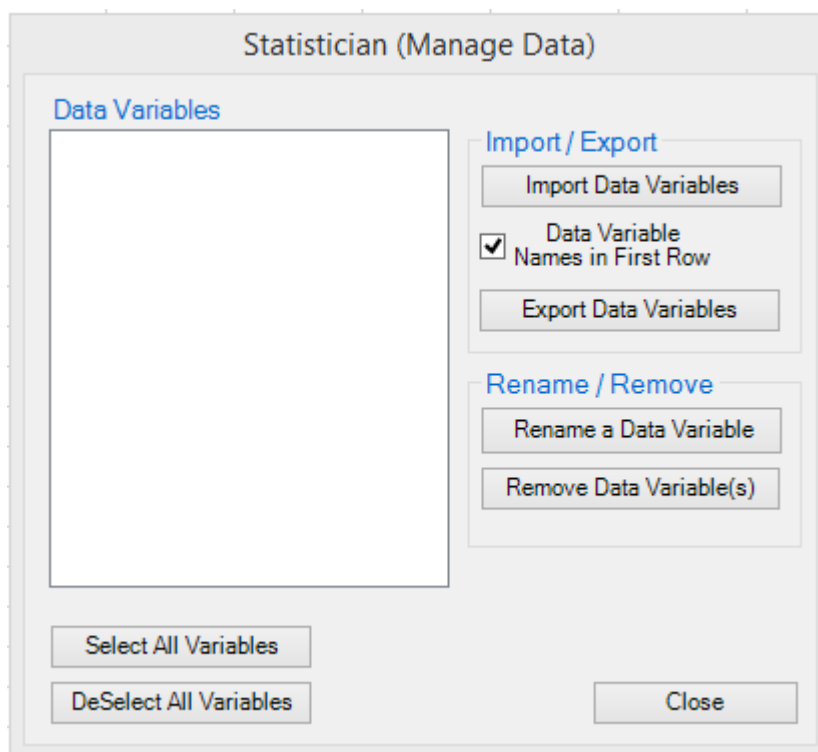
(d) Kolmogorov-Smirnov Test.....	59
(e) Lilliefors test	59
(f) Cramér-von Mises Test.....	60
Section (7) - Non Parametric Tests	61
(a) Runs Test	61
(b) Non Parametric Tests (Two Sample)	63
(c) Mann-Whitney U Test	64
(d) Wilcoxon Rank Sum Test	65
(e) Wilcoxon Signed Rank Sum Test	66
(f) Sign Test	67
(g) Kolmogorov-Smirnov Test (two sample)	68
(g) Non Parametric Tests (Two or More Samples)	70
(h) Kruskal Wallis Test.....	70
(i) Friedman Test	71
Section (8) - χ^2 Tests	75
(a) Multinomial Experiment	75
(b) Contingency Tables.....	77
Section (9) – ANOVA.....	79
(a) Single factor ANOVA.	80
(b) Two factor ANOVA without replication.	83
(c) Two factor ANOVA with Replication.....	85
Section (10) - Regression Analysis.....	89
Section (11) – Binary Models.....	96
Section (12) – Count Models.....	101
Section (13) - Time Series	105
(a) Forecasting	105
(b) Holt-Winters smoothing techniques	112
(c) Hodrick-Prescott filter	115
Section (14) Multivariate Analysis	116
(a) Cluster Analysis.....	116
(i) Hierarchical (or Join) Clustering	119
(ii) K-means Clustering	121

(b) Principal Component Analysis	122
Appendix (A) - Built-in Excel Statistical Functions.....	124
Appendix (B) - Other Excel Functions	128
Math and trigonometry functions	128
Text functions.....	130
Financial functions.....	131
Date and time functions	133
Information functions	134
Logical functions	135
Lookup and reference functions.....	135
Database functions.....	136
Engineering functions.....	136
Add-in and Automation functions.....	138
Cube functions	138

Section (1) - Manage Data

(Note: If you are a first time user of *Statistician* you can get up to speed quickly and easily by going through the Quick Start guide (and data) that is downloadable from www.xlstatistianaddin.com. The User manual that you are now using is more of a reference manual than a learning tool.)

The **Manage Data** form is selected by clicking the **Manage Data** button in the **Tools** groupbox on the *Statistician* ribbon tab.



The image shows a dialog box titled "Statistician (Manage Data)". It features a large empty rectangular area on the left labeled "Data Variables". To the right of this area are two sections of controls. The first section is titled "Import / Export" and contains three buttons: "Import Data Variables", a checked checkbox labeled "Data Variable Names in First Row", and "Export Data Variables". The second section is titled "Rename / Remove" and contains two buttons: "Rename a Data Variable" and "Remove Data Variable(s)". At the bottom of the dialog box, there are three buttons: "Select All Variables", "DeSelect All Variables", and "Close".

In *Statistician*, the analysis of data is not conducted upon data stored in a range of cells in a worksheet. Rather, data is imported from a range within an Excel worksheet and is copied to a hidden worksheet through the **Manage Data** form. This form (like all forms in *Statistician*), is activated from the Excel ribbon by clicking the **Manage Data** button. The **Manage Data** form can also be activated from all other forms in *Statistician*. The **Manage Data** form allows the user to define a *Data Variable*. All of the analysis of data in *Statistician* is conducted upon *Data Variables*.

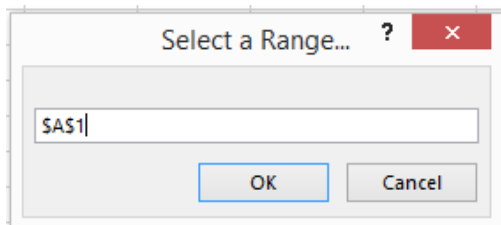
(a) Data Variables

A *Data Variable* in *Statistician* is comprised of a number of *Data Observations* stored in a column which are usually numeric but may be text. To import *Data Variables* into *Statistician* and if **Data Variable Names in First Row** is checked, the data must be stored in columns in a spreadsheet and each column must have a *Data Name* in the first row. (A *Data Name* can be comprised any characters available on the keyboard.) If **Data Variable Names in First Row** is not checked then *Statistician* will assign a *Data Name* to the data.

(b) Importing Data Variables

To import *Data Variables* into *Statistician*, click the **Manage Data** button on the *Statistician* ribbon and then click the **Import Data Variables** button on the **Manage Data** form. An inputbox “**Select a Range ...**” will appear (see figure (1)). Highlight the data range which contains the *Data Name* and *Data Observations* and then click the **OK** button.

Figure (1)



When importing a large set of *Data Variables* with many *Data Observations*, the user can more easily select the import range with the use of the <Shift> and <End> keyboard keys. Firstly select the top and left most cell of the import range. Holding down the <Shift> key and then pressing the <End> key followed by either <↓> or <→> will highlight the right most and bottom most cells in a contiguous range of cells. Then click the **OK** button to import the *Data Variables*.

The **Data Variables** listbox will then display the name of each of the *Data Variables*. The range of numbers in the square brackets before each of the *Data Variable* names in the listbox indicates the range of the lowest and highest numeric *Data Observations* in each column. If any non numeric *Data Observation* lies within this range then an asterisk will appear in the range between the square brackets.

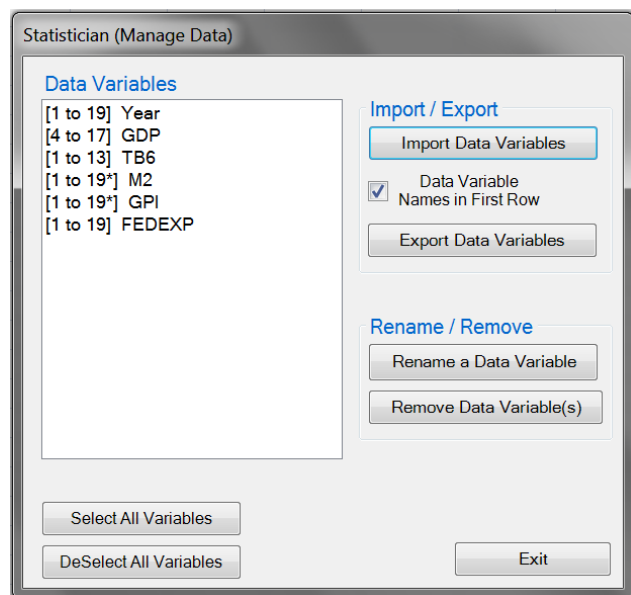
For example, if a spreadsheet containing the data in figure (2) is imported,

Figure (2)

Year	GDP	TB6	M2	GPI	FEDEXP
1970		6.562	626.4	436.2	198.6
1971		4.511	710.1		216.6
1972		4.466	802.1		240
1973	4123.4	7.178	855.2	606.5	259.7
1974	4099	7.926		561.7	291.2
1975	4084.4	6.122			345.4
1976	4311.7	5.266		555.5	371.9
1977	4511.8	5.51	1269.9	639.4	405
1978	4760.6	7.572	1365.5	713	444.2
1979	4912.1	10.017	1473.1	735.4	489.6
1980	4900.9	11.374	1599.1	655.3	576.6
1981	5021	13.776	1754.6	715.6	659.3
1982	4913.3	11.084	1909.5	615.2	732.1
1983	5132.3		2126	673.7	797.8
1984	5505.2		2309.7	871.5	856.1
1985	5717.1		2495.4	863.4	924.6
1986	5912.4		2732.1	857.7	978.5
1987			2831.1	879.3	1018.4
1988			2994.3	902.8	1066.2

then the **Data Variables** listbox will appear as in figure (3).

Figure (3)



Note the asterisk in the sample range between the square brackets for the *Data Variables* M2 and GPI which both contain non numeric data (blank cells).

The above process can be repeated any number of times to load data from various parts of a workbook. When analysis is performed upon *Data Variables*, *Statistician* will often automatically adjust the range of data for analysis to include only valid numerical observations. For example, if the correlation between *Data Variables* GDP and GPI in figure (3) is generated, then the sample range employed will be [4 to 15], (unless overridden by a user defined sample range). If the workbook which contains the *Data Variables* is saved upon exiting Excel, the *Data Variables* will also be saved and available for use the next time the workbook is opened.

(c) Selecting Data Variables

Data Variables can be selected by left clicking the *Data Variable* name in the **Data Variables** listbox. *Data Variables* can also be de-selected by again left clicking the *Data Variable* name. All of the *Data Variables* in the **Data Variables** listbox can be selected (or de-selected) by clicking the **Select All Variables** or **DeSelect All Variables** buttons respectively.

(d) Exporting Data Variables

To export data from *Statistician* to a spreadsheet, select one or more *Data Variables* from the **Data Variables** list box. By clicking the **Export Data Variables** button the user is presented with the “**Select a Range ...**” inputbox which is used select the top left output cell of the output area in a spreadsheet.

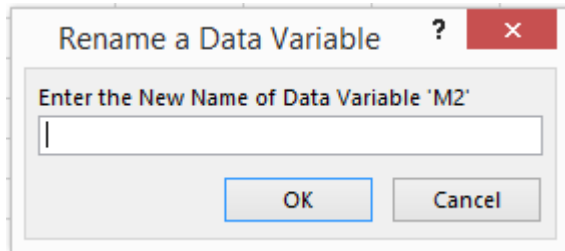
(e) Removing Data Variables

To remove one of more *Data Variables* from *Statistician*, select the *Data Variables* in the **Data Variables** list box that will be removed and then click the **Remove Selected Variable(s)** button.

(f) Renaming a Data Variable

To rename a *Data Variable* select one *Data Variable* from the **Data Variables** list box and then click the **Rename a Data Variable** button. The **Rename a Data Variable** inputbox (see figure (4)) is then presented to the user. Enter the new name for the *Data Variable* into the textbox and then click **OK**.

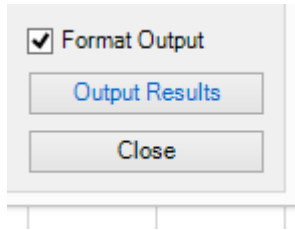
Figure (4)



Section (2) - Controls Common to all Forms

Statistician has a number of controls that are common to many forms. On the bottom right of most forms, users will see two buttons and a checkbox as displayed in figure (1).

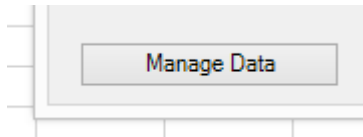
Figure (1)



The bottom button on the form (**Close**), closes the current form and returns the user to Excel. The button second from the bottom of the form (**Output Results**), initiates the statistical analysis and outputs the results to an Excel worksheet after presenting the user with an inputbox which can be used to select the top left output cell with the mouse.

If the **Format Output** checkbox is checked, *Statistician* will align cells, bold face headings and autofit various columns of the output. On the bottom left of all forms is the **Manage Data** button (see figure (2)).

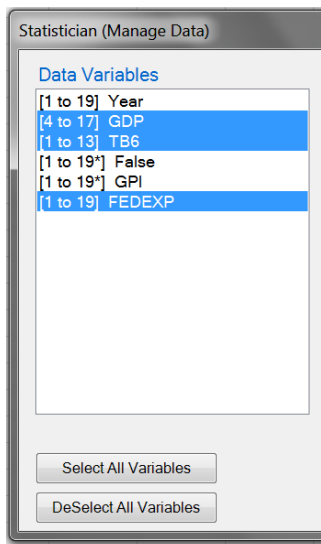
Figure (2)



Clicking the **Manage Data** button opens the **Manage Data** form which allows users to import, export, rename and remove *Data Variables*. Clicking **Return** on the bottom right of the **Manage Data** form returns the user to the current form.

Often the user is required to select two or more *Data Variables* for analysis. These *Data Variables* are selected from the **Data Variables** listbox (see figure (3)).

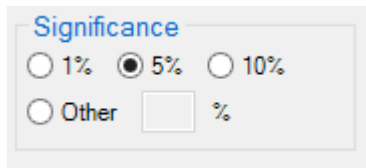
Figure (3)



Data Variables used in a statistical analysis are selected by clicking the name of the *Data Variable* in the listbox. Clicking the *Data Variable* name again will de-select the *Data Variable*. If the **Select All Variables** button is clicked then all of the *Data Variables* will be selected. Clicking the **DeSelect All Variables** button will de-select all of the *Data Variables*. When a *Data Variable* is selected its background will turn blue.

Often the user is required to select a level of significance to perform a statistical test. This level of significance is selected from the **Significance** groupbox. Most statistical tests are conducted at the 1%, 5% or 10% level. These significance levels are selected by clicking the corresponding radiobutton. If some other level of significance is required then the user clicks the **Other** radiobutton and then enters the level of significance in the adjacent textbox (see figure (4)).

Figure (4)



The image shows a software interface element titled "Significance" in blue text. It contains four radio buttons: "1%", "5%", "10%", and "Other". The "5%" radio button is selected, indicated by a black dot in the center. To the right of the "Other" radio button is a small, empty rectangular text input field, followed by a percent sign (%).

Section (3) – Tools

The **Tools** button on the *Statistician* ribbon tab provides the user with five menu items. These are:

- (i) Summary Statistics,
- (ii) Covariance and Correlation,
- (iii) Autocorrelation Function,
- (iv) Statistical Tables,
- (v) Sort and Rank.

Each of these menu items are described in detail below.

(a) Summary Statistics

The **Summary Statistics** form is selected by clicking the **Summary Statistics** menu item from the **Tools** button on the *Statistician* ribbon tab.

Statistician (Summary Statistics)

Data Variables

Extremes

- Count
- Sum
- Minimum
- Maximum

Location

- Mean
- Median
- Mode

Higher Moments

- Skewness
- Std Error (Skewness)
- Excess Kurtosis
- Std Error (Excess Kurtosis)

Spread

- Range
- Interquartile Range
- Standard Deviation (Sample)
- Standard Deviation (Population)
- Variance (Sample)
- Variance (Population)
- Sum Of Squares
- Mean Square Error
- Root Mean Square Error
- Mean Absolute Deviation

Other

- Jacque Bera Test Statistic
- Durban Watson Test Statistic

Format Output

Select All Variables

DeSelect All Variable

Manage Data

Select All Statistics

DeSelect All Statistics

Output Results

Close

The **Summary Statistics** form generates various summary (or descriptive) statistics on the selected *Data Variables*. To generate the statistics, the relevant data variables are selected from the **Data Variables** list box. The statistics to be outputted are selected by clicking (checking) the various checkboxes in the body of the form. If the **Select All Statistics** button is clicked then all of the selected statistics will be outputted. Clicking the **DeSelect All Statistics** button de-selects all of the checkboxes.

Description of the Summary Statistics

(Assume n observations with x_i being the i^{th} observation on a selected *Data Variable*)

Count – The number of observations.

Minimum – The smallest observation.

Maximum – The largest observation.

Sum – The sum of all observations.

Arithmetic Mean – The arithmetic mean of all observations. It is defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Median – The middle observation. The observations are initially sorted in ascending order. If there are an odd number of observations, the median is the middle observation. If there is an even number of observations, then the median is the average of the two middle observations.

Mode – The most common observation. If no mode exists then N/A is reported. If more than one mode exists then the first mode is reported.

Range – The difference of the largest and smallest observation.

Inter-Quartile Range – A measure of spread based upon the difference of the observations at the first and third quartile. The observations are initially sorted in ascending order. If there is an even number of observations, then the data is divided into two groups, a group with the highest numbers and a group with the lowest numbers. If there are an odd number of observations the middle observation is discarded and the two groups of highest and lowest number are then formed. The median of the lowest set of numbers is then subtracted from the median of the highest set of numbers to give the inter-quartile range.

Standard Deviation (Sample) - A measure of the spread of the population based upon a sample of observations. It is defined as:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Standard Deviation (Population) - A measure of the spread of the population based upon all population observations. It is defined as:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Variance (Sample) - A measure of the spread of the population based upon a sample of observations. It is defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Variance (Population) - A measure of the spread of the population based upon all population observations. It is defined as:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Sum of Squares - The sum of the squares of all observations. It is defined as:

$$SS = \sum_{i=1}^n x_i^2$$

Mean Square - The average of the sum of squares. It is defined as:

$$MS = \frac{SS}{n}$$

Root Mean Square - The square root of the average of the sum of squares. It is defined as:

$$RMS = \sqrt{MS}$$

Mean Absolute Deviation – The average of absolute deviations from the mean. It is defined as:

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

where \bar{x} is the mean of x .

Skewness – A measure of the magnitude of observations in the tails of the distribution of observations. It is defined as:

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Standard Error of Skewness – A measure of the significance of the skewness estimate. It is defined as:

$$Std\ Error\ of\ Skewness = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

Excess Kurtosis – A measure of the magnitude of observations in both tails of the distribution relative to the normal distribution. It is defined as:

$$Excess\ Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Standard Error of Excess Kurtosis - A measure of the significance of the skewness estimate. It is defined as:

$$Std\ Error\ of\ Excess\ Kurtosis = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$$

Jacque-Bera Test Statistic – A test statistic to measure the normality of the distribution of a *Data Variable*. The Jacque-Bera statistic has a $\chi^2_{(2)}$ distribution. It is defined as:

$$Jacque\ Bera = n \left(\frac{Skewness^2}{6} + \frac{Excess\ Kurtosis^2}{24} \right)$$

Durban Watson Test Statistic – A test for autocorrelation in the *Data Variable* and is given by:

$$DW = \frac{\sum_{i=2}^n (x_i - x_{i-1})^2}{\sum_{i=1}^n x_i^2}$$

Critical values for the Durban-Watson test statistic are taken from a table of Durban-Watson critical values.

(b) Covariance and Correlation

The **Covariance and Correlation** form is selected by clicking the **Covariance and Correlation** button in the **Tools** groupbox on the *Statistician* ribbon tab.

The screenshot shows the 'Statistician (Covariance and Correlation)' dialog box. It features a 'Data Variables' list on the left, which is currently empty. Below this list are three buttons: 'Select All Variables', 'DeSelect All Variables', and 'Manage Data'. To the right of the list are three sections of options: 'Table Body' with radio buttons for 'Covariance Matrix', 'Correlation' (with 'Pearson' selected), 'Spearman', 'Kendall (tau-a)', and 'Kendall (tau-b)'; 'Output' with radio buttons for 'Correlogram' (selected) and 'Table'; and 'Sample Range' with radio buttons for 'Use All Observations' (selected) and 'User Defined', followed by 'From' and 'to' input fields. At the bottom right, there is a checked 'Format Output' checkbox, an 'Output Results' button, and a 'Close' button.

The **Covariance and Correlation** form creates a variance-covariance matrix or correlation matrix of two or more selected *Data Variables*.

The statistics that are reported are defined as follows:

Covariance

The sample covariance between two *Data Variables* x and y is given as:

$$C_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

where \bar{x} and \bar{y} denote the mean of random variables x and y .

Pearson Product-Moment Correlation Coefficient

The correlation between two variables (often referred to as the Pearson product-moment correlation coefficient) is a measure of association and is in the range $[-1,1]$.

The sample correlation coefficient between two random variables is given as:

$$r_{x,y} = \frac{C_{x,y}}{s_x s_y}$$

where s_x and s_y are the sample standard deviations of random variables x and y respectively. The test statistic to determine the significance of $r_{x,y}$ under the null hypothesis that the population correlation ($\rho_{x,y}$) is equal to zero is given as:

$$t = \frac{r_{x,y} \sqrt{n-2}}{\sqrt{1-r_{x,y}^2}}$$

This test statistic has a student-t distribution with $n - 2$ degrees of freedom.

Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient of a sample of observations from a population is denoted by (r_s). It is calculated as follows. The observations in the sample are sorted in ascending order and 1 is assigned to the number with the lowest value, 2 is assigned to the number with the next lowest value and so on until n is assigned to the observation with the highest value.

When no tied ranks are present in the data, the Spearman rank correlation coefficient can be calculated as:

$$r_s = \frac{1 - 6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the rank of adjacent observations on random variables x and y , that is ($d_i = R(x_i) - R(y_i)$) where $R(x_i)$ and $R(y_i)$ are the ranks of observation x_i and y_i respectively. When no tied ranks are present in the data, the above formulation of the Spearman rank correlation coefficient is equivalent to calculating the Pearson product-moment correlation between the ranks of the two random variables. If there are tied ranks present in the data, the above formulation of the Spearman rank correlation coefficient will inflate the estimate of $|r_s|$.

One method to correctly calculate the Spearman rank correlation coefficient when tied ranks are present in the random variables is to (a), rank the observations on both random variables and then (b), for a set of observations in a random variable with equal value, assign a tied rank which is the average of the corresponding raw ranks. The Pearson product-moment correlation coefficient is then calculated on the tie corrected rankings. An alternative method for calculating the Spearman rank order coefficient when tied ranks are present in the random variables is as follows. The number of distinct tied ranks for each random variable x and y is denoted by n_x and n_y respectively. For random variable x , the number of observations in i^{th} group of tied observations is denoted by $n_{x,i}$. The quantity T_x for random variable x is defined as $T_x = \sum_{i=1}^{n_x} (n_{x,i}^3 - n_{x,i})$. Define S_x as $S_x = \frac{n^3 - n - T_x}{12}$. For random variable y , S_y is similarly defined. The tie corrected Spearman rank correlation coefficient (r_s^*) is then given as:

$$r_s^* = \frac{S_x + S_y - \sum_{i=1}^n d_i^2}{2\sqrt{S_x S_y}}$$

where d_i^2 is the squared difference between the corresponding ranks of observations in variables X and Y .

In the large sample case where $n > 10$, the test statistic under the null hypothesis that the population correlation (ρ_s) is equal to zero is given as:

$$t = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

The test statistic has a student-t distribution with $n - 2$ degrees of freedom.

In the small sample case where $n \leq 10$, the critical value to test the null hypothesis $H_0: \rho_s = 0$, should be taken from a table of Spearman's critical values. If the alternative hypothesis is $H_1: \rho_s \neq 0$, the null hypothesis is rejected if $|r_s|$ is greater than or equal to the two tailed critical value for a given level of significance (α). If the alternative

hypothesis is $H_1: \rho_S > 0$, the null hypothesis is rejected if r_S is positive and $|r_S|$ is greater than or equal to the one tailed critical value. If the alternative hypothesis is $H_1: \rho_S < 0$, the null hypothesis is rejected if r_S is negative and $|r_S|$ is greater than or equal to the one tailed critical value.

Kendall tau

The Kendall correlation coefficient (often referred to as the Kendall tau or Kendall tau-a coefficient) measures the association between the ranks of two ordinal variables and is often denoted by (τ). The Kendall correlation coefficient is given as:

$$\tau = \frac{2(n_C - n_D)}{n(n - 1)}$$

where n_C and n_D denote the number of concordant and discordant pairs respectively and n is the sample size. If tied ranks are present in either of the x or y random variables, the above methodology must be modified. In this case, the tie corrected Kendall tau (or Kendall tau-b coefficient) can be employed. Let the number of distinct tied ranks for each random variable x and y be denoted by t_x and t_y respectively. The tie adjusted Kendall tau (tau-b) is then defined as:

$$\tau = \frac{2(n_C - n_D)}{\sqrt{n(n - 1) - t_x} \sqrt{n(n - 1) - t_y}}$$

In the large sample case when $n > 10$, the null hypothesis $H_0: \tau = 0$ for both tau-a and tau-b can be tested reasonably accurately with the statistic:

$$z = \frac{3\tau\sqrt{n(n - 1)}}{\sqrt{2(2n + 5)}}$$

The z-statistic has a standard normal distribution. In the small sample case where $n \leq 10$, the critical value to test the null hypothesis $H_0: \tau = 0$, should be taken from a table of Kendall critical values.

In the **Output** groupbox users can select either **Correlogram** or **Table**. If **Correlogram** is selected only the correlation estimates are reported. If **Table** is selected a t-statistic is reported under the null hypothesis that the correlation is equal to zero for Pearson and Spearman correlations. If **Table** is selected for the Kendal tau-a or tau-b correlation coefficient a z-statistic is reported under the null hypothesis that the correlation is equal to zero. The number of concordant and discordant pairs is reported and in the case of the Kendal tau-b correlation coefficient the number of ties is reported.

(c) Autocorrelation

The autocorrelation form allows users to identify the autocorrelation and partial autocorrelation function for a time series. The k^{th} autocovariance for data variable y_t is defined as $\gamma_k = \text{Cov}(y_t, y_{t-k})$ and the k^{th} autocorrelation for data variable y_t is then defined as $\rho_k = \gamma_k / \gamma_0$. An estimate of ρ_k is given as:

$$r_k = \frac{\sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

where \bar{y} is the unconditional mean of series $\{y_t\}$ and n is the number of observations.

Each sample autocorrelation estimate is distributed with zero mean and has approximate variance $1/n$ under the null hypothesis that $\{y_t\}$ is a white noise process. Bartlett (1946) proposes that under the null hypothesis that $\{y_t\}$ is a white noise process, the variance of a sample autocorrelation is given as:

$$\text{Var}(r_k) = \frac{1}{n} (1 + 2r_1^2 + \dots + 2r_{k-1}^2)$$

A p-value to test the significance of an autocorrelation estimate is derived from the test statistic given as $r_k / \sqrt{\text{Var}(r_k)} \sim N(0,1)$.

Two tests are commonly employed to test for autocorrelation in the time series $\{y_t\}$. Under the null hypothesis that $\{y_t\}$ is a white noise process, the Box-Pierce (1970) test statistic is given as:

$$Q = n \sum_{k=1}^p r_k^2$$

The Ljung-Box (1979) test statistic is given as:

$$Q' = n(n+2) \sum_{k=1}^p \frac{r_k^2}{n-k}$$

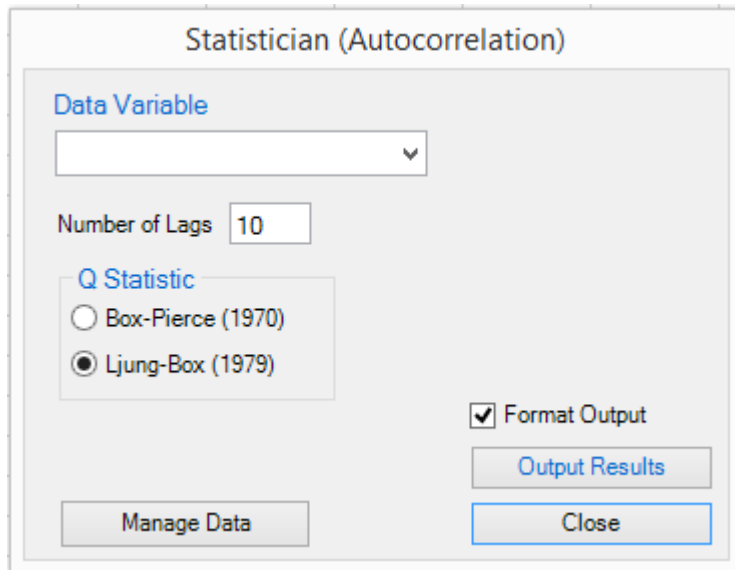
and is generally regarded as having better finite sample properties than the Box-Pierce (1970) statistic. Both the Q and Q' statistics are distributed as a χ_p^2 random variable.

An estimate of the k^{th} partial autocorrelation (ϕ_k) is taken from the linear regression:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \dots + \beta_{k-1} y_{t-(k-1)} + \phi_k y_{t-k} + \varepsilon_t$$

where $E[\varepsilon_t] = 0$.

Using Statistician



The screenshot shows a dialog box titled "Statistician (Autocorrelation)". It contains the following elements:

- A "Data Variable" label above a dropdown menu.
- A "Number of Lags" label next to a text box containing the value "10".
- A "Q Statistic" label above a group box containing two radio buttons: "Box-Pierce (1970)" and "Ljung-Box (1979)". The "Ljung-Box (1979)" option is selected.
- A "Format Output" checkbox, which is checked.
- Four buttons: "Manage Data" (bottom left), "Output Results" (top right), and "Close" (bottom right).

The **Autocorrelation** form is selected by selecting the **Autocorrelation** menu item from the **Tools** button on the *Statistician* ribbon. The user selects the *Data Variable* for analysis from the **Data Variable** combobox.

The user enters the number of lags of the selected *Data Variable* to be tested in the **Number of Lags** textbox (the default is 10). Either the Box-Pierce (1970) or Ljung-Box (1979) test statistic can be selected from the **Q Statistic** groupbox. *Statistician* outputs the autocorrelation, partial autocorrelation, Q statistic and p value at each lag of the selected *Data Variable*.

(d) Statistical Tables

The **Statistical Tables** form is selected by clicking the **Statistical Tables** menu item from the **Tools** button on the *Statistician* ribbon tab.

The image shows a dialog box titled "Statistician (Statistical Tables)". It contains a list of statistical tests, each with a radio button. The "Normal" test is selected. To the right of the list is an "Alpha" dropdown menu with "α =" and a downward arrow. Below the list is a "Format Output" checkbox, which is checked. At the bottom of the dialog are three buttons: "Manage Data", "Output Results", and "Close".

Statistical Test	Selected
Normal	Yes
Student t	No
F	No
Chi Squared	No
Binomial	No
Binomial (Cumulative)	No
Poisson	No
Poisson (Cumulative)	No
Mann Whitney (One Tail)	No
Mann Whitney (TwoTails)	No
Wilcoxon Rank Sum	No
Wilcoxon Signed Rank Sum	No
Runs (Lower Critical Values)	No
Runs (Upper Critical Values)	No
Spearman Rho	No
Studentized Range	No
Durbin Watson	No
Kendall tau	No

Alpha: α = [dropdown]

Format Output

Buttons: Manage Data, Output Results, Close

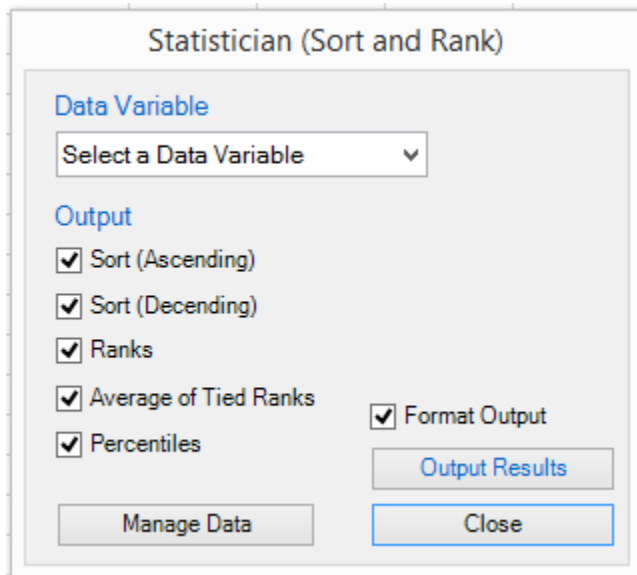
Statistical tables that can be outputted to an Excel spreadsheet are as follows:

- (1) **Normal Distribution** – standard normal probabilities in the range $[0,z]$ where z is a normal critical value.
- (2) **Student t Distribution** – Critical values from the student t distribution for both one and two tailed tests. For a one tailed test, significance values are ($\alpha = 0.1, 0.05, 0.025, 0.01, 0.005, 0.001$) and for a two tailed test, significance values are ($\alpha = 0.2, 0.1, 0.05, 0.02, 0.01, 0.002$).
- (3) **F distribution** – Critical values for the F distribution for a given number of degrees of freedom. The significance level of the table is selected from the **Alpha** textbox and can equal ($\alpha = 0.1, 0.05, 0.01$).
- (4) **χ^2 Distribution** – Critical values for the χ^2 Distribution for a given number of degrees of freedom. Significance values are:
($\alpha = 0.995, 0.99, 0.975, 0.95, 0.9, 0.1, 0.05, 0.025, 0.01, 0.005$).
- (5) **Binomial Distribution** – Table of binomial probabilities for a given probability of success ($p = 0.05, 0.1, \dots, 0.9, 0.95$) and a given number of trials ($n = 1, \dots, 10$).
- (6) **Cumulative Binomial Distribution** – Table of cumulative binomial probabilities for a given probability of success ($p = 0.05, 0.1, \dots, 0.9, 0.95$) and a given number of trials ($n = 1, \dots, 10$).
- (7) **Poisson Distribution** – Table of Poisson distribution probabilities for a given mean ($\lambda = 0.1, 0.2, \dots, 1, 2, \dots, 10$) and a given number of successes ($x = 0, 1, \dots, 20$).
- (8) **Cumulative Poisson Distribution** – Table of cumulative Poisson distribution probabilities for a given mean ($\lambda = 0.1, 0.2, \dots, 1, 2, \dots, 10$) and a given number of successes ($x = 0, 1, \dots, 20$).
- (9) **Mann Whitney (One Tail)** – Critical values for a Mann Whitney one tailed test where $n_1 \leq 20$ and $n_2 \leq 20$. The significance level of the table is selected from the **Alpha** textbox and can equal ($\alpha = 0.05, 0.01$).
- (10) **Mann Whitney (Two Tails)** – Critical values for a Mann Whitney two tailed test where $n_1 \leq 20$ and $n_2 \leq 20$. The significance level of the table is selected from the **Alpha** textbox and can equal ($\alpha = 0.05, 0.01$).
- (11) **Wilcoxon Rank Sum** – Upper and lower critical values for the Wilcoxon rank sum test where ($4 \leq n_1, n_2 \leq 10$). The significance level of the table is selected from the **Alpha** textbox and can equal ($\alpha = 0.1, 0.05, 0.01$).
- (12) **Wilcoxon Signed Rank Sum** – Critical values for the Wilcoxon signed rank sum test where ($5 \leq n \leq 30$). For a one tailed test, significance values are ($\alpha = 0.05, 0.025, 0.01, 0.005$) and for a two tailed test, significance values are ($\alpha = 0.1, 0.05, 0.02, 0.01$).
- (13) **Runs (Lower Critical Values)** – Lower critical values for the Runs test where ($2 \leq n_1, n_2 \leq 20$) and the significance value is ($\alpha = 0.05$).
- (14) **Runs (Upper Critical Values)** – Upper critical values for the Runs test where ($2 \leq n_1, n_2 \leq 20$) and the significance value is ($\alpha = 0.05$).

- (15) **Spearman Rho (ρ)** - Critical values for the Spearman rho correlation coefficient where ($5 \leq n \leq 30$) . For a one tailed test, significance values are ($\alpha = 0.05, 0.025, 0.01, 0.005$) and for a two tailed test, significance values are ($\alpha = 0.1, 0.05, 0.02, 0.01$).
- (16) **Studentized Range** - Critical values for the Studentized range distribution for a given number of degrees of freedom and a given number of treatments ($k = 2, \dots, 20$). The significance level of the table is selected from the **Alpha** textbox and can equal ($\alpha = 0.1, 0.05, 0.01$).
- (17) **Durbin Watson** - Upper and lower critical for the Durbin Watson test for a given number observations in the regression and for a given number of regressors. The significance level of the table is selected from the **Alpha** textbox and can equal ($\alpha = 0.05, 0.01$).
- (18) **Kendall tau (τ)** - Critical values for the Kendall tau correlation coefficient where ($4 \leq n \leq 60$) . For a one tailed test, significance values are ($\alpha = 0.01, 0.05, 0.025, 0.01, 0.005, 0.001$) and for a two tailed test, significance values are ($\alpha = 0.2, 0.1, 0.05, 0.02, 0.01, 0.005$).

(e) Sort and Rank

The **Sort and Rank** form is selected by clicking the **Sort and Rank** menu item from the **Tools** button on the *Statistician* ribbon tab.

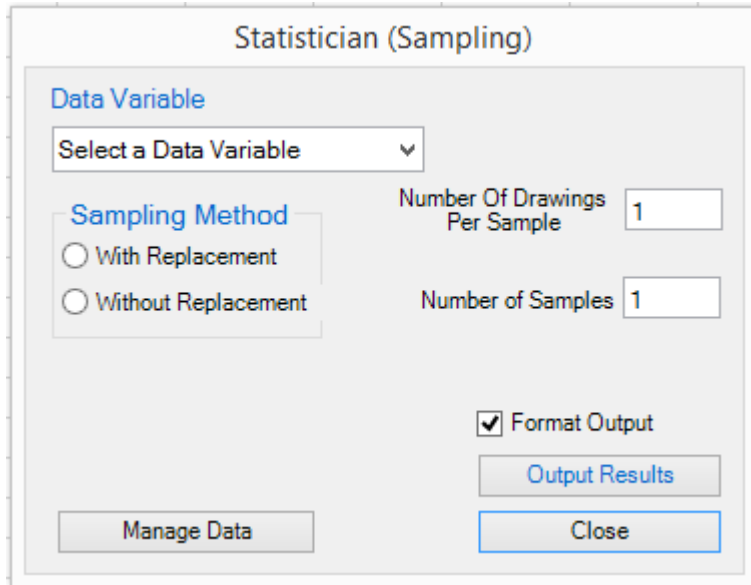


The **Sort and Rank** form allows to user to output the observations in the selected *Data Variable* in various forms. By checking the corresponding check box, the user can output:

- (1) The *Data Variable* sorted in ascending order,
- (2) The *Data Variable* sorted in descending order,
- (3) The rank of each observation in the *Data Variable*,
- (4) The “tied” rank of each observation in the *Data Variable* where all ranks with equal observations are assigned a tied rank which is the average of the rank of all equal observations.
- (5) The percentile of each observation in the *Data Variable*. The percentile of an observation is calculated by counting the number of observations with magnitude greater than the observation (*Number Above*), and counting the number of observations with magnitude less than the observation (*Number Below*). The observation’s percentile is then calculated as $(\text{Number Above} / (\text{Number Above} + \text{Number Below})) \times 100\%$.

(f) Sampling

The **Sampling** form is selected by clicking the **Sampling** menu item from the **Tools** button on the *Statistician* ribbon tab.



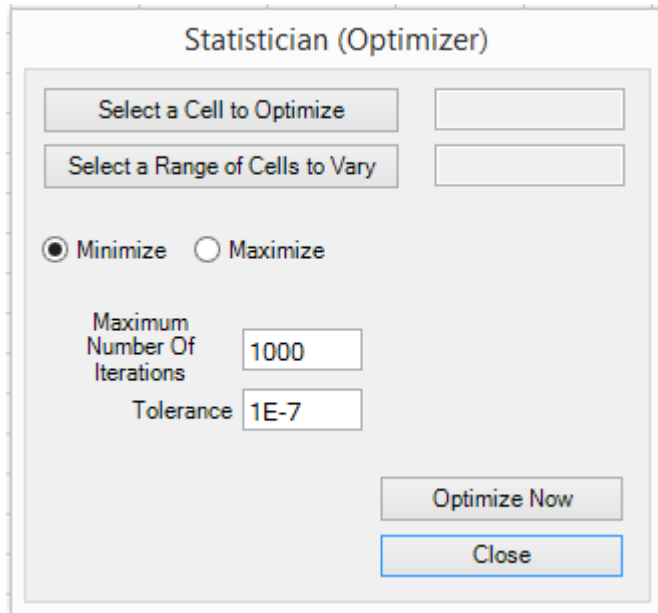
The image shows a dialog box titled "Statistician (Sampling)". It contains the following elements:

- A "Data Variable" section with a dropdown menu labeled "Select a Data Variable".
- A "Sampling Method" groupbox containing two radio buttons: "With Replacement" and "Without Replacement".
- A "Number Of Drawings Per Sample" text box with the value "1".
- A "Number of Samples" text box with the value "1".
- A checked checkbox labeled "Format Output".
- Four buttons: "Manage Data", "Output Results", and "Close".

The user selects a *Data Variable* from which random samples will be taken. The type of sampling, either with or without replacement is selected in the **Sampling Method** groupbox. The number of observations in each sample is written in the **Number Of Drawings Per Sample** textbox. The number of samples to be taken is written in the **Number Of Samples** textbox.

(g) Optimizer

The **Optimizer** form is selected by clicking the **Optimizer** menu item from the **Tools** button on the *Statistician* ribbon tab.



The image shows a dialog box titled "Statistician (Optimizer)". It contains the following elements:

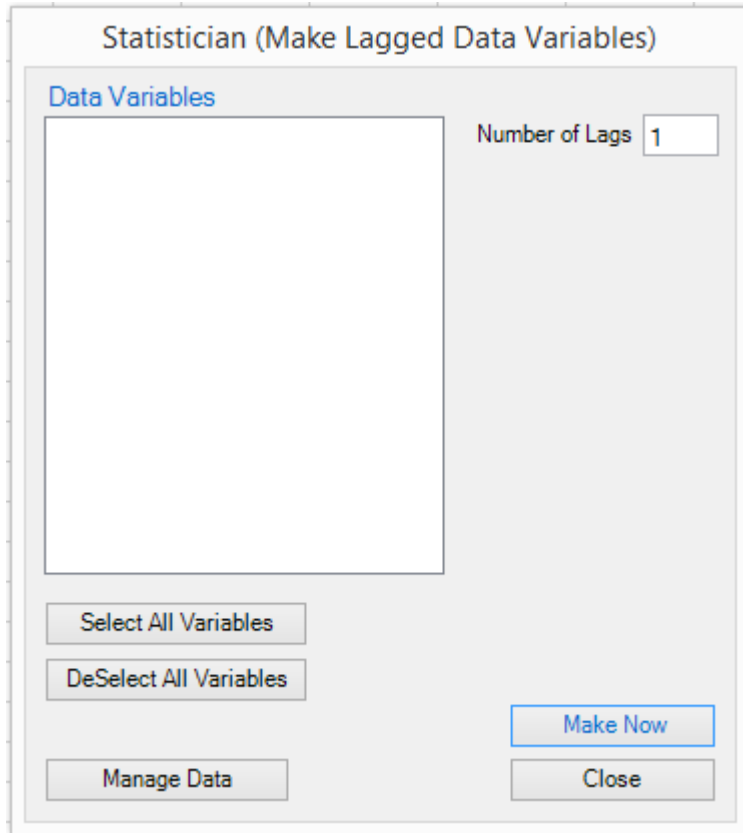
- Two buttons: "Select a Cell to Optimize" and "Select a Range of Cells to Vary", each followed by an empty text input field.
- Two radio buttons: "Minimize" (selected) and "Maximize".
- Two text input fields: "Maximum Number Of Iterations" with the value "1000" and "Tolerance" with the value "1E-7".
- Two buttons at the bottom: "Optimize Now" and "Close".

The Optimizer tool will either minimize or maximize the value of a cell by changing the value in some other cells. The cell to be optimized is selected by clicking the **Select a Cell to Optimize** button and the range of cells to be varied is selected by clicking the **Select a Range of Cells to Vary** buttons. The maximum number of iterations is written in the **Maximum Number of Iterations** textbox and the accuracy or tolerance of the optimization is written in the **Tolerance** textbox.

Optimizer has some advantages and disadvantages in comparison to Excel Solver. The two disadvantages are that Optimizer is slower than Solver and does not do constrained optimization. However the speed of optimization is of little consequence in practice unless highly numerically demanding tasks are performed. On the positive side, Optimizer is simpler to use than Solver and does sometimes arrive at a better solution than Solver because of the inbuilt optimization algorithms. Thus *Statistician* users have the choice of two optimization tools.

(h) Make Lagged Data Variables

The **Make Lagged Data Variables** form is selected by clicking the **Make Lagged Data Variables** menu item from the **Tools** button on the *Statistician* ribbon tab.



The **Make Lagged Data Variables** tool will create lags of a predefined *Data Variable*. To create a set of lagged *Data Variables* the user selects one or more *Data Variables* and then sets then number of desired lags from the **Number of Lags** textbox. Clicking **Make Now** creates the lagged *Data Variables*.

Section (4) – Standard Tests

The **Standard Tests** button on the *Statistician* ribbon tab provides the user with five menu items. These are:

- (i) Test Population Mean,
- (ii) Test Difference in Population Means: Independent Samples,
- (iii) Test Difference in Matched Pairs: Dependent Samples,
- (iv) Test Population Proportion, and
- (v) Test Difference in Population Proportions.

Each of these menu item selections are described in detail below.

(a) Test Population Mean

The **Test Population Mean** form is activated from the **Standard Tests** button on the *Statistician* ribbon tab.

The screenshot shows the "Statistician (Test Population Mean)" dialog box. It is divided into several sections:

- Hypothesis Test ($\alpha = 0.05$)**:
 - Null Hypothesis**: "Mean of" dropdown menu (set to "Select a Data Variable"), followed by an equals sign dropdown and a text input field for the "Hypothesized Mean".
 - Alternative Hypothesis**: "Mean of" text input field, followed by a not-equals sign dropdown and a text input field.
- Standard Deviation**:
 - Unknown
 - Known [text input field]
- Significance**:
 - 1%
 - 5%
 - 10%
 - Other [text input field] %
- Confidence Interval**:
 - Output a 95% Confidence Interval for the Mean
 - Output the Sample Size to Obtain a Confidence Interval Estimate to Within [text input field] of the True Population Mean.
 - Significant Digits: [2] dropdown menu
- Format Output**: Format Output
- Buttons: "Manage Data", "Output Results", and "Close".

The **Test Population Mean** form has two primary functions:

- (a) To test the hypothesis that the population mean (μ) of a selected *Data Variable* is equal to, greater than or equal to, or less than or equal to, a hypothesized value (v).
- (b) To produce a confidence interval for the mean of a *Data Variable*.

The following sets of null and alternative hypotheses can be tested.

- (a) $H_0: \mu = v$
 $H_1: \mu \neq v$ (two tailed test) or,
 $H_1: \mu > v$ or,
 $H_1: \mu < v$
- (b) $H_0: \mu \geq v$
 $H_1: \mu < v$
- (c) $H_0: \mu \leq v$
 $H_1: \mu > v$

The test statistic (z), for a test with a known standard deviation is given as:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where \bar{x} is the sample mean, μ is the hypothesized mean, σ is the known population standard deviation and n is the number of observations in the *Data Variable*. The test statistic (t), for a test with an unknown population standard deviation is given as:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where s is the sample standard deviation.

The critical values for a z test are taken from a standard normal distribution and for a t test are taken from a student-t distribution. For a two tailed test, the critical value for a test with a known standard deviation is given as $z_{\alpha/2}$, and the critical value for a two tailed test calculated with an unknown standard deviation is $t_{\alpha/2, n-1}$, where $n - 1$ is the degrees of freedom of the t distribution. For a one tailed test, the critical value for a test with a known standard deviation is given as z_{α} and the critical value for a one tailed test calculated with an unknown standard deviation is $t_{\alpha, n-1}$. The reported p-value is the probability of rejecting the null hypothesis when it is true.

When the standard deviation of the population is known, the null hypothesis is rejected if:

Alternative Hypothesis	Rejection Criterion
$H_1: \mu \neq v$	$ z > z_{\alpha/2}$
$H_1: \mu < v$	$z < z_{\alpha}$
$H_1: \mu > v$	$z > z_{\alpha}$

When the standard deviation of the population is unknown, the null hypothesis is rejected if:

Alternative Hypothesis	Rejection Criterion
$H_1: \mu \neq v$	$ t > t_{\alpha/2, n-1}$
$H_1: \mu < v$	$t < t_{\alpha, n-1}$
$H_1: \mu > v$	$t > t_{\alpha, n-1}$

The confidence level is given by $(1 - \alpha)$ expressed as a percentage, where α is the significance level. The confidence interval is given as:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

for a known population standard deviation, and

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

for an unknown population standard deviation.

To obtain a sample size required to construct an estimate of the population mean for a given confidence interval width and significance level, the required sample size is the lowest integer greater than:

$$\left(\frac{z_{\alpha/2} \sigma}{w} \right)^2$$

for a known standard deviation and

$$\left(\frac{t_{\alpha/2, n-1} s}{w} \right)^2$$

for an unknown standard deviation, where (w) is half of the confidence interval width.

Using Statistician (Test Population Mean)

The *Data Variable* to be tested is selected from the **Null Hypothesis** combobox in the **Hypothesis Test** groupbox. The hypothesized mean of the *Data Variable* is entered in the **Hypothesized Mean** textbox in the **Hypothesis Test** groupbox. The null hypothesis ($=, \geq, \leq$) and alternative hypothesis ($\neq, <, >$) are selected in the corresponding comboboxes in the **Hypothesis Test** groupbox.

If the standard deviation of the population is known then the **Known** radiobutton is selected in the **Standard Deviation** groupbox and a textbox is displayed where the known standard deviation can be entered. In this case a z test is performed. If the standard deviation of the population is unknown, then the **Unknown** radiobutton is selected. In this case the sample standard deviation of the selected *Data Variable* is calculated and a t test is performed.

When the **Output Results** button is clicked, *Statistician* outputs the:

- null hypothesis,
- alternative hypothesis,
- sample size,
- standard deviation (sample or known),
- test statistic,
- critical value,
- p-value,
- decision rule, and
- conclusion.

To obtain a sample size for a given confidence interval width, the **Output Sample Size ...** checkbox in the **Confidence Interval** groupbox is checked. The user then enters the desired distance from the 'true' population mean in the corresponding textbox. The number of significant digits after the decimal point in the confidence interval output is selected from the **Number of Significant Digits in Output** combobox, (the default is 2).

(b) Test Difference in Population Means: Independent Samples

The **Test Difference in Population Means: Independent Samples** form is activated from the **Standard Tests** button on the *Statistician* ribbon tab.

Statistician (Test Difference In Population Means - Independent Samples)

Hypothesis Test ($\alpha = 0.05$)

Null Hypothesis

Mean of - Mean of = Hypothesized Difference

Alternative Hypothesis

Mean of - Mean of \neq

Confidence Interval

Output a 95% Confidence Interval for the Difference in Means

Output the Sample Size to Obtain a Confidence Interval Estimate of the Difference in Means to Within of the True Population Difference.

Significant Digits

Standard Deviation

Unknown and Unequal Unknown and Equal

Known and Unequal Known and Equal

Std Dev of ??? Std Dev of ???

Output F Test for the Equality of Sample Variances

Significance

1% 5% 10%

Other %

Format Output

The **Test Difference in Population Means: Independent Samples** form has two primary functions:

- To test the hypothesis that the difference in the means (μ_1, μ_2) of two selected *Data Variables* is equal to, greater than or equal to, or less than or equal to, a hypothesized value (D).
- To produce a confidence interval for the difference in the means of two selected *Data Variables*.

The following set of null and alternative hypotheses can be tested.

- (a) $H_0: \mu_1 - \mu_2 = D$
 $H_1: \mu_1 - \mu_2 \neq D$ (two tailed test) or,
 $H_1: \mu_1 - \mu_2 > D$ or,
 $H_1: \mu_1 - \mu_2 < D$
- (b) $H_0: \mu_1 - \mu_2 \geq D$
 $H_1: \mu_1 - \mu_2 < D$
- (c) $H_0: \mu_1 - \mu_2 \leq D$
 $H_1: \mu_1 - \mu_2 > D$

The standard deviation of the two random variables may be known or unknown. They may also be assumed to be equal or unequal. This leaves four possible combinations of assumptions for the standard deviation of the two random variables. The distribution and test statistic for the difference in the means of the two random variables under these four assumptions is displayed in the following table.

Standard Deviation	Distribution	Test Statistic
Unknown and Unequal	Student-t	$\frac{(\bar{x}_1 - \bar{x}_2) - D}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Unknown and equal	Student-t	$\frac{(\bar{x}_1 - \bar{x}_2) - D}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ Where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
Known and unequal	Normal	$\frac{(\bar{x}_1 - \bar{x}_2) - D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Known and equal	Normal	$\frac{(\bar{x}_1 - \bar{x}_2) - D}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$

where s_1^2 and s_2^2 are the sample variances of each random variable, σ_1^2 and σ_2^2 are the known and unequal variances of each random variable and where σ^2 is the known and equal variance of each of the random variables. When the variances are unknown and unequal, the degrees of freedom of the t distribution is given by $d.f. = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left(\left(\frac{s_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2} \right)^2 / (n_2 - 1) \right)$ where $d.f.$ is rounded to the nearest integer. When the variances are unknown and equal the degrees of freedom are given by $d.f. = n_1 + n_2 - 2$. The variable $D = \mu_1 - \mu_2$, is the hypothesized difference between the means where μ_1 and μ_2 are the population means of the two random variables.

The critical values for a z test are taken from a standard normal distribution and for a t test are taken from a student-t distribution. For a two tailed test, the critical value for a test with known standard deviations is given as $z_{\alpha/2}$, and the critical value for a test calculated with unknown standard deviations is $t_{\alpha/2,df}$, where df is the degrees of freedom of the t distribution. For a one tailed test, the critical value for a test with known standard deviations is given as z_{α} and the critical value for a test calculated with unknown standard deviations is $t_{\alpha,df}$. The reported p-value is the probability of rejecting the null hypothesis when it is true.

When the standard deviation of the population is known, the null hypothesis is rejected if:

Alternative Hypothesis	Rejection Criterion
$H_1: \mu_1 - \mu_2 \neq D$	$ z > z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 < D$	$z < z_{\alpha}$
$H_1: \mu_1 - \mu_2 > D$	$z > z_{\alpha}$

When the standard deviation of the population is unknown, the null hypothesis is rejected if:

Alternative Hypothesis	Rejection Criterion
$H_1: \mu_1 - \mu_2 \neq D$	$ t > t_{\alpha/2, n-1}$
$H_1: \mu_1 - \mu_2 < D$	$t < t_{\alpha, n-1}$
$H_1: \mu_1 - \mu_2 > D$	$t > t_{\alpha, n-1}$

The following table displays the formulae for the construction of the confidence intervals where α is the level of significance.

Standard Deviation	Confidence Interval
Unknown and Unequal	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Unknown and equal	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df} \times \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$ <p>where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$</p>
Known and unequal	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Known and equal	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \times \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$

where s_1^2 and s_2^2 are the sample variances of each random variable, σ_1^2 and σ_2^2 are the known and unequal variances of each random variable and where σ^2 is the known and equal variance of each of the random variables. When the variances are unknown and unequal, the degrees of freedom of the t distribution is given by $d.f. = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left(\left(\frac{s_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2} \right)^2 / (n_2 - 1) \right)$ where $d.f.$ is rounded to the nearest integer. When the variances are unknown and equal the degrees of freedom are given by $d.f. = n_1 + n_2 - 2$.

To test the null hypothesis that the variances of the random variables are equal against the alternative hypothesis that the variances are unequal, an F statistic is calculated as s_1^2/s_2^2 . The F statistic has an F distribution with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator. The null hypothesis is rejected if $F < F_{1-\alpha/2, n_1, n_2}$ (lower critical value), or if $F > F_{\alpha/2, n_1, n_2}$ (upper critical value).

To obtain a sample size required to construct an estimate of the difference in population means for a given confidence interval width and significance level, the required sample size is the lowest integer greater than N in the following table.

Standard Deviation	Sample Size
Unknown and Unequal	$N = \frac{(t_{\alpha/2,df})^2 (s_1^2 + s_2^2)}{w^2}$
Unknown and equal	$N = \frac{(t_{\alpha/2,df})^2 (s_1^2 + s_2^2)}{w^2}$
Known and unequal	$N = \frac{(z_{\alpha/2})^2 (\sigma_1^2 + \sigma_2^2)}{w^2}$
Known and equal	$N = \frac{(z_{\alpha/2})^2 (2\sigma^2)}{w^2}$

Where s_1^2 and s_2^2 are the sample variances of each random variable, σ_1^2 and σ_2^2 are the known and unequal variances of each random variable, σ^2 is the known and equal variance of each of the random variables and where (w) is half of the confidence interval width.

Using Statistician (Test Difference in Population Means: Independent Samples)

The *Data Variables* to be tested are selected from the **Null Hypothesis** comboboxes in the **Hypothesis Test** groupbox. The hypothesized difference in the means of the *Data Variables* are entered in the **Hypothesized Difference** textbox in the **Hypothesis Test** groupbox. The null hypothesis ($=, \geq, \leq$) and alternative hypothesis ($\neq, <, >$) are selected in the corresponding comboboxes in the **Hypothesis Test** groupbox.

If the standard deviation of the populations is unknown and unequal or unknown and equal then one of the **Unknown and Unequal** or **Unknown and Equal** radiobuttons are selected in the **Standard Deviation** groupbox. In this case, the sample standard deviation of the selected *Data Variables* are calculated and a t test is performed. If the standard deviation of the populations is known and unequal or known and equal then one of the **Known and Unequal** or **Known and Equal** radiobuttons are selected in the **Standard Deviation** groupbox. The known standard deviations are entered in the

textbox(s) that are enabled upon making one of these selections. In this case a z test is performed.

When the **Output Results** button is clicked, *Statistician* outputs the:

- null hypothesis,
- alternative hypothesis,
- sample size of each *Data Variable* (n_1 and n_2),
- standard deviation each *Data Variable* (sample, known or pooled),
- test statistic,
- critical value,
- p-value,
- decision rule, and
- conclusion.

When the **Unknown and Unequal** or **Unknown and Equal** radiobutton is selected in the **Standard Deviation** groupbox, a checkbox titled **Output Test for the Equality of Sample Variances** is displayed which gives the user the option to additionally test the hypothesis that the variance of both selected *Data Variables* are equal. If this checkbox is checked then *Statistician* also outputs the null hypothesis, alternative hypothesis, sample variances, test statistic, critical value, decision rule and conclusion of the F test.

To obtain the sample size for the difference in means for a given confidence interval width, the **Output Sample Size ...** checkbox in the **Confidence Interval** groupbox is checked. The user then enters the desired distance from the 'true' population mean in the corresponding textbox. The number of significant digits after the decimal point in the confidence interval output is selected from the **Number of Significant Digits in Output** combobox, (the default is 2).

(c) Test Difference in Matched Pairs: Dependent Samples

The **Test Difference in Population Means: Dependent Samples** form is activated from the **Standard Tests** button on the *Statistician* ribbon tab.

Statistician (Test Matched Pairs - Dependent Samples)

Hypothesis Test ($\alpha = 0.05$)

Null Hypothesis
Average Difference of and = Hypothesized Average Difference

Alternative Hypothesis
Average Difference of and \neq

Confidence Interval

Output a 95% Confidence Interval for the Average Difference

Output the Sample Size to Obtain a Confidence Interval Estimate of the Average Difference to within of the Average Population Difference.

Significant Digits

Significance
 1% 5% 10%
 Other %

Format Output

The **Test Difference in Population Means: Dependent Samples** form is activated by selecting **Two Sample (Matched Pairs)** from the **Mean** combobox in the **Standard Tests** groupbox. The **Difference in Population Means: Dependent Samples** form has two primary functions:

- To test the hypothesis that the average difference in the matched pairs (μ_D) of two selected *Data Variables* is equal to, greater than or equal to or less than or equal to a hypothesized value (D).
- To produce a confidence interval for the average difference in the matched pairs of two selected *Data Variables*.

The following set of null and alternative hypotheses can be tested.

- (a) $H_0: \mu_D = D$
 $H_1: \mu_D \neq D$ (two tailed test) or,
 $H_1: \mu_D > D$ or,
 $H_1: \mu_D < D$
- (b) $H_0: \mu_D \geq D$
 $H_1: \mu_D < D$
- (c) $H_0: \mu_D \leq D$
 $H_1: \mu_D > D$

The sample average difference in matched pairs is given as:

$$\bar{X}_D = \frac{1}{n} \sum_{i=1}^n (x_{1,i} - x_{2,i}) = \bar{x}_1 - \bar{x}_2$$

where $x_{1,i}$ and $x_{2,i}$ are the i^{th} observation on the first and second selected *Data Variables* respectively. The quantities \bar{x}_1 and \bar{x}_2 are the means of the first and second selected *Data Variables* respectively. The sample standard deviation of the differences in matched pairs is given as:

$$S_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{1,i} - x_{2,i} - \bar{X}_D)^2}$$

The test statistic under the null hypothesis is given by:

$$t = \frac{\bar{X}_D - D}{S_D / \sqrt{n}}$$

and has a t distribution with $n - 1$ degrees of freedom.

The null hypothesis is rejected if:

Alternative Hypothesis	Rejection Criterion
$H_1: \mu_D \neq D$	$ t > t_{\alpha/2, n-1}$
$H_1: \mu_D < D$	$t < t_{\alpha, n-1}$
$H_1: \mu_D > D$	$t > t_{\alpha, n-1}$

The confidence interval for μ_D is given by:

$$\bar{X}_D \pm t_{\alpha/2, n-1} \frac{S_D}{\sqrt{n}}$$

The required sample size to estimate μ_D to within a particular value (w) is the lowest integer greater than N in the expression.

$$N = \frac{(t_{\alpha/2, n-1})^2 (2S_D^2)}{w^2}$$

Using Statistician (Test Difference in Matched Pairs: Dependent Samples)

The *Data Variables* to be tested are selected from the **Null Hypothesis** comboboxes in the **Hypothesis Test** groupbox. The hypothesized difference in the means of the *Data Variables* are entered in the **Hypothesized Difference** textbox in the **Hypothesis Test** groupbox. The null hypothesis ($=, \geq, \leq$) and alternative hypothesis ($\neq, <, >$) are selected in the corresponding comboboxes in the **Hypothesis Test** groupbox.

When the **Output Results** button is clicked, *Statistician* outputs the:

- null hypothesis,
- alternative hypothesis,
- sample size,
- mean of differences,
- standard deviation of differences,
- test statistic,
- critical value,
- p-value,
- decision rule, and
- conclusion.

To obtain a sample size for a given confidence interval width, the **Output Sample Size ...** checkbox in the **Confidence Interval** groupbox is checked. The user then enters the desired distance from the 'true' population mean in the corresponding textbox. The number of significant digits after the decimal point in the confidence interval output is selected from the **Number of Significant Digits in Output** combobox, (the default is 2).

(d) Test Population Proportion

The **Test Population Proportion** form is activated from the **Standard Tests** button on the *Statistician* ribbon tab.

Statistician (Test Population Proportion)

Hypothesis Test ($\alpha = 0.05$)

Null Hypothesis
Proportion of =

Alternative Hypothesis
Proportion of \neq

Confidence Interval

Output a 95% Confidence Interval for the Proportion

Output the Sample Size to Obtain a Confidence Interval Estimate to Within of the True Population Proportion.

Significant Digits

Significance

1% 5% 10%

Other %

Symbols

Success = ???

Failure = ???

Switch Symbols

Format Output

The **Test Population Proportion** form is activated by selecting **One Sample** from the **Proportion** combobox in the **Standard Tests** groupbox. The **Test Population Proportion** form has two primary functions:

- To test the hypothesis that the population proportion of successes (π) in a selected *Data Variable* is equal to, greater than or equal to, or less than or equal to, a hypothesized value (p).
- To produce a confidence interval for the proportion of successes in a *Data Variable*.

The following set of null and alternative hypotheses can be tested.

- (a) $H_0: \pi = p$
 $H_1: \pi \neq p$ (two tailed test) or,
 $H_1: \pi > p$ or,
 $H_1: \pi < p$
- (b) $H_0: \pi \geq p$
 $H_1: \pi < p$
- (c) $H_0: \pi \leq p$
 $H_1: \pi > p$

A binary variable can take on distinct values, which are classified as either success or failure. Denote $\hat{p} = \frac{x}{n}$ as the sample proportion of successes where (x) is the number of successes in the sample and (n) is the sample size. The sampling distribution of \hat{p} under the null hypothesis is approximately normal with mean p and standard deviation $\sqrt{p(1-p)/n}$. The test statistic:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normally distributed. The normality assumption is reasonably accurate when $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$. The null hypothesis is rejected if:

Alternative Hypothesis	Rejection Criterion
$H_1: \pi \neq p$	$ z > z_{\alpha/2}$
$H_1: \pi < p$	$z < z_{\alpha}$
$H_1: \pi > p$	$z > z_{\alpha}$

where $z_{\alpha/2}$ and z_{α} are the critical values for a two tailed and one tailed test respectively.

The confidence interval for an estimate of the population proportion is given as $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The sample size required to estimate the population proportion to within a particular value (w) for a given confidence level α can be calculated in two ways.

- (1) (Point estimate) If there is reason to believe the 'true' population proportion is reasonably close to the estimated sample proportion then the required sample size is given by the lowest integer greater than:

$$\frac{z_{\alpha/2}^2 \hat{p}(1 - \hat{p})}{w^2}$$

- (2) (Conservative) To obtain a conservative estimate of the required sample size employing no assumptions about the 'true' population proportion then the required sample size is given by the lowest integer greater than:

$$\frac{0.25 z_{\alpha/2}^2}{w^2}$$

Using Statistician (Test Population Proportion)

The *Data Variable* to be tested is selected from the **Null Hypothesis** combobox in the **Hypothesis Test** groupbox. The hypothesized proportion of successes in the *Data Variable* is entered in the **Hypothesized Proportion** textbox in the **Hypothesis Test** groupbox. The null hypothesis ($=, \geq, \leq$) and alternative hypothesis ($\neq, <, >$) are selected in the corresponding comboboxes in the **Hypothesis Test** groupbox.

All *Data Variables* tested must be binary variables, that is they have one of two distinct values referred to as success or failure. These distinct values may be numeric or text. The symbols for success and failure are generated by *Statistician* after analyzing the selected *Data Variable*. *Statistician* enters in the symbols for success and failure in the **Success** and **Failure** labels in the **Symbol** groupbox. Clicking the **Switch Symbols** button switches the symbols for success and failure.

When the **Output Results** button is clicked, *Statistician* outputs the:

- null hypothesis,
- alternative hypothesis,
- sample size,
- number of ones,
- number of zeros,
- proportion of successes,
- $n\hat{p}$,
- $n(1 - \hat{p})$,
- test statistic,
- critical value,
- p-value,
- decision rule and,
- conclusion.

If the user checks the **Estimate Confidence Interval** checkbox a confidence interval for the population proportion is outputted. The number of significant digits after the decimal point in the confidence interval output is selected from the **Number of Significant Digits in Output** combobox, (the default is 3). To obtain the required sample size for a given confidence interval width, the **Output Sample Size ...** checkbox in the **Confidence Interval** groupbox is selected. The user enters the desired distance from the 'true' population proportion in the corresponding textbox.

(e) Test Difference in Population Proportions

The **Test Difference in Population Proportions** form is activated from the **Standard Tests** button on the *Statistician* ribbon tab.

The **Test Difference in Population Proportions** form has two primary functions:

- To test the hypothesis that the difference in the proportion of successes in two selected *Data Variables* (π_1, π_2) is equal to, greater than or equal to or less than or equal to a hypothesized value (p_D).
- To produce a confidence interval for the difference in the proportion of successes of two selected *Data Variables*.

The following set of null and alternative hypotheses can be tested.

- $H_0: \pi_1 - \pi_2 = p_D$
 $H_1: \pi_1 - \pi_2 \neq p_D$ (two tailed test) or,
 $H_1: \pi_1 - \pi_2 > p_D$ or,
 $H_1: \pi_1 - \pi_2 < p_D$
- $H_0: \pi_1 - \pi_2 \geq p_D$
 $H_1: \pi_1 - \pi_2 < p_D$
- $H_0: \pi_1 - \pi_2 \leq p_D$
 $H_1: \pi_1 - \pi_2 > p_D$

Assume the proportion of successes in each variable is denoted by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$ where x_1 and x_2 are the number of successes in each variable and n_1 and n_2 are the number of observations on each variable. The pooled proportion of successes is denoted by $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$.

The test statistic is dependent upon the value of the hypothesized difference in proportions (p_D). If $p_D = 0$ the test statistic is given as:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

If $p_D \neq 0$ the test statistic is given as:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - p_D}{\sqrt{\left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)}}$$

The confidence interval is given by,

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)}$$

where α is the significance level. The z-statistic in each case has approximately a standard normal distribution if $n_1\hat{p}_1 \geq 5$, $n_1(1 - \hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$.

The null hypothesis is rejected if:

Alternative Hypothesis	Rejection Criterion
$H_1: \pi_1 - \pi_2 \neq p_D$	$ z > z_{\alpha/2}$
$H_1: \pi_1 - \pi_2 < p_D$	$z < z_{\alpha}$
$H_1: \pi_1 - \pi_2 > p_D$	$z > z_{\alpha}$

where $z_{\alpha/2}$ and z_{α} are the critical values for a two tailed and one tailed test respectively.

The required sample size to construct a confidence interval for the ‘true’ difference in population proportions can be calculated in two ways.

- (1) (Point estimate) If we have reason to believe the ‘true’ difference in population proportions is close to the estimated difference in sample proportions ($\hat{p}_1 - \hat{p}_2$), then the required sample size is given by the lowest integer greater than:

$$\frac{z_{\alpha/2}^2(\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2))}{w^2}$$

- (2) (Conservative) If we wish to obtain a conservative estimate of the required sample size employing no assumptions about the ‘true’ population difference in proportions, then the required sample size is given by the lowest integer greater than:

$$\frac{0.5z_{\alpha/2}^2}{w^2}$$

where w is half of the required confidence interval width.

Using Statistician (Test Difference in Population Proportions)

The *Data Variables* to be tested are selected from the **Null Hypothesis** combobox in the **Hypothesis Test** groupbox. The hypothesized difference in the proportion of successes in the *Data Variables* is entered in the **Hypothesized Difference** textbox in the **Hypothesis Test** groupbox. The null hypothesis ($=, \geq, \leq$) and alternative hypothesis ($\neq, <, >$) are selected in the corresponding comboboxes in the **Hypothesis Test** groupbox.

All *Data Variables* tested must be binary variables, that is they have one of two distinct values referred to as success or failure. These distinct values may be numeric or text. The symbols for success and failure are generated by *Statistician* after analyzing the selected *Data Variables*. (Both *Data Variables* must use the same symbols for success and failure.) *Statistician* enters in the symbols for success and failure in the **Success** and **Failure** labels in the **Symbol** groupbox. Clicking the **Switch Symbols** button switches the symbols for success and failure.

When the **Output Results** button is clicked, *Statistician* outputs the:

- null hypothesis,
- alternative hypothesis,
- sample size of each *Data Variable* (n_1 and n_2),
- number of ones in each *Data Variable*,
- number of zeros in each *Data Variable*,
- proportion of successes in each *Data Variable* ($\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$), where x_1 and x_2 are the number of successes in each *Data Variable*,
- pooled sample size ($n = n_1 + n_2$),
- pooled proportion of successes ($p = \frac{x_1 + x_2}{n_1 + n_2}$),
- $n_1\hat{p}_1$, $n_2\hat{p}_2$,
- $n_1(1 - \hat{p}_1)$, $n_2(1 - \hat{p}_2)$,
- test statistic,
- critical value,
- p-value,
- decision rule, and
- conclusion.

To obtain the sample size for the difference in proportions for a given confidence interval width, the **Output Sample Size ...** checkbox in the **Confidence Interval** groupbox is checked. The user then enters the desired distance from the ‘true’ difference in population proportion in the corresponding textbox. The number of significant digits after the decimal point in the confidence interval output is selected from the **Number of Significant Digits in Output** combobox, (the default is 3).

Section (5) – Variance Tests **(not available in *Statistician (Lite)*)**

The **Variance Tests** button on the *Statistician* ribbon tab provides the user with four menu items. These are:

- (i) One Sample,
- (ii) Two Samples,
- (iii) Bartlett, and
- (iv) Levene .

Each of these menu item selections are described in detail below.

Note that in the following discussion, the notation $\chi^2_{(\alpha,n)}$ and $F_{(\alpha,n_1,n_2)}$ denote values of the χ^2 and F distribution, where α is the probability (area), in the right tail of the distribution.

(a) One Sample

Statistician (Variance Test - One Sample)

Hypothesis Test ($\alpha = 0.05$)

Null Hypothesis

Variance of =

Alternative Hypothesis

Variance of \neq

Output Confidence Interval

Significant Digits

Significance

1% 5% 10%

Other %

Format Output

The **One Sample** menu item has two primary functions:

- (a) To test the hypothesis that the population variance (σ^2) of a selected *Data Variable* is equal to, greater than or equal to, or less than or equal to, a hypothesized value (σ_0^2).
- (b) To produce a confidence interval for the variance of a *Data Variable*.

The following sets of null and alternative hypotheses can be tested.

- (a) $H_0: \sigma^2 = \sigma_0^2$
 $H_1: \sigma^2 \neq \sigma_0^2$ (two tailed test) or,
 $H_1: \sigma^2 > \sigma_0^2$ or,
 $H_1: \sigma^2 < \sigma_0^2$
- (b) $H_0: \sigma^2 \geq \sigma_0^2$
 $H_1: \sigma^2 < \sigma_0^2$
- (c) $H_0: \sigma^2 \leq \sigma_0^2$
 $H_1: \sigma^2 > \sigma_0^2$

The test statistic is given as $T = \frac{(n-1)s^2}{\sigma_0^2}$ where n is the number of observations in the *Data Variable*, s^2 is the sample variance of the *Data Variable* and σ_0^2 is the hypothesized variance of the *Data Variable*. The T statistic has a χ^2 distribution with $n - 1$ degrees of freedom.

If the alternate hypothesis is $H_1: \sigma < \sigma_0$ then the null hypothesis is rejected if $T < \chi^2_{(1-\alpha, n-1)}$ where α is the level of significance and $n - 1$ is the degrees of freedom. If the alternate hypothesis is $H_1: \sigma > \sigma_0$ then the null hypothesis is rejected if $T > \chi^2_{(\alpha, n-1)}$. If the alternate hypothesis is $H_1: \sigma = \sigma_0$ then the null hypothesis is rejected if $T > \chi^2_{(\alpha/2, n-1)}$ or $T < \chi^2_{(1-\alpha/2, n-1)}$. A confidence interval for the population variance at confidence level $1 - \alpha$ is given as $\frac{(n-1)s^2}{\chi^2_{(\alpha/2, n-1)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2, n-1)}}$.

(b) Two Samples

Statistician (Variance Test - Two Samples)

Hypothesis Test ($\alpha = 0.05$)

Null Hypothesis

Variance of Select a Data Variable = Variance of Select a Data Variable

Alternative Hypothesis

Variance of \neq Variance of

Output Confidence Interval

Significant Digits 2

Significance

1% 5% 10%

Other

Format Output

Output Results
Close

Manage Data

The primary function **Two Sample** groupbox is to test the hypothesis that the population variances (σ_1^2, σ_2^2) of two selected *Data Variables* are equal to, greater than or equal to, or less than or equal to each other.

The following sets of null and alternative hypotheses can be tested.

- (a) $H_0: \sigma_1^2 = \sigma_2^2$
 $H_1: \sigma_1^2 \neq \sigma_2^2$ (two tailed test) or,
 $H_1: \sigma_1^2 > \sigma_2^2$ or,
 $H_1: \sigma_1^2 < \sigma_2^2$
- (b) $H_0: \sigma_1^2 \geq \sigma_2^2$
 $H_1: \sigma_1^2 < \sigma_2^2$
- (c) $H_0: \sigma_1^2 \leq \sigma_2^2$
 $H_1: \sigma_1^2 > \sigma_2^2$

The test statistic is given as $= s_1^2/s_2^2$. The F statistic has an F distribution with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator where n_1 and n_2 are the sample sizes of the first and second *Data Variables* respectively.

If the alternate hypothesis is $H_1: \sigma_1^2 < \sigma_2^2$ then the null hypothesis is rejected if $F < F_{(1-\alpha, n_1-1, n_2-1)}$ where α is the level of significance. If the alternate hypothesis is $H_1: \sigma_1^2 > \sigma_2^2$ then the null hypothesis is rejected if $F > F_{(\alpha, n_1-1, n_2-1)}$. If the alternate hypothesis is $H_1: \sigma_1^2 \neq \sigma_2^2$ then the null hypothesis is rejected if $F < F_{(1-\alpha/2, n_1-1, n_2-1)}$ or $F > F_{(\alpha/2, n_1-1, n_2-1)}$. A confidence interval for the ratio of the two population variances σ_1^2/σ_2^2 is given as $\frac{s_1^2}{s_2^2} F_{(1-\alpha/2, n_1-1, n_2-1)} \leq \sigma_1^2/\sigma_2^2 \leq \frac{s_1^2}{s_2^2} F_{(\alpha/2, n_1-1, n_2-1)}$.

(c) Many Samples

Two tests are available to test if k samples have equal variances, they are the Bartlett test and the Levine test. The null and alternative Hypotheses are given as:

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1: \text{At least one pair of variances are unequal, i.e. } \sigma_i \neq \sigma_j (i \neq j)$$

Bartlett Test

Statistician (Variance Test - Bartlett)

Data Variables

Summary Statistics

Significance

1% 5% 10%

Other %

Select All Variables

DeSelect All Variables

Manage Data

Significant Digits 4

Format Output

Output Results

Close

The Bartlett test is constructed as follows. Let n_i denote the size of the i^{th} sample and let $n = \sum_{i=1}^k n_i$ be the pooled sample size. Let s_i^2 be the variance of the i^{th} sample and denote the weighted average of the variance of all samples as $s_p^2 = \frac{1}{(n-k)} \sum_{i=1}^k (n_i - 1) s_i^2$. The test statistic is given as:

$$T = \frac{(n-k) \ln(s_p^2) - \sum_{i=1}^k (n_i - 1) \ln(s_i^2)}{1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{(n_i - 1)} - \frac{1}{(n-k)} \right)}$$

The T statistic has a $\chi_{(k-1)}^2$ distribution. The null hypothesis is rejected if $T > \chi_{(\alpha, k-1)}^2$.

Levene Test

Statistician (Variance Test - Levene)

Data Variables

Summary Statistics

Significance

1% 5% 10%

Other %

Levene Test Type

Mean

Median

Trimmed Mean

Significant Digits

Format Output

[Output Results](#)

[Close](#)

[Select All Variables](#)

[DeSelect All Variables](#)

[Manage Data](#)

The Levene test is constructed as follows. Let n_i denote the size of the i^{th} sample and let $n = \sum_{i=1}^k n_i$ be the pooled sample size. The test statistic is given as:

$$W = \frac{(n - k) \sum_{i=1}^k n_i (\bar{z}_i - \bar{z}_{..})^2}{(k - 1) \sum_{i=1}^k n_i (z_{ij} - \bar{z}_i)^2}$$

where z_{ij} can take one of the following three definitions.

- (a) Mean: $z_{ij} = |y_{ij} - \bar{y}_i|$ where \bar{y}_i is the mean of the i^{th} sample.
- (b) Median: $z_{ij} = |y_{ij} - \bar{y}_i|$ where \bar{y}_i is the median of the i^{th} sample.
- (c) Trimmed Mean: $z_{ij} = |y_{ij} - \bar{y}'_i|$ where \bar{y}'_i is the 10% trimmed mean of the i^{th} sample.

\bar{z}_i denotes the sample means of z_{ij} and $\bar{z}_{..}$ denotes the overall mean of z_{ij} .

The W statistic has a $F_{(k-1, n-k)}$ distribution. The null hypothesis is rejected if $F > F_{(\alpha, k-1, n-k)}$.

Section (6) Normality Tests (not available in *Statistician (Lite)*)

The **Normality Tests** button on the *Statistician* ribbon tab provides the user with ability to perform six different normality tests. These are the:

- (i) Jacque-Bera test,
- (ii) Anderson-Darling test,
- (iii) Shapiro-Wilk test,
- (iv) Kolmogorov-Smirnov test,
- (v) Lilliefors test, and the
- (vi) Cramér-von Mises test.

The screenshot shows the 'Statistician (Normality Tests)' dialog box. It features a 'Data Variable' dropdown menu, a 'Significance' section with radio buttons for 1%, 5% (selected), and 10%, and an 'Other' option with a percentage input field. The 'Distribution Parameters' section includes input fields for 'Mean' (0) and 'Std Dev' (1). A 'Tests' section contains radio buttons for Jacque-Bera, Anderson-Darling, Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Cramer-von Mises. There are also 'Format Output' (checked), 'Output Results', and 'Close' buttons, and a 'Manage Data' button at the bottom left.

Each of these tests are described in detail below.

The Null and Alternative hypothesis in a normality test is given as:

H_0 : The data is normally distributed.

H_1 : The data is not normally distributed.

(a) Jacque-Bera Test

The Jacque-Bera test statistic (JB) has a $\chi^2_{(2)}$ distribution. It is defined as:

$$JB = n \left(\frac{Skewness^2}{6} + \frac{Excess\ Kurtosis^2}{24} \right)$$

$$\text{where } Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

$$\text{and } Excess\ Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

The null hypothesis is rejected if the JB test statistic is greater than the $\chi^2_{(2)}$ critical value.

(b) Anderson-Darling Test

The Anderson-Darling test statistic (often denoted as A^2), is given as:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\ln(F(z_i^*)) + \ln(1 - F(z_{n+1-i}^*))]$$

where $F(\cdot)$ is the normal distribution, z_i^* is the z-score from the i^{th} sorted (ascending) observation and n is the number of observations. The null hypothesis is rejected if A^2 is greater than the critical value (CV) where:

$$CV = a / \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right)$$

The value of a is dependent on the level of significance required.

Significance	a
10%	0.631
5%	0.752
2.5%	0.873
1%	1.035
0.5%	1.159

The p-value reported is computed from the modified statistic

$$Z = A^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right) \text{ as given in Stephens (1986).}$$

(c) Shapiro-Wilk Test

The Shapiro-Wilk (1965) test statistic (W) is given as:

$$W = \frac{b^2}{(n-1)s^2}$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, n is the number of observations, x_i is the i^{th} observation and \bar{x} is the mean of these observations. The term b^2 is constructed as follows:

- (1) All observations are sorted into ascending order (let y_i denote the i^{th} sorted observation).
- (2) k is defined as $k = n/2$ if n is even and $k = n/2 + \frac{1}{2}$ if n is odd.
- (3) b is then calculated as $b = \sum_{i=1}^k a_i (y_{n-i+1} - y_i)$ where the weights a_i are taken from a table given in Shapiro and Wilk (1965).

(d) Kolmogorov-Smirnov Test

Assume $F(X)$ is the empirical distribution from which a sample is derived and $F_0(X)$ is the hypothesized normal distribution from which the sample is derived with mean and standard deviation denoted by (μ, σ) respectively. The Kolmogorov-Smirnov test statistic (D) is the maximum distance between the cumulative probability distribution of the empirical data and the hypothesised normal distribution. It is defined as:

$$D = \max_i [\sup(F(x_i) - F_0(x_i)), \sup(F(x_i) - F_0(x_{i-1}))]$$

The null hypothesis is rejected if D is greater than or equal to the Kolmogorov-Smirnov critical value.

(e) Lilliefors test

The Lilliefors test is almost identical to the Kolmogorov-Smirnov test but differs in two key ways:

- (i) The sample mean and standard deviation are employed in the test rather than employ a hypothesised mean and standard deviation.
- (ii) Different critical values are used (see Abdi and Molin (2007))

(f) Cramér-von Mises Test

The Cramér-von Mises test statistic (T) is given as:

$$T = \frac{1}{12n} + \sum_{i=1}^n \left[\frac{2i-1}{2n} - F(x_i) \right]^2$$

where $F(x_i)$ is the cumulative normal distribution with the mean and standard deviation estimated from the data.

References:

Abdi., Hervé and Molin., Paul, (2007), "Lilliefors/Van Soest's test of normality", in Neil Salkind (Ed.), (2007), "Encyclopedia of Measurement and Statistics", Thousand Oaks (CA): Sage.

Thode Jr., H.C., (2002), "Testing for Normality", Marcel Dekker, New York.

Shapiro., S and Wilk., B, (1965), "An analysis of variance test for normality (complete samples)", *Biometrika*, 52 , 3 and 4, p. 691.

Sheskin., David, (2007), "Handbook of parametric and non-parametric statistics", 4th edition, Chapman and Hall.

M. A. Stephens (1986). "Tests Based on EDF Statistics". In D'Agostino, R.B. and Stephens, M.A.. "Goodness-of-Fit Techniques". New York: Marcel Dekker.

Section (7) - Non Parametric Tests **(not available in *Statistician (Lite)*)**

The **Non Parametric Tests** button on the *Statistician* ribbon tab provides the user with a number menu items which activate various forms to perform a number of non parametric tests. These are:

- (i) Runs,
- (ii) Mann Whitney,
- (iii) Wilcoxon Rank Sum,
- (iv) Wilcoxon Signed Rank Sum,
- (v) Sign,
- (vi) Kolmogorov Smirnov (two sample),
- (vii) Kruskal Wallis,
- (viii) Friedman,
- (ix) Chi square goodness-of-fit,
- (x) Kolmogorov Smirnov goodness-of-fit.

Each of these menu items are described in detail below.

(a) Runs Test

The runs test is employed to test if data is serially related. The null and alternative hypotheses to be tested are:

H_0 : *The data is serially independent*

H_1 : *The data is serially dependent*

The test is conducted on a series of binary data, that is, data that can only take two distinct values. Under the null hypothesis, it can be expected that there is no particular pattern in consecutive runs of observations in each of the two categories of data in the series. Let n_1 and n_2 denote the number of observations in each of the two categories of data in the series. The statistic (R) is the number of runs in the series.

In the small sample case (where $n_1 \leq 20$ and $n_2 \leq 20$), if R is above an upper critical value or below a lower critical value for a given level of significance, the null hypothesis is rejected. In the large sample case (where $n_1 > 20$ or $n_2 > 20$), the distribution of runs is approximately normal with mean $\mu_R = \frac{2n_1n_2}{n_1+n_2} + 1$ and standard

deviation $\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$. The test statistic in the large sample case is given as $z = \frac{R - \mu_R}{\sigma_R}$ and can be tested with a standard two tailed z-test.

Using Statistician

The **Non Parametric Tests – Runs** form is selected by clicking the **Runs** button in the **Non Parametric Tests** groupbox on the *Statistician* ribbon. The user selects the *Data Variable* for analysis from the **Data Variable** combobox. If the data is a small sample (defined in the preceding section), then only a significance level of 0.05 can be selected.

Statistician outputs the sample size of the *Data Variable*, test statistic, critical value(s), significance, null and alternative hypothesis, decision rule and conclusion for the test. All *Data Variables* tested must be binary variables, that is they have one of two distinct values referred to as Category 1 or Category 2. These distinct values may be numeric or text. The symbols for Category 1 and Category 2 are generated by *Statistician* after analyzing the selected *Data Variable*. *Statistician* enters in the symbols for Category 1 and Category 2 in the **Category 1** and **Category 2** labels in the **Symbols for Categories** groupbox. Clicking the **Switch Symbols** button switches the symbols for success and failure.

Statistician automatically determines if the user has selected a small or large sample based upon the criteria defining a small and large sample outlined in the preceding section.

(b) Non Parametric Tests (Two Sample)

Non parametric testing techniques involve testing hypothesis about ordinal (or ranked data). When working with ordinal data, the concept of a mean makes little sense and consequently many statisticians prefer to refer to the 'location' of the data as a measure of central tendency. In the following tests, two samples from two populations are drawn. Let L_1 and L_2 denote the location of the first and second populations respectively. The following set of null and alternative hypothesis can be tested.

- (a) $H_0: L_1 = L_2$
 $H_1: L_1 \neq L_2$ (two tailed test) or,
 $H_1: L_1 > L_2$ or,
 $H_1: L_1 < L_2$
- (b) $H_0: L_1 \geq L_2$
 $H_1: L_1 < L_2$
- (c) $H_0: L_1 \leq L_2$
 $H_1: L_1 > L_2$

Four location tests are available. They are the Mann-Whitney U test, the Wilcoxon Rank Sum test, the Wilcoxon Signed Rank Sum test and the Sign test.

(c) Mann-Whitney U Test

The Mann-Whitney U test, tests the relative location of two independent samples of ordinal data. Let n_1 and n_2 be the sample size of the observations drawn from both populations respectively and let $n = n_1 + n_2$. The data from both samples are pooled and then ranked from 1 to n . Observations with a tied rank are assigned a rank which is the average of the corresponding raw ranks. Define S_1 as the sum of the ranks from the first sample and define S_2 as the sum of the ranks from the second sample. Let $A_1 = S_1/n_1$ and $A_2 = S_2/n_2$ denote the average of the ranks of the samples from the first and second populations respectively. Define the U_1 and U_2 statistics as $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - S_1$ and $U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - S_2$. The U test statistic is defined as, $U = \min(U_1, U_2)$.

Let a small sample be defined as the case where $n_1 \leq 10$ and $n_2 \leq 10$. In the small sample case, critical values are taken from a table of Mann Whitney critical values. (*Statistician* has Mann Whitney critical values in the small sample case for one and two tailed tests where $\alpha = 0.05$ and 0.01). Let $U_{Crit,\alpha}^1$ and $U_{Crit,\alpha}^2$ denote the critical value for a one and two tailed test respectively with significance level α . If the alternative hypothesis is $H_1: L_1 \neq L_2$, the null hypothesis is rejected if $U \leq U_{Crit,\alpha}^2$. For a one tailed test with alternative hypothesis $H_1: L_1 < L_2$, the null hypothesis is rejected if $U \leq U_{Crit,\alpha}^1$ and $A_1 < A_2$. For a one tailed test with alternative hypothesis $H_1: L_1 > L_2$, the null hypothesis is rejected if $U \leq U_{Crit,\alpha}^1$ and $A_1 > A_2$.

A large sample is the case where $n_1 > 10$ or $n_2 > 10$. The test statistic is given as $z = \left(U - \frac{n_1 n_2}{2} \right) / \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$. The z test statistic has a standard normal distribution and can be tested with a standard z -test. If the alternative hypothesis is $H_1: L_1 \neq L_2$, the null hypothesis is rejected if $|z| > z_{\alpha/2}$ where $z_{\alpha/2}$ is the critical value. For a one tailed test with alternative hypothesis $H_1: L_1 < L_2$, the null hypothesis is rejected if $z < -z_\alpha$ and $A_1 < A_2$. For a one tailed test with alternative hypothesis $H_1: L_1 > L_2$, the null hypothesis is rejected if $z > z_\alpha$ and $A_1 > A_2$.

(d) Wilcoxon Rank Sum Test

The Wilcoxon Rank Sum test tests if the location of a sample is significantly different from the location of another sample when both of the samples are independent.

Let n_1 and n_2 be the sample size of the observations drawn from both populations and let $n = n_1 + n_2$. (Also, assign the sample with the least number of observations as the first sample, so that $n_1 \leq n_2$) The data from both samples are pooled and then ranked from 1 to n . Observations with a tied rank are assigned a rank which is the average of the corresponding raw ranks. Define T_1 as the sum of the ranks from the first sample and define T_2 as the sum of the ranks from the second sample.

The Wilcoxon Rank Sum test is conducted differently for small and large samples. Let a small sample be defined as the case where $n_1 \leq 10$ and $n_2 \leq 10$. In the small sample case, the test statistic is T_1 . For a two tailed test, if T_1 is greater than an upper critical value (T_U) or less than a lower critical value (T_L), then the null hypothesis is rejected for a given significance level (α). For a one tailed test where the alternative hypothesis is ($H_1: T_1 > T_2$), the null hypothesis is rejected if $T_1 > T_U$. For a one tailed test where the alternative hypothesis is ($H_1: T_1 < T_2$), the null hypothesis is rejected if $T_1 < T_L$. In the small sample case, acceptable values of alpha are 0.1, 0.05 and 0.01.

The large sample case is defined when $n_1 > 10$ or $n_2 > 10$. The test statistic is given as
$$z = \left(T_1 - \frac{n_1(n_1+n_2+1)}{2} \right) / \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$
. In the large sample case, the test statistic has a standard normal distribution and can be tested with a standard z-test.

(e) Wilcoxon Signed Rank Sum Test

The Wilcoxon signed rank sum test tests if the location of a set of ordinal matched pairs are different. This test is employed when the matched pairs are not independent.

The test statistic is constructed by taking the difference of the matched pairs ($d_i = x_{1,i} - x_{2,i}$) where $x_{1,i}$ and $x_{2,i}$ are the i^{th} observation on the first and second selected *Data Variables* respectively and d_i is the difference between the matched pairs ($1 \leq i \leq n$). If $d_i = 0$ then the i^{th} matched pair is ignored. Let the number of remaining matched pairs with a non zero difference be denoted by n^* . The absolute value of the non zero differences are calculated and then ranked in ascending order. Observations with a tied rank are assigned a rank which is the average of the corresponding raw ranks. Let (T^+) be the sum of the ranks where $d_i > 0$ and let (T^-) be the sum of the ranks where $d_i < 0$. The test statistic (T) is defined as $T = \min(T^+, T^-)$. The Wilcoxon signed rank sum test is conducted differently for small and large samples.

Let a small sample be defined as the case where $n^* \leq 30$. The critical value ($T_{Crit,\alpha}$) is taken from a table of Wilcoxon signed rank sum test critical values for a given significance level α . (*Statistician* has Wilcoxon Signed Rank Sum critical values for one and two tailed tests where $\alpha = 0.1, 0.05$ and 0.01 in the small sample case.) For a two tailed test with alternative hypothesis $H_1: L_1 \neq L_2$, the null hypothesis is rejected if $T < T_{Crit,\alpha}$. For a one tailed test with alternative hypothesis $H_1: L_1 < L_2$, the null hypothesis is rejected if $T < T_{Crit,\alpha}$ and $T^+ < T^-$. For a one tailed test with alternative hypothesis $H_1: L_1 > L_2$, the null hypothesis is rejected if $T < T_{Crit,\alpha}$ and $T^+ > T^-$.

The large sample case is defined when $n^* > 30$. The test statistic is given as $z = \left(U - \frac{n_1 n_2}{2} \right) / \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$. The z test statistic has a standard normal distribution. For a two tailed test with alternative hypothesis $H_1: L_1 \neq L_2$, the null hypothesis is rejected if $|z| \geq z_{\alpha/2}$ where $z_{\alpha/2}$ is the critical value and α is the level of significance. For a one tailed test with alternative hypothesis $H_1: L_1 < L_2$, the null hypothesis is rejected if $|z| \geq z_\alpha$ and $T^+ < T^-$. For a one tailed test with alternative hypothesis $H_1: L_1 > L_2$, the null hypothesis is rejected if $|z| \geq z_\alpha$ and $T^+ > T^-$.

(f) Sign Test

The sign test tests if the location of matched pairs are different where the matched pairs may not be independent. The test statistic is constructed by taking the difference of the matched pairs ($d_i = x_{1,i} - x_{2,i}$) where $x_{1,i}$ and $x_{2,i}$ are the i^{th} observation on the first and second selected *Data Variables* respectively and d_i is the difference between the matched pairs ($1 \leq i \leq n$). If $d_i = 0$ then the i^{th} matched pair is ignored. Let the number of remaining matched pairs with a non zero difference be denoted by n^* . The number of positive (n^+) and negative (n^-) differences are then recorded. If the location of both samples is the same then it is expected that the population proportion of successes (p) is given by $p = 0.5$ The null hypothesis is given as:

$$H_0: p = 0.5$$

The alternative hypotheses are given as:

$$H_1: p \neq 0$$

$$H_1: p > 0$$

$$H_1: p < 0$$

The number of positive differences in the sample is modeled with a binomial distribution. If $n^*p \geq 5$ and $n^*p(1 - p) \geq 5$ (which implies $n \geq 10$ when $p = 0.5$), then this distribution can be closely approximated with a normal distribution with mean $0.5n^*$ and standard deviation $\sigma_p = \sqrt{0.25n^*}$. The test statistic is given as $z = \frac{n^+ - 0.5n^*}{0.5\sqrt{n^*}}$.

(g) Kolmogorov-Smirnov Test (two sample)

The Kolmogorov-Smirnov test tests if a sample from two random variables are drawn from the same distribution. Let $F_1(X)$ and $F_2(X)$ denote the cumulative probability distribution from which the data from the first and second sample are drawn respectively. The null hypothesis for the test is given as:

$$H_0: F_1(X) = F_2(X)$$

The alternative hypotheses are given as:

$$H_1: F_1(X) \neq F_2(X)$$

$$H_1: F_1(X) > F_2(X)$$

$$H_1: F_1(X) < F_2(X)$$

The test statistic is constructed as follows. Assume we have two samples of data with values $X_1 = \{0, 1, 2, 2, 4, 5\}$ and $X_2 = \{1, 2, 4, 4, 6\}$. All distinct values are sorted and placed in a table (column 1). The number of values of X_1 and X_2 that correspond to the distinct values in column 1 are placed in columns 2 and 3.

1	2	3	4	5	6
Value	Number in X_1	Number in X_2	Cum Prob X_1	Cum Prob X_2	Difference
0	1	0	1/6 = 0.167	0/5 = 0.0	0.167
1	1	1	2/6 = 0.333	1/5 = 0.2	0.133
2	2	1	4/6 = 0.667	2/5 = 0.4	0.267
4	1	2	5/6 = 0.833	4/5 = 0.8	0.033
5	1	0	6/6 = 1.000	4/5 = 0.8	0.2
6	0	1	6/6 = 1.000	5/5 = 1.0	0.0

The cumulative probability of obtaining a particular value for X_1 and X_2 is placed in columns 4 and 5. The difference in cumulative probabilities is placed in column 6. The Kolmogorov-Smirnov test statistic (KS) is the largest of all of the absolute differences in cumulative probabilities. In the above example $KS = 0.267$. Let n_1 and n_2 denote the size of the first and second sample. In the large sample case where $n_1 \geq 10$ and $n_2 \geq 10$ the critical value for the test can be constructed as follows. Define K as $K = \sqrt{\frac{n_1+n_2}{n_1n_2}}$.

α (one tailed test)	0.1	0.05	0.025	0.01	0.005
α (two tailed test)	0.2	0.1	0.05	0.02	0.01
Critical Value	1.07K	1.22K	1.36K	1.52K	1.63K

In the small sample case where $n_1 \leq 10$ or $n_2 \leq 10$, critical values are taken from a table of Kolmogorov-Smirnov critical values.

If the alternative hypothesis is $H_1:F_1(X) \neq F_2(X)$, the null hypothesis is rejected if the absolute value of the test statistic is greater than or equal to the critical value. If the alternative hypothesis is $H_1:F_1(X) > F_2(X)$, the null hypothesis is rejected if the absolute value of the test statistic is greater than or equal to the critical value and the test statistic is positive. If the alternative hypothesis is $H_1:F_1(X) < F_2(X)$, the null hypothesis is rejected if the absolute value of the test statistic is greater than or equal to the critical value and the test statistic is negative.

Using Statistician (Two Sample Tests)

The user selects the two *Data Variables* for analysis from the two **Null Hypothesis** comboboxes in the **Hypothesis Test** groupbox. The null hypothesis ($=, \geq, \leq$) and alternative hypothesis ($\neq, <, >$) are selected in the corresponding comboboxes in the **Hypothesis Test** groupbox. If the **Output Ranks** checkbox is checked then *Statistician* will output the rank of each *Data Variable* observation for the Mann-Whitney U test, the Wilcoxon Rank Sum test and the Wilcoxon Signed Rank Sum test. For the Kolmogorov-Smirnov test, the user can output the cumulative probabilities of the selected *Data Variables* by checking the **Output Cumulative Probabilities** checkbox.

Statistician outputs the sample size of each *Data Variable*, test statistic, critical value, significance, null and alternative hypothesis, decision rule and conclusion for each of the five tests. *Statistician* automatically determines if the user has selected a small or large sample based upon (a), the number of observations in each *Data Variable* and (b), the criteria defining a small and large sample outlined in the preceding discussions on each of the tests.

For each of the four location non parametric tests, the following additional information for each selected *Data Variable* is also outputted.

Mann Whitney test – Rank Sum, Average Rank and U statistic.

Wilcoxon Rank Sum test – Rank Sum.

Wilcoxon Signed Rank Sum test - Number of positive, negative, zero and non-zero differences, Rank sum of positive (T⁺) and negative (T⁻) differences.

Sign test – Number of positive, negative and zero and differences and p-value.

(g) Non Parametric Tests (Two or More Samples)

When comparing two or more samples of ordinal data the Kruskal Wallis and Friedman tests can be employed. Assuming we have k populations of ordinal data. The null and alternative hypothesis to be tested is:

H_0 : *The location of the k populations are the same*

H_1 : *The location of at least one population differs from the others*

(h) Kruskal Wallis Test

The Kruskal Wallis test is employed to compare two or more populations of ordinal data which may have a non-normal distribution and are independent.

Let n_1, \dots, n_k denote the sample size of each of the k samples let $n = n_1 + \dots + n_k$. All of the observations from each of the k samples are pooled and ranked from 1 to n . Observations with a tied rank are assigned a rank which is the average of the corresponding raw ranks. The sum of the ranks from each of the k samples are denoted by T_1, \dots, T_k . The test statistic (H) is given as:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^k \frac{T_j^2}{n_j} \right] - 3(n+1)$$

Let a large sample size be defined as the case where the sample size from each of the populations is greater than or equal to five (ie $n_i \geq 5$). The H statistic has a χ^2 distribution with $k-1$ degrees of freedom. If $H > \chi_{\alpha, k-1}^2$ the null hypothesis is rejected, where α is the selected significance level.

(i) Friedman Test

The Friedman is employed when either ordinal or numerical data is generated from a randomized block experiment (as is the case with an ANOVA table without replication). Assume we have k samples (treatments) and b blocks within each sample. The test statistic is constructed by initially ranking all observations within each block. Observations with a tied rank are assigned a rank which is the average of the corresponding raw ranks. The sum of all ranks for each treatment is denoted by T_1, \dots, T_k . The Friedman test statistic (F) is defined as:

$$F = \left[\frac{12}{bk(k+1)} \sum_{j=1}^k T_j^2 \right] - 3b(k+1)$$

The large sample case is defined as the case where either k or b is greater than or equal to five. The F statistic has a χ^2 distribution with $k-1$ degrees of freedom. If $H > \chi_{\alpha, k-1}^2$ the null hypothesis is rejected, where α is the selected significance level. In the small sample case, critical values are taken from a table of Friedman critical values.

Using Statistician (Two or More Samples)

The user selects two or more *Data Variables* for analysis from the **Data Variable** listbox. If the **Output Ranks** checkbox is checked then *Statistician* will output the rank of each *Data Variable* observation. *Statistician* outputs the sample size of each *Data Variable*, test statistic, critical value, significance, p-value, rank sum, null and alternative hypothesis, decision rule and conclusion for each of the two tests.

In the case of the Friedman test, each selected *Data Variable* represents a treatment and each observation within a treatment represents a block. The number of observations (blocks) within each *Data Variable* (treatments) must be equal.

(j) Goodness-Of-Fit Tests – Chi square

The chi-square and goodness-of-fit test tests if a *Data Variable* has a specific distribution. The chi-square goodness-of-fit test can be applied to both discrete and continuous distributions. The null and alternative hypothesis for a goodness-of-fit test is given as:

$$\begin{aligned}H_0: F(X) &= F_0(X) \\H_1: F(X) &\neq F_0(X)\end{aligned}$$

where $F(X)$ is the population distribution from which the sample is derived and $F_0(X)$ is the hypothesized theoretical distribution from which the sample is derived.

With the chi-square goodness-of-fit test, the data is divided into k bins and the test statistic is defined as:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i . The expected frequency is calculated by:

$$E_i = N(F_0(X_{Ui}) - F_0(X_{Li}))$$

where X_{Ui} is the upper limit for bin i , X_{Li} is the lower limit for bin i , and N is the sample size. The parameters of the theoretical distribution can be either estimated from the sample data or can be assumed to have a specific value. If the parameters of the theoretical distribution are estimated from the sample data then the test statistic follows, approximately, a chi-square distribution with $(k - 1 - c - b)$ degrees of freedom where k is the number of bins, c = the number of estimated parameters for the distribution and b is the number of empty bins. If the parameters are assumed to have a specific value then $c = 0$.

Using Statistician

To perform a chi-square goodness-of-fit test, the user selects a *Data Variable* from the **Distribution of** combobox in the **Hypothesis Test** groupbox. The hypothesized distribution of the sample is selected from the **Distribution** combobox. If the parameters of the hypothesized distribution are to be estimated from the sample data then **Assigned** radiobutton in the **Distribution Parameters** groupbox is checked. To assign specific parameter values to the hypothesized distribution the **Assigned** radiobutton is checked. When the **Assigned** radiobutton is checked users can enter the hypothesized distribution parameters in the corresponding textboxes in the

Distribution Parameters groupbox. To specify the bins for the test users enter the lower bound, upper bound and bin width in the **Bins** groupbox. If the **Output Bin Frequencies** checkbox is checked then the expected and observed frequencies of observations within each bin is outputted. Distributions which can be tested are as follows:

Distribution	Parameters	Type
Normal	Mean Standard Deviation	Decimal Positive Decimal
Uniform	Lower Bound Upper Bound	Decimal Decimal
Log Normal	Mean Standard Deviation	Decimal Positive Decimal
Exponential	Mean	Positive Decimal

References:

David J Sheskin, 2007, Handbook of parametric and nonparametric statistical procedures, 4th ed, Chapman & Hall/CRC.

(k) Goodness-Of-Fit Tests - Kolmogorov-Smirnov

The Kolmogorov-Smirnov goodness-of-fit test tests if a *Data Variable* has a specific distribution. The Kolmogorov-Smirnov goodness-of-fit test can only be applied to continuous distributions. The null and alternative hypothesis for the Kolmogorov-Smirnov goodness-of-fit test is given as:

$$H_0: F(X) = F_0(X)$$

$$H_1: F(X) \neq F_0(X)$$

where $F(X)$ is the empirical distribution from which the sample is derived and $F_0(X)$ is the hypothesized distribution from which the sample is derived. The Kolmogorov-Smirnov test statistic is defined as:

$$D = \max_{0 \leq i \leq N} \left(F_0(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F_0(Y_i) \right)$$

The null hypothesis is rejected if D is greater than or equal to the Kolmogorov-Smirnov critical value.

Using Statistician

To perform a Kolmogorov-Smirnov goodness-of-fit test, the user selects a *Data Variable* from the **Distribution of** combobox in the **Hypothesis Test** groupbox. The hypothesized distribution of the sample is selected from the **Distribution** combobox. The parameters of the hypothesized distribution are entered in the **Distribution Parameters** groupbox which is presented to the user after a distribution is selected. Distributions which can be tested are as follows:

Distribution	Parameters	Type
Normal	Mean Standard Deviation	Decimal Positive Decimal
Uniform	Lower Bound Upper Bound	Decimal Decimal
Log Normal	Mean Standard Deviation	Decimal Positive Decimal
Exponential	Mean	Positive Decimal
Weibull	Scale Shape	Positive Decimal Positive Decimal

References:

David J Sheskin, 2007, Handbook of parametric and nonparametric statistical procedures, 4th ed, Chapman & Hall/CRC.

Section (8) - χ^2 Tests

(not available in *Statistician (Lite)*)

(a) Multinomial Experiment

A multinomial experiment is an extension of a binomial experiment where two or more outcomes are possible in each trial. Assume there are n trials of an experiment with k possible outcomes. Let p_i ($1 \leq i \leq k$) be the probability of outcome i . The null and alternative hypotheses are:

$$H_0: p_1 = p_2 = \dots = p_k$$

H_1 : *At least one probability is different from the others*

Let O_j denote the observed number of successes for the j^{th} outcome. The expected number of successes under the null hypothesis is given as $e_j = np_j$. The test statistic is given as:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

The test statistic has a χ^2 distribution with $k-1$ degrees of freedom. If the test statistic is greater than the critical value then the null hypothesis is rejected.

Using Statistician (Multinomial Experiment)

Statistician (Multinomial Experiment)

Data Variable

Significance

1% 5% 10%

Other %

Format Output

Output Results

Manage Data Close

The **Non Parametric Tests – Multinomial Experiment** form is selected by clicking the **Multinomial** button in the **Chi 2 Tests** groupbox on the *Statistician* ribbon. The user selects the *Data Variable* for analysis from the Data Variable combobox. Each outcome of the experiment is an observation in the selected *Data Variable*. *Statistician* outputs the null and alternative hypothesis, number of outcomes, expected outcome, test statistic, critical value, significance, p-value, decision rule and conclusion of the test.

(b) Contingency Tables

A contingency table is used to test if evidence exists to infer that two nominal variables are related. The method involves classifying data according to two different criteria. The null and alternative hypotheses to be tested are:

H_0 : The two variables are independent

H_1 : The two variables are dependent

The test is initially set up by placing the observations in a table with r rows and c columns. The observations typically represent the frequency of a particular event. Rows and columns each represent the two classifications (criteria) of the two nominal variables. Let $n = r \times c$ denote the number of cells in the contingency table. Let $S_{c,i}$ ($1 \leq i \leq c$) denote the sum of the i^{th} column, let $S_{r,j}$ ($1 \leq j \leq r$) denote the sum of the j^{th} row and let S denote the sum of all cells. If the two nominal variables are independent, the expected value of the cell in row i and column j is given by $e_{i,j} = \frac{S_{r,i}S_{c,j}}{n}$. Denoting the observation in cell in row i and column j as $O_{i,j}$, the test statistic is given as:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - e_{i,j})^2}{e_{i,j}}$$

The test statistic has a χ^2 distribution with $(r - 1)(c - 1)$ degrees of freedom. If the test statistic is greater than the critical value then the null hypothesis is rejected.

Using Statistician (Contingency Tables)

The Non Parametric Tests – Contingency Table form is selected by clicking the **Contingency Table** button in the **Chi 2 Tests** groupbox on the *Statistician* ribbon. The user selects the *Data Variables* for analysis from the **Data Variables** listbox. Each selected *Data Variable* represents the values of the first nominal variable (or columns in the contingency table). Each observation within a selected *Data Variable* represents the observed value for each category of the second nominal variable (or rows in the contingency table).

To enhance the readability of the output, the user has the option of selecting a *Data Variable* from the **Row Classifications** combobox. This *Data Variable* should contain the description of each row of the second nominal variable. If no selection is made from the **Row Classifications** combobox, *Statistician* will simply refer to Row(1),

Row(2) etc in the output. *Statistician* uses the selected *Data Variable* names as the description for each of the columns of the contingency table. The user also has the option of entering a general title for the rows and columns in the contingency table by entering the corresponding titles in the **Row Title** and **Column Title** textboxes. If these textboxes are left blank, *Statistician* will refer to Rows and Columns in the output.

The screenshot shows the 'Statistician (Contingency Table)' dialog box. It features a 'Column Variables' section with a large empty text area. The 'Row Classifications' section includes a dropdown menu, 'Row Title', and 'Column Title' textboxes. The 'Significance' section has radio buttons for 1%, 5% (selected), and 10%, along with an 'Other' checkbox and a percentage input field. The 'Table Output' section contains checkboxes for 'Output Contingency Table' and 'Output Expected Values Table'. The 'Significant Digits' dropdown is set to 2. At the bottom right, there is a checked 'Format Output' checkbox, an 'Output Results' button, and a 'Close' button. At the bottom left, there is a 'Manage Data' button.

By checking either the Output Contingency Table and/or **Output Expected Values Table** checkboxes, *Statistician* will output the contingency table and/or a table of expected values within the contingency table.

Statistician outputs the null and alternative hypothesis, number of column classifications, number of row classifications, test statistic, critical value, significance, p-value, decision rule and conclusion of the test.

Section (9) – ANOVA

The **ANOVA** form is selected by clicking the **ANOVA** button in the *Statistician* ribbon tab.

Statistician (ANOVA)

Data Variables

ANOVA Table Type

One Factor

Two Factor (Without Replication)

Two Factor (With Replication)

Number Of Replications

Significance

1% 5% 10%

Other %

Output Summary Statistics

Significant Digits

Pairwise Tests

None

Fisher LSD

Tukey Kramer

Scheffe

Pairwise Alpha

1%

5%

10%

Select All Variables

DeSelect All Variables

Manage Data

Format Output

Output Results

Close

The analysis of variance form (**ANOVA**), allows users to test the difference in the mean of a number of populations on the basis of samples. Three forms of ANOVA can be implemented with the **ANOVA** form. These are:

- (1) Single factor ANOVA,
- (2) Two factor ANOVA (without replication),
- (3) Two factor ANOVA with interaction (ie. with replication).

Each of these three forms of ANOVA analysis is discussed separately.

(a) Single factor ANOVA.

Assume there are a number of populations of interest each of which is comprised of a number of experimental observations. Each population is referred to as a treatment. Assume there are a treatments, and let μ_1, \dots, μ_a be the mean of the experimental observations in each of the a treatments. Assume there are n_i experimental observations in treatment i .

The technique of ANOVA involves testing a null and alternative hypothesis of the form:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_a$$

H_1 : at least two of the means are not equal

The total number of observations in all treatments is given as:

$$n = \sum_{i=1}^a n_i$$

The sample mean of treatment i is given as:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j}$$

where $x_{i,j}$ is the j^{th} observation in treatment i . The sample mean of all treatments is given as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^a n_i \bar{x}_i$$

The sample variance of observations within each treatment is given as:

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2}{n_i - 1}$$

The total sum of squares ($SS(Total)$), treatment sum of squares (SST) and the error sum of squares (SSE) are given as:

$$SS(Total) = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2$$

$$SST = \sum_{i=1}^a n_i (\bar{x}_i - \bar{x})^2$$

$$SSE = \sum_{i=1}^a (n_i - 1) s_i^2 = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2$$

The three sums of squares are related by:

$$SS(Total) = SST + SSE$$

The mean square for treatments (*MST*) and the mean square for error (*MSE*) are given as:

$$MST = \frac{SST}{a - 1}$$

$$MSE = \frac{SSE}{n - a}$$

To test the null hypothesis, the test statistic is defined as $= \frac{MST}{MSE}$. This ratio is *F*-distributed with $a - 1$ degrees of freedom in the numerator and $n - a$ degrees of freedom in the denominator.

The user can test the hypothesis that any pairs of treatments have significantly different means. Three tests are available, the Fisher least significant difference (LSD) test, the Tukey Kramer test and the Scheffe test. Critical values for these tests are as follows:

Test	Critical Value	Notes
Fisher (LSD)	$t_{\alpha/2, n-a} \sqrt{MSE \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$	
Tukey-Kramer	$q_{\alpha}(a, n - a) \sqrt{\frac{MSE}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$	$q_{\alpha}(a, n - a)$ is the critical value of the Studentised range distribution.
Scheffe	$\sqrt{\frac{2(a - 1)F_{(\alpha/2, a-1, n-a)}MSE}{n}}$	

If $|\bar{x}_1 - \bar{x}_2| < \text{Critical Value}$ then it is concluded that there is no significant difference between the means of the two treatments being tested.

Using Statistician (Single Factor ANOVA)

After selecting the **One Factor** radiobutton in the **ANOVA Table Type** groupbox, the user then selects two or more *Data Variables* in the **Data Variables** listbox. Each of these *Data Variables* are the observations on a treatment and may be of unequal sample size.

The number of observations, sum, average and variance of all of the observations within each treatment group is outputted. A table describing the source of variation as displayed below is then reported.

Source of Variation	d.f.	Sum of Squares	Mean Squares	F Statistic	F critical	p-value
Treatments:	$a-1$	SST	MST	$f = MST / MSE$	$F_{(\alpha, a-1, n-1)}$	$P(F > f)$
Error:	$n-a$	SSE	MSE			
Total:	$n-1$	$SS(Total)$				

The null and alternate hypothesis, the test significance level, the F statistic, critical value, p-value, the decision rule and the conclusion are also reported. If the **Output Summary Statistics** checkbox is checked then summary statistics for each of the treatments are outputted.

The user can also test the hypothesis that any pairs of treatments have significantly different means. Three tests are available, the Fisher's least significant difference test, Tukey Kramer test and the Scheffe. The results of these tests are outputted by selecting the **Fisher LSD**, **Tukey Kramer** or **Scheffe** radiobuttons respectively in the **Pairwise Tests** groupbox. Statistician then outputs the null and alternative hypothesis, the difference in the means of the *Data Variables*, the critical value and the decision to reject or not reject the null hypothesis

(b) Two factor ANOVA without replication.

When two factors impact upon an experimental observation, a two factor ANOVA table can be used to test the effect of both of these factors. The treatment effect is usually considered as the first factor and a block (or replicate) effect is taken as the second factor. Two sets of hypothesis are of interest:

- (a) A test of the hypothesis that the means of all of the treatments are equal.
- (b) A test of the hypothesis that the means of all of the blocks are equal.

Let μ_i^T be the mean of treatment i where $1 \leq i \leq a$ and let μ_j^B be the mean of block j where $1 \leq j \leq b$. The null and alternative hypothesis for both sets of tests can be stated as:

- (a) $H_0: \mu_1^T = \mu_2^T = \dots = \mu_a^T$
 $H_1: \text{not all of the means of the treatments are equal}$
- (b) $H_0: \mu_1^B = \mu_2^B = \dots = \mu_b^B$
 $H_1: \text{not all of the means of the blocks are equal}$

Let $\bar{x}_{i,j}$ be the sample mean of all observations in the i^{th} treatment and j^{th} block. Let \bar{x}_i^T be the sample mean of all of the block means in the i^{th} treatment and let \bar{x}_j^B be the sample mean of all the treatment means in the j^{th} block. Let $\bar{x} = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b \bar{x}_{i,j}$ be the sample mean of all means.

The total sum of squares ($SS(Total)$), the treatment sum of squares (SST), the block sum of squares (SSB) and the error sum of squares (SSE) are given as:

$$SS(Total) = \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{i,j} - \bar{x})^2$$
$$SST = b \sum_{i=1}^a (\bar{x}_i^T - \bar{x})^2$$
$$SSB = a \sum_{j=1}^b (\bar{x}_j^B - \bar{x})^2$$
$$SSE = \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{i,j} - \bar{x}_i^T - \bar{x}_j^B + \bar{x})^2$$

The four of sum of squares are related by:

$$SS(Total) = SST + SSB + SSE$$

The mean square of treatments (MST), the mean square of blocks (MSB) and the mean square error (MSE) are defined as:

$$MST = \frac{SST}{a - 1}$$

$$MSB = \frac{SSB}{b - 1}$$

$$MSE = \frac{SSE}{(a - 1)(b - 1)}$$

Under the null hypothesis that the mean of all treatments are equal, the test statistic is given as $f_T = \frac{MST}{MSE}$ where f_T is F-distributed with $a - 1$ degrees of freedom in the numerator and $(a - 1)(b - 1)$ degrees of freedom in the denominator. Under the null hypothesis that the mean of all blocks are equal, the test statistic is given as $f_B = \frac{MSB}{MSE}$ where f_B is F-distributed with $b - 1$ degrees of freedom in the numerator and $(a - 1)(b - 1)$ degrees of freedom in the denominator.

Using Statistician (Two Factor Table without replication)

When the user selects the **Two Factor Table without replication** radiobutton from the **ANOVA Table Type** groupbox, user then selects two or more *Data Variables* from the **Data Variables** listbox. Each of these *Data Variables* are the treatments. Each observation in a treatment is the mean of a block. (Hence, the first observation of each *Data Variable*, is the mean of the first block, the second observation of each *Data Variable*, is the mean of the second block and so on.) Each treatment can be thought of as a column of data. Each block can be considered to be a row of data.

The number of observations, sum, average and variance of the means of treatments and blocks is outputted. A table describing the source of variation as displayed below is then reported.

Source of Variation	d.f.	Sum of Squares	Mean Squares	F Statistic	F critical	p-value
Treatments:	$a-1$	SST	MST	$f_T = MST / MSE$	$F_{(\alpha, a-1, n-1)}$	$P(F > f_T)$
Blocks:	$b-1$	SSB	MSB	$f_B = MSB / MSE$	$F_{(\alpha, b-1, n-1)}$	$P(F > f_B)$
Error:	$(a-1)(b-1)$	SSE	MSE			
Total:	$ab-1$	$SS(Total)$				

The null and alternate hypothesis for the test on treatments and blocks, the test significance level, the F statistic, critical value, p-value, the decision rule and the

conclusion are also reported. If the **Output Summary Statistics** checkbox is checked then summary statistics for each of the treatments and blocks are outputted.

(c) Two factor ANOVA with Replication.

The two factor ANOVA model is now extended to the case where there are two experimental factors (treatments and blocks), and these two factors have an interaction effect. An interaction effect takes place if the response of one factor is dependent on the level of the other factor. Three sets of hypothesis are of interest:

- (a) A test of the hypothesis that the means of all of the treatments are equal.
- (b) A test of the hypothesis that the means of all of the blocks are equal.
- (c) A test of the hypothesis that the interaction effect between factors is zero.

Let μ_i^T be the mean of treatment i where $1 \leq i \leq a$ and let μ_j^B be the mean of block j where $1 \leq j \leq b$. Let $\mu_{i,j}^I$ be the interaction effect between treatment i and block j . There will be $a \times b$ interaction terms. It is also assumed that the experiment is replicated r times.

The null and alternative hypothesis for the sets of tests can be stated as:

- (a) $H_0: \mu_1^T = \mu_2^T = \dots = \mu_a^T$
 $H_1: \text{not all of the means of the treatments are equal}$
- (b) $H_0: \mu_1^B = \mu_2^B = \dots = \mu_b^B$
 $H_1: \text{not all of the means of the blocks are equal}$
- (c) $H_0: \mu_{1,1}^I = \mu_{1,2}^I = \dots = \mu_{a,b}^I = 0$
 $H_1: \text{not all of the interaction effects are equal to zero}$

The following notation will be used in this section.

- $\bar{x}_{i,j,k}$ is the observation in the i^{th} treatment and j^{th} block in the k^{th} replication.
- \bar{x}_i^T is the sample mean of all observations in the the i^{th} treatment.
- \bar{x}_j^B is the sample mean of all observations in the the j^{th} block.
- \bar{x}_k^I is the sample mean of all observations in the the k^{th} replication.
- $\bar{x}_{i,j}^{TB}$ is the sample mean of all observations in the i^{th} treatment and j^{th} block
- $\bar{x} = \frac{1}{abr} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r \bar{x}_{i,j,k}$ is the sample mean of all observations.

The total sum of squares ($SS(\text{Total})$), the treatment sum of squares (SST), the block sum of squares (SSB), the interaction sum of squares (SSI) and the error sum of squares (SSE) are given as:

$$SS(Total) = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (\bar{x}_{i,j,k} - \bar{x})^2$$

$$SST = br \sum_{i=1}^a (\bar{x}_i^T - \bar{x})^2$$

$$SSB = ar \sum_{j=1}^b (\bar{x}_j^B - \bar{x})^2$$

$$SSI = r \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{i,j}^{TB} - \bar{x}_i^T - \bar{x}_j^B + \bar{x})^2$$

$$SSE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (\bar{x}_{i,j,k} - \bar{x}_{i,j}^{TB})^2$$

The five sum of squares are related by:

$$SS(Total) = SST + SSB + SSI + SSE$$

The mean square of treatments (MST), the mean square of blocks (MSB), the mean square of interaction (MSI) and the mean square error (MSE) are defined as:

$$MST = \frac{SST}{a - 1}$$

$$MSB = \frac{SSB}{b - 1}$$

$$MSI = \frac{SSI}{(a - 1)(b - 1)}$$

$$MSE = \frac{SSE}{ab(r - 1)}$$

Under the null hypothesis that the mean of all treatments are equal, the test statistic is given as $f_T = \frac{MST}{MSE}$ where f_T is F-distributed with $a - 1$ degrees of freedom in the numerator and $ab(r - 1)$ degrees of freedom in the denominator. Under the null hypothesis that the mean of all blocks are equal, the test statistic is given as $f_B = \frac{MSB}{MSE}$ where f_B is F-distributed with $b - 1$ degrees of freedom in the numerator and $ab(r - 1)$ degrees of freedom in the denominator. Under the null hypothesis that the mean of all replications are equal, the test statistic is given as $f_I = \frac{MSI}{MSE}$ where f_I is F-distributed

with $(a - 1)(b - 1)$ degrees of freedom in the numerator and $ab(r - 1)$ degrees of freedom in the denominator.

Using Statistician (Two factor ANOVA with Replication)

When the user selects the **Two Factor Table with replication** radiobutton from the **ANOVA Table Type** groupbox, user then selects two or more *Data Variables* from the **Data Variables** listbox. All treatment data is stored in a *Data Variable*. Each treatment *Data Variable* is structured as follows. The set of observations for the first block (replications 1 to r) are stored at the top of the *Data Variable*. The next set of observations for the second block is then stored below the first set of block observations. This process is continued until all replication data is stored in the treatment *Data Variable*. A schematic representation of the data is as follows:

	Treatment 1	Treatment 2	...	Treatment a	
Block 1	29	63	...	55	Replication 1
Block 1	37	57	...	64	Replication 2
			...		
Block 1	23	68	...	69	Replication r
Block 2	73	80	...	84	Replication 1
Block 2	78	74	...	79	Replication 2
			...		
Block 2	65	83	...	82	Replication r
Block b	61	79	...	88	Replication 1
Block b	65	74	...	96	Replication 2
			...		
Block b	57	69	...	95	Replication r

The number of observations, sum, average and variance of the means of treatments and blocks is outputted. A table describing the source of variation as displayed below is then reported.

Source of Variation	d.f.	Sum of Squares	Mean Squares	F Statistic	F critical	p-value
Treatments:	$a-1$	SST	MST	$f_T = MST / MSE$	$F_{(\alpha, a-1, ab(r-1))}$	$P(F > f_T)$
Blocks:	$b-1$	SSB	MSB	$f_B = MSB / MSE$	$F_{(\alpha, b-1, ab(r-1))}$	$P(F > f_B)$
Interaction:	$(a-1)(b-1)$	SSI	MSI	$f_I = MSI / MSE$	$F_{(\alpha, (a-1)(b-1), ab(r-1))}$	$P(F > f_I)$
Error:	$ab(r-1)$	SSE	MSE			
Total:	$abr-1$	$SS(Total)$				

The null and alternate hypothesis for the test on treatments and blocks, the test significance level, the F statistic, critical value, p-value, the decision rule and the conclusion are also reported. If the **Output Summary Statistics** checkbox is checked then summary statistics for each of the treatments and blocks are outputted.

Section (10) - Regression Analysis

The **Regression** form is selected by clicking the **Regression** button on the *Statistician* ribbon tab.

Statistician (Regression Analysis)

Dependent Variable
Select a Data Variable

Independent Variables

Constant
 With Constant
 Without Constant

Standard Errors
 OLS
 White (1980)
 Newey West (1987)

Sample Range
 Use All Observations
 User Defined
From
to

Output
 Regression Statistics
 Variance Inflation Factors
 Residual Diagnostics
 Variance Covariance Matrix
 Fitted Values and Residuals
 Mallows Cp Analysis

Format Output

Output Results

Close

Manage Data

Technical Discussion

The standard linear regression model with k independent variables (one of which may be a constant), with n observations is given as:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i$$

where y_i is the i^{th} observation on the dependent variable, $x_{j,i}$ is the i^{th} observation on the j^{th} independent variable, β_i is the i^{th} coefficient of the i^{th} independent variable that is to be estimated and ε_i is the i^{th} residual (or error, or disturbance) term. The expression in (*) can be written more compactly as:

$$Y = X\beta + \varepsilon.$$

where Y is a $(n \times 1)$ vector of observations on the dependent data variable, X is a $(n \times k)$ matrix of observations on the independent data variables, β is a $(k \times 1)$ vector of fixed coefficients and ε is a $(n \times 1)$ vector of residuals. An estimate of the regression coefficients is given as:

$$b = (X'X)^{-1}X'Y,$$

where b is a $(k \times 1)$ vector of coefficient estimates. The estimated residuals from the regression are given as:

$$\hat{\varepsilon} = Y - Xb$$

where $\hat{\varepsilon}$ is a $(n \times 1)$ vector of residual estimates. The sum of squares of the residuals is given as:

$$SSR = \hat{\varepsilon}'\hat{\varepsilon}.$$

The total sum of squares of the regression is given as:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

where \bar{y} is the mean of the dependent variable y . The regression sum of squares is then given by:

$$RSS = TSS - SSR$$

An estimate of the variance of the residuals is given as:

$$s^2 = \frac{SSR}{n - k}$$

The OLS standard errors of the coefficient estimates are given by the square root of diagonal elements of the OLS variance-covariance matrix defined as:

$$\hat{\Sigma}_{OLS} = s^2(X'X)^{-1}$$

where $\hat{\Sigma}_{OLS}$ is a $(k \times k)$ matrix. The i^{th} OLS t-statistic of the coefficient estimates is given as:

$$t_{stat,i} = \frac{b_i}{se_i}$$

where b_i is the i^{th} coefficient estimate and se_i is the i^{th} standard error of the coefficient estimate. The OLS p-value of the coefficient estimate is taken from a t-distribution with $n - k$ degrees of freedom and is the probability that the coefficient is equal to zero.

The White (1980) estimate of the parameter variance-covariance matrix corrects for heteroscedasticity in the residuals and is given as:

$$\hat{\Sigma}_{White} = \frac{n}{n-k} (X'X)^{-1} \hat{\Omega} (X'X)^{-1}$$

where

$$\hat{\Omega} = \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i'$$

and x_i' is the i^{th} row of X .

The Newey-West (1987) estimate of the parameter variance-covariance matrix corrects for both autocorrelation and heteroscedasticity in the residuals and is given as:

$$\hat{\Sigma}_{Newey West} = \frac{n}{n-k} (X'X)^{-1} \hat{\Omega} (X'X)^{-1}$$

where

$$\hat{\Omega} = \left\{ \sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i' + \sum_{v=1}^q \left[\left(1 - \frac{v}{q+1} \right) \sum_{i=v+1}^n (x_i \hat{\varepsilon}_i \hat{\varepsilon}_{i-v} x_{i-v}' + x_{i-v} \hat{\varepsilon}_{i-v} \hat{\varepsilon}_i x_i') \right] \right\}$$

and q is set equal to largest integer lower than $4 \left(\frac{n}{100} \right)^{2/9}$. When White (1980) or Newey-West (1987) standard errors are selected, the t-ratios and p-values of the parameter estimates are adjusted accordingly.

The regression R^2 and adjusted R^2 (\bar{R}^2) are defined as:

$$R^2 = \frac{SSR}{TSS}$$

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k}$$

respectively. The log likelihood of the regression is given as:

$$l = -\frac{n}{2} \left(1 + \log(2\pi) + \log \left(\frac{SSR}{n} \right) \right)$$

The Akaike and Schwarz information criterion are given as:

$$AIC = -\frac{2l}{n} + \frac{2k}{n}$$

$$SIC = -\frac{2l}{n} + \frac{k \times \log(k)}{n}$$

respectively. The F statistic is given as:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

and has an F distribution with $k-1$ degrees of freedom in the numerator and $n-k$ degrees of freedom in the denominator. The probability of F is the probability that all coefficients are equal to zero.

The Skewness statistic is a measure of the skewness of the estimated residuals and is given by:

$$Skew = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}}{s} \right)^3$$

Where $\bar{\hat{\varepsilon}}$ is the mean of the estimated residuals. The mean of the estimated residuals is equal to zero if a constant is included in the regression and is usually non zero otherwise. The standard error of the skewness estimate is given as:

$$se \text{ of Skewness} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

The Excess Kurtosis statistic is a measure of the kurtosis of the estimated residuals and is given by:

$$Excess \text{ Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

The standard error of the excess kurtosis estimate is given as:

$$se \text{ of Kurtosis} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$$

The Jacque-Bera test statistic tests for the normality of the residuals and has a $\chi^2_{(2)}$ distribution. It is given as:

$$JB = n \left(\frac{Skew^2}{6} + \frac{Excess\ Kurtosis^2}{24} \right)$$

The Durban Watson test statistic tests for autocorrelation in the residuals and is given by:

$$DW = \frac{\sum_{i=2}^n (\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^n \hat{\epsilon}_i^2}$$

The Variance Inflation Factor (VIF) statistic in a regression is employed as an indicator of the contribution to multicollinearity of an individual regressor (or independent variable). The VIF of regressor k is given as:

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2 computed from regressing the k^{th} regressor upon all other regressors. (Note: a regression must have at least two independent variables for the VIF statistic to make any sense.)

Mallows' C_p Statistic

The Mallows C_p statistic is often employed as a criterion to select a subset of p regressors from k potential regressors ($p \leq k$) in multiple linear regressions. (The intercept is included in p .) The Mallows C_p statistic with p regressors is defined as:

$$C_p = \frac{SSR_p}{\hat{\sigma}^2} + 2p - n$$

where SSR_p is the residual sums of squares for the model with p regressors, $\hat{\sigma}^2$ is the estimated variance of the error term of the full model with k regressors and n is the sample size. When an important regressor has been omitted from the regression it is expected that $C_p > p$. The favoured model is one where C_p is close to or less than p .

Using Statistician (Regression)

To perform a regression the user selects a dependent variable from the **Dependent Variable** combobox and then selects one or more variables from the **Independent Variables** listbox. Checking either **With Constant** or **Without Constant** in the **Constant** groupbox determines if a constant term is estimated in the regression. The regression can be performed with all observations by clicking the **Use All**

Observations option in the **Sample Range** frame. If the **User Defined** option is selected in the **Sample Range** frame, the regression can be restricted to a subset of all of the observations by entering the starting number and ending number of the range of observations in the **From** and **To** textbox that will be enabled.

The user can select the type of standard errors and variance-covariance matrix of the parameter estimates that will be reported in the **Standard Errors** groupbox. At least one type of standard error (OLS, White or Newey West) must be selected.

The user can select the type of output from the regression in the **OutPut** groupbox. If no option is selected in the **OutPut** frame then *xlStatistician* reports:

- dependent variable name,
- sample range,
- number of observations in the regression,
- estimation method,
- parameter estimates,
- parameter estimate standard errors (OLS, White and/or Newey West),
- parameter estimate t-ratio's,
- parameter estimate p-values.

Output from each of the four options in the **OutPut** groupbox are as follows:

(1) Regression Statistics

- mean of the dependent variable,
- standard error of the dependent variable,
- R^2 ,
- adjusted R^2 ,
- standard error of the regression,
- total sums of squares,
- regression sums of squares,
- residual sums of squares,
- F statistic,
- P-value of F statistic,
- Log likelihood,
- Akaike information criterion,
- Scwharz information criterion.

(2) Variance Inflation Factors

The Variance Inflation Factors (VIF) are reported alongside the table of parameter estimates, standard errors, t ratios and p-values.

(3) Residual Diagnostics

- skewness,
- standard error of skewness,
- kurtosis,
- standard error of kurtosis,
- Jacque-Bera test statistic,
- Durbin-Watson test statistic.

(3) Variance–Covariance Matrix

This may include OLS, White or Newey West variances and covariances depending on the type of standard errors that are selected.

(4) Fitted Values and Residuals

The estimated fitted values and residuals from the regression are numbered and reported.

(5) Mallows C_p Analysis

A Mallows C_p analysis is conducted with output p , C_p , R^2 , adjusted R^2 and a list of the relevant Data variables. Output is sorted by the value of C_p . (Note, no more than 9 regressors can be selected in a Mallows C_p analysis.)

Section (11) – Binary Models (not available in *Statistician (Lite)*)

The **Binary Models** form is selected by clicking the **Binary Model** button on the *Statistician* ribbon tab.

Statistician (Binary Models)

Dependent Variable
Select a Data Variable

Dependent Variable Data
Success = ???
Failure = ???
Switch

Independent Variables

Model
 Logit
 Probit

Constant
 With Constant
 Without Constant

Sample Range
 Use All Observations
 User Defined
From
to

Output
 Model Statistics
 Variance Covariance Matrix
 Estimation Information
 Fitted Values and Residuals

Format Output

Output Results

Close

Manage Data

Technical Discussion

The Probit and Logit (Logistic) models model the probability that a binary dependent variable equals one. This probability is given as a function $F(\cdot)$ of the explanatory variables (X) and a vector of parameters β . The models are specified as:

$$P(Y = 1|X) = F(X, \beta)$$
$$P(Y = 0|X) = 1 - F(X, \beta)$$

Binary dependent variable models have the dependent variable assigned the value of zero (0) or one (1). Let y_i be the i^{th} binary observation on the dependent variable (where $1 \leq i \leq n$), let Y be a $(n \times 1)$ vector of the dependent variable observations, let $x_{j,i}$ be the i^{th} observation on the j^{th} independent variable (where $1 \leq j \leq k$), let X be a $(n \times k)$ matrix of the independent variable observations, let x'_i be a $(1 \times k)$ vector which is the i^{th} row of X , let β_i be the i^{th} coefficient of the i^{th} independent variable that is to be estimated, let β be a $(k \times 1)$ vector of independent variable coefficients and ε_i is the i^{th} residual (or error, or disturbance) term.

For the Probit model, the function $F(\cdot)$ as specified as:

$$F(X, \beta) = \int_{-\infty}^{x'\beta} \phi(t)dt = \Phi(X'\beta)$$

where $\phi(\cdot)$ is the standard normal distribution function and $\Phi(\cdot)$ is the cumulative standard normal distribution function. For the Logit model, the function $F(\cdot)$ as specified as:

$$F(X, \beta) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

The parameters of the Probit and Logit models are estimated by maximising a log likelihood function of the form:

$$\ln(L) = \sum_{i=1}^n [y_i \ln(F(x'_i)) - (1 - y_i) \ln(1 - F(x'_i))]$$

The parameter estimates which maximise the log likelihood function are held in a $(k \times 1)$ vector denoted by (b) .

For the Logit model, the Hessian matrix (H) is given as:

$$H = - \sum_{i=1}^n \Lambda_i(1 - \Lambda_i)x_i x'_i$$

where $\Lambda_i = \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}}$. For the Probit model, the Hessian matrix (H) is given as:

$$H = \sum_{i=1}^n -\lambda_i(\lambda_i - x'_i \beta)x_i x'_i$$

where $\lambda_i = \frac{q_i \phi(q_i x_i' \beta)}{\Phi(q_i x_i' \beta)}$ and $q_i = 2y_i - 1$. The square root of the diagonal elements of the inverse of the Hessian matrix provides maximum likelihood standard errors for the parameter estimates.

For the Logit model, the first partial derivative with respect to β is given as:

$$\frac{\partial \ln(L)}{\partial \beta} = \sum_{i=1}^n (y_i - \Lambda_i) x_i$$

and for the Probit model the first partial derivative with respect to β is given as:

$$\frac{\partial \ln(L)}{\partial \beta} = \sum_{i=1}^n \lambda_i x_i$$

The estimated residuals from the models are given as:

$$\hat{\varepsilon} = Y - F(Xb)$$

where $\hat{\varepsilon}$ is a $(k \times 1)$ vector of residual estimates and $F(\cdot)$ is the functional form for either the Probit or Logit model.

Let $l(b)$ denote the maximum of the log likelihood function for a Probit or Logit model. The average log likelihood is given as $l(b)/n$. The restricted log likelihood is estimated with only a constant as the independent variable and is denoted by $l(\tilde{b})$. The LR statistic tests the joint hypothesis that all of the independent variables, (except the constant), are equal to zero. It is given as:

$$LR = -2(l(\tilde{b}) - l(b))$$

The asymptotic distribution of the LR statistic has a χ_{k-1}^2 distribution from which the probability of the LR statistic can be obtained. The McFadden R^2 is a likelihood ratio index and is defined as:

$$McFadden R^2 = 1 - \frac{l(b)}{l(\tilde{b})}$$

The Akaike, Bayesian and Hannan-Quinn information criterion are given as:

$$AIC = -\frac{2l(b)}{n} + \frac{2k}{n}$$

$$BIC = -\frac{2l(b)}{n} + \frac{k \ln(k)}{n}$$

$$HIC = -\frac{2l(b)}{n} + \frac{2k \ln(\ln(k))}{n}$$

respectively.

Using Statistician (Binary Models)

To estimate a Binary regression model, the model to be estimated is selected by selecting either **Probit** or **Logit** option in the **Model** groupbox. The user selects a binary dependent variable from the **Dependent Variable** combobox and then selects one or more independent variables from the **Independent Variables** listbox. The binary dependent variable must have one of two distinct values. These distinct values may be numeric or text and represent either a '1' or '0'. The symbols for '1' and '0' are generated by *Statistician* after analyzing the selected *Data Variable*. *Statistician* enters in the symbols for '1' or '0' in the **1 =** and **0 =** labels in the **Dependent Variable Data** groupbox. Clicking the **Switch** button switches the symbols for '1' and '0'.

Checking either **With Constant** or **Without Constant** in the **Constant** frame determines if a constant term is estimated in the model. The regression can be performed with all observations by clicking the **Use All Observations** option in the **Sample Range** groupbox. If the **User Defined** option is selected in the **Sample Range** groupbox, the regression can be restricted to a subset of all of the observations by entering the starting number and ending number of the range of observations in the **From** and **To** textbox that will be displayed.

The user can select the type of output from the regression in the **OutPut** groupbox. If no option is selected in the **OutPut** frame then *Statistician* reports:

- dependent variable name,
- sample range,
- number of observations in the regression,
- model estimated,
- parameter estimates,
- parameter estimate standard errors,
- parameter estimate t-ratio's,
- parameter estimate p-values.

Output from each of the four options in the **OutPut** groupbox are as follows:

(1) Model Statistics

- count of dependent variable = 0,
- count of dependent variable = 1,
- mean of the dependent variable,
- standard error of the dependent variable,
- residual sums of squares,
- standard error of the regression,
- log likelihood,
- average log likelihood,
- Akaike information criterion,
- Schwarz information criterion,
- Hannan Quinn information criterion,
- McFadden R^2 ,
- restricted log likelihood,
- LR statistic,
- probability of LR statistic.

(2) Variance–Covariance Matrix

Maximum likelihood variances and covariances are reported.

(3) Estimation Information

- estimation algorithm,
- tolerance,
- iterations required,
- function evaluations required,
- starting values.

(4) Fitted Values and Residuals

The fitted values estimated residuals from the regression are numbered and reported.

Section (12) – Count Models (not available in *Statistician (Lite)*)

A count model is employed to estimate the mean number of times an event occurs conditional upon a number of independent variables. A commonly used count model is the Poisson regression.

If a random variable Y has a Poisson distribution, then the probability that Y has a specific value (y) is given as:

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where $E[Y] = \mu$ and $Var[Y] = \mu$. Let x_i denote the i^{th} independent variable and let β_i denote the coefficient of the i^{th} independent variable. A linear combination of independent variables is given as $x'\beta = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_p$. Let Y_i denote the number of events observed from n_i repetitions for the i^{th} covariate pattern. The expected value of Y_i in a Poisson regression is given as:

$$E[Y_i] = \mu_i = n_i e^{x_i'\beta}$$

or equivalently:

$$\log(\mu_i) = \log(n_i) + x_i'\beta$$

The parameters of the Poisson model are estimated with maximum likelihood techniques. The log-likelihood of the Poisson model is given as:

$$\ln(L) = \sum_{i=1}^N (-\exp(x_i'\beta) + y_i x_i'\beta - \ln(y_i!))$$

where N is the number of observations. The gradient function is given as:

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^N (x_i'(y_i - \exp(x_i'\beta)))$$

The Hessian (H) of the Poisson log likelihood function is given as:

$$H = \frac{\partial^2 L}{\partial \beta \partial \beta'} = - \sum_{i=1}^n (\exp(x_i'\beta) x_i x_i')$$

The square root of the diagonal elements of the inverse of the Hessian matrix provides maximum likelihood standard errors for the parameter estimates.

Fitted values from a Poisson regression are given as:

$$\hat{Y}_i = \hat{\mu}_i = n_i e^{x_i' b}$$

where b is the maximum likelihood estimate of β . Residuals from the regression are given as:

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Pearson residuals are given as:

$$r_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

The chi-squared goodness of fit statistic is given as:

$$X^2 = \sum_i r_i^2$$

The deviance for a Poisson model is given as:

$$D = 2 \sum_i \left[Y_i \log \left(\frac{Y_i}{\hat{\mu}_i} \right) - (Y_i - \hat{\mu}_i) \right]$$

X^2 and D are approximately equal with a chi squared distribution with $N - p$ degrees of freedom where N is the number of observations and p is the number of parameters in the model.

Let $l(b)$ denote the maximum of the log likelihood function for a Poisson model. The restricted log likelihood is estimated with only a constant as the independent variable and is denoted by $l(\tilde{b})$. The LR statistic tests the joint hypothesis that all of the independent variables, (except the constant), are equal to zero. It is given as:

$$LR = -2 \left(l(\tilde{b}) - l(b) \right)$$

The asymptotic distribution of the LR statistic has a χ_{k-1}^2 distribution from which the probability of the LR statistic can be obtained. The pseudo R^2 is a likelihood ratio index and is defined as:

$$pseudo R^2 = 1 - \frac{l(b)}{l(\tilde{b})}$$

The Akaike information criterion, corrected Akaike information criterion and Bayesian information criterion are given as:

$$AIC = -\frac{2l(b)}{n} + \frac{2k}{n}$$

$$\text{Corrected AIC} = -\frac{2l(b)}{n} + \frac{2k}{n - k - 1}$$

$$BIC = -\frac{2l(b)}{n} + \frac{k \ln(k)}{n}$$

Using Statistician (Count Models)

To estimate a Poisson regression, the user selects a binary dependent variable from the **Dependent Variable** combobox and then selects the independent variables from the **Independent Variables** listbox. If no independent variables are selected then a constant term must be included in the regression. Checking either **With Constant** or **Without Constant** in the **Constant** frame determines if a constant term is estimated in the model. The regression can be performed with all observations by clicking the **Use All Observations** option in the **Sample Range** groupbox. If the **User Defined** option is selected in the **Sample Range** groupbox, the regression can be restricted to a subset of all of the observations by entering the starting number and ending number of the range of observations in the **From** and **To** textbox that will be displayed.

The user can select the type of output from the regression in the **OutPut** groupbox. If no option is selected in the **OutPut** frame then *Statistician* reports:

- dependent variable name,
- sample range,
- number of observations in the regression,
- model estimated,
- parameter estimates,
- parameter estimate standard errors,
- parameter estimate t-ratio's,
- parameter estimate p-values.

Output from each of the four options in the **OutPut** groupbox are as follows:

(1) Model Statistics

- mean of the dependent variable,
- variance of the dependent variable,
- deviance,
- log likelihood,
- restricted log likelihood,
- Akaike information criterion,

- corrected Akaike information criterion,
- Schwarz information criterion,
- Pseudo R^2 ,
- LR statistic,
- Chi square statistic.

(2) Variance–Covariance Matrix

Maximum likelihood variances and covariances are reported.

(3) Estimation Information

- estimation algorithm,
- tolerance,
- iterations required,
- function evaluations required,
- starting values.

(4) Fitted Values and Residuals

The fitted values, estimated residuals and estimated Pearson residuals from the regression are numbered and reported.

Section (13) - Time Series (not available in *Statistician (Lite)*)

Any variable that changes in value over time is referred to as a time series. The **Time Series** groupbox offers the user a number of forms to investigate the behavior of a number of economic and business time series. These forms include:

- (a) Forecasting – users can smooth a time series, estimate seasonal indexes and produce forecasts,
- (b) Holt-Winters – includes the one, two and three factor model, and the
- (c) Hodrick-Prescott filter.

(a) Forecasting

The screenshot shows the 'Statistician (Time Series - Forecasting)' dialog box. It is organized into several sections:

- Data Variable:** A dropdown menu with 'Select a Data Variable' selected.
- Smoothing Method:** Includes radio buttons for 'Average', 'Moving Average Methods' (Contemporaneous, Lagged, Centered, Weighted), and 'Trend Methods' (Linear, Quadratic, Exponential, Autoregressive). The 'Lagged' method has an 'Interval' field, and the 'Weighted' method has a 'Weighting Data Variable' dropdown. The 'Autoregressive' method has a 'Lags' field set to 1.
- Season Type:** Includes radio buttons for 'Multiplicative' and 'Additive'. Below it is a 'Seasons' section with radio buttons for 'Weekdays (All)', 'Weekdays (Excluding Sun)', 'Weekdays (Excluding Sat and Sun)', 'Monthly', 'Quarterly', 'Numeric - Number of Seasons', and 'From Data Variable'. There is also a 'Seasonal Data Variable' dropdown and a 'Data Begins at Season' dropdown.
- Output:** Includes checkboxes for 'Original Series', 'Smoothed Series', 'Percentage Of Trend', 'DeSeasonalized Series', 'Fitted Series', and 'Forecast to -'. There is also a 'Format Output' checkbox.
- Buttons:** 'Manage Data', 'Output Results', and 'Close'.

If the **Forecasting** menu item is selected from the the **Time Series** button the user is offered a total of nine smoothing methods. There is a simple averaging method, four moving average methods and four trend methods to smooth the random effects out of a time series. (Holt-Winters methods are implemented on the Holt-Winters form). The

nine methods are defined as follows. In the following discussion, let S_t be the smoothed observation on the time series observation y_t .

(1) Average

Each value in the smoothed series is simply the arithmetic average of all observations in the original series.

(2) Moving average smoothing techniques

Initially the user enters a moving average interval in the **Interval** textbox that is displayed. This interval must be a positive integer ($I \geq 2$).

(i) Contemporaneous Moving Average – The smoothed observation is the arithmetic average of the current and past ($I-1$) observations. If for example, a three point moving average is selected, that is ($I=3$), then the smoothed estimate for the current observation is given as $S_t = (y_t + y_{t-1} + y_{t-2})/3$. In this example, the first two observations for the smoothed series are not defined.

(ii) Lagged Moving Average –The smoothed observation is the arithmetic average of the past (I) observations. If for example, a three point moving average is selected, that is ($I=3$), then the smoothed estimate for the current observation is given as $S_t = (y_{t-1} + y_{t-2} + y_{t-3})/3$. In this example, the first three observations for the smoothed series are not defined.

(iii) Centered Moving Average – If the interval (I) is an odd number, then the smoothed observation is the average of an equal number of leads and lags of the original observation. For example, if $I=3$, then $S_t = (y_{t+1} + y_t + y_{t-1})/3$. If the interval is an even number then the smoothed observation is a two point moving average of two moving averages, where the first moving average has one more lead term than the second moving average. For example, if $I=4$ then $S_t = \frac{(y_{t+2} + y_{t+1} + y_t + y_{t-1})/4 + (y_{t+1} + y_t + y_{t-1} + y_{t-2})/4}{2}$. Using a centered moving average involves $int(I/2)$ smoothed observations being undefined at the start and at the end of the smoothed time series. (The $int(.)$ function removes any values after the decimal point).

(iv) Weighted Moving Average – The current value of a smoothed series is the weighted average of the past (l) observations, that is, $S_t = \sum_{j=1}^l y_{t-j} w_j$. The weights must be in the range $0 < w_j < 1$ and sum to one, that is $\sum_{j=1}^l w_j = 1$. The weights are defined by selecting a *Data Variable* containing the weights from the **Select a Weighting Data Variable** combobox. When using the Weighted Moving Average technique users can readily find the optimal value of the weights that minimize the MAD, MAPE or RMSE (defined later), with the use of Excel solver.

(3) Trend smoothing techniques

(i) Linear Trend – The time series is modeled as a linear function of a time variable $t = \{1, \dots, n\}$ where n is the number of observations. Using regression analysis, the following model is estimated.

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

Where β_0 and β_1 are fixed parameters and $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. Let b_0 and b_1 be least squares estimates of β_0 and β_1 respectively. The smoothed values of y_t are the fitted values from the regression, that is $S_t = \hat{y}_t = b_0 + b_1 t$ for any given value of t . An estimate of ε_t is given as $\hat{\varepsilon}_t = y_t - (b_0 + b_1 t)$.

(ii) Quadratic Trend - The time series is modeled as a quadratic function of the time variable t . Using regression analysis, the following model is estimated.

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$

where β_0 , β_1 and β_2 are fixed parameters and $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. Let b_0 , b_1 and b_2 be least squares estimates of β_0 , β_1 and β_2 respectively. The smoothed values of y_t are the fitted values from the regression, that is $S_t = \hat{y}_t = b_0 + b_1 t + b_2 t^2$ for any given value of t . An estimate of ε_t is given as $\hat{\varepsilon}_t = y_t - (b_0 + b_1 t + b_2 t^2)$.

(iii) Exponential Trend - The time series is modeled as an exponential function of the time variable t . Using regression analysis, the following model is estimated.

$$\ln(y_t) = \beta_0 + \beta_1 t + \varepsilon_t$$

where β_0 and β_1 are fixed parameters and $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. Let b_0 , b_1 be least squares estimates of β_0 and β_1 respectively. The smoothed values of y_t is given by $S_t = \hat{y}_t = e^{(b_0 + b_1 t)}$ for any given value of t . An estimate of ε_t is given as $\hat{\varepsilon}_t = y_t - e^{(b_0 + b_1 t)}$.

(iv) Autoregressive Trend - The time series is modeled as a function of p past values of the dependent variable. Using regression analysis, the following model is estimated.

$$y_t = \alpha_0 + \sum_{i=1}^p \alpha_i y_{t-i} + \varepsilon_t$$

where α_i is a fixed parameter and $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. Let a_i be least squares estimates of α_i . The smoothed values of y_t is given by $S_t = \hat{y}_t = a_0 + \sum_{i=1}^p a_i y_{t-i}$. An estimate of ε_t is given as $\hat{\varepsilon}_t = y_t - a_0 - \sum_{i=1}^p a_i y_{t-i}$.

Cyclical effects can be isolated in a time series using the following methodology.

- (1) For each time period, calculate the smoothed value S_t using one of the methods outlined above.
- (2) Calculate the percentage of trend (P_t) as $P_t = (y_t / S_t) \times 100\%$.
- (3) A visual inspection of the graph of P_t over time will be centered on 100% in the horizontal direction. If P_t is consistently above or below the 100% for an extended period of time then this may indicate the existence of a cyclical effect.

Statistician outputs the percentage of trend series when the **Percentage of Trend** checkbox is checked.

Seasonal Indexes

A time series $\{y_t\}$ is often modeled with four components:

- (a) A trend component (T_t), which is a long term pattern or direction that the time series exhibits,
- (b) A cyclical component (C_t), which is a wavelike pattern around the trend which typically becomes apparent over a number of years,
- (c) A seasonal component (S_t), which is a repetitive pattern which occurs typically in weekly, monthly, quarterly or annual cycles,
- (d) A random component (R_t), which are irregular and unpredictable patterns not associated with any of the other components.

Statistician models two common time series models, the additive model and the multiplicative model. The additive model is defined as:

$$y_t = T_t + C_t + S_t + R_t$$

The multiplicative model is defined as:

$$y_t = T_t \times C_t \times S_t \times R_t$$

Seasonal indexes are calculated by initially smoothing the data. Data can be smoothed by using any method discussed previously. Steps are as follows:

Step 1.

Remove the effect of seasonal and random variations. This is done by creating a new series with a lagged or centered moving average process on the original series where the length of the moving average interval is set to the number of seasons. Alternately, a linear, quadratic, exponential or autoregressive trend model can be estimated from the original series. The fitted values from the estimated trend model is taken to be the new smoothed series. The new smoothed series ($Smoothed_t$) has only trend and cyclical components. Thus for the additive model $Smoothed_t = T_t + C_t$ and for the multiplicative model $Smoothed_t = T_t \times C_t$.

Step 2.

For the additive take the difference of the original series and the smoothed series $y_t - Smoothed_t = S_t + R_t$. For the multiplicative take the ratio of the original series and the moving average series $\frac{y_t}{Smoothed_t} = S_t \times R_t$. In both cases we are left with a new series that contains only seasonal and random components. Denote the new additive series as $A_t = y_t - Smoothed_t = S_t + R_t$ and the new multiplicative series as $M_t = \frac{y_t}{Smoothed_t} = S_t \times R_t$.

Step 3.

In each season, calculate the average of A_t or M_t . The result is the unadjusted seasonal index in each season. This averaging removes most but not all of the random effects.

Step 4.

The unadjusted seasonal indexes derived in step 3 are adjusted so that the average of the additive seasonal indexes is 0 and the average of the multiplicative indexes is 1. Assume that there are s seasons (and seasonal indexes). The i^{th} adjusted seasonal index for the additive model is given as $SI_t = A_t - \frac{\sum_{j=1}^s A_j}{s}$. The i^{th} adjusted seasonal index for the multiplicative model is given as $SI_t = \left(\frac{M_t}{\sum_{j=1}^s A_j} \right) \times s$.

The seasons can be defined in a number of different ways in the **Seasons** groupbox depending upon the selected radiobutton as follows.

Radiobutton	Seasons	Number of Seasons
Weekdays (All)	Mon, Tue Sat, Sun	7
Weekdays (Excluding Sun)	Mon, Tue Fri, Sat	6
Weekdays (Excluding Sat and Sun)	Mon, Tue, Wed, Thu, Fri	5
Monthly	Jan, Feb Dec	12
Quarterly	Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec	4
Numeric – Number of Seasons	User enters the number of seasons as an integer in an input box.	User defined
From Data Variable	User selects a Data Variable which contains the names of all of the seasons.	User defined

Users can select the season at which the data begins by making a selection in the **Data Begins at Season** combobox. By default, the data begins at the first season that is defined.

The t^{th} deseasonalized observation for an additive model is given as $y_t - SI_t$ and for a multiplicative model is given by y_t/SI_t , where SI_t is the adjusted seasonal index corresponding to time t . The t^{th} fitted observation (f_t) for an additive model is given as $f_t = S_t + SI_t$ and for a multiplicative model is given by $f_t = S_t \times SI_t$, where SI_t is the adjusted seasonal index corresponding to time t .

The Mean Absolute Deviation (*MAD*), Mean Absolute Percentage Error (*MAPE*) and the Root Mean Square Error (*RMSE*) are reported and defined as:

$$MAD = \frac{1}{s} \sum_{i=1}^s |y_t - S_t|$$

$$MAPE = \frac{100\%}{s} \sum_{i=1}^s \left| \frac{y_t - S_t}{y_t} \right|$$

$$RMSE = \sqrt{\frac{1}{s} \sum_{t=1}^s (y_t - S_t)^2}$$

where y_t is the t^{th} observation in the original series, S_t is the t^{th} observation in smoothed series and s is the number of smoothed observations (which will be different from the number of observations in the original series if a moving average method is employed). If seasonal indexes are not calculated for the original time series then the MAD, MAPE and RMSE are reported for the smoothed series, otherwise they are reported for the fitted series which is adjusted for seasonal effects.

If the user checks the **Forecast To** checkbox, a forecast horizon is then entered into the corresponding textbox. This forecast horizon must be an integer that is greater than the number of original observations. However no forecasts can be obtained if the time series has been smoothed with a Contemporaneous or Centered moving average method. (This is because these smoothing methods employ current or future values to construct a smoothed series.)

(b) Holt-Winters smoothing techniques

Statistician (Time Series - Holt Winters)

Data Variable
Select a Data Variable

Method
 Single (Simple)
 Double (Simple with Trend)
 Triple (Simple with Trend and Seasonals)

Number of Seasons

Season Type
 Multiplicative
 Additive

Factor Value(s)
Simple Factor (Alpha)
Trend Factor (Beta)
Seasonal Factor (Gamma)

Forecasts
 Forecast to -

Format Output

(i) Single (or Simple) exponential smoothing – The forecast F_{t+1} of a time series $\{y_1, \dots, y_T\}$ is given by:

$$F_{t+1} = \alpha y_t + (1 - \alpha)F_t$$

where the damping (or smoothing) factor α is a fixed parameter. The damping factor α is restricted to the range $0 < \alpha < 1$. The forecast series is produced by initially by setting $F_2 = y_1$. (Note that F_1 is undefined). Subsequent forecasts are calculated iteratively from the preceding forecast and preceding observation. The first out-of-sample forecast beginning after the final observation is given by $F_{T+1} = \alpha y_T + (1 - \alpha)F_T$. Thereafter the level of the series is flat (or constant) and given as:

$$F_{T+h} = F_{T+1} \quad (h = 1, 2, \dots)$$

In the single factor Holt-Winters model, the level of the series is identical to the forecast series.

(ii) Double exponential smoothing – If the time series contains a trend the Holt-Winters two factor model is defined by the equations:

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1}$$

$$F_{t+m} = L_t + mb_t$$

where α and β are fixed parameters in the range [0,1] and $\{L_t\}$, $\{b_t\}$ and $\{F_t\}$ are the level, trend and forecast series respectively. The forecast series is produced by initially by setting $F_1 = y_1$. The default initial value for the trend series is given as $b_1 = y_2 - y_1$. The m step ahead out-of-sample forecast is given by:

$$F_{t+m} = S_T + mb_T$$

(iii) Triple exponential smoothing - If the time series contains a trend and seasonal effects then the Holt-Winters three factor model is employed. The equations are dependent upon whether the seasonal indexes are additive or multiplicative. If the seasonal indexes are additive the equations are given as:

$$L_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1}$$

$$S_t = \gamma(y_t - L_t) + (1 - \gamma)S_{t-s}$$

where α , β and γ are fixed parameters in the range [0,1] and $\{L_t\}$, $\{b_t\}$, $\{F_t\}$ and $\{S_t\}$ are the level, trend, forecast and seasonal index series respectively. Parameter s is the number of seasons and S_{t-s} is the seasonal index for the time series at time t . The in-sample forecast for the three factor model with additive seasonal indexes is given as $F_t = L_{t-1} + b_{t-1} + S_{t-s}$. The m step ahead out-of-sample forecast beginning after the final observation in the series is calculated as:

$$F_{t+m} = L_T + mb_T \times S^*$$

where S^* is the final in-sample calculated seasonal index that corresponds to the forecast F_{t+m} . If the seasonal indexes are multiplicative then the equations are given as:

$$L_t = \alpha \frac{y_t}{S_{t-s}} + (1 - \alpha)(L_{t-1} + b_{t-1})$$

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1}$$

$$S_t = \gamma \frac{y_t}{L_t} + (1 - \gamma)S_{t-s}$$

The in-sample forecast for the three factor model with multiplicative seasonal indexes is given as $F_t = (L_{t-1} + b_{t-1}) \times S_{t-s}$. The m step ahead out-of-sample forecast beginning after the final observation in the series is calculated as:

$$F_{t+m} = (L_T + mb_T) \times S^*$$

The triple factor Holt Winters requires initial values for the seasonal indices and also for the initial trend value (b_1). These can be estimated in a variety of ways. Two seasons of data (or $2s$ observations) are employed in this task. The initial value of the trend at time s is given as:

$$b_L = \frac{1}{s} \left(\frac{y_{L+1} - y_1}{s} + \frac{y_{L+2} - y_2}{s} + \dots + \frac{y_{2s} - y_s}{s} \right) = \frac{1}{s^2} \sum_{i=1}^s (y_{L+i} - y_i) = \frac{1}{s^2} \left(\sum_{i=1}^s y_{s+i} - \sum_{i=1}^s y_i \right)$$

The initial value for the i^{th} seasonal index is the seasonal calculated from the first two seasons of data. For additive seasonal indexes it is given as:

$$s_i = \frac{y_i + y_{s+i}}{2} - A$$

where A is the arithmetic average of all observations in the first two seasons. For multiplicative seasonal indexes, the i^{th} seasonal index is given as:

$$s_i = \left(\frac{y_i + y_{s+i}}{2} \right) / A$$

The initial value of the smoothed series at time s is given as:

$$S_s = \frac{1}{s} \sum_{i=1}^s y_i$$

From these initial values of S_s , b_s and s_i the calculation of the series is given from time $s+1$.

Of course some users will wish to employ other initial values for the level, trend and seasons. This can be easily implemented by manually changing the formulas within Excel itself. Users may also wish to select the alpha, beta and gamma parameters based upon the criteria of minimizing the MAD, RMSE or MAPE. This is an optimization problem that is not trivial as the curve to be minimized often contains many local minima and maxima. Excel solver is a fast optimization tool that can be employed in this task. However it does not always find the global minimum. *Statistician Optimizer* (found in the **Tools** groupbox), can also be employed in the optimization task. It is slower than Excel Solver but does sometimes find a global minimum that Excel Solver does not find.

(c) Hodrick-Prescott filter

The Hodrick-Prescott is employed to separate the cyclical component (c_t) from the growth component (g_t) in a time series (y_t) ($t=1, \dots, T$). The time series is written as:

$$y_t = g_t + c_t$$

Typically y_t is the logarithm of a macroeconomic variable such as GDP. The growth components are found by minimizing:

$$\sum_{t=1}^T (y_t - g_t)^2 + \lambda \sum_{t=2}^{T-1} [(g_{t+1} - g_t) - (g_t - g_{t-1})]^2$$

Typically, for annual data $\lambda = 100$, for quarterly data $\lambda = 1,600$ and for monthly data $\lambda = 14,400$. The solution to this equation is given by:

$$\hat{g} = [I + \lambda K]^{-1} y$$

where \hat{g} is the estimated growth component, $y = [y_1, \dots, y_T]'$, I is a $T \times T$ identity matrix and K is a $T \times T$ symmetric matrix of the form:

$$K = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & \dots & \dots & 0 \\ -2 & 5 & -4 & 1 & 0 & \dots & \dots & \dots \\ 1 & -4 & 6 & -4 & 1 & 0 & \dots & \dots \\ 0 & 1 & -4 & 6 & \dots & \dots & \dots & \dots \\ \dots & 0 & 1 & \dots & \dots & -4 & 1 & 0 \\ \dots & \dots & \dots & \dots & -4 & 6 & -4 & 1 \\ \dots & \dots & \dots & 0 & 1 & -4 & 5 & -2 \\ 0 & \dots & \dots & \dots & 0 & 1 & -2 & 1 \end{bmatrix}$$

Section (14) Multivariate Analysis (not available in *Statistician (Lite)*)

Multivariate analysis in *Statistician* consists of two techniques, cluster analysis and principle component analysis.

(a) Cluster Analysis

The screenshot shows the 'Statistician (Cluster Analysis)' dialog box. It is divided into several sections:

- Data Variables (Objects):** A large empty rectangular box for selecting data variables.
- Measure Names (optional):** A dropdown menu labeled 'Select A Data Variable'.
- Standardization:** A checked checkbox for 'Standardize Variables' and an empty text box for 'Number Of Clusters'.
- Method:** Two radio buttons: 'Hierarchical' (selected) and 'K-Means'.
- Distance Metric:** A dropdown menu set to 'Euclidian' and a 'Power' text box set to '2'.
- Linkage Method:** A dropdown menu set to 'NearestNeighbour'.
- Output:** A group of checkboxes: 'Objects', 'Standardized Objects', 'Distance Matrix', 'Clusters (Vertically)', 'Clusters (Horizontally)', and 'Dendrogram'. The 'Format Output' checkbox is checked.
- Buttons:** 'Select All Variables', 'DeSelect All Variables', 'Manage Data', 'Output Results', and 'Close'.

Cluster analysis refers to a group of techniques that groups objects based upon the characteristics they possess. Cluster analysis groups objects so that similar objects with respect to some set of predetermined selection criteria (or attributes or measures), will be assigned to the same cluster. Fundamental to the use of any clustering technique is the computation of a measure of similarity or distance between the objects. Before any analysis takes place the measurements for each object are often standardized by subtracting their average and then dividing by the standard deviation. The distance between any two objects can be defined in a number of ways. Each object in *Statistician* is defined as a *Data Variable* and each observation in that *Data Variable* is a measure on that *Data Variable*. Let m and n be vectors of measures on two objects where m_i is the i^{th} measure on the first object and n_i is the i^{th} measure

on the second object. Let d_{mn} be the distance between the two objects and let k be the number of measures on each object.

The full set of distance measures available in *Statistician* are:

Distance Measure	Definition
Euclidian	$d_{mn} = \sqrt{\sum_{i=1}^k (m_i - n_i)^2}$
Squared Euclidian	$d_{mn} = \sum_{i=1}^k (m_i - n_i)^2$
Manhattan (City Block)	$d_{mn} = \sum_{i=1}^k m_i - n_i $
Chebychev (Maximum)	$d_{mn} = \sup_{1 \leq i \leq k} \{ m_i - n_i \}$
Correlation 1	$d_{mn} = 1 - r_{mn}$ <p>where r_{mn} is the sample correlation between m and n.</p>
Correlation 2	$d_{mn} = 1 - r_{mn} $ <p>where r_{mn} is the sample correlation between m and n.</p>
Correlation 3	$d_{mn} = \sqrt{1 - r_{mn}^2}$ <p>where r_{mn} is the sample correlation between m and n.</p>
Cosine	$d_{mn} = 1 - \frac{\sum_{i=1}^k m_i \times n_i}{\sum_{i=1}^k m_i^2 \times \sum_{i=1}^k n_i^2}$
Minkowski	$d_{mn} = \left(\sum_{i=1}^k m_i - n_i ^p \right)^{1/p}$ <p>where $p \geq 1$ (default is $p = 2$)</p>

Bray-Curtis	$d_{mn} = \sum_{i=1}^k \frac{ m_i - n_i }{ m_i + n_i }$
Canberra	$d_{mn} = \sum_{i=1}^k \frac{ m_i - n_i }{ m_i + n_i }$
Mahalanobis	$d_{mn} = \sqrt{(m - n)' S^{-1} (m - n)}$ where S^{-1} is the inverse of the covariance matrix between all objects. Note that distances cannot be standardized when using the Mahalanobis distance measure.

Statistician also offers a number of distance measures that are suitable for working with binary data (0 and 1). Define a , b , c and d as:

a = count of i such that $m_i = n_i = 1$

b = count of i such that $m_i = 1$ and $n_i = 0$

c = count of i such that $m_i = 0$ and $n_i = 1$

d = count of i such that $m_i = n_i = 0$

The binary distance measures are defined as follows:

Distance Measure	Definition
Jaccard	$d_{mn} = \frac{b + c}{a + b + c}$
Simple Matching	$d_{mn} = \frac{b + c}{k}$
Russel and Rao	$d_{mn} = \frac{b + c + d}{k}$
Sokal and Sneath 1	$d_{mn} = \frac{b + c}{2a + b + c + 2d}$
Sokal and Sneath 2	$d_{mn} = \frac{b + c}{a + 2(b + c)}$
Rogers and Tanimoto	$d_{mn} = \frac{2(b + c)}{a + 2(b + c) + d}$

Dice	$d_{mn} = \frac{b + c}{2a + b + c}$
Yule	$d_{mn} = \frac{2bc}{ad + bc}$

Statistician offers two forms of Cluster analysis, a hierarchical technique and the K-means technique.

(i) Hierarchical (or Join) Clustering

Agglomerative hierarchical clustering performs successive fusions of the data into clusters where each object initially starts out as its own cluster. Agglomerative hierarchical clustering techniques differ to the extent that different measures are employed to measure the distance between clusters (often referred to as the linkage method). In the following discussion the following notation is employed.

- C_X - cluster X
- D_{XY} - the distance between cluster X and Y
- n_X - the number of objects in cluster X
- \bar{x} - the centroid of cluster X

When two clusters (C_R and C_S) are merged to form a new cluster C_T , a combinatorial formula can be employed to calculate the distance between the new merged cluster and all other clusters C_K .

Linkage methods available in *Statistician* are as follows:

Linkage Method	Description
Nearest Neighbour (Single)	The distance between two clusters is determined by the distance of the two closest objects (nearest neighbours) in the different clusters. The combinatorial formula is $D_{TK} = \min(D_{RK}, D_{SK})$.
Furthest Neighbour (Complete)	The distance between two clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors"). The combinatorial formula is $D_{TK} = \max(D_{RK}, D_{SK})$.

Group Average	Also called weighted pair-group method using averages (WPGMA). The distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters weighted by the number of objects in each cluster. The combinatorial formula is $D_{TK} = \frac{n_R}{n_R+n_S}D_{RK} + \frac{n_S}{n_R+n_S}D_{SK}$.
McQuitty	Also called unweighted pair-group method using averages (UPGMA). The distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. This method is identical to the Group Average method, except that equal weighting is to the clusters to be merged. The combinatorial formula is $D_{TK} = \frac{1}{2}D_{RK} + \frac{1}{2}D_{SK}$.
Centroid	The distance between two clusters is the squared Euclidian distance between the centroids of the clusters. The combinatorial formula is $D_{TK} = \frac{n_R}{n_R+n_S}D_{RK} + \frac{n_S}{n_R+n_S}D_{SK} - \frac{n_R n_S}{(n_R+n_S)^2}D_{RS}$.
Median	Also called weighted pair-group method using centroid averages (UPGMC). This method is identical to the Centroid method, except that equal weighting is to the clusters to be merged. The combinatorial formula is $D_{TK} = \frac{1}{2}D_{RK} + \frac{1}{2}D_{SK} - \frac{1}{4}D_{RS}$.
Ward	The cluster to be merged is the one which will produce the least increase in the within cluster sums of squares error. The sums of squares error within a cluster is from the centroid of a cluster and the distances between objects in that cluster. The combinatorial formula is $D_{TK} = \frac{n_R+n_K}{n_R+n_S+n_K}D_{RK} + \frac{n_S+n_K}{n_R+n_S+n_K}D_{SK} - \frac{n_K}{n_R+n_S+n_K}D_{RS}$.

Cluster analysis often uses standardized data as unstandardized data yields inconsistent results when different scales are employed to form clusters.

(ii) K-means Clustering

K-means clustering is a simple non-hierarchical technique. For a specified number of clusters K , the clustering algorithm proceeds as follows:

- (1) Arbitrarily assign each object to one of K -clusters.
- (2) Calculate the centroid of each cluster.
- (3) For each object in each cluster calculate its Euclidian distance from itself to the centroid of all clusters. If the object is closer to another cluster's centroid then move it into that cluster.
- (4) Recalculate the centroid of each cluster.
- (5) Continue looping through steps (3) and (4) until no further changes are made.

(b) Principal Component Analysis

Statistician (Principle Component Analysis)

Data Variables

Standardize Variables

Sample Names (Optional)

Select a Data Variable

Output

Covariance Matrix

Correlation Matrix

Components

Correlogram (Variables and Compo)

Significant Digits 4

Select All Variables

DeSelect All Variables

Manage Data

Format Output

Output Results

Close

Principal Component Analysis (PCA) is a commonly used as a variable reduction technique that expresses a set of variables in terms of a smaller set of composite variables (or principle components). These principle components are a linear combination of the original variables. Principal Component Analysis is a methodology that estimates those components that contribute most to the variation in the data. These components are uncorrelated with each other. The first principle component extracts the maximum variance from the variables. The second principle component is constructed from the residual correlations and extracts the maximum variance from a linear function of the random variables that is orthogonal to the first principle component. Subsequent principle components are similarly formed.

Principle components are calculated as follows. Let X be a $(n \times m)$ matrix where m is the number of measurement types (or characteristics, factors, variables) and n is the number of samples recorded on each of the measurement types. Let ρ denote the correlation matrix of X and let Σ denote the covariance matrix of X . The principle components of X are the eigenvectors of ρ or Σ . The variance of the principle components λ_i ($1 \leq i \leq p$), are the eigenvalues of matrix ρ or Σ .

After calculating the eigenvalues and eigenvectors Statistician reports:

- The eigenvalue for each component,
- The percentage of variance for each component,
- The cumulative percentage of variance for each component,
- The correlation matrix,
- The covariance matrix,
- The eigenvectors (loadings) for each component,
- The principle components,
- A correlogram between the principle components and the original variables.

Note that when **Standardize** is checked the analysis is performed upon the correlation matrix. When **Standardize** is not checked the analysis is performed upon the covariance matrix.

Appendix (A) - Built-in Excel Statistical Functions

Function	Description
AVEDEV	Returns the average of the absolute deviations of data points from their mean
AVERAGE	Returns the average of its arguments
AVERAGEA	Returns the average of its arguments, including numbers, text, and logical values
AVERAGEIF	Returns the average (arithmetic mean) of all the cells in a range that meet a given criteria
AVERAGEIFS	Returns the average (arithmetic mean) of all cells that meet multiple criteria.
BETADIST	Returns the beta cumulative distribution function
BETAINV	Returns the inverse of the cumulative distribution function for a specified beta distribution
BINOMDIST	Returns the individual term binomial distribution probability
CHIDIST	Returns the one-tailed probability of the chi-squared distribution
CHIINV	Returns the inverse of the one-tailed probability of the chi-squared distribution
CHITEST	Returns the test for independence
CONFIDENCE	Returns the confidence interval for a population mean
CORREL	Returns the correlation coefficient between two data sets
COUNT	Counts how many numbers are in the list of arguments
COUNTA	Counts how many values are in the list of arguments
COUNTBLANK	Counts the number of blank cells within a range
COUNTIF	Counts the number of cells within a range that meet the given criteria
COUNTIFS	Counts the number of cells within a range that meet multiple criteria
COVAR	Returns covariance, the average of the products of paired deviations
CRITBINOM	Returns the smallest value for which the cumulative binomial distribution is less than or equal to a criterion value
DEVSQ	Returns the sum of squares of deviations
EXPONDIST	Returns the exponential distribution

FDIST	Returns the F probability distribution
FINV	Returns the inverse of the F probability distribution
FISHER	Returns the Fisher transformation
FISHERINV	Returns the inverse of the Fisher transformation
FORECAST	Returns a value along a linear trend
FREQUENCY	Returns a frequency distribution as a vertical array
FTEST	Returns the result of an F-test
GAMMADIST	Returns the gamma distribution
GAMMAINV	Returns the inverse of the gamma cumulative distribution
GAMMALN	Returns the natural logarithm of the gamma function, $\Gamma(x)$
GEOMEAN	Returns the geometric mean
GROWTH	Returns values along an exponential trend
HARMEAN	Returns the harmonic mean
HYPGEOMDIST	Returns the hypergeometric distribution
INTERCEPT	Returns the intercept of the linear regression line
KURT	Returns the kurtosis of a data set
LARGE	Returns the k-th largest value in a data set
LINEST	Returns the parameters of a linear trend
LOGEST	Returns the parameters of an exponential trend
LOGINV	Returns the inverse of the lognormal distribution
LOGNORMDIST	Returns the cumulative lognormal distribution
MAX	Returns the maximum value in a list of arguments
MAXA	Returns the maximum value in a list of arguments, including numbers, text, and logical values
MEDIAN	Returns the median of the given numbers
MIN	Returns the minimum value in a list of arguments
MINA	Returns the smallest value in a list of arguments, including numbers, text, and logical values
MODE	Returns the most common value in a data set
NEGBINOMDIST	Returns the negative binomial distribution
NORMDIST	Returns the normal cumulative distribution
NORMINV	Returns the inverse of the normal cumulative distribution

NORMSDIST	Returns the standard normal cumulative distribution
NORMSINV	Returns the inverse of the standard normal cumulative distribution
PEARSON	Returns the Pearson product moment correlation coefficient
PERCENTILE	Returns the k-th percentile of values in a range
PERCENTRANK	Returns the percentage rank of a value in a data set
PERMUT	Returns the number of permutations for a given number of objects
POISSON	Returns the Poisson distribution
PROB	Returns the probability that values in a range are between two limits
QUARTILE	Returns the quartile of a data set
RANK	Returns the rank of a number in a list of numbers
RSQ	Returns the square of the Pearson product moment correlation coefficient
SKEW	Returns the skewness of a distribution
SLOPE	Returns the slope of the linear regression line
SMALL	Returns the k-th smallest value in a data set
STANDARDIZE	Returns a normalized value
STDEV	Estimates standard deviation based on a sample
STDEVA	Estimates standard deviation based on a sample, including numbers, text, and logical values
STDEVP	Calculates standard deviation based on the entire population
STDEVPA	Calculates standard deviation based on the entire population, including numbers, text, and logical values
STEYX	Returns the standard error of the predicted y-value for each x in the regression
TDIST	Returns the Student's t-distribution
TINV	Returns the inverse of the Student's t-distribution
TREND	Returns values along a linear trend
TRIMMEAN	Returns the mean of the interior of a data set
TTEST	Returns the probability associated with a Student's t-test
VAR	Estimates variance based on a sample
VARA	Estimates variance based on a sample, including numbers, text, and logical values

VARP	Calculates variance based on the entire population
VARPA	Calculates variance based on the entire population, including numbers, text, and logical values
WEIBULL	Returns the Weibull distribution
ZTEST	Returns the one-tailed probability-value of a z-test

Appendix (B) - Other Excel Functions

Math and trigonometry functions

Function	Description
ABS	Returns the absolute value of a number
ACOS	Returns the arccosine of a number
ACOSH	Returns the inverse hyperbolic cosine of a number
ASIN	Returns the arcsine of a number
ASINH	Returns the inverse hyperbolic sine of a number
ATAN	Returns the arctangent of a number
ATAN2	Returns the arctangent from x- and y-coordinates
ATANH	Returns the inverse hyperbolic tangent of a number
CEILING	Rounds a number to the nearest integer or to the nearest multiple of significance
COMBIN	Returns the number of combinations for a given number of objects
COS	Returns the cosine of a number
COSH	Returns the hyperbolic cosine of a number
DEGREES	Converts radians to degrees
EVEN	Rounds a number up to the nearest even integer
EXP	Returns e raised to the power of a given number
FACT	Returns the factorial of a number
FACTDOUBLE	Returns the double factorial of a number
FLOOR	Rounds a number down, toward zero
GCD	Returns the greatest common divisor
INT	Rounds a number down to the nearest integer
LCM	Returns the least common multiple
LN	Returns the natural logarithm of a number
LOG	Returns the logarithm of a number to a specified base
LOG10	Returns the base-10 logarithm of a number
MDETERM	Returns the matrix determinant of an array
MINVERSE	Returns the matrix inverse of an array
MMULT	Returns the matrix product of two arrays

MOD	Returns the remainder from division
MROUND	Returns a number rounded to the desired multiple
MULTINOMIAL	Returns the multinomial of a set of numbers
ODD	Rounds a number up to the nearest odd integer
PI	Returns the value of pi
POWER	Returns the result of a number raised to a power
PRODUCT	Multiplies its arguments
QUOTIENT	Returns the integer portion of a division
RADIANS	Converts degrees to radians
RAND	Returns a random number between 0 and 1
RANDBETWEEN	Returns a random number between the numbers you specify
ROMAN	Converts an arabic numeral to roman, as text
ROUND	Rounds a number to a specified number of digits
ROUNDDOWN	Rounds a number down, toward zero
ROUNDUP	Rounds a number up, away from zero
SERIESSUM	Returns the sum of a power series based on the formula
SIGN	Returns the sign of a number
SIN	Returns the sine of the given angle
SINH	Returns the hyperbolic sine of a number
SQRT	Returns a positive square root
SQRTPI	Returns the square root of (number * pi)
SUBTOTAL	Returns a subtotal in a list or database
SUM	Adds its arguments
SUMIF	Adds the cells specified by a given criteria
SUMIFS	Adds the cells in a range that meet multiple criteria
SUMPRODUCT	Returns the sum of the products of corresponding array components
SUMSQ	Returns the sum of the squares of the arguments
SUMX2MY2	Returns the sum of the difference of squares of corresponding values in two arrays
SUMX2PY2	Returns the sum of the sum of squares of corresponding values in two arrays

SUMXMY2	Returns the sum of squares of differences of corresponding values in two arrays
TAN	Returns the tangent of a number
TANH	Returns the hyperbolic tangent of a number
TRUNC	Truncates a number to an integer

Text functions

Function	Description
ASC	Changes full-width (double-byte) English letters or katakana within a character string to half-width (single-byte) characters
BAHTTEXT	Converts a number to text, using the ฿ (baht) currency format
CHAR	Returns the character specified by the code number
CLEAN	Removes all nonprintable characters from text
CODE	Returns a numeric code for the first character in a text string
CONCATENATE	Joins several text items into one text item
DOLLAR	Converts a number to text, using the \$ (dollar) currency format
EXACT	Checks to see if two text values are identical
FIND, FINDB	Finds one text value within another (case-sensitive)
FIXED	Formats a number as text with a fixed number of decimals
JIS	Changes half-width (single-byte) English letters or katakana within a character string to full-width (double-byte) characters
LEFT, LEFTB	Returns the leftmost characters from a text value
LEN, LENB	Returns the number of characters in a text string
LOWER	Converts text to lowercase
MID, MIDB	Returns a specific number of characters from a text string starting at the position you specify
PHONETIC	Extracts the phonetic (furigana) characters from a text string
PROPER	Capitalizes the first letter in each word of a text value
REPLACE, REPLACEB	Replaces characters within text
REPT	Repeats text a given number of times
RIGHT, RIGHTB	Returns the rightmost characters from a text value

SEARCH, SEARCHB	Finds one text value within another (not case-sensitive)
SUBSTITUTE	Substitutes new text for old text in a text string
T	Converts its arguments to text
TEXT	Formats a number and converts it to text
TRIM	Removes spaces from text
UPPER	Converts text to uppercase
VALUE	Converts a text argument to a number

Financial functions

Function	Description
ACCRINT	Returns the accrued interest for a security that pays periodic interest
ACCRINTM	Returns the accrued interest for a security that pays interest at maturity
AMORDEGRC	Returns the depreciation for each accounting period by using a depreciation coefficient
AMORLINC	Returns the depreciation for each accounting period
COUPDAYBS	Returns the number of days from the beginning of the coupon period to the settlement date
COUPDAYS	Returns the number of days in the coupon period that contains the settlement date
COUPDAYSNC	Returns the number of days from the settlement date to the next coupon date
COUPNCD	Returns the next coupon date after the settlement date
COUPNUM	Returns the number of coupons payable between the settlement date and maturity date
COUPPCD	Returns the previous coupon date before the settlement date
CUMIPMT	Returns the cumulative interest paid between two periods
CUMPRINC	Returns the cumulative principal paid on a loan between two periods
DB	Returns the depreciation of an asset for a specified period by using the fixed-declining balance method
DDB	Returns the depreciation of an asset for a specified period by using the double-declining balance method or some other method that you specify
DISC	Returns the discount rate for a security

DOLLARDE	Converts a dollar price, expressed as a fraction, into a dollar price, expressed as a decimal number
DOLLARFR	Converts a dollar price, expressed as a decimal number, into a dollar price, expressed as a fraction
DURATION	Returns the annual duration of a security with periodic interest payments
EFFECT	Returns the effective annual interest rate
FV	Returns the future value of an investment
FVSCHEDULE	Returns the future value of an initial principal after applying a series of compound interest rates
INTRATE	Returns the interest rate for a fully invested security
IPMT	Returns the interest payment for an investment for a given period
IRR	Returns the internal rate of return for a series of cash flows
ISPMT	Calculates the interest paid during a specific period of an investment
MDURATION	Returns the Macauley modified duration for a security with an assumed par value of \$100
MIRR	Returns the internal rate of return where positive and negative cash flows are financed at different rates
NOMINAL	Returns the annual nominal interest rate
NPER	Returns the number of periods for an investment
NPV	Returns the net present value of an investment based on a series of periodic cash flows and a discount rate
ODDFPRICE	Returns the price per \$100 face value of a security with an odd first period
ODDFYIELD	Returns the yield of a security with an odd first period
ODDLPRICE	Returns the price per \$100 face value of a security with an odd last period
ODDLYIELD	Returns the yield of a security with an odd last period
PMT	Returns the periodic payment for an annuity
PPMT	Returns the payment on the principal for an investment for a given period
PRICE	Returns the price per \$100 face value of a security that pays periodic interest
PRICEDISC	Returns the price per \$100 face value of a discounted security
PRICEMAT	Returns the price per \$100 face value of a security that pays interest at maturity

PV	Returns the present value of an investment
RATE	Returns the interest rate per period of an annuity
RECEIVED	Returns the amount received at maturity for a fully invested security
SLN	Returns the straight-line depreciation of an asset for one period
SYD	Returns the sum-of-years' digits depreciation of an asset for a specified period
TBILLEQ	Returns the bond-equivalent yield for a Treasury bill
TBILLPRICE	Returns the price per \$100 face value for a Treasury bill
TBILLYIELD	Returns the yield for a Treasury bill
VDB	Returns the depreciation of an asset for a specified or partial period by using a declining balance method
XIRR	Returns the internal rate of return for a schedule of cash flows that is not necessarily periodic
XNPV	Returns the net present value for a schedule of cash flows that is not necessarily periodic
YIELD	Returns the yield on a security that pays periodic interest
YIELDDISC	Returns the annual yield for a discounted security; for example, a Treasury bill
YIELDMAT	Returns the annual yield of a security that pays interest at maturity

Date and time functions

Function	Description
DATE	Returns the serial number of a particular date
DATEVALUE	Converts a date in the form of text to a serial number
DAY	Converts a serial number to a day of the month
DAYS360	Calculates the number of days between two dates based on a 360-day year
EDATE	Returns the serial number of the date that is the indicated number of months before or after the start date
EOMONTH	Returns the serial number of the last day of the month before or after a specified number of months
HOUR	Converts a serial number to an hour
MINUTE	Converts a serial number to a minute
MONTH	Converts a serial number to a month

NETWORKDAYS	Returns the number of whole workdays between two dates
NOW	Returns the serial number of the current date and time
SECOND	Converts a serial number to a second
TIME	Returns the serial number of a particular time
TIMEVALUE	Converts a time in the form of text to a serial number
TODAY	Returns the serial number of today's date
WEEKDAY	Converts a serial number to a day of the week
WEEKNUM	Converts a serial number to a number representing where the week falls numerically with a year
WORKDAY	Returns the serial number of the date before or after a specified number of workdays
YEAR	Converts a serial number to a year
YEARFRAC	Returns the year fraction representing the number of whole days between start_date and end_date

Information functions

Function	Description
CELL	Returns information about the formatting, location, or contents of a cell
ERROR.TYPE	Returns a number corresponding to an error type
INFO	Returns information about the current operating environment
ISBLANK	Returns TRUE if the value is blank
ISERR	Returns TRUE if the value is any error value except #N/A
ISERROR	Returns TRUE if the value is any error value
ISEVEN	Returns TRUE if the number is even
ISLOGICAL	Returns TRUE if the value is a logical value
ISNA	Returns TRUE if the value is the #N/A error value
ISNONTEXT	Returns TRUE if the value is not text
ISNUMBER	Returns TRUE if the value is a number
ISODD	Returns TRUE if the number is odd
ISREF	Returns TRUE if the value is a reference
ISTEXT	Returns TRUE if the value is text

N	Returns a value converted to a number
NA	Returns the error value #N/A
TYPE	Returns a number indicating the data type of a value

Logical functions

Function	Description
AND	Returns TRUE if all of its arguments are TRUE
FALSE	Returns the logical value FALSE
IF	Specifies a logical test to perform
IFERROR	Returns a value you specify if a formula evaluates to an error; otherwise, returns the result of the formula
NOT	Reverses the logic of its argument
OR	Returns TRUE if any argument is TRUE
TRUE	Returns the logical value TRUE

Lookup and reference functions

Function	Description
ADDRESS	Returns a reference as text to a single cell in a worksheet
AREAS	Returns the number of areas in a reference
CHOOSE	Chooses a value from a list of values
COLUMN	Returns the column number of a reference
COLUMNS	Returns the number of columns in a reference
HLOOKUP	Looks in the top row of an array and returns the value of the indicated cell
HYPERLINK	Creates a shortcut or jump that opens a document stored on a network server, an intranet, or the Internet
INDEX	Uses an index to choose a value from a reference or array
INDIRECT	Returns a reference indicated by a text value
LOOKUP	Looks up values in a vector or array
MATCH	Looks up values in a reference or array
OFFSET	Returns a reference offset from a given reference
ROW	Returns the row number of a reference

ROWS	Returns the number of rows in a reference
RTD	Retrieves real-time data from a program that supports COM automation (Automation: A way to work with an application's objects from another application or development tool. Formerly called OLE Automation, Automation is an industry standard and a feature of the Component Object Model (COM).)
TRANSPOSE	Returns the transpose of an array
VLOOKUP	Looks in the first column of an array and moves across the row to return the value of a cell

Database functions

Function	Description
DAVERAGE	Returns the average of selected database entries
DCOUNT	Counts the cells that contain numbers in a database
DCOUNTA	Counts nonblank cells in a database
DGET	Extracts from a database a single record that matches the specified criteria
DMAX	Returns the maximum value from selected database entries
DMIN	Returns the minimum value from selected database entries
DPRODUCT	Multiplies the values in a particular field of records that match the criteria in a database
DSTDEV	Estimates the standard deviation based on a sample of selected database entries
DSTDEVP	Calculates the standard deviation based on the entire population of selected database entries
DSUM	Adds the numbers in the field column of records in the database that match the criteria
DVAR	Estimates variance based on a sample from selected database entries
DVARP	Calculates variance based on the entire population of selected database entries

Engineering functions

Function	Description
BESSELI	Returns the modified Bessel function $I_n(x)$

BESSELJ	Returns the Bessel function $J_n(x)$
BESSELK	Returns the modified Bessel function $K_n(x)$
BESSELY	Returns the Bessel function $Y_n(x)$
BIN2DEC	Converts a binary number to decimal
BIN2HEX	Converts a binary number to hexadecimal
BIN2OCT	Converts a binary number to octal
COMPLEX	Converts real and imaginary coefficients into a complex number
CONVERT	Converts a number from one measurement system to another
DEC2BIN	Converts a decimal number to binary
DEC2HEX	Converts a decimal number to hexadecimal
DEC2OCT	Converts a decimal number to octal
DELTA	Tests whether two values are equal
ERF	Returns the error function
ERFC	Returns the complementary error function
GESTEP	Tests whether a number is greater than a threshold value
HEX2BIN	Converts a hexadecimal number to binary
HEX2DEC	Converts a hexadecimal number to decimal
HEX2OCT	Converts a hexadecimal number to octal
IMABS	Returns the absolute value (modulus) of a complex number
IMAGINARY	Returns the imaginary coefficient of a complex number
IMARGUMENT	Returns the argument theta, an angle expressed in radians
IMCONJUGATE	Returns the complex conjugate of a complex number
IMCOS	Returns the cosine of a complex number
IMDIV	Returns the quotient of two complex numbers
IMEXP	Returns the exponential of a complex number
IMLN	Returns the natural logarithm of a complex number
IMLOG10	Returns the base-10 logarithm of a complex number
IMLOG2	Returns the base-2 logarithm of a complex number
IMPOWER	Returns a complex number raised to an integer power
IMPRODUCT	Returns the product of complex numbers
IMREAL	Returns the real coefficient of a complex number

IMSIN	Returns the sine of a complex number
IMSQRT	Returns the square root of a complex number
IMSUB	Returns the difference between two complex numbers
IMSUM	Returns the sum of complex numbers
OCT2BIN	Converts an octal number to binary
OCT2DEC	Converts an octal number to decimal
OCT2HEX	Converts an octal number to hexadecimal

Add-in and Automation functions

Function	Description
CALL	Calls a procedure in a dynamic link library or code resource
EUROCONVERT	Converts a number to euros, converts a number from euros to a euro member currency, or converts a number from one euro member currency to another by using the euro as an intermediary (triangulation)
GETPIVOTDATA	Returns data stored in a PivotTable report
REGISTER.ID	Returns the register ID of the specified dynamic link library (DLL) or code resource that has been previously registered
SQL.REQUEST	Connects with an external data source and runs a query from a worksheet, then returns the result as an array without the need for macro programming

Cube functions

Function	Description
CUBEKPIMEMBER	Returns a key performance indicator (KPI) name, property, and measure, and displays the name and property in the cell. A KPI is a quantifiable measurement, such as monthly gross profit or quarterly employee turnover, used to monitor an organization's performance.
CUBEMEMBER	Returns a member or tuple in a cube hierarchy. Use to validate that the member or tuple exists in the cube.
CUBEMEMBERPROPERTY	Returns the value of a member property in the cube. Use to validate that a member name exists within the cube

and to return the specified property for this member.

CUBERANKEDMEMBER

Returns the nth, or ranked, member in a set. Use to return one or more elements in a set, such as the top sales performer or top 10 students.

CUBESET

Defines a calculated set of members or tuples by sending a set expression to the cube on the server, which creates the set, and then returns that set to Microsoft Office Excel.

CUBESETCOUNT

Returns the number of items in a set.

CUBEVALUE

Returns an aggregated value from a cube.