Guide to Using

StatTools

Statistics Add-In for Microsoft® Excel

Version 5.7
September, 2010

Palisade Corporation
798 Cascadilla St
Ithaca, NY 14850
(607) 277-8000
http://www.palisade.com

Welcome to StatTools for Excel

Welcome

StatTools gives Microsoft Excel - the industry-standard data analysis and modeling tool - a new, powerful statistics toolset! StatTools is a Microsoft Excel statistics add-in, allowing you to analyze data in Excel worksheets and work in the familiar Microsoft Office environment. By combining a powerful data manager, along with analyses that rival the best statistics packages available, StatTools brings you the best of two worlds: Microsoft Office ease-of-use and reporting, and robust statistical power.

Work Where You're Comfortable

If you know Excel, you'll know StatTools! StatTools works just as Excel does, with toolbars, menus and custom worksheet functions, all inside of Excel. Unlike stand-alone statistics software, there's no steep learning curve and upfront training costs with StatTools, because you work just as you are used to working in Excel. Your data and variables are in Excel spreadsheets. You can utilize standard Excel formulas for calculations and transformations, along with Excel sorting and pivot tables. Reports and charts from your statistical analyses are in standard Excel format and can utilize all of Excel's built-in formatting capabilities.

Robust Statistics Inside Excel

StatTools replaces Excel's built-in statistics with its own robust and fast calculations. The accuracy of Excel's built-in statistics calculations has often been questioned, and StatTools uses none of them! Even Excel's worksheet statistics functions – such as STDEV() – are replaced by new, robust StatTools versions – such as StatSTDEV(). StatTools statistics calculations meet the highest tests for accuracy, with performance optimized through the use of C++ .DLLs, not macro calculations.

StatTools Analyses

StatTools covers the range of the most commonly used statistical procedures, and offers unprecedented capabilities for adding new, custom analyses. A total of 36 wide-ranging statistical procedures plus 8 built-in data utilities cover the most widely used statistical analyses. Statistical functions provided include descriptive statistics, normality tests, group comparisons, correlation, regression analysis, quality control, forecasts and more. Add to this a library of custom procedures (written by your staff or other experts in the field) and you've got a comprehensive and customizable statistics toolset, right inside of Excel!

StatTools features live, "hot-linked" statistics calculations! If you change a value in Excel, you expect your worksheet to recalculate and give you a new answer. Well, the same thing happens in StatTools! Change a value in your dataset and your statistics report automatically updates. StatTools uses a powerful set of custom worksheet functions to insure that the statistics displayed in your reports are always up-to-date with your current data.

StatTools Data Management

StatTools provides a comprehensive dataset and variable manager right in Excel, just as you would expect from a stand-alone statistics package. You can define any number of datasets, each with the variables you want to analyze, directly from your data in Excel. StatTools intelligently assesses your blocks of data, suggesting variable names and data locations for you. Your datasets and variables can reside in different workbooks and worksheets, allowing you to organize your data as you see fit. Then, you run statistical analyses that refer to your variables, instead of re-selecting your data over and over again in Excel. And StatTools variables aren't limited in size to a single column of data in an Excel worksheet – you can use the same column across up to 255 worksheets for a single variable! (65,535 X 255, or over 16 million cases in StatTools Industrial; 10,000 cases in StatTools Professional Edition)

StatTools Reporting

Excel is great for reports and graphs, and StatTools makes the most of this. StatTools uses Excel-format graphs, which can be easily customized for new colors, fonts and added text. Report titles, number formats and text can be changed just as is any standard Excel worksheet. Drag and drop tables and charts from StatTools reports straight into your own documents in other applications. Charts and tables stay linked to your data in Excel, so whenever your analysis reports change, your document is automatically updated.

Data Access and Sharing

Excel has great data import features, so bringing your existing data into StatTools is easy! Use standard Excel capabilities to read in data from Microsoft SQL Server, Oracle, Microsoft Access, or any other ODBC compliant database. Load data from text files or other applications – if you can read it into Excel, you can use it with StatTools!

StatTools saves all its results and data in Excel workbooks. Just like any other Excel file, you can send your StatTools results and data to colleagues anywhere. Sharing couldn't be easier!

StatTools Industrial

StatTools Industrial includes a complete, object-oriented, programming interface, where custom statistical procedures may be added using Excel's built-in VBA programming language. These custom procedures can utilize StatTools' built-in data management, charting and reporting tools, all accessible via StatTools custom controls, functions and methods. Your custom procedures can even be displayed on the StatTools menu for easy access!

So what if you're not going to write your own statistical procedures? StatTools Industrial still provides a great benefit to you, as you can use custom procedures that are written by others, right off the standard StatTools menu! Experts in the field are constantly adding to the library of new, custom procedures that are built with StatTools. Simply copy a workbook with a new procedure into your StatTools directory on your PC and it instantly shows up on the StatTools menu. Run it and you'll see all the standard StatTools data management tools, combined with the new statistical analysis you need!

Table of Contents

Chapter 1: Getting Started

Introduction

This introduction describes the contents of your StatTools package and shows you how to install StatTools and attach it to your copy of Microsoft Excel 2000 for Windows or higher.

Checking Your Package

Your StatTools package should contain:

Learning Statistics with StatTools, a statistics book that uses StatTools written by Dr. S. Christian Albright of Indiana University

The StatTools or DecisionTools Suite CD-ROM including:

- *StatTools Program*
- *StatTools Tutorial*
- *The StatTools Users Guide (this book) in .PDF format*

The StatTools Licensing Agreement

If your package is not complete, please call your StatTools dealer or supplier or contact Palisade Corporation directly at (607) 277-8000.

What the Package Includes

StatTools may be purchased on its own and also ships with the DecisionTools Suite Professional and Industrial versions. The StatTools CD-ROM contains the StatTools Excel add-in, several StatTools examples, and a fully-indexed StatTools on-line help system. The DecisionTools Suite Professional and Industrial versions contain all of the above plus additional applications.

About This Version

This version of StatTools can be installed as a 32-bit program for Microsoft Excel 2000 or higher.

Working with your Operating Environment

This User's Guide assumes that you have a general knowledge of the Windows operating system and Excel. In particular:

- *You are familiar with your computer and using the mouse.*
- *You are familiar with terms such as icons, click, double-click, menu, window, command and object.*
- *You understand basic concepts such as directory structures and file naming.*

If You Need Help

Technical support is provided free of charge for all registered users of StatTools with a current maintenance plan, or is available on a per incident charge. To ensure that you are a registered user of StatTools, **please register online at http://www.palisade.com/support/register.asp**.

If you contact us by telephone, please have your serial number and User's Guide ready. We can offer better technical support if you are in front of your computer and ready to work.

Before Calling

Before contacting technical support, please review the following checklist:

- *Have you referred to the on-line help?*
- *Have you checked this User's Guide and reviewed the on-line multimedia tutorial?*
- *Have you read the README.WRI file? It contains current information on* StatTools *that may not be included in the manual.*
- *Can you duplicate the problem consistently? Can you duplicate the problem on a different computer or with a different model?*
- *Have you looked at our site on the World Wide Web? It can be found at* ***http://www.palisade.com****. Our Web site also contains the latest FAQ (a searchable database of tech support questions and answers) and* StatTools *patches in our Technical Support section. We recommend visiting our Web site regularly for all the latest information on* StatTools *and other Palisade software.*

Contacting Palisade

Palisade Corporation welcomes your questions, comments or suggestions regarding StatTools. Contact our technical support staff using any of the following methods:

- *Email us at support@palisade.com.*
- *Telephone us at (607) 277-8000 any weekday from 9:00 AM to 5:00 PM, EST. Follow the prompt to reach technical support.*
- *Fax us at (607) 277-8001.*
- *Mail us a letter at:*

  Technical Support
  Palisade Corporation
  798 Cascadilla St.
  Ithaca, NY 14850 USA

If you want to contact Palisade Europe:

- *Email us at support@palisade-europe.com.*
- *Telephone us at +44 1895 425050 (UK).*
- *Fax us at +44 1895 425051 (UK).*
- *Mail us a letter at:*

  Palisade Europe
  31 The Green
  West Drayton
  Middlesex
  UB7 7PN
  United Kingdom

If you want to contact Palisade Asia-Pacific:

- *Email us at support@palisade.com.au*
- *Telephone us at + 61 2 9252 5922 (AU).*
- *Fax us at + 61 2 9252 2820 (AU).*
- *Mail us a letter to:*

  Palisade Asia-Pacific Pty Limited
  Suite 404, Level 4
  20 Loftus Street
  Sydney NSW 2000
  Australia

Regardless of how you contact us, please include the product name, version and serial number. The exact version can be found by selecting the Help About command on the StatTools menu in Excel.

Student Versions

Telephone support is not available with the student version of StatTools. If you need help, we recommend the following alternatives:

- *Consult with your professor or teaching assistant.*
- *Log on to http://www.palisade.com for answers to frequently asked questions.*
- *Contact our technical support department via e-mail or fax.*

StatTools System Requirements

System requirements for StatTools for Microsoft Excel for Windows include:

- *Pentium PC or faster with a hard disk.*
- *Microsoft Windows 2000 SP4, Windows XP or higher.*
- *Microsoft Excel 2000 or higher.*

Installation Instructions

General Installation Instructions

The Setup program copies the StatTools system files into a directory you specify on your hard disk. To run the Setup program in Windows 2000 or higher:

1) *Insert the StatTools or DecisionTools Suite CD-ROM in your CD-ROM drive*
2) *Click the Start button, click Settings and then click Control Panel*
3) *Double-click the Add/Remove Programs icon*
4) *On the Install/Uninstall tab, click the Install button*
5) *Follow the Setup instructions on the screen*

If you encounter problems while installing StatTools, verify that there is adequate space on the drive to which you're trying to install. After you've freed up adequate space, try rerunning the installation.

Removing StatTools from Your Computer

If you wish to remove StatTools from your computer, use the Control Panel's Add/Remove Programs utility and select the entry for StatTools.

Setting Up the StatTools Icons or Shortcuts

Creating the Shortcut in the Windows Taskbar

In Windows, setup automatically creates a StatTools command in the Programs\Palisade DecisionTools menu of the Taskbar. However, if problems are encountered during Setup, or if you wish to do this manually another time, follow these directions. Note that the directions given below are for Windows XP Professional. Instructions for other operating systems may vary.

1) *Click the Start button, and then point to Settings.*
2) *Click Taskbar and Start Menu, and then click the Start Menu tab.*
3) *Click Customize, click Add, and then click Browse.*
4) *Locate the file StatTools.EXE, click it and then click OK.*
5) *Click Next, and then double-click the menu on which you want the program to appear.*
6) *Type the name "StatTools", and then click Finish.*
7) *Click OK on all opened dialogs.*

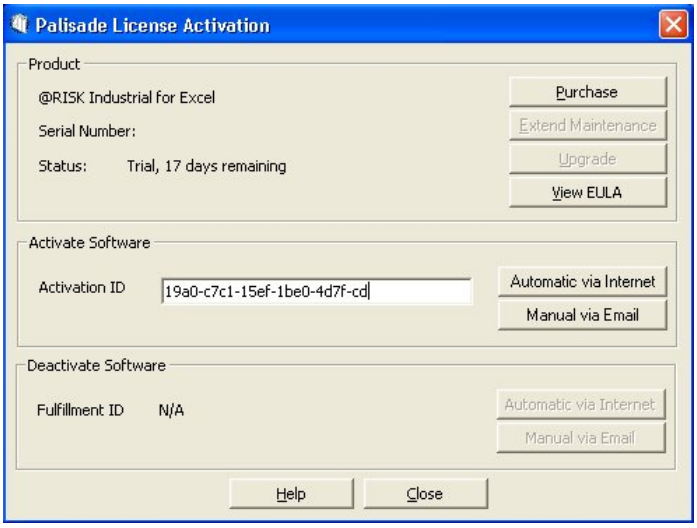
The DecisionTools Suite

StatTools is part of the DecisionTools Suite, a set of products for risk and decision analysis available from Palisade Corporation. The default installation procedure of StatTools puts StatTools in a subdirectory of a main "Program Files\Palisade" directory. This is quite similar to how Excel is often installed into a subdirectory of a "Microsoft Office" directory.

One subdirectory of the Program Files\Palisade directory will be the StatTools directory (by default called StatTools5). This directory contains the Evolver add-in program file (STATTOOLS.XLA) plus example models and other files necessary for StatTools to run. Another subdirectory of Program Files\Palisade is the SYSTEM directory which contains files needed by every program in the DecisionTools Suite, including common help files and program libraries.

Software Activation

Activation is a one time license verification process that is required in order for your StatTools software to run as a fully licensed product. An **activation code** is on your printed/emailed invoice and may resemble a dash separated sequence like "19a0-c7c1-15ef-1be0-4d7f-cd". If you enter your Activation code during installation, then your software is activated the first time the software is run and no further user action is required. If you wish to activate your software after installation, select the StatTools Help menu License Activation command and enter your activation code in the displayed **Palisade License Activation** dialog box.



Frequently Asked Questions

1) What if my software is not activated?

If you do not enter an activation code during installation or you are installing a trial version, your software will run as a trial version with time and/or number of uses limitations and must be activated with an activation code in order to run as a fully licensed product.

2) How long can I use the product before I have to activate it?

Software that is not activated may be run for fifteen days. All of the product's features are present, but the License Activation dialog will appear each time the program is launched to remind you to activate and to indicate the time remaining. If the 15 day trial period expires, the software will require activation in order to run.

3) How do I check my activation status?

The License Activation dialog box is viewed through the StatTools Help menu License Activation command. Activated software shows a status of **Activated** and trial version software shows a status of **Not Activated**. If the software is not activated, the remaining time that the software is allowed to run is displayed.

4) How do I activate my software?

If you do not have an activation code you may obtain one by clicking the Puchase button in the License Activation dialog. An online purchase will be immediately given an activation code and an optional link to download the installer should reinstallation become necessary. To purchase by phone call the local Palisade office given in the **Contacting Palisade** section of this chapter.

Activation may be done over the Internet or via email:

- **Activation if you have Internet Access**

In the Palisade License Activation dialog box, type or paste the activation code and press "Automatic via Internet". A success message should appear after a few seconds and the License Activation dialog box will reflect the software's activated status.

- **Activation if you do not have Internet Access**

Automated activation by email requires a few steps:

1. **Click "Manual via Email"** to display the request.xml file which you may save to disk or copy to the Windows clipboard. (It is recommended you note the location on your computer of the request.xml file.)
2. **Copy or attach the XML file** to an email and send it to *activation@palisade.com*. You should receive an automatic response to the return address in your email shortly.
3. **Save the response.xml attachment** in the response email to your hard drive.
4. **Click on the Process button** that is now in the Palisade License Activation dialog box and navigate to the response.xml file. Select the file and click OK.

A success message should appear and the License Activation dialog will reflect the software's activated status.

5) How do I transfer my software license to another machine?

Transfer of a license, or **rehosting**, may be peformed through the Palisade License Activation dialog box as a two step procedure: *deactivation* on the first machine and *activation* on the second machine. A typical use of rehosting is to transfer your copy of StatTools from your office PC to your laptop. To rehost a license from *Machine1* to *Machine2*, make sure both machines have the software installed and are connected to the Internet during the deactivation/activation rehosting.

1. On *Machine1*, click deactivate **Automatic via Internet** in the License Activation dialog. Wait for the success message.
2. On *Machine2*, click activate **Automatic via Internet.** Wait for the success message.

If the machines do not have Internet access then you may follow the similar instructions above for rehosting by the automated email process.

6) I have Internet Access but I am still unable to Activate/Deactivate automatically.

Your firewall must be set to allow TCP access to the licensing server. For single user (non network installations) ths is http://service.palisade.com:8888 (TCP port 8888 on http://service.palisade.com).

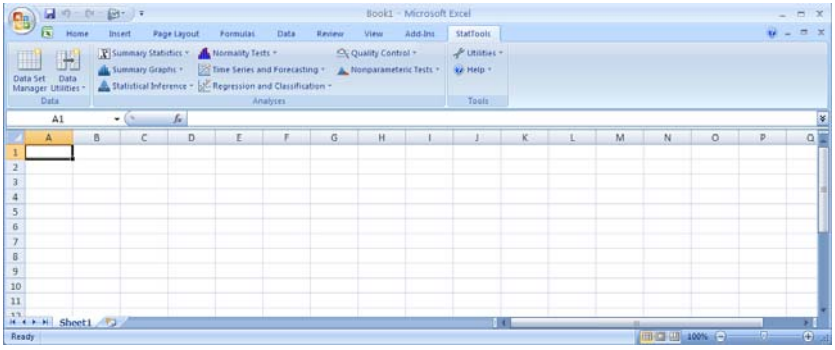
Chapter 2: An Overview to StatTools

Overview

StatTools provides you with powerful statistical capabilities in an environment that you are familiar with - Microsoft Excel. StatTools procedures - such as creating scatterplots, testing a variable for normality, and running a regression analysis - can be run on your data in Excel and the reports and charts from your analyses are created in Excel.

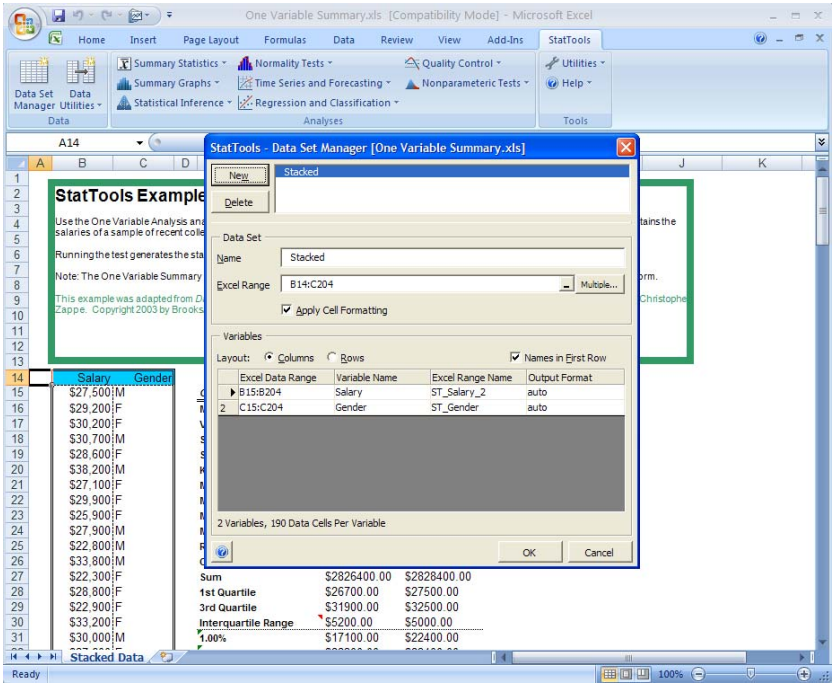
StatTools Menu and Toolbar

Once you have installed StatTools in Excel 2003 and earlier, its menu and commands will be included as part of the Excel menu bar. There will also be a StatTools toolbar displayed. In Excel 2007 and later, a StatTools ribbon will be displayed.



Data Sets and the Data Manager

StatTools is analogous to most stand-alone statistical software packages in that it is structured around variables. For most analyses it is required that you work with a data set, or a set of statistical variables, often located in contiguous columns with variable names in the first row of the data set. The StatTools **Data Set Manager** allows you to define your data sets and variables. You can then use these predefined variables in your statistical analyses, without re-selecting the data you wish to analyze over and over.



Each variable in a data set has a name and a range of Excel cells associated with it. A typical variable layout is **One Variable Per Column**, but variables may also be laid out by row. A data set can include multiple blocks of cells, allowing you to put data on different sheets in the same workbook.

When you are defining a data set, StatTools attempts to identify the variables in a block of cells surrounding the current selection in Excel. This makes it quick and easy to set up a data set with variable names in the top row and variables laid out by column.

The lengths of the columns in your data set do not necessarily have to be equal. For example, you could have two variables, *Weight_Men* and *Weight_Women*, with different numbers of observations. However, for many analyses, StatTools will treat the blank cells in the shorter columns as missing data.

Multi-Range Data

A single column in an Excel 2003 or earlier worksheet can hold up to 65536 data points for a variable. If your variables have more values than this and you choose not to adopt Excel 2007, StatTools allows multiple cell ranges to be assigned to a single data set. For example, you could "repeat" a data set across multiple sheets, assigning the same columns in different worksheets to hold all the values for a data set. You could also use this capability to assign different blocks of cells on the same worksheet to a single data set. This is helpful if your data is scattered about a single worksheet, but you want to combine it all into a single data set.

Stacked and Unstacked Data

StatTools supports both Stacked and Unstacked data. With some statistical procedures it easier to work with Stacked data, and with others, Unstacked data. For example, if we are comparing mean household incomes in several different neighborhoods, then in unstacked form there would be a separate *Income* variable (or column) for each neighborhood. In stacked form there would be a value variable *Income* and a category variable *Neighborhood* indicating which neighborhood each household is in.

StatTools Variable Stacking utility allows you to "stack" your variables into two columns: a value column, *Income*, and a category column, *Neighborhood*. Depending on the type of analysis, the stacked data set might be easier to work with than the unstacked version.

Treatment of Missing Values

If your data set has missing values (a common occurrence in statistical analysis) StatTools deals with them in an appropriate way, depending on the task. For example, summary measures such as means and standard deviations ignore missing values. As another example, a regression analysis involving three variables uses only the rows of the data set that have no missing values for any of the three variables. (This is called "listwise", or "casewise", deletion.) As a third example, a scatterplot of two variables plots only those points where both of the variables have nonmissing values.

Note: Not all StatTools procedures allow missing values. Check the Reference section of this manual to see how each procedure deals with missing values.

StatTools Reports and Charts

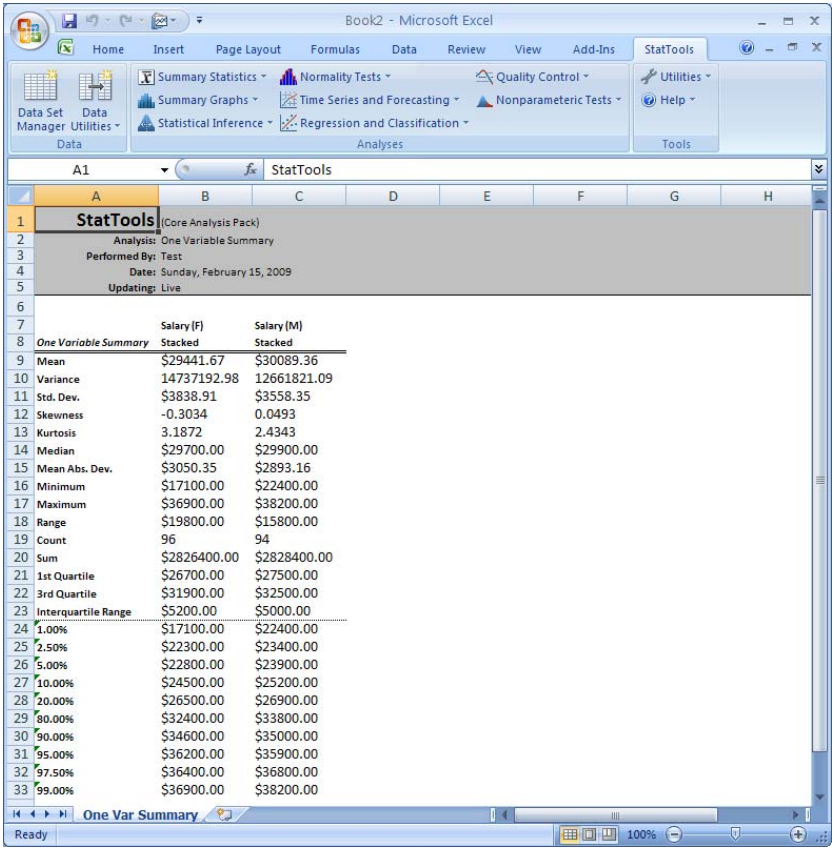
Whenever StatTools creates numerical output, such as a report from a regression analysis or a table of summary statistics, it gives a set of options for the placement of the report. These include:

- **In a New Workbook**, where a new workbook is created (if necessary) and each report is placed on a sheet in that workbook.
- **In New Worksheet in the Active Workbook**, where each report is placed in a new sheet in the active workbook.
- **Starting After Last Used Column**, where each report is placed on the active sheet to the right of the last used column.
- **Starting at Cell,** where you have the opportunity to select a cell where the top-left corner of the report or graph will be placed.

Whenever StatTools creates one or more charts, it places them with the reports. Charts are created in Excel format and may be customized using standard Excel chart commands.

Use of Formulas Versus Values

By default, StatTools tries to make the results as "live" as possible. That is, whenever it is practical, reports have formulas that link to the original data. For example, suppose you have a variable *Weight* and you want summary measures on *Weight*, such as its mean and standard deviation. The Summary Statistics procedure names the range of weights as *Weight*, and then it enters formulas in the output cells: **=StatMean(Weight)** and **=StatStdDev(Weight).** StatMean and StatStdDev are built in StatTools functions for calculating mean and standard deviation. These replace the standard built-in Excel functions for the same statistics.



Formulas are used in reports for two reasons. First, it helps you to learn statistical procedures and the StatTools functions in Excel. You don't just see a numerical result; you see how it is formed. Second, it has the practical advantage that if your data change, the results change automatically, so that you don't have to rerun the procedure.

There are times when it is not practical to do this. The prime example is regression. StatTools does not provide the formulas that are used to create regression output; it provides only the numerical results. In such cases, if your data change, you will have to rerun the procedures.

StatTools also gives you the option to turn off live updating. This is useful if Excel recalculation time becomes an issue as data changes.

Use of Comments in Cells

A feature of Excel is the ability to include a "pop-up" comment in any cell. You can tell that there is a comment in a cell by noticing a small red triangle in the upper right corner of the cell. You can read the comment by placing the cursor over the cell. StatTools has taken advantage of these comments to insert some context-sensitive help. You can think of these as the most "online" of all online help!

By the way, if you ever have a spreadsheet where the comments won't go away, that is, they always appear in front of your data, select the Tools/Options menu item, click on the View tab, and click on the Comment Indicator Only button.

Chapter 3: StatTools Reference Guide

Introduction

The StatTools Reference Guide chapter describes the icons, commands, and statistics functions used by StatTools. This chapter is divided into four sections:

1) *Reference: StatTools Icons*
2) *Reference: Summary of StatTools Procedures*
3) *Reference: StatTools Commands*
4) *Reference: StatTools Functions*

StatTools VBA Macro Language and Developer's Toolkit

StatTools also includes a powerful VBA-based macro language which can be used for:

1) **Automating StatTools analyses**
2) **Developing new statistical analyses which use the StatTools Data Set Manager, reports and graphs.** These custom calculations can provide analyses not found in the built-in StatTools procedures. These analyses can show up on the StatTools menu and toolbars if desired.

For more information on StatTools VBA Macro Language and Developer's Toolkit, see the on-line documentation supplied with the product.

Reference: StatTools Icons

StatTools Toolbar

StatTools icons are used to define data sets and variables and then run statistical procedures on those variables. StatTools icons appear on the Excel toolbar (i.e., as a custom toolbar in Excel) in Excel 2003 and earlier and on a ribbon in Excel 2007. This section briefly describes each icon, outlining the functions they perform and the menu command equivalents associated with them.

The following icons are shown on the StatTools toolbar in Excel 2003 and earlier.

Icon	*Function Performed and Command Equivalent*
	**Define a data set and variables, or edit or delete an existing data set and variables** *Command equivalent: Data Set Manager command*
	**Run a data utility** *Command equivalent: Data Utilities command*
	**Run a summary statistics procedure** *Command equivalent: Summary Statistics command*
	**Create summary graphs for variables** *Command equivalent: Summary Graphs command*
	**Run a statistical inference procedure** *Command equivalent: Statistical Inference command*
	**Run a normality test on variables** *Command equivalent: Normality Tests command*

Run a time series or forecasting procedure

Command equivalent: Time Series & Forecasting command

Run a regression or classification procedure

Command equivalent: Regression & Classification command

Run a quality control procedure

Command equivalent: Quality Control command

Run a nonparametric test

Command equivalent: Nonparametric Tests command

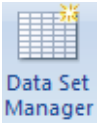
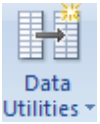
Display StatTools Utilities

Command equivalent: Utilities commands

Display StatTools help file

Command equivalent: Help command

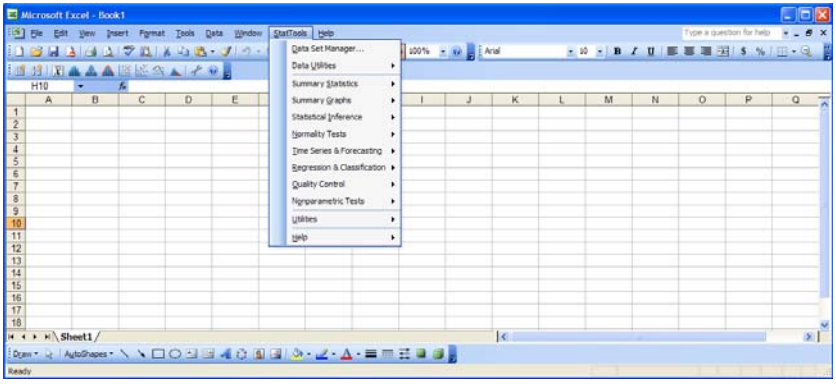
The following icons are shown on the StatTools ribbon in Excel 2007.

Icon	Function Performed and Command Equivalent
	**Define a data set and variables, or edit or delete an existing data set and variables** *Command equivalent: Data Set Manager command*
	**Run a data utility** *Command equivalent: Data Utilities command*
	**Run a summary statistics procedure** *Command equivalent: Summary Statistics command*
	**Create summary graphs for variables** *Command equivalent: Summary Graphs command*
	**Run a statistical inference procedure** *Command equivalent: Statistical Inference command*
	**Run a normality test on variables** *Command equivalent: Normality Test command*
	Run a time series or forecasting procedure *Command equivalent: Time Series & Forecasting command*
	Run a regression or classification procedure *Command equivalent: Regression & Classification command*
	Run a quality control procedure *Command equivalent: Quality Control command*
	Run a nonparametric test *Command equivalent: Nonparametric Tests command*
	Display StatTools Utilities *Command equivalent: Utilities commands*
	Display StatTools help file *Command equivalent: Help command*

Reference: StatTools Menu Commands

Introduction

This section of the Reference Guide details the available StatTools commands as they appear on the StatTools menu in Excel 2003 or earlier and on the StatTools ribbon in Excel 2007 and later. Commands are discussed as they appear on the menu, starting with the Data Set Manager command and subsequently moving down. StatTools icons can be used to perform many of the available commands. The **Reference: StatTools Icons** section of this chapter gives the command equivalents for each **StatTools** icon.



Several **StatTools** commands are also available in a pop-up floating menu that is displayed when the right mouse button is clicked in Excel.

Add-in Analysis Packs

StatTools Industrial can use "analysis packs", or add-in packages of new analyses not found in the core StatTools product. Commands found in these packages will be added to the StatTools menu. Thus, if you are using add-in analysis packs your menu may not look exactly as is shown here. For more information on developing add-on procedures for StatTools, see the on-line documentation for the StatTools VBA Macro Language and Developer's Toolkit.

Command Listing

The procedures available in StatTools come in natural groups. For each group there is a item on a StatTools menu. If a group has more than one item, there is a submenu listing the items in this group. This section provides a brief description of each procedure in each group. More detailed information on each procedure is provided in the section of this chapter titled **Reference: StatTools Commands**.

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Summary Statistics							
One Variable Summary Command	Generates Summary Statistics including the usual measures, such as average, median, and standard deviation, plus options such as quartiles and percentiles.	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-100
Correlations and Covariance Command	Creates a table of correlations and/or covariances for a set of variables that you select.	Allowed at beginning, middle and end of data	Live	Unstacked data only Up to 16m cases allowed	Yes	Ignored	1-250
Summary Graphs							
Histogram Command	Creates a Histogram for each variable you select. It gives you the option of defining the histogram's categories or "bins".	Allowed at beginning, middle and end of data	Partially Live – data changes update graph when data is within the graph's X-axis range	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-100
Scatter Plot Command	Creates a Scatter plot for each pair of variables you select.	Allowed at beginning, middle and end of data	Live	Unstacked data only Up to 32,000 cases allowed	No	Not allowed	1-10
Box-Whisker Plot Command	Creates a single boxplot (if you select a single variable) or side-by-side boxplots (if you select several variables).	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	No	Ignored	1-10

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Statistical Inference							
Confidence Interval - Mean/ Std. Deviation Command	Calculates a confidence interval for the mean and standard deviation of single variables, or the differences between the means for pairs of variables. The confidence intervals can be calculated using a One-Sample Analysis, a Two-Sample Analysis, or a Paired Sample Analysis.	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired sample analysis)
Confidence Interval - Proportion Command	Analyzes the proportion of items in a sample that belong to a given category (One-Sample Analysis), or compares two samples with regard to the proportion of items in a given category (Two-Sample Analysis).	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired sample analysis)
Hypothesis Test - Mean/ Std. Deviation Command	Performs hypothesis tests for the mean and standard deviation of single variables, or the differences between the means for pairs of variables. The hypothesis tests can be performed using a One-Sample Analysis, a Two-Sample Analysis, or a Paired Sample Analysis.	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired sample analysis)
Hypothesis Test - Proportion Command	Analyzes the proportion of items in a sample that belong to a given category (One-Sample Analysis), or compares two samples with regard to the proportion of items in a given category (Two-Sample Analysis).	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired sample analysis)

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Sample Size Command	Determines the sample size (or sample sizes) required to obtain a confidence interval with a prescribed half-length. It does this for confidence intervals for a mean, a proportion, the difference between two means, and the difference between two proportions.	n/a	n/a	n/a	n/a	n/a	n/a
One-Way ANOVA Command	An extension of the two-sample analysis for comparing two population means. It tests whether two or more means are all equal.	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	2-50
Two-Way ANOVA Command	Performs a two-way analysis of variance. This is usually done in the context of an experimental design where there are two "factors" that are each set at several "treatment levels.	Not allowed	Live	Stacked data Up to 16m cases allowed Experiment must be balanced	Yes	Not allowed	2 category variables, 1 value variable
Chi-square Independence Test Command	Uses a chi-square test to check whether row and column attributes in a contingency table are statistically independent.	No	Live (as long as table size does not change)	n/a	n/a	n/a	n/a

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Normality Tests							
Chi-Square Normality Test Command	Runs a chi-square test of normality for any variable you select.	Allowed at beginning, middle and end of data	Partially Live (bin positioning will not change but occupation and graphs will)	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1
Lilliefors Test Command	Provides a more powerful test for normality than the chi-square goodness-of-fit test. (More powerful means that it is more likely to detect nonnormality if it exists)	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-10
Q-Q Normal Plot Command	Creates a quantile-quantile (Q-Q) plot for a selected variable. It provides an informal test of normality.	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	No	Ignored	1
Time Series & Forecasting							
Time Series Graph Command	Creates a time series plot of one or more time series variable(s), all on the same chart.	Allowed at beginning, middle and end of data	Live	Unstacked data Up to 32,000 cases allowed	No	Not allowed	1-100

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Autocorrelation Command	Calculates any number of autocorrelations for a time series variable, indicates which (if any) are significantly nonzero, and (optionally) provides a bar chart (called a correlogram) of the autocorrelations.	Allowed at beginning or end of data	Live	Unstacked data Up to 32,000 cases allowed	No	Not allowed	1-10
Runs Test for Randomness Command	Performs a runs test to check whether a variable (usually a time series variable) is random.	Allowed at beginning or end of data	Live	Unstacked data Up to 16m cases allowed	Yes	Not allowed	1 or more
Forecasting Command	Forecasts time series data using the moving averages method, simple exponential smoothing, Holt's exponential smoothing method for capturing trend, and Winters' exponential smoothing method for capturing seasonality.	Allowed at beginning of data only	Live	Unstacked data Up to 32,000 cases allowed	Yes	Not allowed	1 or more
Regression & Classification							
Regression Command	Runs a variety of regression analyses including Simple Multiple, Stepwise, Forward, Backward and Block.	Allowed at beginning, middle and end of data	Static	Unstacked data Up to 16m cases allowed	Yes	Not allowed	1 dependent; 1-250 independent
Logistic Regression Command	Performs a logistic regression analysis on a data set. This is essentially a nonlinear type of regression analysis where the response variable is binary: 0 or 1.	Allowed at beginning, middle and end of data	Static	Stacked data Up to 16m cases allowed	Yes	Not allowed	1 dependent; 1-250 independent
Discriminant Analysis Command	Performs a discriminant analysis on a data set. There should be a "category" variable that specifies which of two or more groups each observation is in, plus one or more explanatory variables that can be used to predict group membership.	Allowed at beginning, middle and end of data	Static	Stacked data Up to 16m cases allowed	Yes	Not allowed except in dependent variables	1 dependent; 1-250 independent

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Quality Control							
Pareto Chart Command	Produces Pareto chart that allows you to see the relative importance of categorized data.	Allowed at beginning, middle and end of data	Static	Unstacked data	Yes	Ignored	1 category, or 1 value and 1 category
X/R Charts Command	Produces X and R charts that allow you to see whether a process is in statistical control.	Not allowed	Static	Unstacked data Up to 32,000 cases allowed	No	Not allowed	2-25
P Chart Command	Produces P charts that allow you to see whether a process is in statistical control.	Not allowed	Static	Unstacked data Up to 32,000 cases allowed	No	Not allowed	1 variable 1 size variable
C Chart Command	Produces C charts that allow you to see whether a process is in statistical control.	Not allowed	Static	Unstacked data Up to 32,000 cases allowed	No	Not allowed	1
U Chart Command	Produces U charts that allow you to see whether a process is in statistical control.	Not allowed	Static	Unstacked data Up to 32,000 cases allowed	No	Not allowed	1
Non-Parametric Tests							
Sign Test Command	Performs hypothesis tests for the median of a single variable or for the median of differences for a pair of variables	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired sample analysis)

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Wilcoxon Signed-Rank Test command	Performs hypothesis tests as does the Sign Test but assumes that the probability distribution is symmetric	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired sample analysis)
Mann-Whitney Test Command	Performs a hypothesis test on two samples	Allowed at beginning, middle and end of data	Live	Stacked and unstacked data Up to 16m cases allowed	Yes	Ignored	1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired sample analysis)
Data Utilities							
Stack Command	Takes a data set with separate variables for each group in separate columns, and allows you to "stack" them into two columns: a "category" column, and a "value" column. Depending on the type of analysis, the stacked data set might be easier to work with than the unstacked version.	Yes – anywhere in variable	Static	Unstacked data only Up to 65535 cases allowed	No	n/a	1-100
Unstack Command	Does the exact opposite of the Stack procedure.	Yes – anywhere in variable	n/a	Stacked data only Up to 16m cases allowed	Yes	n/a	1-32
Dummy Command	Creates dummy (0-1) variables based on existing variables.	Yes – anywhere in variable	Live	Unstacked data only Up to 16m cases allowed	Yes	n/a	1

Procedure	Description	Missing Data?	Live vs. Static Reports	Data Requirement	Multi-Range Data?	Invalid Data	# of Vars.
Combination Command	Creates a new variable from a pair of numeric variables, a category and numeric variable or two category variables	Yes – anywhere in variable	Live	Unstacked data only Up to 16m cases allowed	Yes	Not allowed	2-32 from same dataset
Interaction Command	Creates a new variable by taking the product, sum, average, min, max or min-max range from one or more variables.	Yes – anywhere in variable	Live	Unstacked data only Up to 16m cases allowed	Yes	n/a	2 value vars., or 1 value and 1 category var, or 2 category var
Lag Command	Creates a new lagged variable based on an existing variable. A lagged variable is simply a version of the original variable, "pushed down" by a number of rows equal to the lag.	Yes – anywhere in variable	Live	Unstacked data only Up to 16m cases allowed	Yes	Ignored	1
Transform Command	Applies any of four nonlinear transformations to any variables you select - natural logarithm, square, square root, or reciprocal - to create a new variable	Yes – anywhere in variable	Live or static	Unstacked data only Up to 16m cases allowed	Yes	Ignored	1-100
Difference Command	Creates any number of difference variables from an original variable.	Yes – anywhere in variable	Live	Unstacked data only Up to 16m cases allowed	Yes	n/a	1
Random Sample Command	Allows you to generate any number of random samples from a given data set, where sampling is with or without replacement.	Yes – anywhere in variable	Static	Stacked data only Up to 16m cases allowed	Yes	Ignored	1-32

StatTools Menu – DataSets

Data Set Manager Command

Defines StatTools data sets and variables, or edits or deletes an existing data set and variables

The **Data Set Manager** command allows you to define your data sets and variables. Once data sets and variables are defined, they may be analyzed in StatTools procedures. The Data Set Manager dialog box allows you to add or remove data sets, name a data set, specify the layout of the variables in a data set, and name the variables in a data set.

What Are Data Sets and Variables?

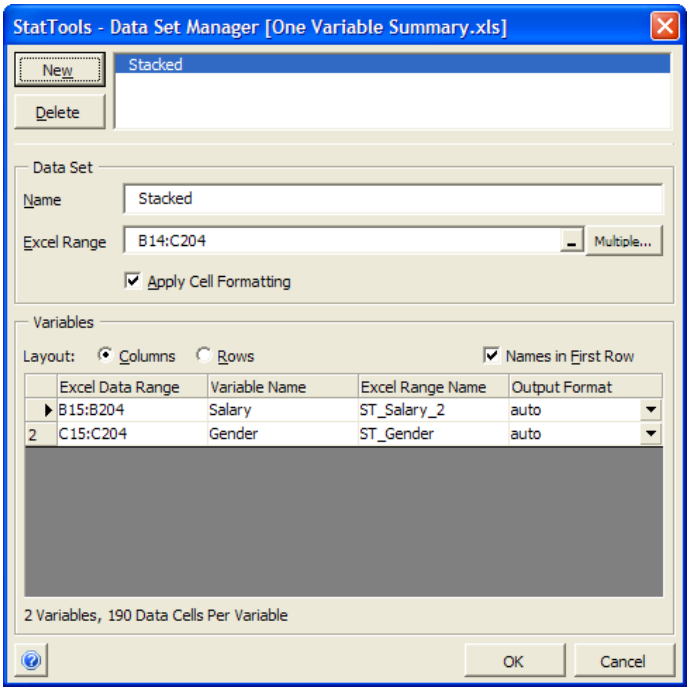
StatTools is analogous to most stand-alone statistical software packages in that it is structured around variables. For most analyses it is required that you work with a data set, or a set of statistical variables, often located in contiguous columns with variable names in the first row of the data set. You can then use these predefined variables when you run statistical analyses, without re-selecting the data you wish to analyze over and over.

Each variable in a data set has a name and a range of Excel cells associated with it. The selected **Layout** specifies how variables are located within a data set. A typical variable layout is **Columns** with one variable per column, but variables may also be laid out by **Rows**. A data set can include multiple blocks of cells, allowing you to put data on different sheets in the same workbook.

When you are defining a data set, StatTools attempts to identify the variables in a block of cells surrounding the current selection in Excel. This can make it quick and easy to set up a data set with variable names in the top row and variables laid out by column.

The lengths of the columns in your data set do not necessarily have to be equal. For example, you could have two variables, *Weight_Men* and *Weight_Women*, with different numbers of observations. However, for many analyses, StatTools will treat the blank cells in the shorter columns as missing data.

Data Set Manager Dialog Box



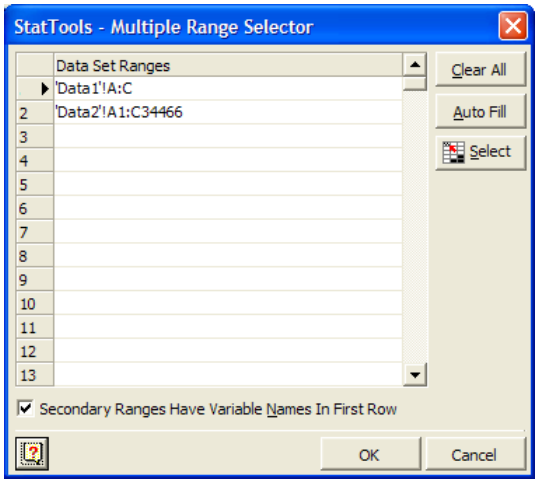
The **Data Set** options in the Data Set Manager dialog box include:

- **New, Delete –** adds a new data set, or deletes an existing one.
- **Name –** specifies the name of the data set.
- **Excel Range –** specifies the Excel Range associated with a data set. If multiple cell ranges have been assigned to a data set this entry will be prefaced by the label **Multiple**.
- **Apply Cell Formatting** - adds a grid and colors that identify your data sets.
- **Multiple** - clicking the **Multiple** button in the Data Set Manager dialog box displays the **Multiple Range Selector** dialog. This dialog allows the entry of individual cell ranges that comprise the multiple cell range data set.

Multiple Range Data Sets

StatTools allows multiple cell ranges to be assigned to a single data set. A multiple range data set can be used when:

1) Each variable in a data set has more than 65536 data points (in Excel 2003 or earlier), requiring the data set to extend across multiple worksheets in the same workbook,
2) The data for a variable is located in multiple blocks scattered throughout the worksheets in a workbook.

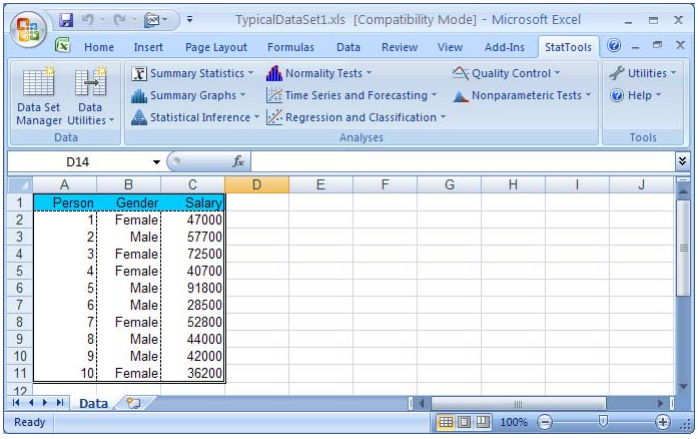


The options in the Multiple Range Selector dialog include:

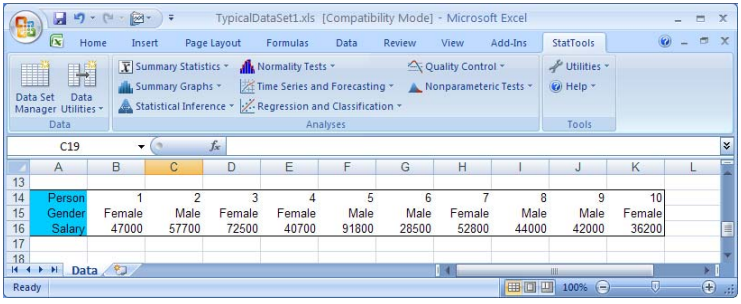
- **Clear All –** Clears all entered ranges.
- **Auto Fill –** Applies the first range entered (in row 1) to all visible worksheets in the active workbook, and enters these *SheetName!CellRange* references in the grid
- **Select –** Displays a selector for highlighting a block of cells to be used as a Data Set Range.
- **Secondary Ranges Have Variable Names in the First Column (Row)** - Multiple range data sets can have variable names labeling each column (or row, depending on the variable layout selected) in each range listed in the dialog, or variable names labeling the column or row in just the first selected range. The first selected range is the range entered in row 1 of the Multiple Range Selector dialog.

The **Variables** options in the Data Set Manager dialog box include:

- **Layout –** specifies how variables are structured in the Excel range that holds the data set. The options for Layout include:
  - **Columns**. This is the typical layout where each column in the data set's Excel range has the data for a variable. Often the names of the variables will be entered at the top of each column.



  - **Rows**. With this layout, each row of the data set holds the data for a variable. This layout often is used for time series data in Excel.

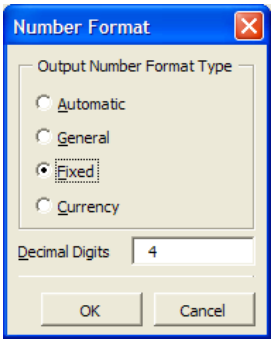


- **Names in First Column (or Row) –** select this when you have included the names of the variables in a data set in cells at the top of columns (or in the leftmost cells when variable layout is **Rows**).

Variables Options

Each row in the grid in the Data Set Manager dialog box lists the variables in a data set, including the name of each variable, the Excel range that holds the data points for a variable, and the Excel range name used to identify the data for the variable in Excel formulas.

- **Excel Range Name** - The shown range name will be used in Excel formulas that are created in StatTools reports and graphs. These formulas allow your reports to be "live" – that is, update automatically when a variable's data changes. Having range names that are understandable to you helps to make formulas more readable.
- **Output Format** – specifies the format for values shown for a variable in reports from StatTools analyses. The **Auto** entry specifies that StatTools will select a "best" format based on the numeric formatting applied to the cells containing the variable's values in Excel. By clicking the arrow next to the Output Format entry, you can alternatively select a specific format to be used:



General is equivalent to Excel's General numeric format. **Fixed** displays precision using the entered number of **Decimal Digits**. **Currency** is equivalent to Excel's Currency numeric format, and displays precision using the entered number of **Decimal Digits.**

Note: The desired output format can also be entered directly in the Data Set Manager dialog, using the notation *selectedFormat(#decimalDigits)*, such as *Currency(4).*

Data Set and Variable Capacities

In a single session, StatTools allows:

- Up to 256 data sets, located in a single workbook.
- Up to 256 variables per data set. All the data for a single data set must be located in the same workbook.
- Up to 16,777,216 data points per variable.

Actual data capacities may be less than shown above depending on the system configuration and version of Excel in use. Specific StatTools analyses may have different limitations. Memory limitations of Excel itself may also affect data capacities.

Note: the Data Set Manager dialog box lists all data sets and variables in the active workbook (this is the workbook listed in the caption of the Data Set Manager dialog). To list data sets in other workbooks, activate the desired workbook in Excel and display the Data Set Manager dialog.

Data Utilities Menu

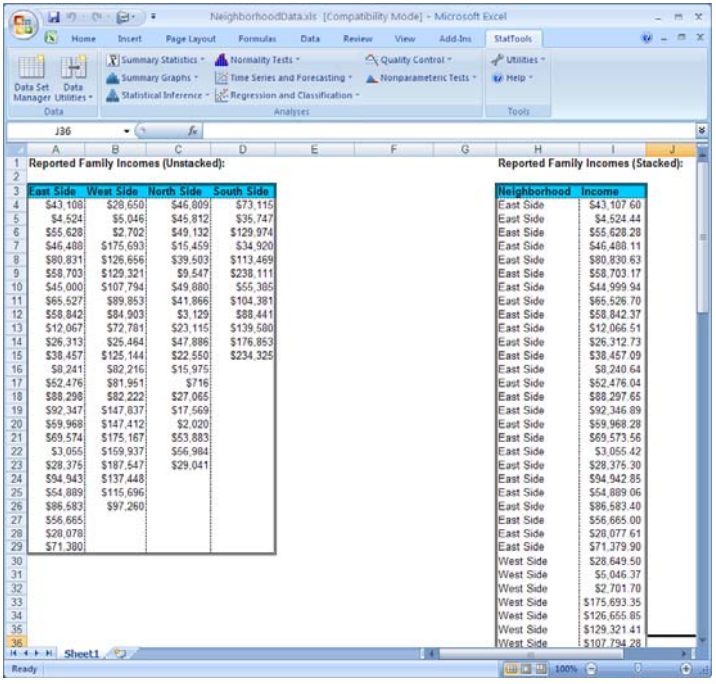
Stack Command

Converts a set of variables to stacked format from unstacked format

The Stack command allows you to convert data from "unstacked" form, where a data set includes at least two **value** variables, to a "stacked" form, in where the data set includes a **category** variable and a value variable. For example, if we are comparing mean household incomes in several different neighborhoods, then in unstacked form there would be a separate *Income* variable (or column) for each *Neighborhood*. These columns would not need to be of equal length, that is, each neighborhood could have a different sample size. In stacked form there would be a value variable *Income* and a category variable *Neighborhood* indicating which neighborhood each household is in.

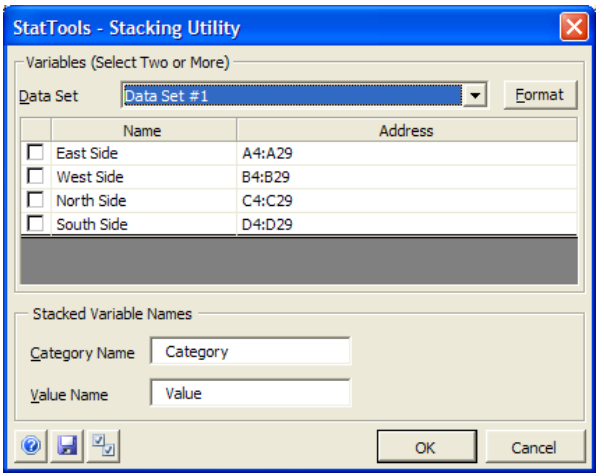
Essentially, this procedure allows you to "stack" your variables into two columns: a value column, *Income*, and a category column, *Neighborhood*. Depending on the type of analysis, the stacked data set might be easier to work with than the unstacked version.

Stacked and Unstacked Variables



Stacking Utility Dialog Box

Variables are stacked using the **Stacking Utility** dialog box:



At least two or more variables must be selected for stacking. The selected data set is always initially treated as unstacked data. Variables can be from different data sets.

The options in the Variable Stacking Utility dialog include:

- **Stacked Variable Names** – specifies the name of the **category** and **value** variables that will comprise the two variable stacked data set. These names will appear at the top of the columns for the category and value variables.

When OK is clicked, the variables are stacked and a new data set is created for the stacked data.

What are Category and Value Variables?

Category and **Value** variables are required for a stacked data set. The Category variable (sometimes referred to as the "code" variable) is simply a descriptive identifier for a related set of value variable(s). The Category variable is often a text label. Value variables, on the other hand, (sometimes referred to as "measurement" variables), are standard numeric variables which can be analyzed in statistical procedures.

Number of Variables for Analysis in Stacked Format

If a StatTools procedure imposes limits on the number of variables to select for analysis, in stacked format that limit applies to the number of categories in the category variable. In stacked format one typically selects a single category and a single value variable, thereby specifying multiple variables for analysis, one corresponding to each category in the category variable.

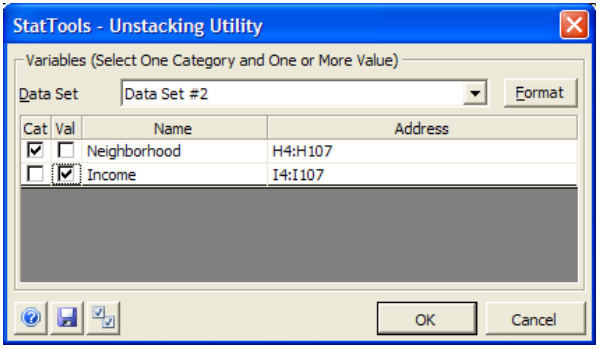
Unstack Command

Converts a set of variables from stacked format to unstacked format

The Unstack command does the exact opposite of the Stack command. For example, if you start with a category variable *Gender* and a value variable *Weight,* this command unstacks them into separate *Weight_Men* and *Weight_Women* columns.

Unstacking Utility Dialog Box

Variables are unstacked using the **Unstacking Utility** dialog box:



The selected data set is always initially treated as stacked data. At least two or more variables must be selected for unstacking. One of these variables is identified as the **Category** variable (by checking **Cat**) and one or more variables are identified as **Value** variables (by checking **Val**). Variables can be from different data sets.

When OK is clicked, the variables are unstacked and one or more new data set(s) are created for the unstacked data.

Transform Command

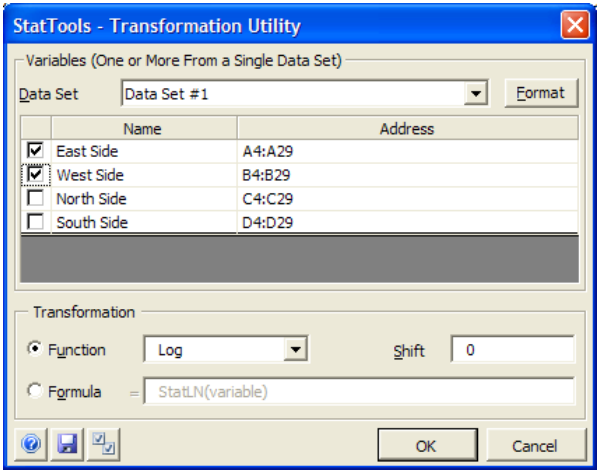
Transforms one or more variables to new variables and values based on an entered transformation function

The Transform command allows you to transform any variable with one of four possible transformations: natural logarithm, square, square root, or reciprocal. In addition, it allows you to enter a formula which will be used to calculate a transformed variable value.

If there are missing values for the variable on which the transformed variable is based, there will be corresponding missing values in the transformed variable.

Transformation Utility Dialog Box

Variables are transformed using the **Transformation Utility** dialog box:



The selected data set is always unstacked data. Only variables from one data set at a time can be transformed.

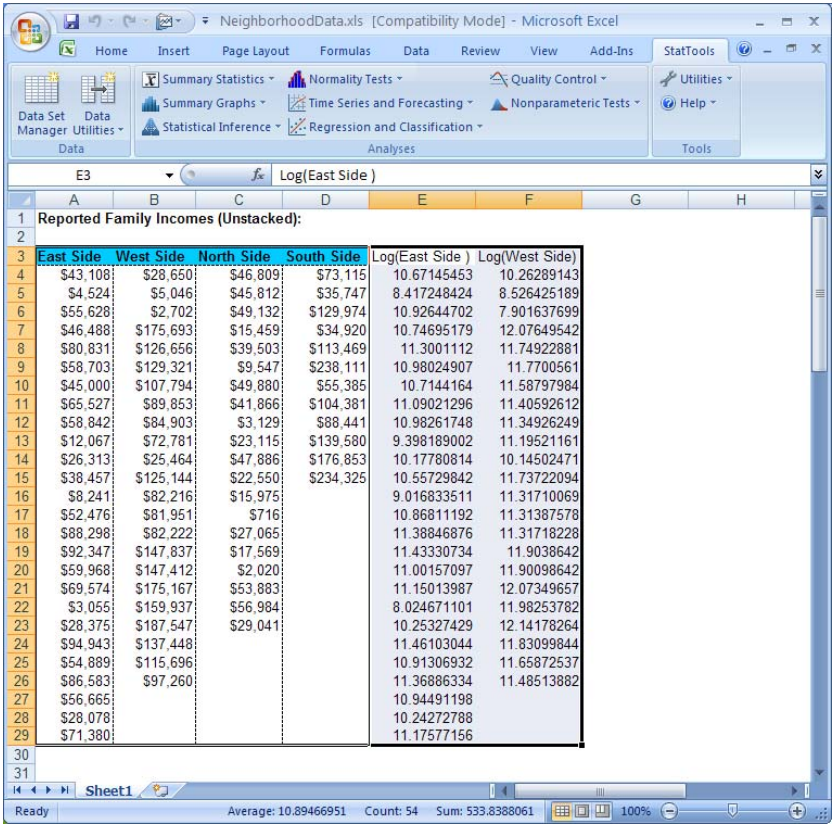
The options in the Transformation Utility dialog box include:

- **Transformation Function** – or the mathematical operation that will be performed on each value for the selected variables when generating the new transformed value. Built-in transformation functions include natural logarithm, square, square root, and reciprocal. A custom **Formula** can also be entered that calculates a new variable value based on a mathematical expression, such as:

  *(Variable*1.5)^2*

  Note that in the equation the keyword "Variable" is used as a placeholder for the actual value of the variable to be transformed.

Data Set with New Transformed Variables



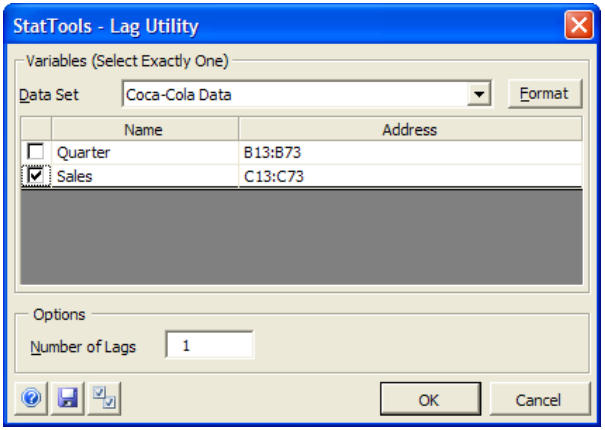
Lag Command

Creates a new lagged variable based on an existing variable

The Lag command allows you to create a new lagged variable based on an existing variable. A lagged variable is simply a version of the original variable, "pushed down" by a number of rows equal to the lag. For example, the lag 3 version of sales in November 1998 is sales three months earlier, in August 1998.

Lag Utility Dialog Box

Variables are lagged using the **Lag Utility** dialog box:

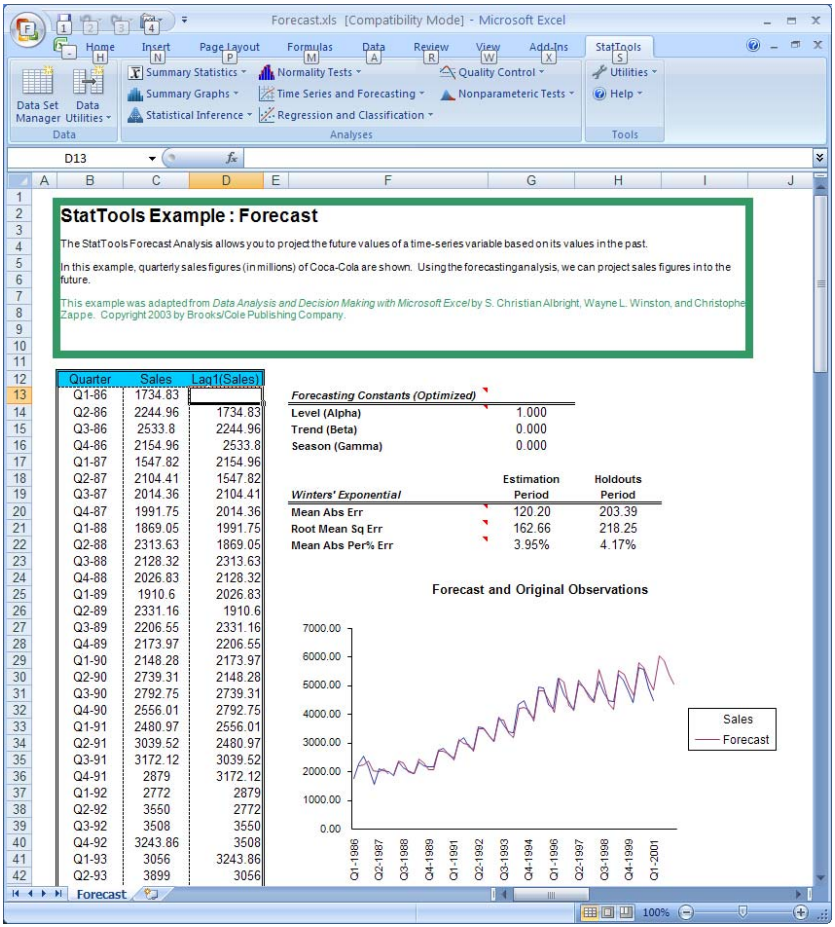


The selected data set is always unstacked data. Only one variable at a time can be lagged.

The options in the Variable Lag Utility dialog box include:

- **Number of Lags** – or the number of time periods to lag values when creating the new variables. A new variable is created for each of the lags up to the entered Number of Lags.

Lagged Variable to Right of Source Data Set



Difference Command

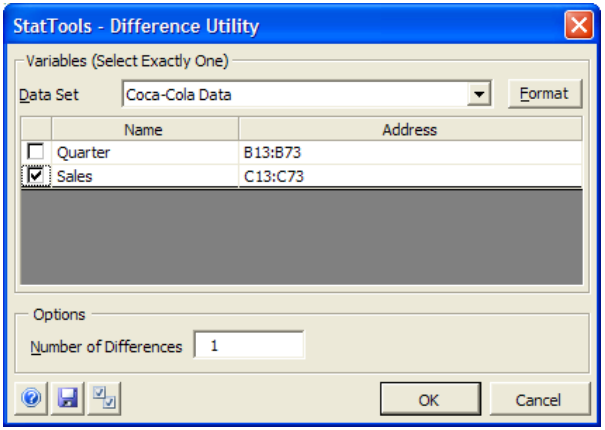
Creates any number of difference variables from an original variable

The Difference command allows you to create any number of difference variables from an original variable. It is used primarily for time series variables. A variable to be differenced is selected along with the number of differences (usually 1 or 2). The procedure then creates this many new difference variables. Each difference variable contains differences of the selected variable. For example (for monthly data), the March 1997 difference value is the original March 1997 value minus the original February 1997 value. Similarly, the second difference variable (if requested) contains the differences of the first differences.

Differencing is often used in time series analysis when the original variable is not "stationary" through time. For example, a time series with an upward trend is not stationary. Differencing often achieves stationarity. Sometimes second differencing is useful, but it is less common. Third differencing (or differencing beyond the third) is almost never necessary.

Difference Utility Dialog Box

Difference variables are created using the **Differences Utility** dialog box:

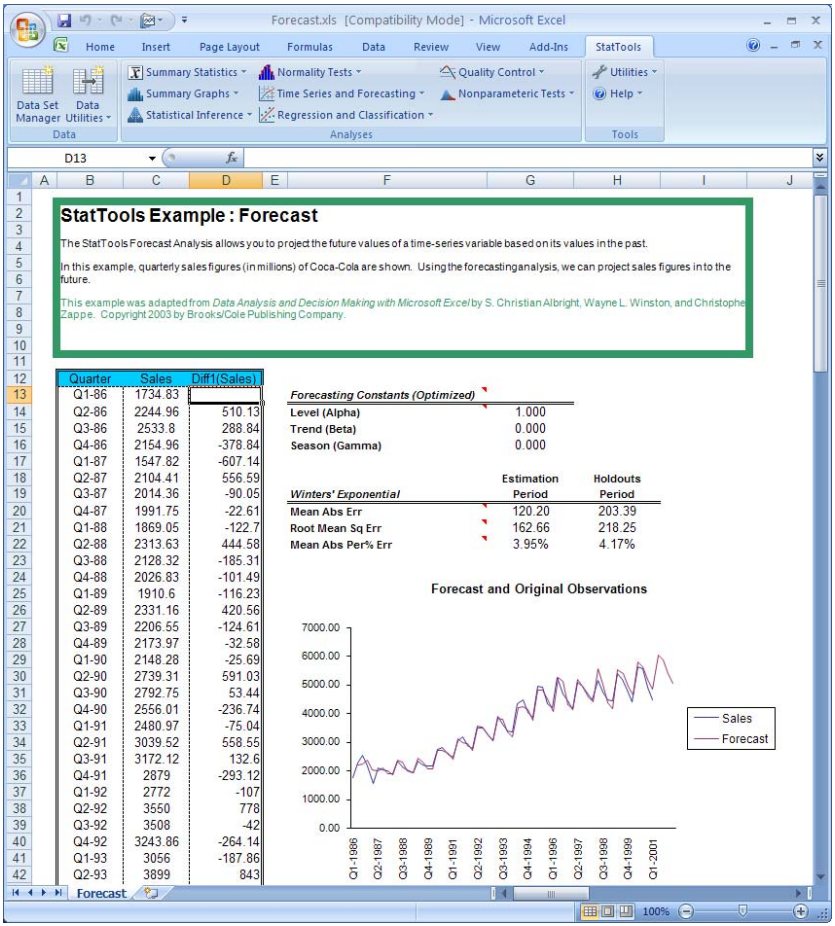


The selected data set is always unstacked data. Only one variable at a time can be used for creating difference variables.

The options in the Differences Utility dialog box include:

- **Number of Differences** – or the number of differences to create.

Data Set with Difference Variable



Interaction Command

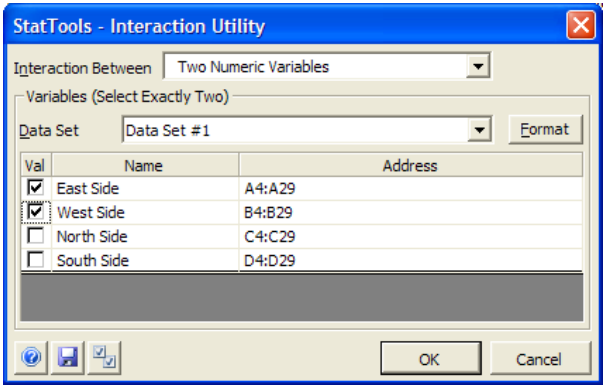
Creates an interaction variable from one or more original variables

The Interaction command allows you to create an interaction variable from one or more original variables. Interaction variables can be created using two numeric variables, one numeric and one category variable and two category variables.

If both variables are numerical (non-categorical), it creates their product. If one variable is numerical and the other is categorical, it creates the products of the numerical variable with each dummy corresponding to the categories of the categorical variable. Finally, if both variables are categorical, it creates products of all pairs of dummies from the two categorical variables.

Interaction Utility Dialog Box

Interaction variables are created using the **Interaction Utility** dialog box:



The selected data set is always unstacked data. One or more variables at a time can be used for creating an interaction variable.

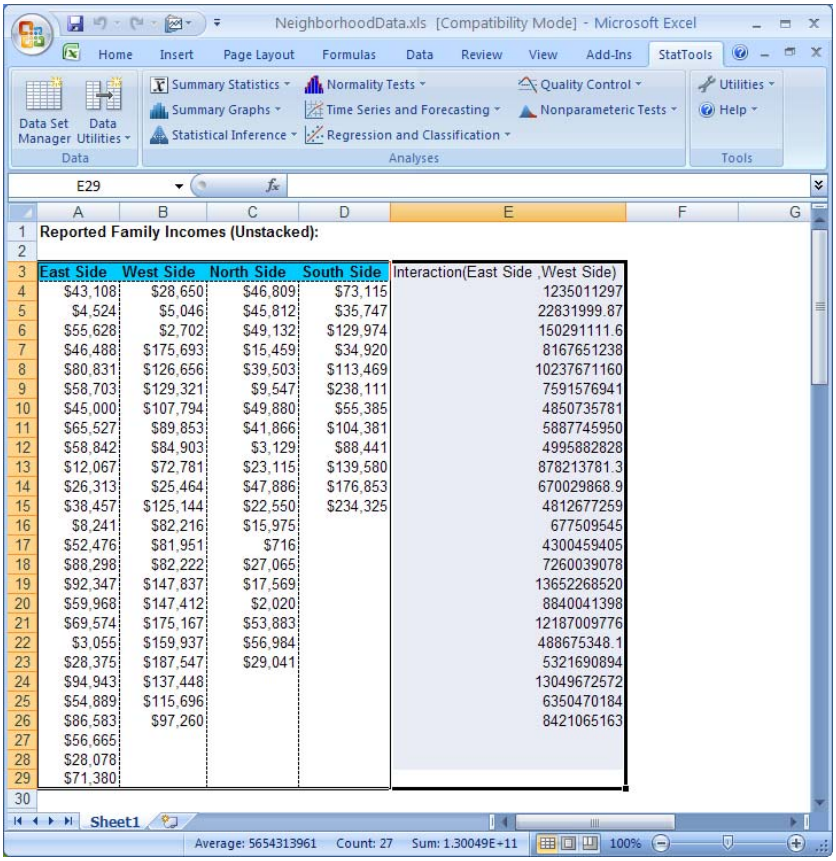
The options in the Interaction Utility dialog box include:

- **Interaction Between** – selects the type of each variable to be selected; **Two Numeric Variables, One Numeric and One Category Variable** or **Two Category Variables**.

How is an Interaction Variable Created?

An interaction variable is formed from the two variables you select in the dialog. There are three basic options for these two variables. First, they can both be numerical "measurement" variables. Then the interaction variable is their product. Second, one variable can be a numerical "measurement" variable and the other can be a categorical variable. Then StatTools internally creates dummy variables for each category of the categorical variable and multiplies *each* dummy by the numerical variable. Third, both variables can be categorical variables. Then StatTools internally creates dummy variables for each category of each categorical variable and multiplies each dummy for the first by each dummy for the second. For example, if the two categorical variables have 2 and 5 categories, respectively, then StatTools will create 2x5=10 interaction variables.

Data Set with Interaction Variable



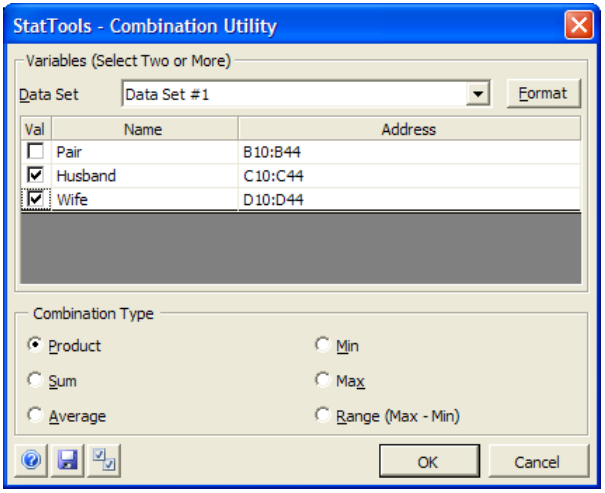
Combination Command

Creates a combination variable from one or more original variables

The Combination command allows you to create a combination variable from one or more original variables. Taking the product, sum, average, min, max or min-max range from one or more variables creates a combination variable.

Combination Utility Dialog Box

Combination variables are created using the **Combination Utility** dialog box:

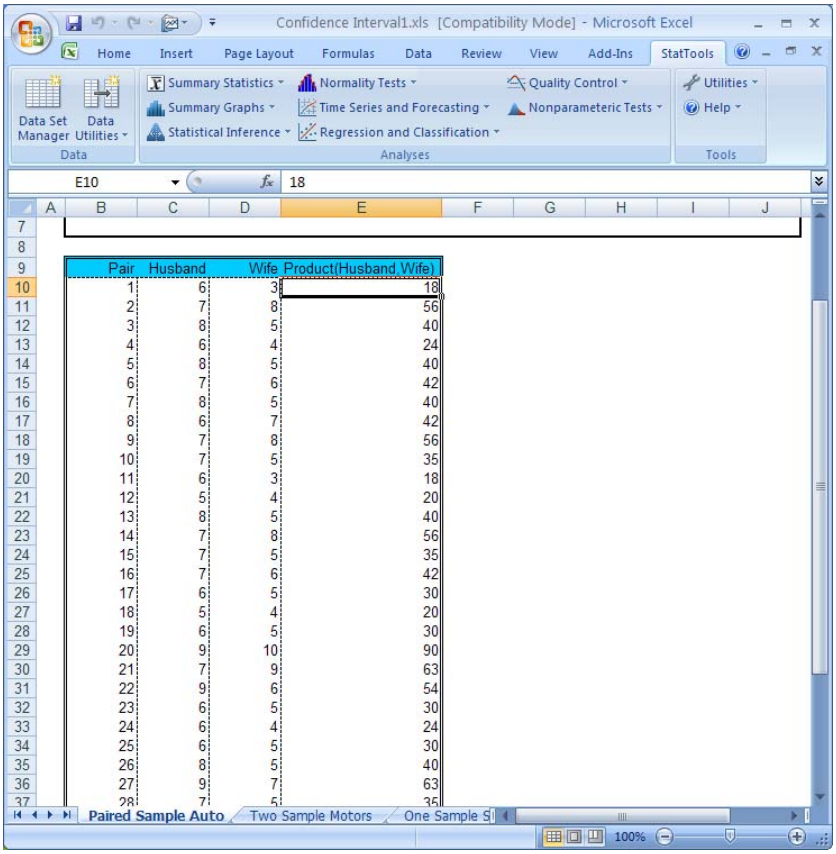


The selected data set is always unstacked data. One or more variables at a time can be used for creating a combination variable.

The options in the Combination Utility dialog box include:

- **Options** – or mathematical operation to be performed on the selected variables when creating the interaction variable. This can be a product, sum, average, min, max or min-max range.

Data Set with Combination Variable



Dummy Command

Creates dummy (0-1) variables based on existing variables

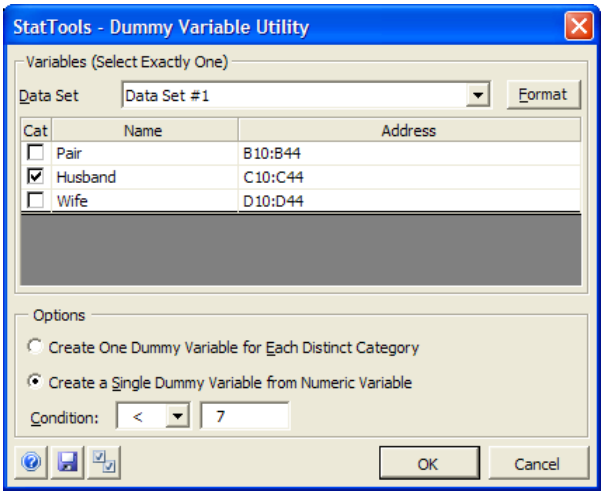
The Dummy command creates dummy (0-1) variables based on existing variables. There are two options:

1) You can create a dummy variable for each category of a categorical variable. For example, if you have a categorical variable *Origin* (for automobiles) with categories *US*, *Europe*, and *Asia*, it creates three dummies with variable names *Origin_US*, *Origin_Europe*, and *Origin_Asia*.

2) You can create a single dummy variable from a numerical variable, based on a cutoff value you select. For example, if you have a variable *Weight*, you could create dummies for the condition *Weight* <= *160*. In this case, a new variable would be added that had the value 0 when *Weight* > *160* and 1 when *Weight* <= *160*.

Dummy Utility Dialog Box

Dummy variables are created using the **Dummy Utility** dialog box:

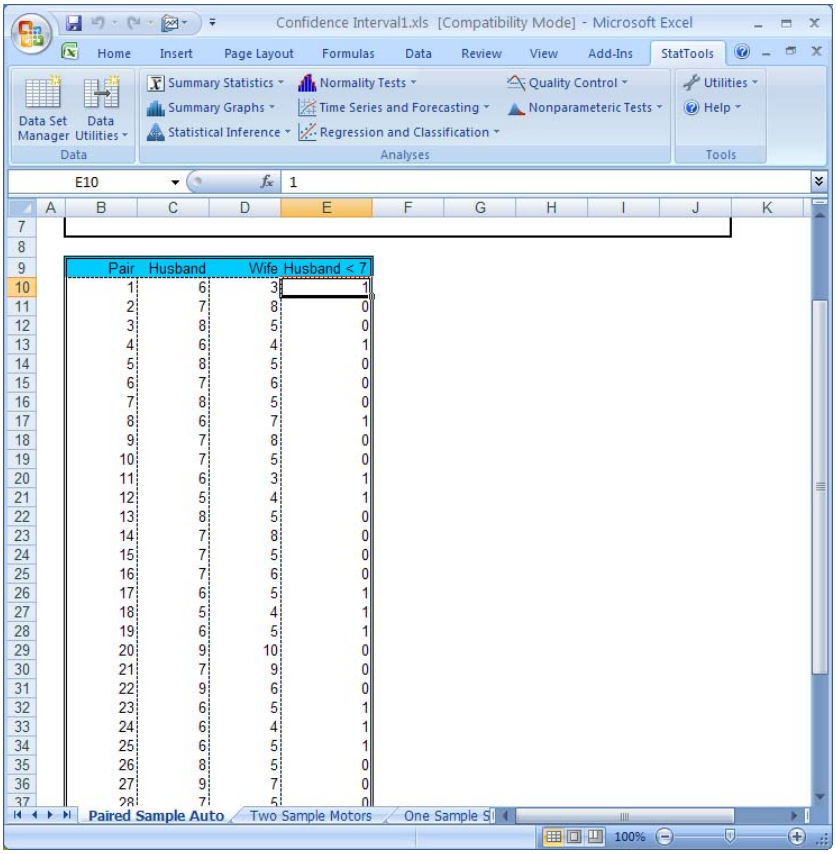


The selected data set must be unstacked data. One variable at a time can be used for creating dummy variables.

The options in the Dummy Utility dialog box include:

- **Options** – selects the manner in which dummy variables will be created – either 1) a dummy variable for each category of a categorical variable or 2) single dummy variable from a numerical variable. The **Cutoff value** specifies the cutoff to use when assigning a numeric variable to a 0-1 dummy variable.

Data Set with Dummy Variable



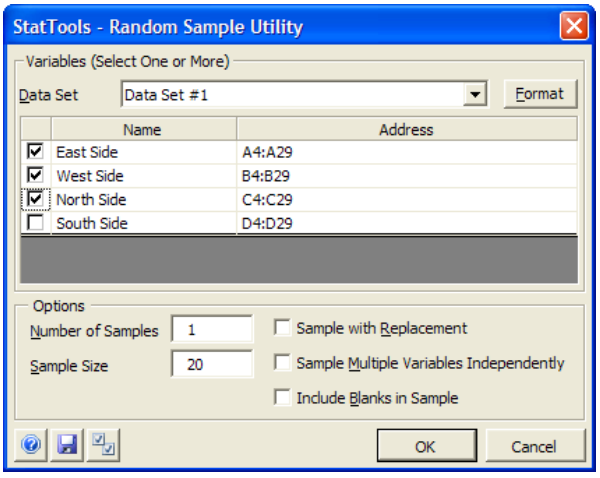
Random Sample Command

Generates any number of random samples from selected variables

The Random Samples command allows you to generate any number of random samples from selected variable(s). You specify the number of samples and the sample size for each sample, and StatTools generates the samples from the selected variables. Multiple variables may be sampled independently or dependently, and sampling may be done with or without replacement.

Random Sample Utility Dialog Box

Random Samples are generated using the **Random Sample Utility** dialog box:

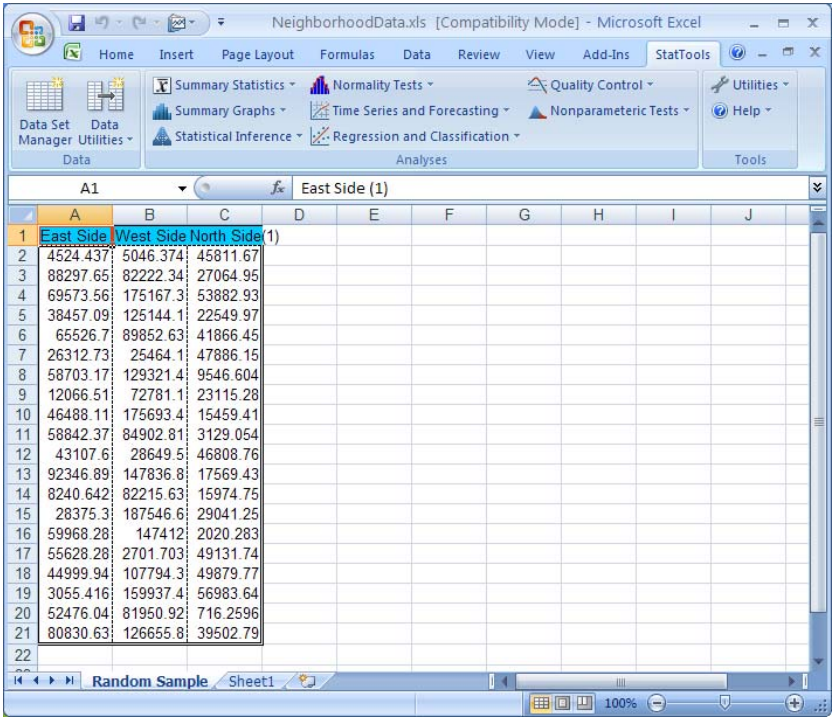


The selected data set is always unstacked data. One or more variables at a time can be used for generating random samples.

The options in the Random Sample Utility dialog box include:

- **Number of Samples** and **Sample Size**. The selected number of samples will be generated for each selected variable, and each sample will have a number of elements equal to the sample size.
- **Sample with Replacement.** Indicates that a value "goes back" to the original population after it is sampled, allowing it to be sampled again. Otherwise, if **Sampling with Replacement** is not selected (i.e., Sampling Without Replacement is used) a value is not returned and cannot be sampled again.
- **Sample Multiple Variables Independently.** Selects to have an independent draw used for each sampled value for each variable. Otherwise, for each sampled value, the same sampled index (a number between 1 to # of values in the variable) will be used for all variables.

Generated Random Samples



	A	B	C
1	East Side	West Side	North Side (1)
2	4524.437	5046.374	45811.67
3	88297.65	82222.34	27064.95
4	69573.56	175167.3	53882.93
5	38457.09	125144.1	22549.97
6	65526.7	89852.63	41866.45
7	26312.73	25464.1	47886.15
8	58703.17	129321.4	9546.604
9	12066.51	72781.1	23115.28
10	46488.11	175693.4	15459.41
11	58842.37	84902.81	3129.054
12	43107.6	28649.5	46808.76
13	92346.89	147836.8	17569.43
14	8240.642	82215.63	15974.75
15	28375.3	187546.6	29041.25
16	59968.28	147412	2020.283
17	55628.28	2701.703	49131.74
18	44999.94	107794.3	49879.77
19	3055.416	159937.4	56983.64
20	52476.04	81950.92	716.2596
21	80830.63	126655.8	39502.79
22			

Summary Statistics Menu

The commands on the Summary Statistics Menu allow you to calculate several numerical summary measures for single variables or pairs of variables. Note that there are no contingency tables in StatTools' procedures. Excel already provides this capability with pivot tables.

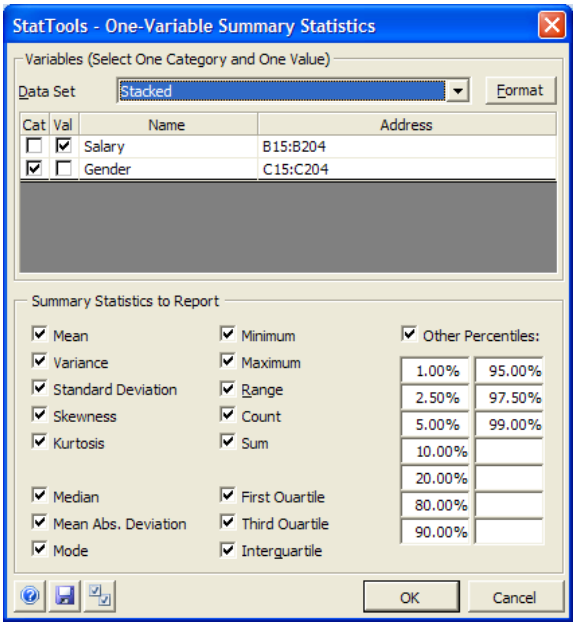
One Variable Summary Command

Calculates summary statistics for variables

The **One Variable Summary** command provides summary data for any number of selected numerical variables. These include the mean, median, standard deviation, variance, minimum, maximum, range, first quartile, third quartile, interquartile range, mean absolute deviation, skewness, kurtosis, count, sum, and selected percentiles.

One Variable Summary Statistics Dialog Box

This analysis is set up using the **One Variable Summary Statistics** dialog box:



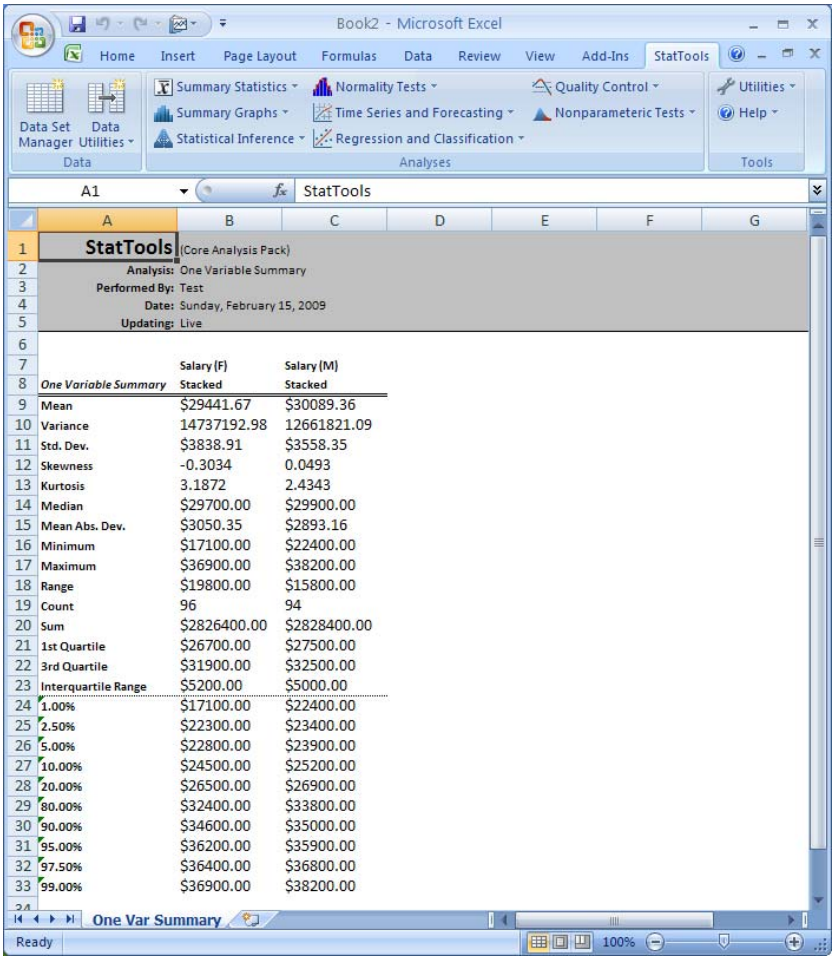
One or more variables can be selected for analysis. The selected data set can be stacked or unstacked data. Variables can be from different data sets.

The options in the One Variable Summary Statistics dialog include:

- **Summary Statistics to Report** – Selects the desired statistics to include in the report. Percentiles can be added by typing in the desired values.

One Variable Summary Report

The One Variable Summary Report uses StatTools Stat functions (such as **StatSkewness**) to allow hot-linking to data. The report is placed in the location specified using the Settings command.



Missing Data and Link to Data

- **Missing Data -** This procedure allows missing data in a casewise manner. That is, for each variable, the missing data for that variable are ignored when calculating the summary measures. (This is Excel's default method anyway. For example, if you use the AVERAGE function on a range, it will average only the *numerical* values in the range.)
- **Link to Data** - All of the summary measures are calculated by formulas that are linked to the data. Therefore, if any of the data change, the summary measures change automatically.

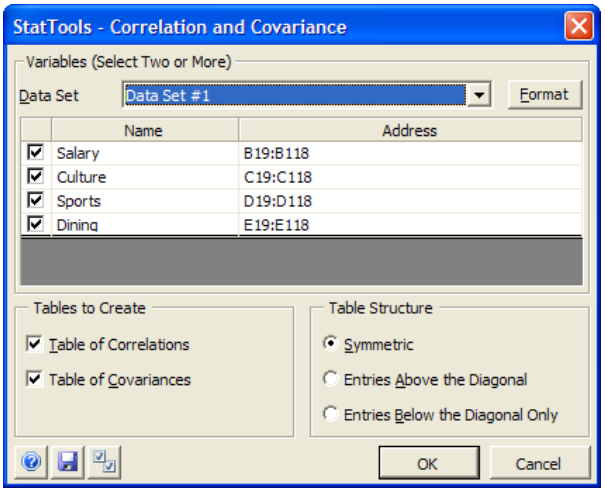
Correlations and Covariance Command

Produces a table of correlations and/or a table of covariances between variables

The **Correlations and Covariance** command produces a table of correlations and/or a table of covariances between any set of selected numerical variables. Because both of these tables are symmetric (e.g., the correlation between X and Y is the same as the correlation between Y and X), you can choose to have (1) only the correlations (or covariances) *below* the diagonal show, (2) only those *above* the diagonal show, or (3) those below *and* above the diagonal show.

Correlations and Covariance Dialog Box

This analysis is set up using the **Correlations and Covariance** dialog box:



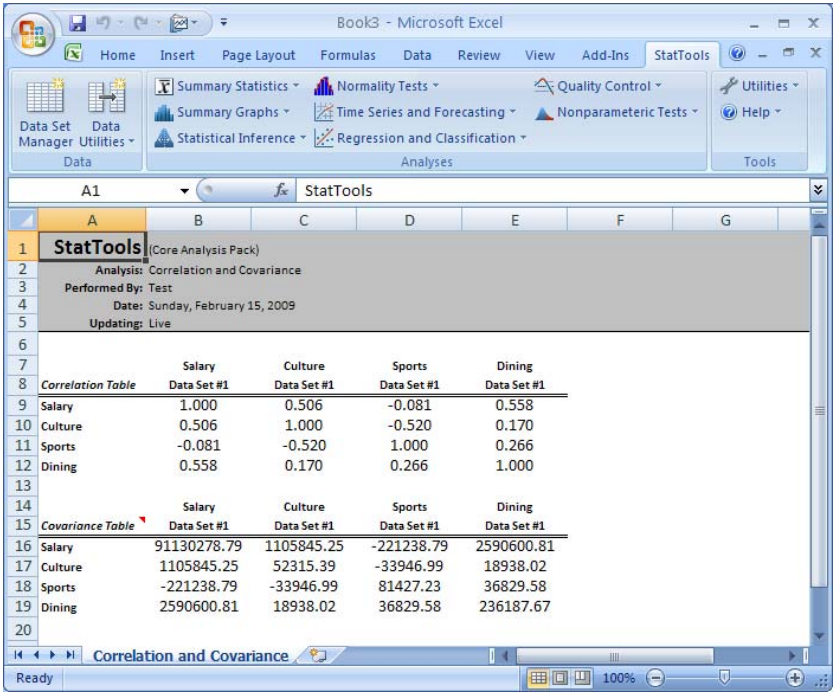
Two or more variables must be selected for analysis. The selected data set must be unstacked data. Variables can be from different data sets.

The options in the Correlations and Covariance dialog include:

- **Tables to Create** – Selects the desired correlation and/or covariance table
- **Table Structure** – Specifies the structure of the table(s) to be generated:
  - **Symmetric** correlations (or covariances) below *and* above the diagonal show
  - **Entries Above the Diagonal Only** correlations (or covariances) *above* the diagonal show
  - **Entries Below the Diagonal Only** correlations (or covariances) *below* the diagonal show

Correlations and Covariance Report

The Correlations and Covariance Report uses StatTools Stat functions (such as **StatCorrelationCoeff**) to allow hot-linking to data. The report is placed in the location specified using the Settings command.



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed, and they are treated in a pairwise manner. That is, to obtain the correlation (or covariance) between any pair of variables, all cases with missing data on *either* of the two variables are ignored.

- **Link to Data** - The correlations and covariances are calculated by formulas that are linked to the data. Therefore, if any of the data change, these summary measures update automatically.

Summary Graphs Menu

The commands on the Summary Graphs Menu enable you to create charts that are very useful in statistical analysis and not terribly easy (or possible) to produce with Excel's chart wizard. Of course, Excel's charting capabilities are extensive, so StatTools tries not to duplicate things that Excel already does well.

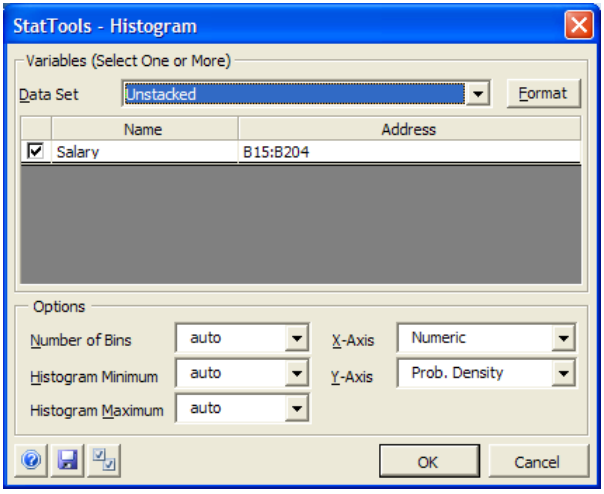
Histogram Command

Creates histograms for variables

The Histogram command creates a histogram for each variable you select. It gives you the option of defining the histogram's categories (often called "bins"), and it shows these clearly on the chart. It also creates a frequency table that each histogram is based on.

Histogram Dialog Box

This graph type is set up using the **Histogram** dialog box:

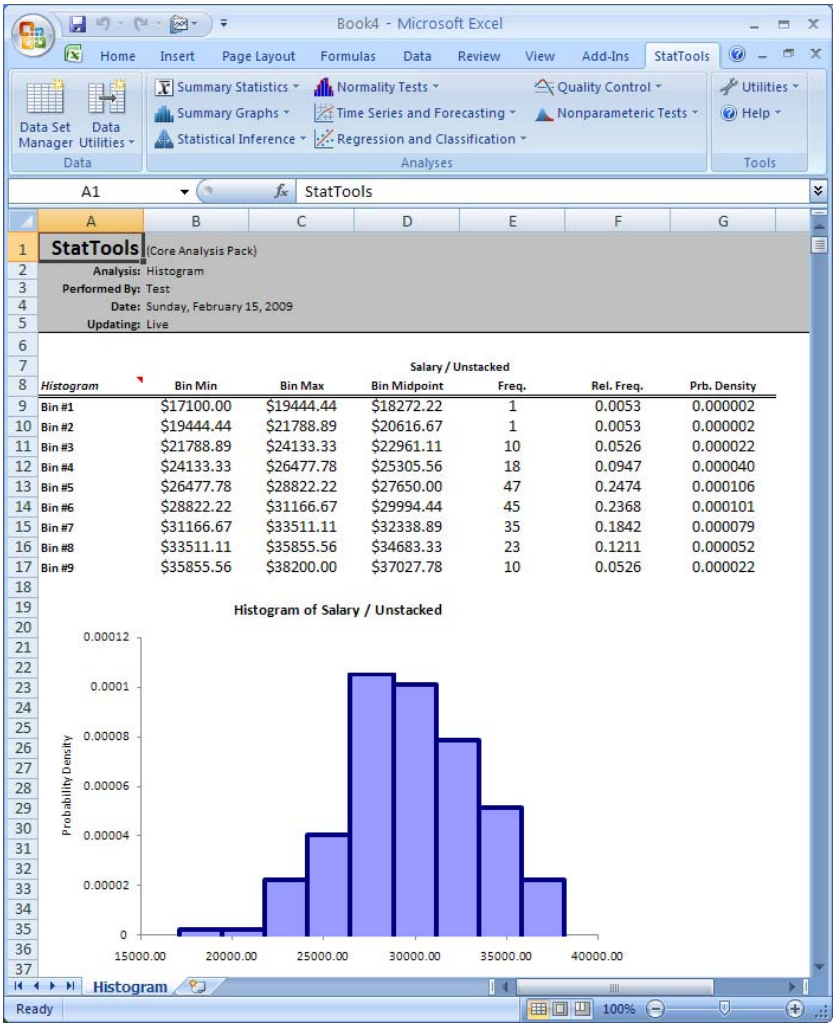


One or more variables can be selected for graphing. The selected data set can be stacked or unstacked data. Variables can be from different data sets.

The options in the **Histogram** dialog include:

- **Number of Bins**. Sets the number of histogram intervals calculated across the range of a graph. The value entered must be in the range 1 to 200. The setting **Auto** calculates the best number of bins to use for your data based on an internal heuristic.
- **Histogram Minimum**. Sets the minimum value where histogram bins start. **Auto** specifies that StatTools will start the histogram bins based on the minimum of the data graphed.
- **Histogram Maximum**. Sets the maximum value where histogram bins end. **Auto** specifies that StatTools will end the histogram bins based on the maximum of the data graphed.
- **X-axis**. Selects **Categorical** or **Numeric**. A categorical x-axis simply labels each bin with the midpoint of the bin. A numeric x-axis has a "readable" x-axis minimum and maximum and can be rescaled using Excel's standard rescaling options.
- **Y-axis.** Selects **Frequency, Rel. Frequency** or **Prob. Density** as the unit of measure reported on the Y-axis. Frequency is the actual number of observations in a bin. Relative Frequency is the probability of a value in the range of a bin occurring (observations in a bin/total observations). Density is the relative frequency value divided by the width of the bin, insuring that Y-axis values stay constant as the number of bins is changed.

Histogram of a Single Variable



Histogram	Bin Min	Bin Max	Bin Midpoint	Freq.	Rel. Freq.	Prb. Density
Bin #1	$17100.00	$19444.44	$18272.22	1	0.0053	0.000002
Bin #2	$19444.44	$21788.89	$20616.67	1	0.0053	0.000002
Bin #3	$21788.89	$24133.33	$22961.11	10	0.0526	0.000022
Bin #4	$24133.33	$26477.78	$25305.56	18	0.0947	0.000040
Bin #5	$26477.78	$28822.22	$27650.00	47	0.2474	0.000106
Bin #6	$28822.22	$31166.67	$29994.44	45	0.2368	0.000101
Bin #7	$31166.67	$33511.11	$32338.89	35	0.1842	0.000079
Bin #8	$33511.11	$35855.56	$34683.33	23	0.1211	0.000052
Bin #9	$35855.56	$38200.00	$37027.78	10	0.0526	0.000022

Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** Histograms are partially linked to data. When data changes and new values fall within the range of the bins of the original histogram, the graph will automatically update. Data changes that require re-binning will not update on the graph.

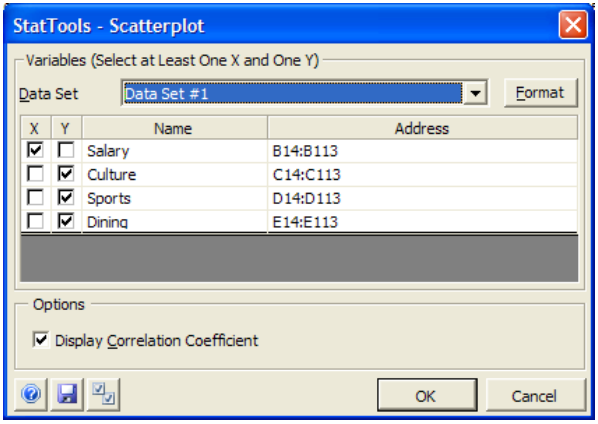
Scatter Plot Command

Creates scatter plots between pairs of variables

The Scatter Plot command creates a scatterplot for each pair of variables you select. Scatter plots can be created with Excel's XY-plot option but Excel automatically puts the first (or leftmost) variable on the horizontal axis, which might not be what you want. StatTools allows you to choose which of the variables will be placed on the horizontal axis. Each scatterplot shows the correlation between the two variables with the corresponding plot.

Scatter Plot Dialog Box

This graph type is set up using the **Scatter Plot** dialog box:

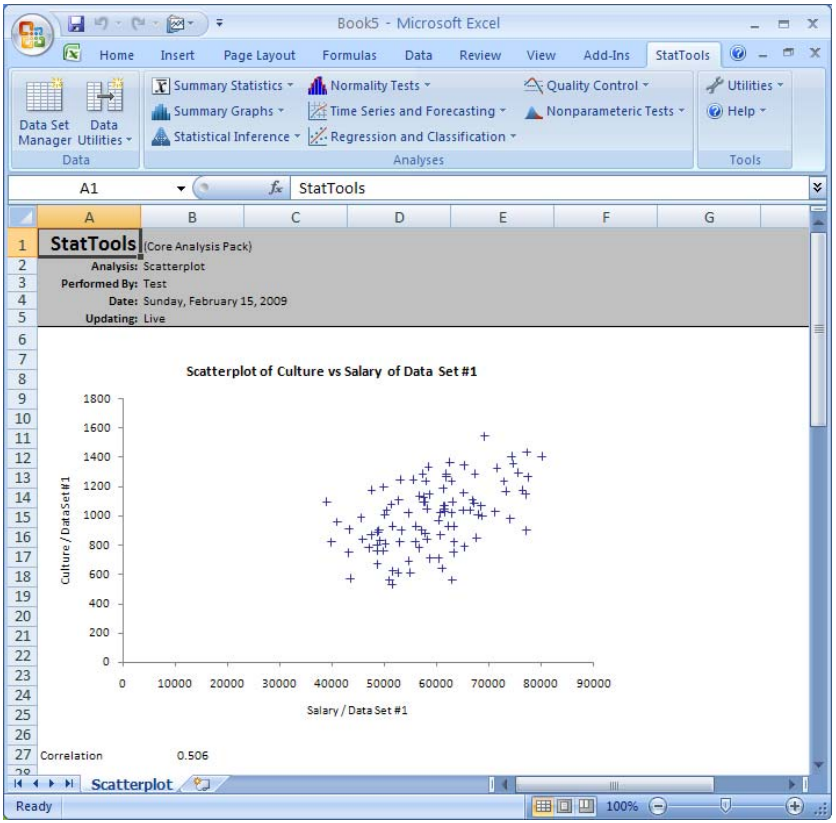


Two or more variables can be selected for graphing. At least one X-axis and one Y-axis variable are required. If more than two variables are selected multiple scatter plots are created. The selected data set must be unstacked data. Variables can be from different data sets.

The options in the **Scatter Plot** dialog include:

- **Display Correlation Coefficient**. Specifies that the correlation coefficient between the graphed variables will be displayed

Scatter Plot Example



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on either of the two selected variables in any given pair are ignored.
- **Link to Data** – The scatterplots are linked to the original data. If the data change, so do the scatterplots. However, the scales of the axes might need to be updated manually if the ranges of the selected variables change significantly.

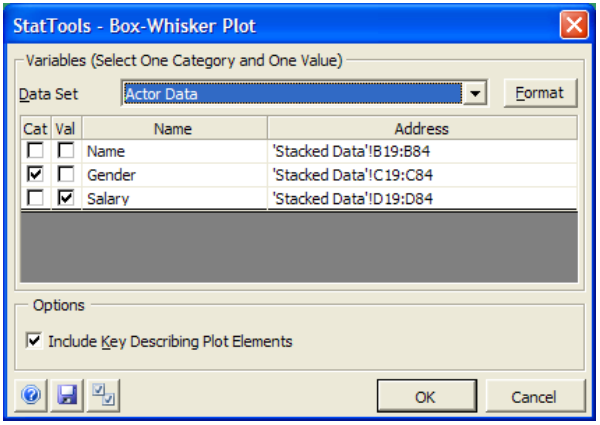
Box-Whisker Plot Command

Creates Box-Whisker plots for variables

The Box-Whisker plot command creates a single Box-Whisker plot (if you select a single variable) or side-by-side Box-Whisker plots (if you select several variables). It also creates a sheet that shows the summary statistics (quartiles, interquartile range, etc.) that are used to form the Box-Whisker plot(s).

Box-Whisker Plot Dialog Box

This graph type is set up using the **Box-Whisker Plot** dialog box:

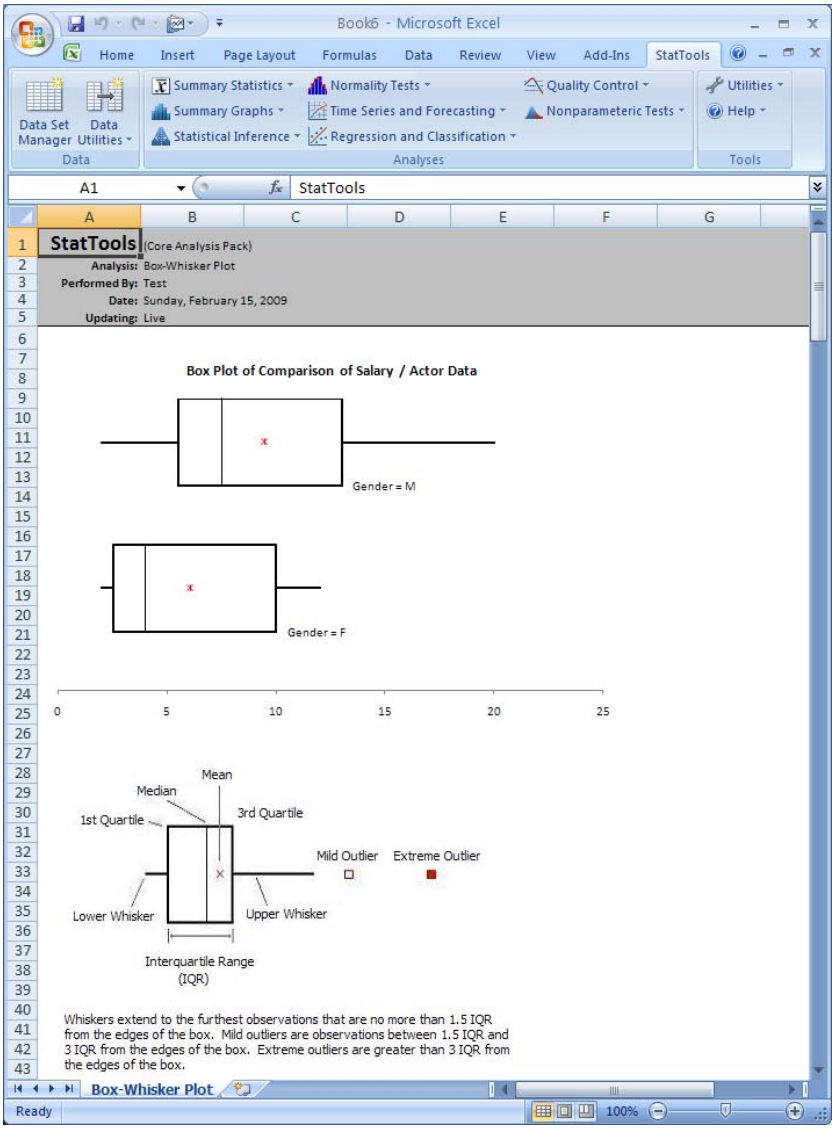


One or more variables can be selected for graphing. The selected data set can be stacked or unstacked data. Variables can be from different data sets.

The options in the **Box-Whisker Plot** dialog include:

- **Include Key Describing Plot Elements**. Specifies that a separate key describing chart elements will be displayed below the graph.

Box-Whisker Plot Example



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on any of the selected variables for the chart are ignored.
- **Link to Data** – The Box-Whisker plots that are created are linked to the original data; if the data change, so do the Box-Whisker plots. However, the scale of the horizontal axis might need to be updated manually if the scale of the data changes significantly

Statistical Inference Menu

The commands on the Statistical Inference Menu perform the most common statistical inference analyses: confidence intervals and hypothesis tests, along with one and two-way ANOVA.

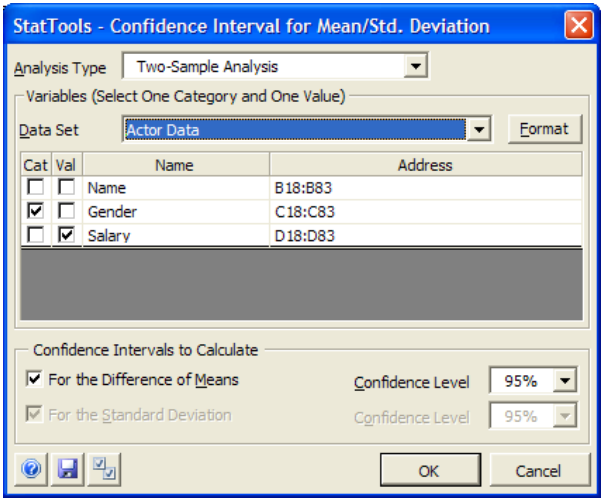
Confidence Interval - Mean/ Std. Deviation Command

Calculates confidence intervals for mean and standard deviation of variables

The Confidence Interval for Mean/ Std. Deviation command calculates a confidence interval for the mean and standard deviation of single variables, or the differences between the means for pairs of variables. The confidence intervals can be calculated using a **One-Sample Analysis**, a **Two-Sample Analysis**, or a **Paired Sample Analysis**.

Confidence Interval for Mean/ Std. Deviation Dialog Box

This analysis is set up using the **Confidence Interval for Mean/ Std. Deviation** dialog box:

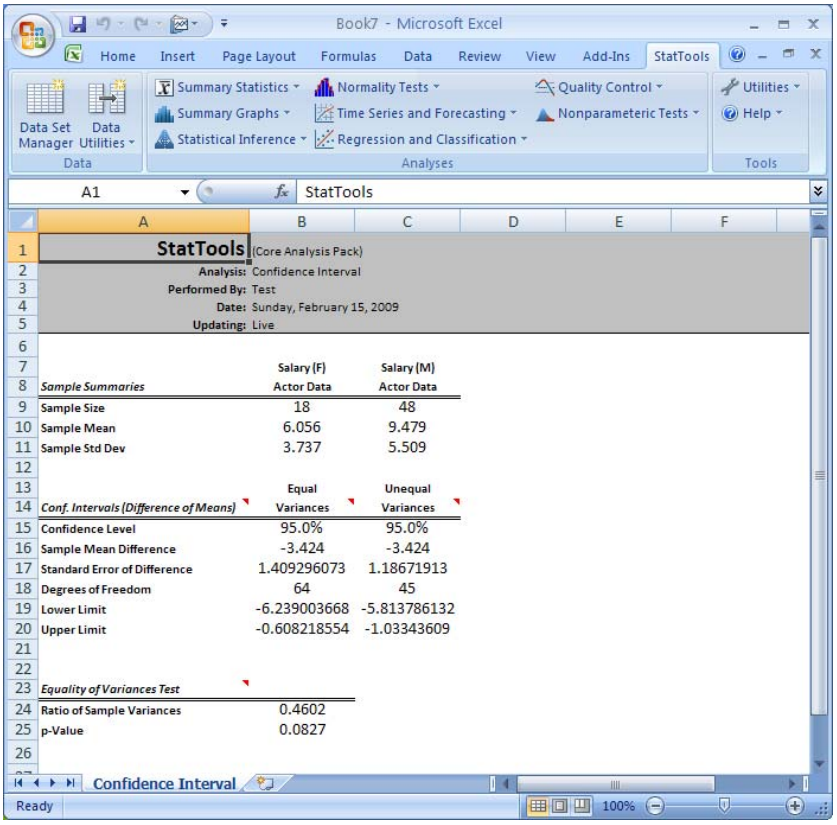


The number of variables selected depends on the Analysis Type used. A One-Sample Analysis requires one or more variables, while a Two-Sample Analysis and a Paired Sample Analysis require two variables. The selected data set can be stacked or unstacked data. Variables can be from different data sets.

The options in the **Confidence Interval** dialog include:

- **Analysis Type.** Selects the type of analysis performed. Options include:
  - **One-Sample Analysis**. Calculates confidence intervals for a single numerical variable.
  - **Two-Sample Analysis**. Calculates confidence interval for the difference between means from two independent populations.
  - **Paired Sample Analysis**. This is basically the same as the two-sample analysis, but it is appropriate when the two variables are naturally paired in some way. It essentially runs a one-sample analysis on the differences between pairs.
- **Confidence Intervals to Calculate**. Specifies the confidence intervals that will be calculated on the selected variables. Options change based on the analysis type selected:
  - **One-Sample Analysis**. Selects to calculate confidence intervals on the mean and/or standard deviation, and the confidence level (0 to 100%) for each.
  - **Two-Sample Analysis** or **Paired Sample Analysis**. Selects to calculate the confidence interval for the difference between means for two variables, and specifies the confidence level (0 to 100%) desired.

Confidence Interval Report



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

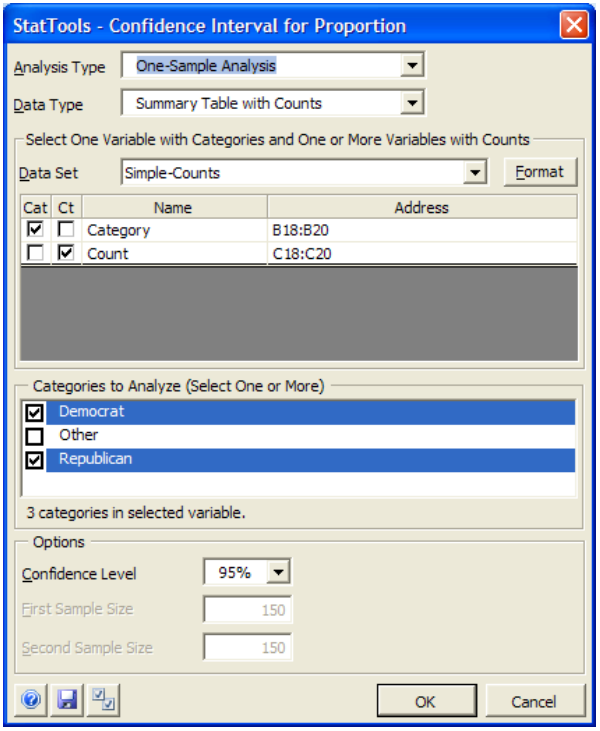
Confidence Interval - Proportion Command

Calculates confidence intervals for proportions

The Confidence Interval for Proportion command allows you to analyze the proportion of items in a sample that belong to a given category (**One-Sample Analysis**), or to compare two samples with regard to the proportion of items in a given category (**Two-Sample Analysis**). There are three Data Types supported by this procedure: **Population Sample**, **Summary Table with Counts**, and **Summary Table with Proportions**.

Confidence Interval for Proportion Dialog Box

This analysis is set up using the **Confidence Interval - Proportions** dialog box:



The number of variables selected depends on the Analysis Type used. A **One-Sample** analysis requires one or more samples for analysis; a **Two-Sample** analysis requires two samples. One selects variables with sample information using column labeled **Ct** (Count), % (Proportion), or column with no label for the **Population Sample** Data Type. (If the Population Sample data is stacked, the samples are selected in columns labeled **C1** and **C2**, where C1 contains the stacked

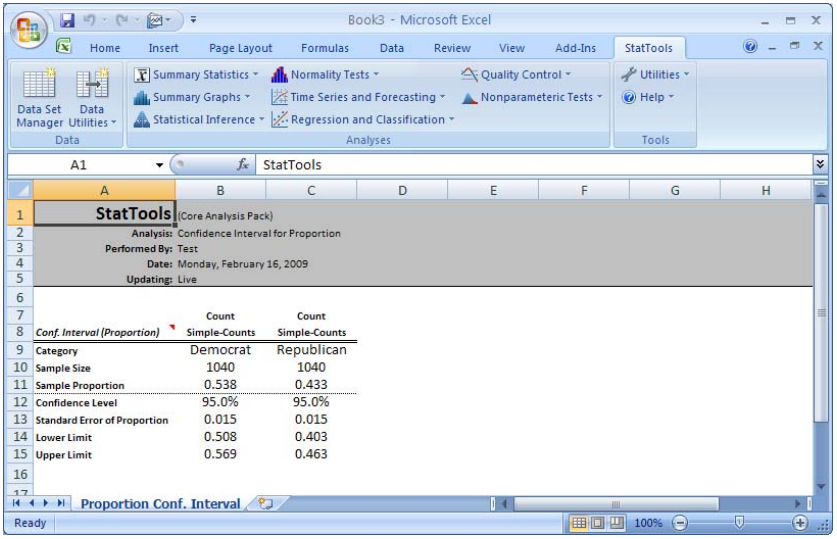
categories.) If the data is in the form of a table with counts or proportions, there is an additional **Cat** column for selecting one variable with category names.

With the **Population Sample** Data Type variables can be from different data sets.

The options in the **Confidence Interval** dialog include:

- **Analysis Type.** Selects the type of analysis performed. Options include:
  - **One-Sample Analysis**. Calculates confidence intervals for the proportion of items in a sample that belong to a given category.
  - **Two-Sample Analysis**. Calculates confidence interval for two samples with regard to the proportion of items in a given category
- **Data Type**. Specifies the type of data to be analyzed, either **Population Sample**, **Summary Table with Counts**, or **Summary Table with Proportions**.
- **Options**. Options available change with the Analysis Type and Data Type. Options include:
  - **Confidence Level**. Selects the confidence level (0 to 100%) for the analysis.
  - **First Sample Size and Second Sample Size**. For the **Summary Table with Proportions Data Type**, selects the size of the first sample and second sample (Two-Sample Analysis only).

Confidence Interval Report



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data** – All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

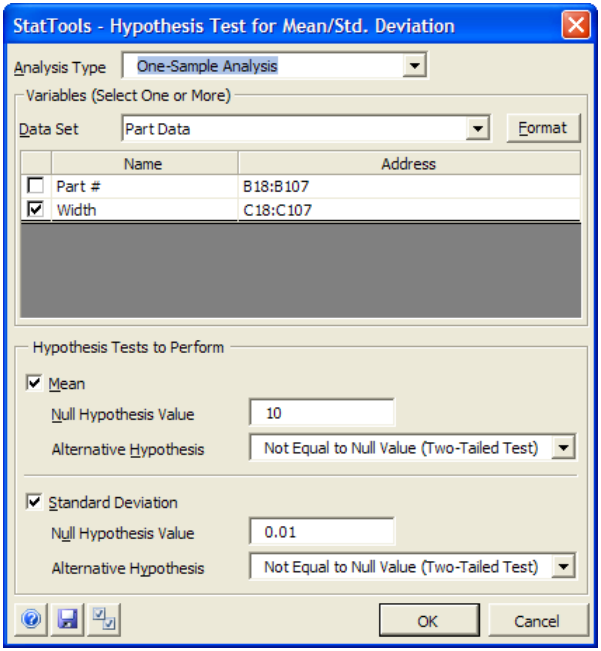
Hypothesis Test - Mean/ Std. Deviation Command

Performs a hypothesis test for mean and standard deviation of variables

The Hypothesis Test for Mean/ Std. Deviation command performs hypothesis tests for the mean and standard deviation of single variables, or the differences between the means for pairs of variables. The hypothesis tests can be performed using a **One-Sample Analysis**, a **Two-Sample Analysis**, or a **Paired Sample Analysis**.

Hypothesis Test for Mean/ Std. Deviation Dialog Box

This analysis is set up using the **Hypothesis Test for Mean/ Std. Deviation** dialog box:

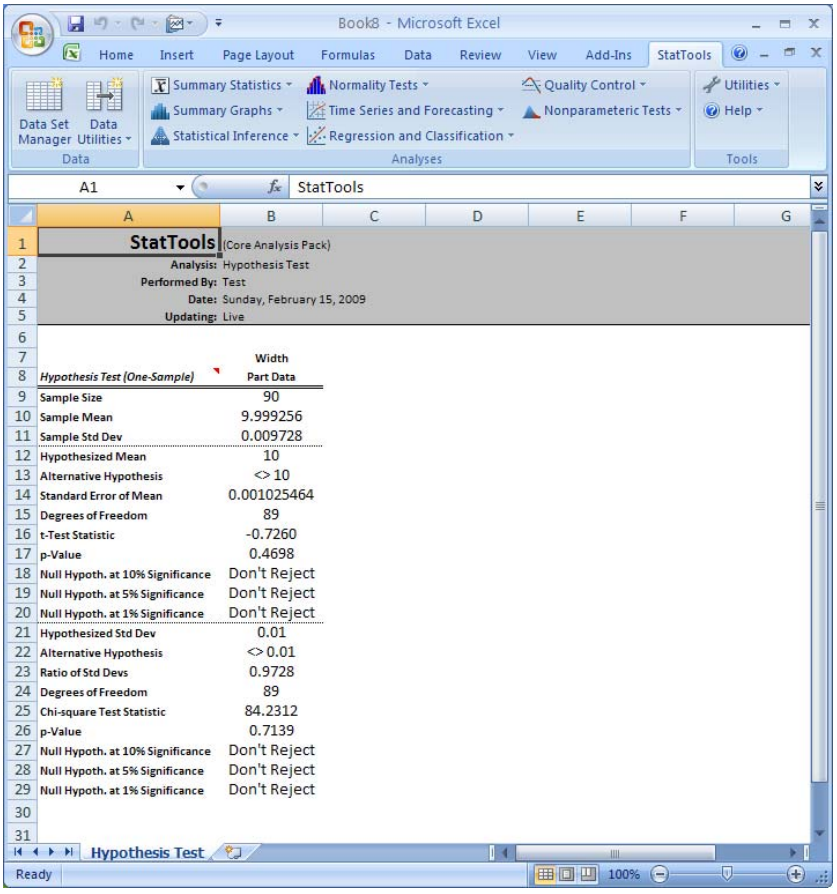


The number of variables selected depends on the Analysis Type used. A One-Sample Analysis requires one or more variables, while a Two-Sample Analysis and a Paired Sample Analysis require two variables. The selected data set can be stacked or unstacked data. Variables can be from different data sets.

The options in the **Hypothesis Test** dialog include:

- **Analysis Type.** Selects the type of analysis performed. Options include:
  - **One-Sample Analysis**. Performs hypothesis tests for a single numerical variable.
  - **Two-Sample Analysis**. Performs hypothesis tests for the difference between means from two independent populations.
  - **Paired Sample Analysis**. This is basically the same as the two-sample analysis, but it is appropriate when the two variables are naturally paired in some way. It essentially runs a one-sample analysis on the differences between pairs.
- **Hypothesis Tests to Perform**. Specifies the hypothesis tests that will be performed on the selected variables. Options change based on the analysis type selected. A **One-Sample Analysis** selects hypothesis tests on the mean and/or standard deviation. A **Two-Sample Analysis** or **Paired Sample Analysis** selects hypothesis tests for the difference between means for two variables. For each hypothesis test selected, options include:
  - **Null Hypothesis Value**, or the value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis Type**, or the alternative to the Null Hypothesis Value that will be evaluated during the analysis. The Alternative Hypothesis Type can be either **"one-tailed"** (that is, greater or less than the null hypothesis) or **"two-tailed"** (that is, not equal to the null hypothesis).

Hypothesis Tests Report



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

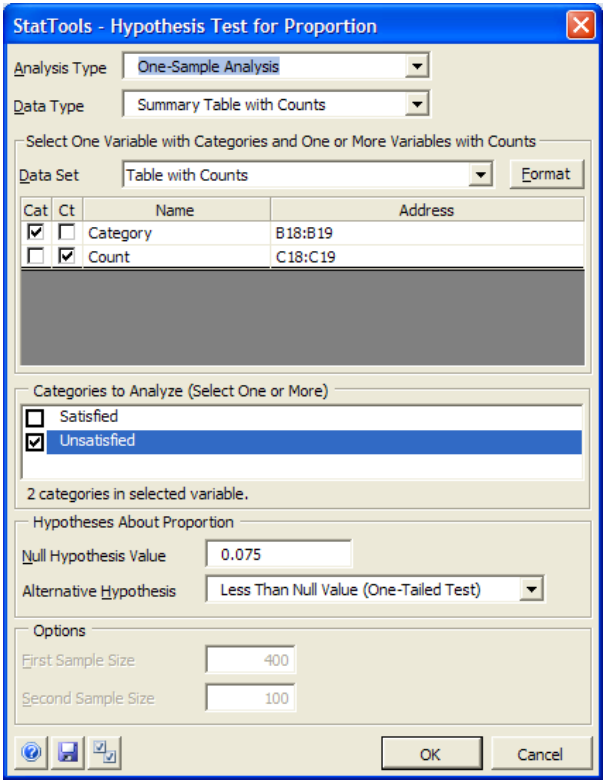
Hypothesis Test - Proportion Command

Performs a hypothesis test for proportions

The Hypothesis Test for Proportion command analyzes the proportion of items in a sample that belong to a given category (**One-Sample Analysis**), or compares two samples with regard to the proportion of items in a given category (**Two-Sample Analysis**). There are three Data Types supported by this procedure: **Population Sample**, **Summary Table with Counts**, and **Summary Table with Proportions**.

Hypothesis Test for Proportion Dialog Box

This analysis is set up using the **Hypothesis Test for Proportion** dialog box:



The number of variables selected depends on the Analysis Type used. A **One-Sample** analysis requires one or more samples for analysis; a **Two-Sample** analysis requires two samples. One selects variables with sample information using column labeled **Ct** (Count), % (Proportion), or column with no label for the **Population Sample** Data Type. (If the Population Sample data is stacked, the samples are

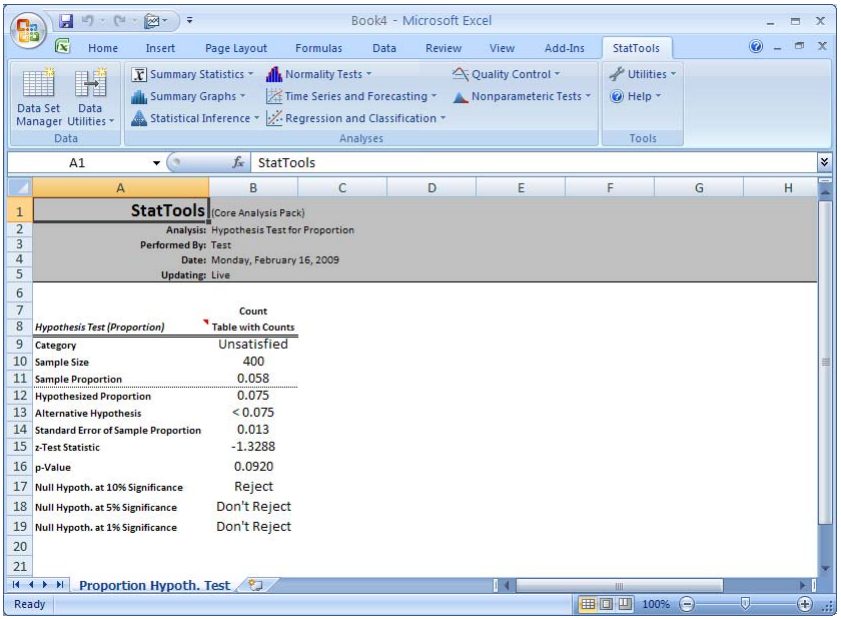
selected in columns labeled **C1** and **C2**, where C1 contains the stacked categories.) If the data is in the form of a table with counts or proportions, there is an additional **Cat** column for selecting one variable with category names.

With the **Population Sample** Data Type variables can be from different data sets.

The options in the **Hypothesis Test for Proportion** dialog include:

- **Analysis Type.** Selects the type of analysis performed. Options include:
  - **One-Sample Analysis**. Performs hypothesis test for the proportion of items in a sample that belong to a given category.
  - **Two-Sample Analysis**. Performs hypothesis test for two samples with regard to the proportion of items in a given category
- **Data Type**. Specifies the type of data to be analyzed, either **Population Sample**, **Summary Table with Counts**, or **Summary Table with Proportions**.
- **Hypothesis Tests to Perform**. Specifies the hypothesis tests that will be performed on the selected proportion. Options change based on the analysis type selected. Options include:
  - **Null Hypothesis Value**, or the value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis Type**, or the alternative to the Null Hypothesis Value that will be evaluated during the analysis. The Alternative Hypothesis Type can be either **"one-tailed"** (that is, greater or less than the null hypothesis) or **"two-tailed"** (that is, not equal to the null hypothesis).
- **First Sample Size and Second Sample Size**. For the **Summary Table with Proportions Data Type**, selects the size of the first sample and second sample (Two-Sample Analysis only).

Hypothesis Tests Report



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

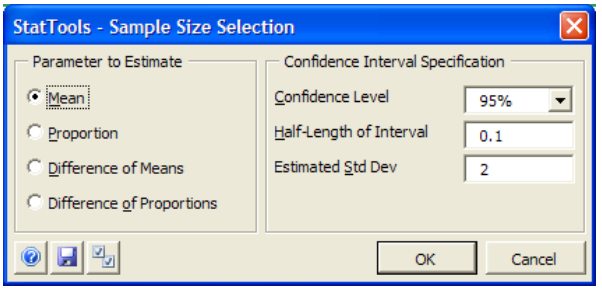
Sample Size Selection Command

Determines the sample size required to calculate confidence intervals

The Sample Size Selection command determines the sample size (or sample sizes) required to obtain a confidence interval with a prescribed half-length. It does this for confidence intervals for a mean, a proportion, the difference between two means, and the difference between two proportions. No data sets or variables are required, as the sample size is information you typically need *before* you collect data. You need to specify the confidence level, the desired half-length, and any other parameters necessary for determining sample size.

Sample Size Selection Dialog Box

This analysis is set up using the **Sample Size Selection** dialog box:

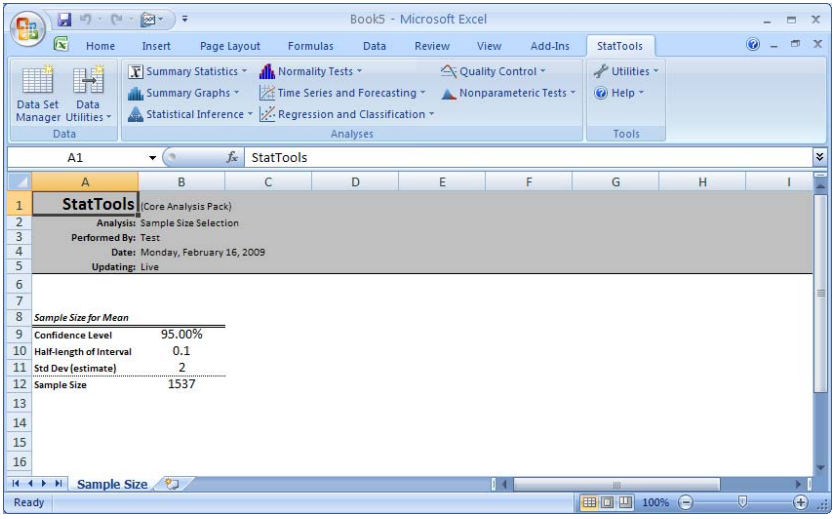


The options in the **Sample Size Selection** dialog include:

- **Parameter to Estimate -** Selects the type of parameter that will be estimated from the sample (whose size you are determining). Options include **Mean**, **Proportion** (values between 0 and 1), **Differences of Means**, and **Differences of Proportions.**
- **Confidence Interval Specification –** These options vary by the selected Parameter to Estimate, as shown:
  - When Parameter to Estimate is **Mean** and **Difference of Means**. First, enter the desired **Confidence Level** (usually between 90% and 100%), the **Half Length of the Interval** (the "plus or minus" component of the interval) and the **Estimated Standard Deviation** of the population. Note: Confidence Level and Interval Length are related, as a higher confidence level requires a longer confidence interval length.

- When Parameter to Estimate is **Proportion** and **Difference of Proportions.** First, enter the desired **Confidence Level** (usually between 90% and 100%), the **Half Length of the Interval** (the "plus or minus" component of the interval) and the **Estimated Proportion** (a value between 0 and 1). If **Difference of Proportions** is being estimated, provide an **Estimated Proportion** for each population.

Sample Size Estimate Report



Missing Data and Link to Data

- **Missing Data -** Not relevant.
- **Link to Data –** Not relevant.

One-Way ANOVA Command

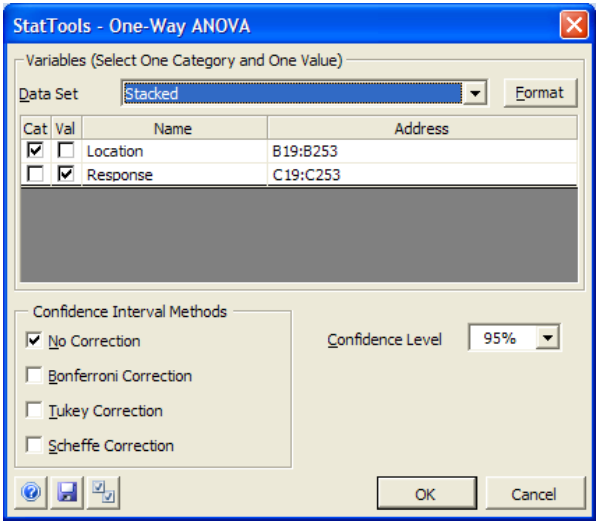
Performs a One-Way ANOVA on variables

The One-Way ANOVA command is a generalization of the two-sample procedure for comparing means between two populations. With One-Way ANOVA the means from *at least two* (usually more than two) populations are compared. This is done with an ANOVA (analysis of variance) table. This table actually compares two sources of variation: the variation *within* each population against the variation *among* sample means from the different populations. If the latter variation is large relative to the former, as measured by an F test, then there is evidence of differences between population means.

The key value in the ANOVA table is the p-value. A small p-value is evidence of different population means. Besides the ANOVA table, it is informative to look at confidence intervals for all differences between pairs of means. Confidence intervals that do *not* include 0 are evidence of means that are *not* equal. StatTools provides the option of several types of confidence intervals, each based on a slightly different method.

ANOVA Dialog Box

This analysis is set up using the **One-Way ANOVA** dialog box:

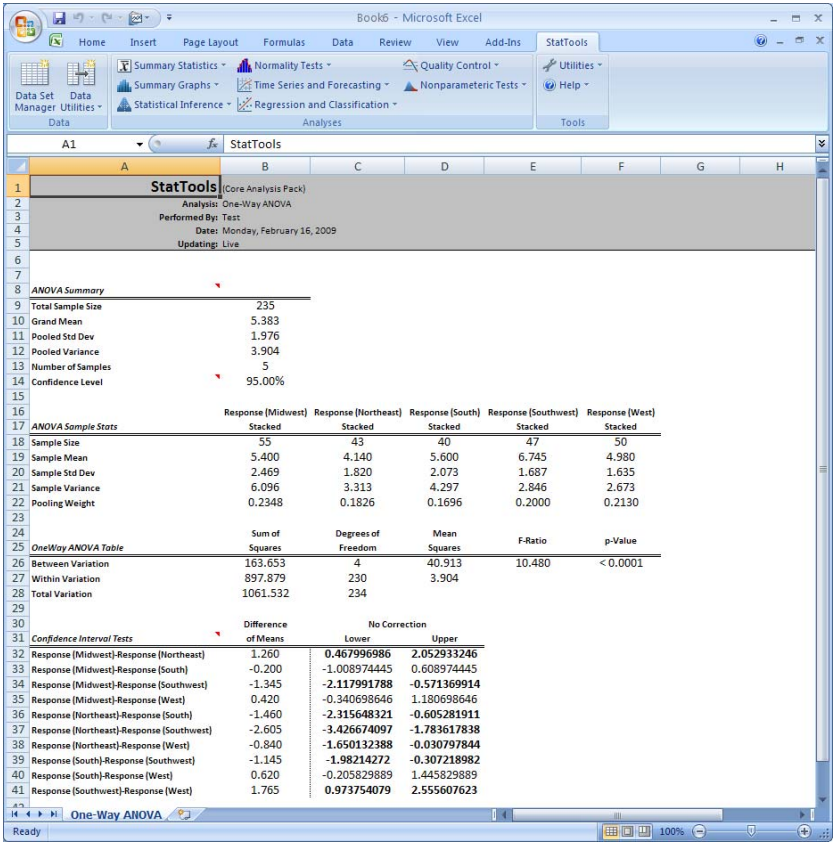


Two or more variables need to be selected for analysis. The selected data sets can be stacked or unstacked data. Variables can be from different data sets.

The options in the One-Way ANOVA dialog include:

- **Confidence Interval Methods -** Select one or more from the following methods for correcting confidence intervals for individual variables. **No Correction**, or no correction performed, and **Bonferroni**, **Tukey**, and **Scheffe** correction methods.
- **Confidence Level -** This is a "simultaneous" confidence level for the results for all variables. That is, it is the confidence you want to have that *all* of the confidence intervals contain their respective population mean differences. For technical reasons, the *actual* overall confidence level will typically be less than this specified level for the "no correction" confidence intervals. This is the reason for the "correction" methods above. They correct (expand) the lengths of the confidence intervals so that the overall confidence level is the one specified.

One-Way ANOVA Report



Book6 - Microsoft Excel

Home | Insert | Page Layout | Formulas | Data | Review | View | Add-Ins | StatTools

Summary Statistics | Normality Tests | Quality Control | Utilities
Summary Graphs | Time Series and Forecasting | Nonparametric Tests | Help
Statistical Inference | Regression and Classification
Data Set Manager | Data Utilities
Data | Analyses | Tools

A1 | StatTools

StatTools (Core Analysis Pack)
Analysis: One-Way ANOVA
Performed By: Test
Date: Monday, February 16, 2009
Updating: Live

ANOVA Summary

Total Sample Size	235
Grand Mean	5.383
Pooled Std Dev	1.976
Pooled Variance	3.904
Number of Samples	5
Confidence Level	95.00%

ANOVA Sample Stats	Response (Midwest) Stacked	Response (Northeast) Stacked	Response (South) Stacked	Response (Southwest) Stacked	Response (West) Stacked
Sample Size	55	43	40	47	50
Sample Mean	5.400	4.140	5.600	6.745	4.980
Sample Std Dev	2.469	1.820	2.073	1.687	1.635
Sample Variance	6.096	3.313	4.297	2.846	2.673
Pooling Weight	0.2348	0.1826	0.1696	0.2000	0.2130

OneWay ANOVA Table	Sum of Squares	Degrees of Freedom	Mean Squares	F-Ratio	p-Value
Between Variation	163.653	4	40.913	10.480	< 0.0001
Within Variation	897.879	230	3.904		
Total Variation	1061.532	234			

Confidence Interval Tests	Difference of Means	No Correction Lower	No Correction Upper
Response (Midwest)-Response (Northeast)	1.260	**0.467996986**	**2.052933246**
Response (Midwest)-Response (South)	-0.200	-1.008974445	0.608974445
Response (Midwest)-Response (Southwest)	-1.345	**-2.117991788**	**-0.571369914**
Response (Midwest)-Response (West)	0.420	-0.340698646	1.180698646
Response (Northeast)-Response (South)	-1.460	**-2.315648321**	**-0.605281911**
Response (Northeast)-Response (Southwest)	-2.605	**-3.426674097**	**-1.783617838**
Response (Northeast)-Response (West)	-0.840	**-1.650132388**	**-0.030797844**
Response (South)-Response (Southwest)	-1.145	**-1.98214272**	**-0.307218982**
Response (South)-Response (West)	0.620	-0.205829889	1.445829889
Response (Southwest)-Response (West)	1.765	**0.973754079**	**2.555607623**

One-Way ANOVA

Ready | 100%

In the One-Way ANOVA report here, summary statistics for each population (in this case, each plant) appear at the top. After the summary statistics, a table of sample statistics for each variable is included. The ANOVA table appears next. In this example, the very small p-value indicates without a doubt that the mean scores from the five plants are not all equal. To see which means are different from which others, we look at the confidence intervals at the bottom of the report. Those pairs with values in bold have significantly different means.

Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

Two-Way ANOVA Command

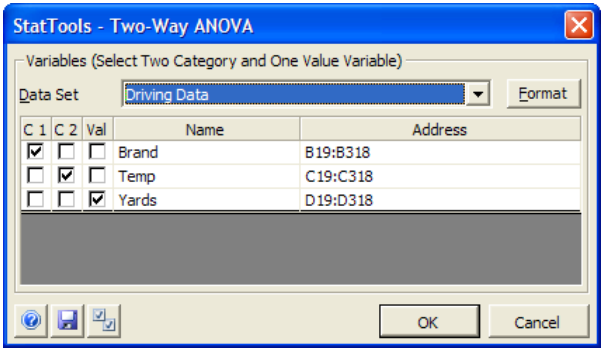
Performs a Two-Way ANOVA on variables

The Two-Way ANOVA command performs a two-way analysis of variance. This is usually done in the context of an experimental design where there are two "factors" that are each set at several "treatment levels." For example, in a study of golf ball performance, the two factors might *Brand* and *Outside Temperature*. Then the treatment levels for *Brand* would be "*A*" through "*E*", and the treatment levels for Temperature would be the "*Cool*", "*Mild*" and "*Warm*". The value variable would be *Yards Driven*, and observations on this variable would be collected for a number of balls of each *Brand /Temperature* combination. The purpose of the study is to see whether there are significant mean differences among the various treatment level combinations.

The data for two-way ANOVA *must* be in stacked form. That is, there must be two "category" variables (corresponding to *Brand* and *Outside Temperature* in the above example), and there must be a "value" variable (corresponding to *Yards Driven* above). Also, the data set must be "balanced," meaning that there should be an equal number of observations in each treatment level combination. It is certainly possible to analyze an unbalanced design, but this design is best analyzed with regression (with dummy variables).

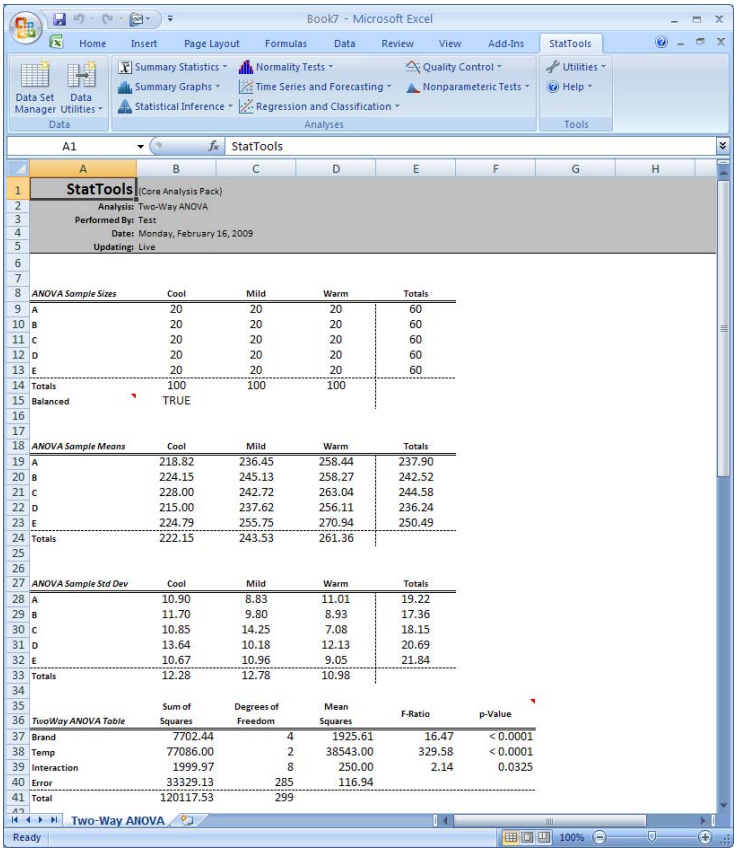
ANOVA Dialog Box

This analysis is set up using the **Two-Way ANOVA** dialog box:



Two category variables (**C1** and **C2**) and one value variable (**Val**) need to be selected for analysis. The selected data set needs to be stacked data.

Two-Way ANOVA Report



The top three items are summary measures (sample sizes, sample means, and sample standard deviations) for the various treatment level combinations. The bottom part of the output shows the ANOVA table. There are three important p-values in this table: two for "main effects" and one for "interactions." The main effects indicate whether there are significant mean differences across levels of either factor, averaged over the levels of the other factor. For example, the Temperature main effect indicates whether the values in cells B17 to D17 are significantly different. (They are, as indicated by the very small p-value for *Temperature* in the ANOVA table.)

Missing Data and Link to Data

- **Missing Data -** Because of the requirement for a balanced design, there should not be any missing data.
- **Link to Data –** All of the ANOVA formulas are linked to the data. If the data change, the results change automatically. The balance of the experiment is verified when the analysis is run. Changes in the data may affect the balance and cause invalid results.

Chi-square Independence Test Command

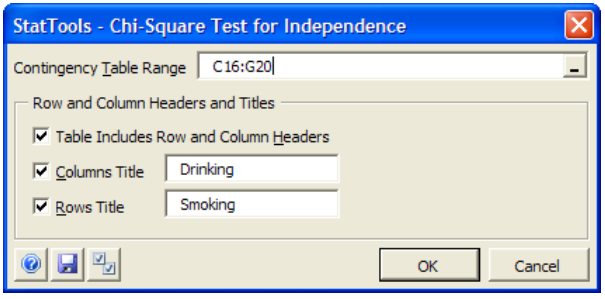
Tests for independence between the row and column attributes of a contingency table

The Chi-square Independence Test command tests for independence between the row and column attributes of a contingency table. For example, if the contingency table lists counts of people in different drinking and smoking categories, the procedure tests whether smoking habits are independent of drinking habits. The contingency table (also called cross-tabs) could be an Excel pivot table.

This procedure is somewhat different from most StatTools procedures. For this analysis all that is required is a rectangular contingency table. Each cell in this table should be a count of observations in a particular row/column combination (nondrinkers and heavy smokers, for example). The table can have row and column labels (headings) and/or row and column totals, but these are not necessary and are only used for clarity in StatTools reports.

Chi-square Independence Test Dialog Box

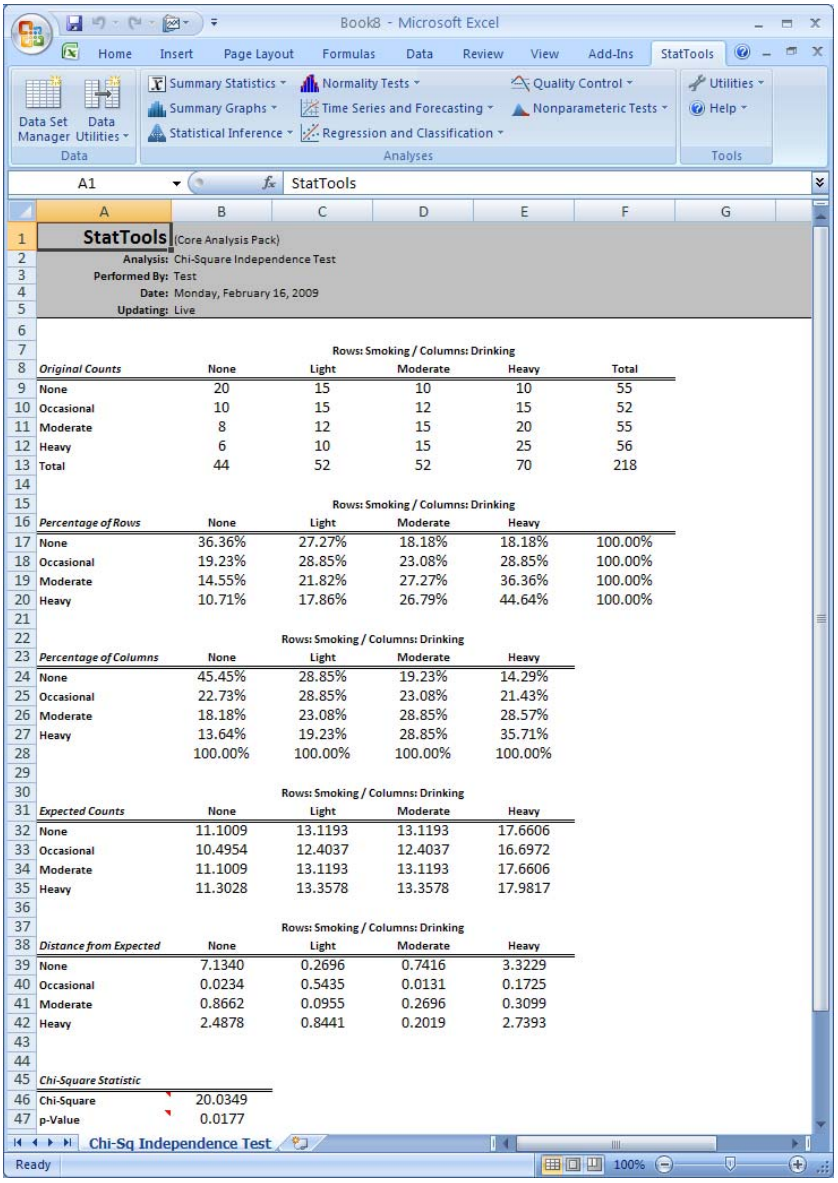
This analysis is set up using the **Chi-square Independence Test** dialog box:



The options in the Chi-square Independence Test dialog include:

- **Row and Column Headers and Titles -** Select one or more from the following: **Table Includes Row and Column Headers** or headers in the leftmost column and topmost row of the table, **Columns Title** or the title you want to use to represent the columns in the table, **Rows Title** or the title you want to use to represent the rows in the table.

Chi-square Independence Test Report



The report above shows the basic result of the test, a p-value. If this p-value is small (as here) we can conclude that the row and column attributes are *not* independent. We can study the numbers on this sheet to understand better how smoking and drinking are related.

Missing Data and Link to Data

- **Missing Data -** There should not be any missing data in the cells of the contingency table.
- **Link to Data –** The formulas in the Chi-square Independence Test Report are linked to the data. So if the counts in the original contingency table change, the outputs on this sheet change also.

Normality Tests Menu

Because so many statistical procedures assume that a set of data is normally distributed, it is useful to have methods for checking this assumption. StatTools provides three commonly used checks, as described in this section.

Chi-Square Normality Test Command

Tests if observed data for a variable is normally distributed

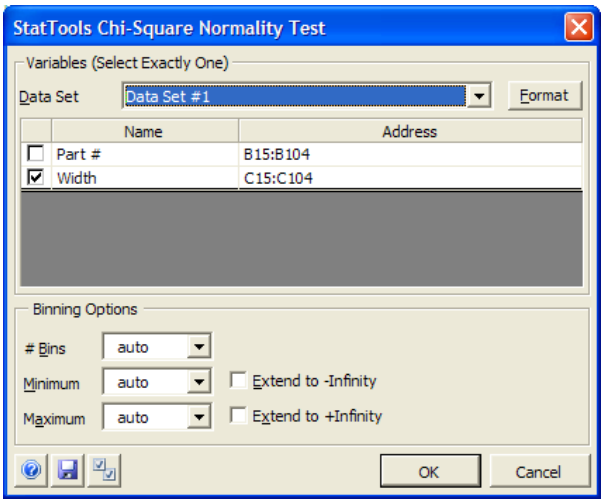
The Chi-Square Normality Test procedure uses a chi-square goodness-of-fit test to test whether the observed data in a specified variable could have come from a normal distribution. To do so, it creates a histogram of this variable, using the categories you specify, and it superimposes a histogram for a normal distribution on the histogram from the data. If the two histograms have essentially the same shape, we cannot reject the null hypothesis of a normal fit.

The formal test is performed by comparing the observed counts in the various categories with the expected counts that are based on a normality assumption. Actually, the procedure allows you to test several variables (separately) for normality. A histogram is created for each variable you select, and the chi-square test is run on each of them.

The only requirement for the Chi-Sq Normality Test is that there must be at least one numerical variable. Beyond this, most analysts suggest that there should be at least 100 observations -- the more, the better.

Chi-Sq Normality Test Dialog Box

This analysis is set up using the **Chi-Sq Normality Test** dialog box:

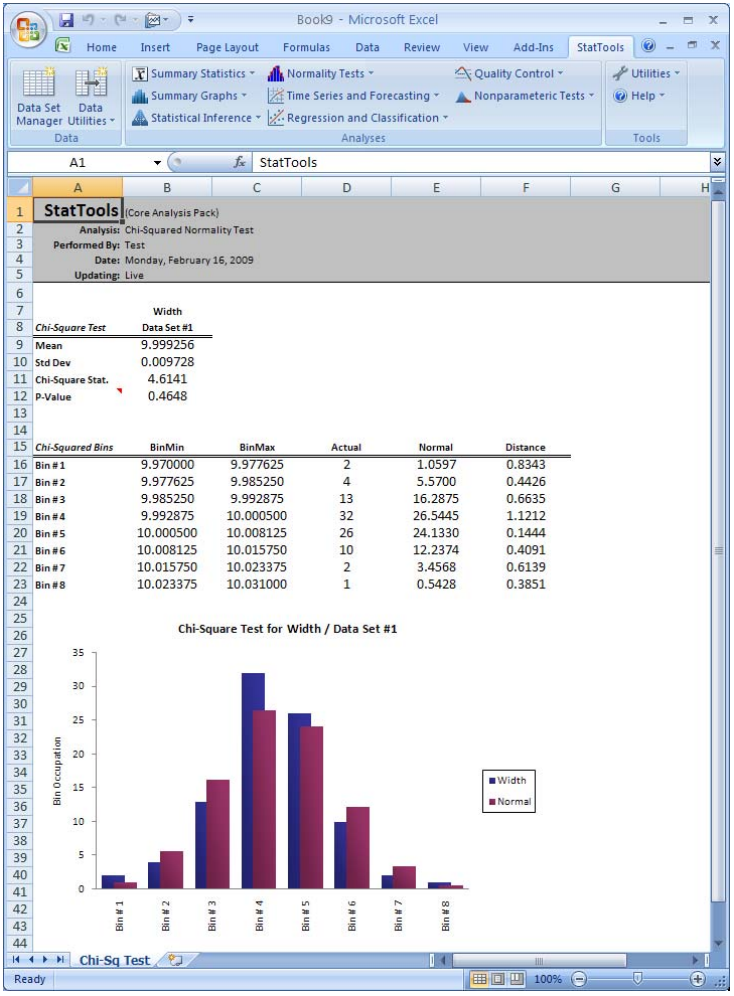


One variable can be selected for testing. The selected data set must be unstacked data.

The options in the **Chi-Sq Normality Test** dialog include:

- **# Bins -** Specifies a fixed number of bins or, alternatively, specifies that the number of bins will be automatically calculated for you.
- **Minimum and Maximum - auto** specifies that the minimum and maximum of your data set will be used to calculate the minimum and maximum of equal interval bins. First and last bins, however, may be added using the **Extend to -Infinity** and **Extend to +Infinity** options. If Auto is not selected, you can enter a specific **Minimum** and **Maximum** value where your bins will start and end. This allows you to enter a specific range where binning will be performed without regard to the minimum and maximum values in your data set.
- **Extend to –Infinity** indicates that the first bin used will stretch from the specified minimum to -Infinity. All other bins will be of equal length. In certain circumstances, this improves testing for data sets with unknown lower bounds.
- **Extend to +Infinity** indicates that the last bin used will stretch from the specified maximum to +Infinity. All other bins will be of equal length. In certain circumstances, this improves testing for data sets with unknown upper bounds.

Chi-Sq Normality Test Report



The results of the test are shown in the report above. The p-value of 0.4776 is good evidence that the amounts are normally distributed. More evidence to this effect appears in the histograms in Figure 4 and the frequency data. However, be aware of two things. First, if there are too few observations (well less than 100, say), then the chi-square test is not good at distinguishing normality from nonnormality. The effect is that the p-value is usually not small enough to reject the normality hypothesis. In essence, almost *everything* tends to look normal with small data sets. On the other hand, if the data set is really large (several hundred observations, say), then the p-value will usually be small, indicating nonnormality. The reason is that with large data sets, every little "bump" in the curve is likely to create a small p-value. In this case, the real test is a *practical* one: Do the histograms really differ that much for all practical purposes?

Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** The histogram and all formulas for the test are linked to the original data. So if the data change, the histogram and the test results change automatically.

Lilliefors Test Command

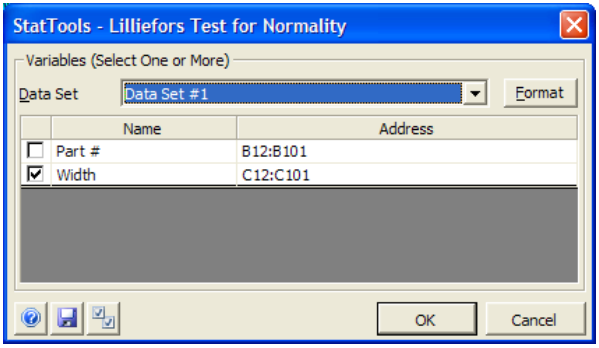
Tests if observed data for a variable is normally distributed

The Lilliefors Test procedure provides a more powerful test for normality than the more familiar chi-square goodness-of-fit test. (More powerful means that it is more likely to detect non-normality if it exists.) It is based on a comparison of the "empirical cdf" and a normal cdf, where "cdf" stands for cumulative distribution function, showing the probability of being less than or equal to any particular value.

The empirical cdf is based on the data. For example, if there are 100 observations and the 13th smallest is 137, then the empirical cdf, evaluated at 137, is 0.13. The Lilliefors test finds the maximum vertical distance between the empirical cdf and the normal cdf, and it compares this maximum to tabulated values (that are based on sample size). If the observed maximum vertical distance is sufficiently large, then we have evidence that the data do *not* come from a normal distribution.

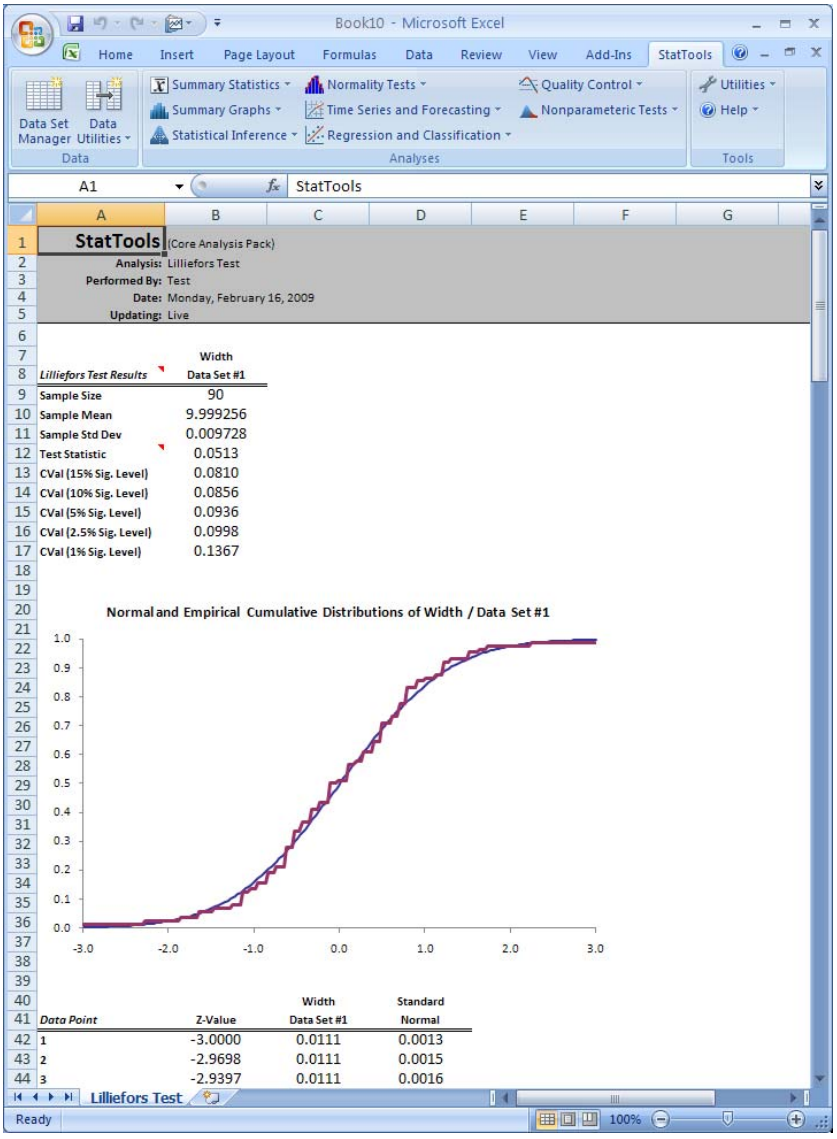
Lilliefors Test Dialog Box

This analysis is set up using the **Lilliefors Test** dialog box:



One or more variables can be selected for testing. The selected data set needs to be unstacked data. Variables can be from different data sets.

Lilliefors Test Report



The results of the test are shown in the report above. There is no p-value (as in most hypothesis tests), but we see from the statement that the maximum vertical distance is sufficiently large to cast doubt on the normality assumption. More evidence to this effect appears in the cdf's in the included chart. Actually, the fit between the two curves appears to be "pretty good," and it might be good enough for all practical purposes. That is, we might conclude that these data are "close enough" to being normally distributed for our purposes.

Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** The CDFs and all formulas for the test are linked to the original data. So if the data change, the graph and the test results change automatically.

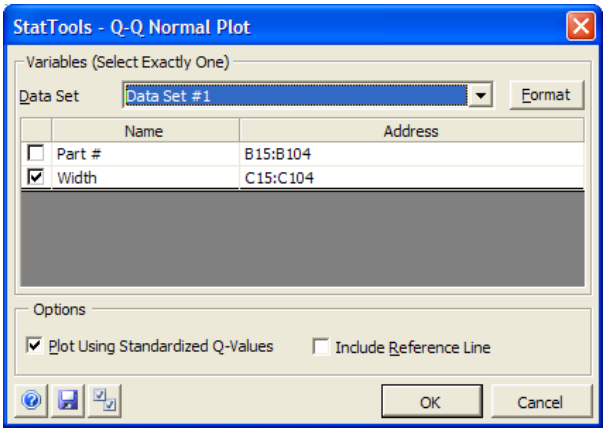
Q-Q Normal Plot Command

Tests if observed data for a variable is normally distributed

The Q-Q Normal Plot command creates a quantile-quantile (Q-Q) plot for a single variable. It provides an informal test of normality. Although the details are somewhat complex, the objective is fairly simple: to compare the quantiles (or percentiles) for the data to the quantiles from a normal distribution. If the data are essentially normal, then the points on the Q-Q plot should be close to a 45-degree line. However, obvious curvature in the plot is an indication of some form of non-normality (skewness, for example).

Q-Q Normal Plot Dialog Box

This analysis is set up using the **Q-Q Normal Plot** dialog box:

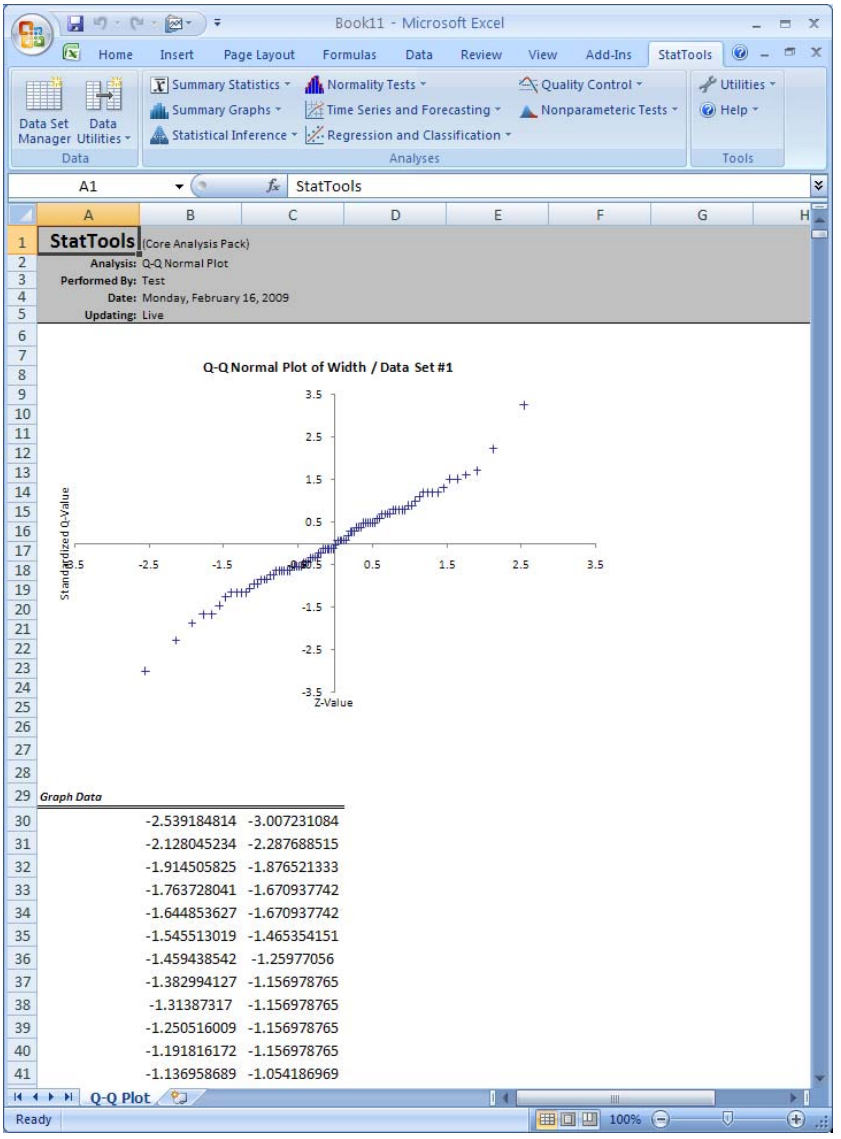


One variable can be selected for plotting. The selected data set needs to be unstacked data.

This options in the **Q-Q Normal Plot** dialog box include:

- **Plot Using Standardized Q-Values** - Specifies to use a standardized Q-Value, instead of Q-Q data, on the Y-axis of the graph. This makes comparisons of the Y-axis values between Q-Q Normal plots possible.

Q-Q Normal Plot Report



As stated earlier, this is an *informal* test of normality. It is difficult to say "how close" to a 45-degree line the plot should be to accept a normality assumption. Typically, we look for obvious curvature in the plot, and none is apparent in the plot here.

Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** The plots and all formulas are linked to the original data. So if the data change, the plot changes automatically.

Time Series and Forecasting Menu

The procedures on the Time Series and Forecasting menu deal with the analysis of data collected over time, with applications for forecasting and quality control. The forecasting methods provided include the moving averages method, simple exponential smoothing, Holt's exponential smoothing method for capturing trend, and Winters' exponential smoothing method for capturing seasonality.

Time Series Graph Command

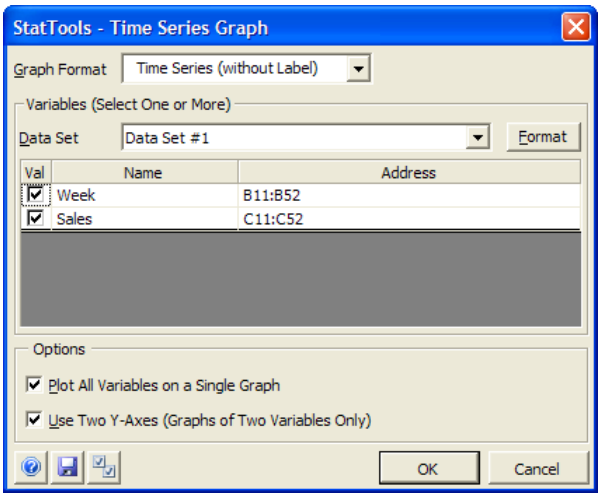
Creates a time series graph for variables

The Time Series Graph command plots one or more time series variables all on the same plot. If two variables are selected, you have the option of using the same or different scales on the Y-axis for the two variables. The latter option is useful when the ranges of values for the two variables are considerably different. However, if more than two variables are plotted, they must all share the same vertical scale.

There must be at least one numerical variable in the data set. There can also be a "date" variable, but if it is to be used to label the horizontal axis of the chart, it must be selected as the "label" variable.

Time Series Graph Dialog Box

This graph type is set up using the **Time Series Graph** dialog box:

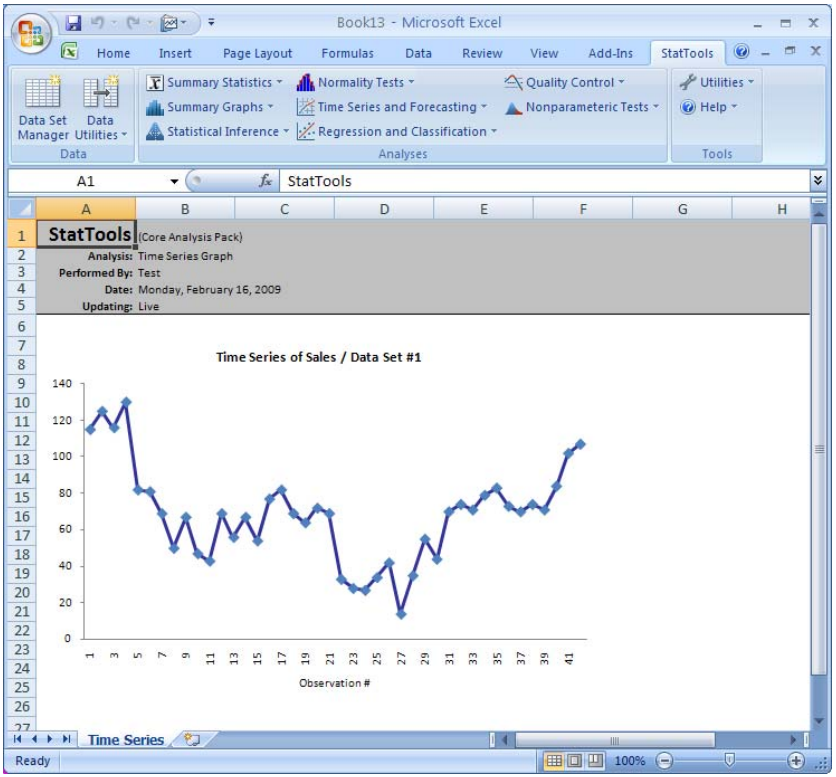


One or more variables can be selected for graphing. The selected data set must be unstacked data. Variables can be from different data sets. The Label variable (Lbl checkbox) appears on the X-axis.

The options in the **Time Series Graph** dialog include:

- **Plot All Variables on a Single Graph**. Selects to plot all variables in one graph.
- **Use Two Y-Axes**. Select to display a separate Y-axis for each variable in a two variable graph. Units and values for each variable can then be displayed on the graph.

Time Series Graph of Two Variables



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data** – Graphs are linked to data so if the data changes the graph automatically updates.

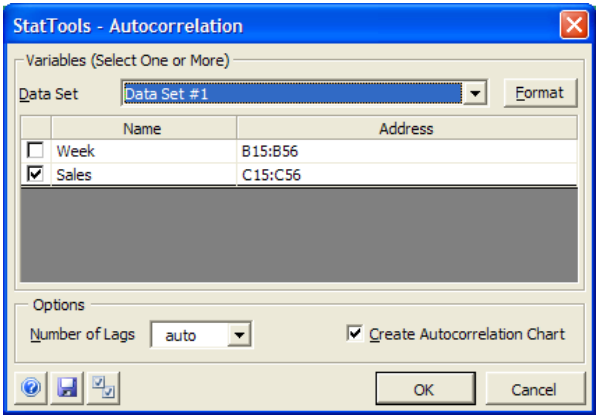
Autocorrelation Command

Calculates the autocorrelations for variables

The Autocorrelation Command calculates the autocorrelations for any selected numeric variable. Typically, this variable will be a time series variable, although StatTools will do the calculations for any variable. You can select the number of autocorrelations you want (i.e., the number of lags). You can also request a chart of the autocorrelations, called a correlogram. The output shows which of the autocorrelations, if any, are significantly different from 0.

Autocorrelation Dialog Box

This graph type is set up using the **Autocorrelation** dialog box:

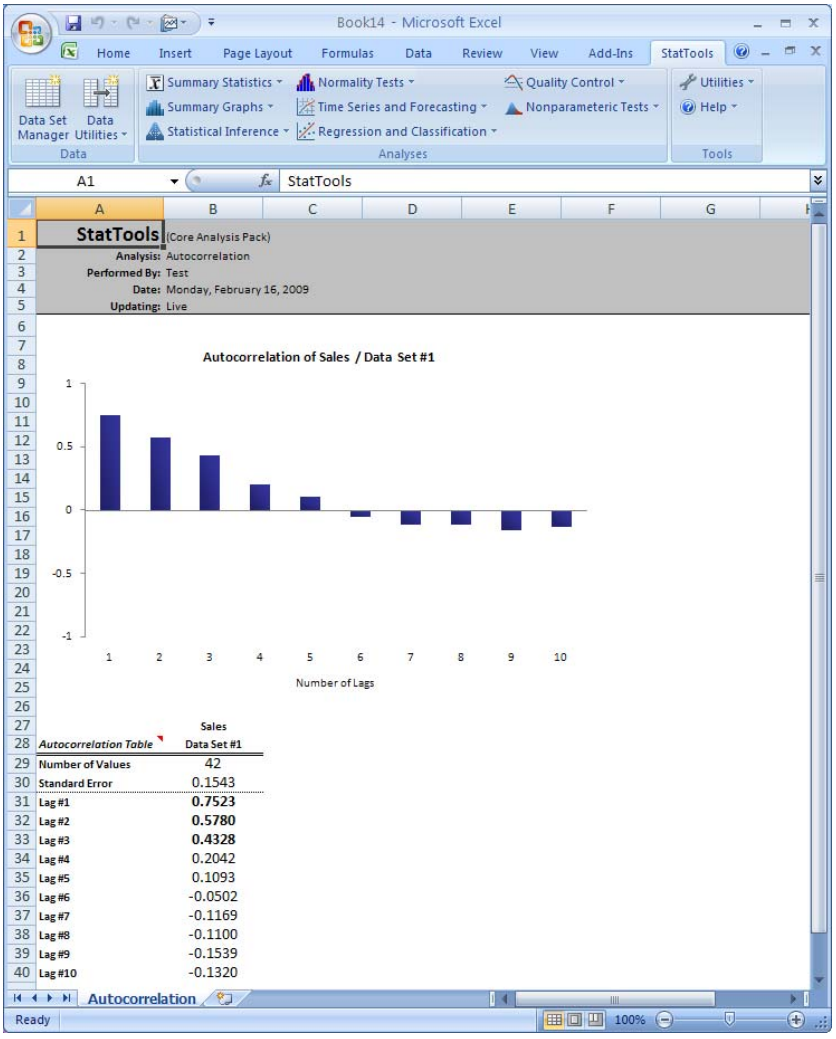


One or more variables can be selected for analysis. The selected data set must be unstacked data.

The options in the **Autocorrelation** dialog include:

- **Number of Lags**. The number of time periods to lag when calculating autocorrelations. If **Auto** is selected, StatTools determines the appropriate number of lags to test. If you enter a specific # of lags, the maximum number of lags you can request is 25% of the number of observations in the series. For example, if you have 80 monthly values, you can request up to 20 lags.
- **Create Autocorrelation Chart.** Creates a bar chart with the height of each bar equal to the corresponding autocorrelation.

Autocorrelation Report



The Autocorrelation report appears above. For each lag, the corresponding autocorrelation appears, along with an *approximate* standard error.

Missing Data and Link to Data

- **Missing Data -** This procedure allows missing data at the beginning of the time series, but none in the middle or at the end of the series.
- **Link to Data** – StatTools ties the output to the data. Therefore, if the data change, the autocorrelations (and the correlogram) change automatically.

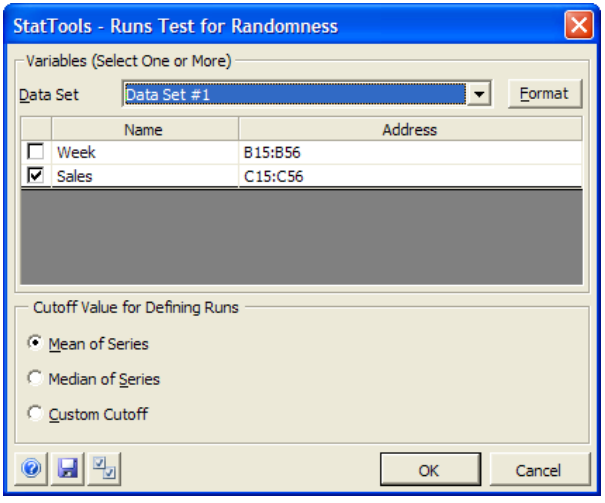
Runs Test for Randomness Command

Performs a runs test to check whether a variable is random

The Runs Test for Randomness command allows you to check the "randomness" of a sequence of values in a variable, usually a time series variable. It indicates how many "runs" there are in the sequence, where a run is a consecutive number of values on one side or the other of some cutoff point (such as the mean or median of the sequence). For a random sequence, we would expect neither too few runs nor too many. The runs test counts the number of runs and then reports a p-value for the test. If this p-value is small, we can conclude that sequence is probably not random, i.e., there are either too many or too few runs.

Runs Test for Randomness Dialog Box

This analysis is set up using the **Runs Test for Randomness** dialog box:

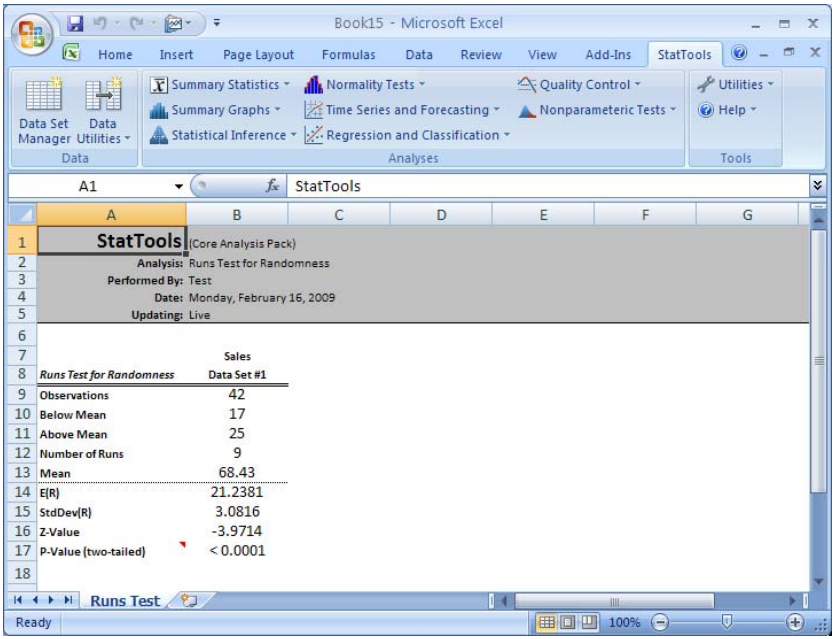


One or more variables can be selected for analysis. The selected data set must be unstacked data. Variables can be from different data sets.

The options in the **Runs Test for Randomness** dialog include:

- **Cutoff Value for Defining Runs -** A runs test is always based on runs above or below some Cutoff Value. This can be the **Mean of Series**, the **Median of Series**, or any other **Custom Cutoff Value** you want to enter.

Runs Test for Randomness Report



The report above shows the number of runs and the expected number of runs under randomness E(R). As 9 is significantly below 21.2381, this series is not entirely random; i.e., sales do not "zigzag" as much as a random series.

Missing Data and Link to Data

- **Missing Data -** This procedure allows missing data at the beginning and at the end of the time series, but none in the middle of the series.
- **Link to Data –** StatTools ties the output to the data. Therefore, if the data change, the reports change automatically.

Forecasting Command

Generates forecasts for time series variables

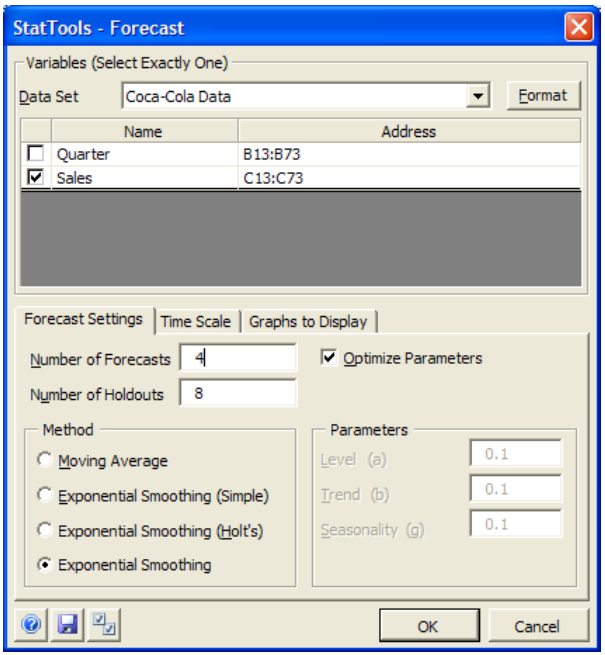
The Forecasting command provides you with a number of methods for forecasting a time series variable. These methods include the moving averages method, simple exponential smoothing, Holt's exponential smoothing method for capturing trend, and Winters' exponential smoothing method for capturing seasonality. The Forecasting command also allows you to deseasonalize the data first, using the ratio-to-moving-averages method and a multiplicative seasonality model. Then you can use any of the forecasting methods (other than Winters' method) to forecast the deseasonalized data, and finally "reseasonalize" the forecasts to get back to original units.

The forecast reports include a set of columns to show the various calculations (for example, the smoothed levels and trends for Holt's method, the seasonal factors from the ratio-to-moving-averages method, and so on), the forecasts, and the forecast errors. Summary measures are also included (MAE, RMSE, and MAPE) for tracking the fit of the model to the observed data. (When you use exponential smoothing methods, you have the option of using optimization to find the smoothing constant(s) that minimize RMSE.)

Finally, several time series plots are available, including a plot of the original series, a plot of the series with forecasts superimposed, and a plot of the forecast errors. In case of deseasonalizing, these plots are available for the original series and the deseasonalized series.

Forecasting Dialog Box

Forecasts are set up using the **Forecasting** dialog box:



One variable can be selected for analysis. The selected data set must be unstacked data.

Forecasting Options - Forecasting Dialog Box

The **Forecasting** options in this dialog specify the forecasting method used and the settings for the selected method. Options include:

- **Number of Forecasts**. Specifies the number of *future* periods to provide forecasts for.
- **Number of Holdouts**. Specifies the number of observations to "hold out", or not use in, the forecasting model. You can choose to use all of the observations for estimating the forecasting model (0 Holdouts), or you can hold out a few for validation. Then the model is estimated from the observations not held out, and it is used to forecast the held-out observations.
- **Optimize Parameters (exponential smoothing methods only)**. Finds the smoothing constant that minimizes the RMSE (for the non-holdout period). Optimization requires Parameters shown in the dialog to be between 0 and 1. If you edit parameter values directly in a Forecasting report, make sure to enter values in this range.
- **Deseasonalize**. Selects to deseasonalize data before forecasting. For seasonal data, that is, data suspected of having a seasonal pattern, there are two options. You can use Winters' method,

which deals with seasonality directly, or you can select this option to deseasonalize the data first, using the ratio-to-moving-averages method for deseasonalizing. Then *any* method can be used to forecast the deseasonalized series.

- **Method –** selects the forecasting method to be used; either **Moving Average** or **Simple**, **Holt's**, or **Winters'** exponential smoothing methods.
- **Parameters –** specifies the parameters to be used for the selected forecasting method:
  - **Span parameter** (Moving Average method only), or the number of consecutive observations used in each moving average.
  - **Level parameter** (all Exponential Smoothing methods), a smoothing parameter that can take any value between 0 and 1 (the default value is 0.1).
  - **Trend parameter** (Holt's and Winter's Exponential Smoothing methods), a second smoothing parameter that can take any value between 0 and 1 (the default value is 0.1).
  - **Seasonality parameter** (Winter's Exponential Smoothing method only), a third smoothing parameter that can take any value between 0 and 1 (the default value is 0.1).

Note: If Optimize Parameters is selected, Level, Trend and Seasonality parameters cannot be set, as these are the parameters whose values are being optimized.

Time Scale Options - Forecasting Dialog Box

The Time Scale options specify the timing and time scale labeling for the analyzed variable. Options include:

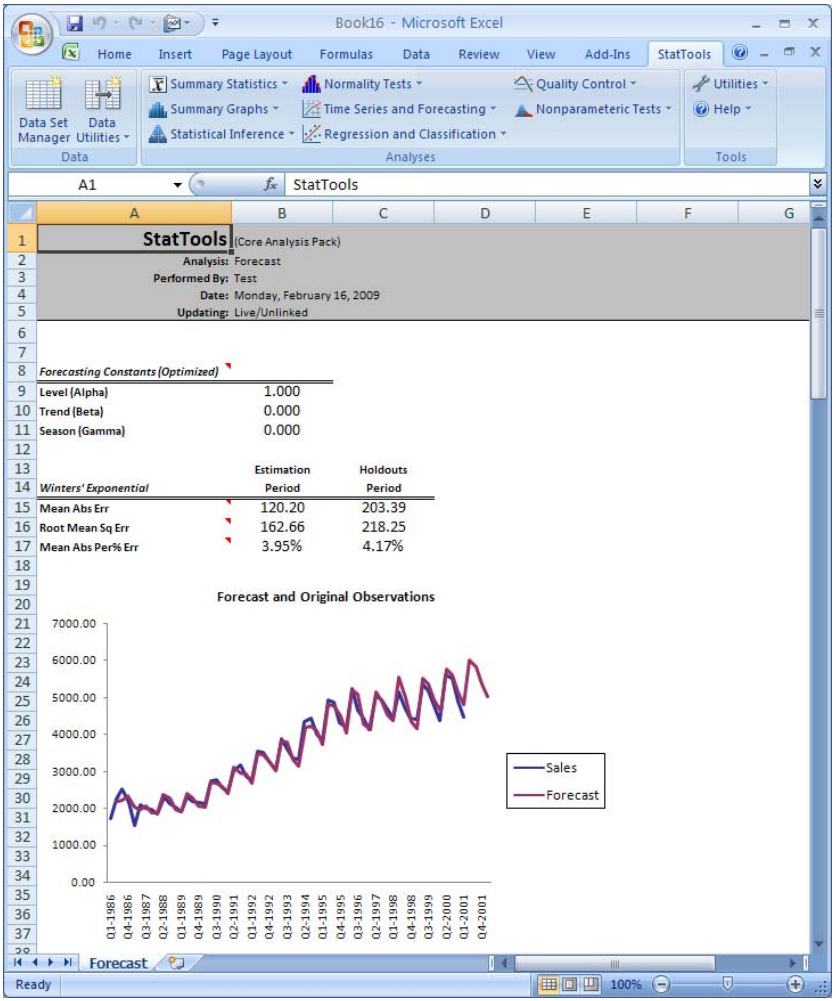
- **Seasonal Period**. Specifies the type of time series data; either Annual, Quarterly, Monthly, Weekly, Daily or None. This is used for seasonalizing data and for labeling.
- **Label Style.** Specifies how the time scale will be labeled on any generated graphs.
- **Starting Label**. Specifies the entry for the first time scale label on the graph.

Graph Options - Forecasting Dialog Box

The Graph options specify the forecast graphs that will be generated. Available graphs include:

1) **Forecast Overlay**, or the time series graph of the data values generated by the forecast
2) **Original Series**, or the time series graph of the actual data
3) **Forecast Errors**, or the error between the forecast and actuals
4) **Deseasonalized Forecast Overlay**, or the time series graph of the data values generated by the forecast after the original data is deseasonalized
5) **Deseasonalized Original Series**, or the time series graph of the actual data after it is deseasonalized
6) **Deseasonalized Forecast Errors**, or the error between the forecast and actuals after the original data is deseasonalized

Example Forecasting Report



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed at the *beginning* of the time series, but not in the middle or at the end.
- **Link to Data –** Due to the lengthy calculations required, forecasts are not linked to data. If changes are made to the original data, the procedure should be rerun.

Regression and Classification Menu

The commands on the Regression and Classification Menu perform regression and classification analyses. Available regression analyses include Simple Multiple, Stepwise, Forward, Backward and Block. Other analyses on the Regression and Classification menu include discriminant analysis and logistic regression.

Regression Command

Runs regression analyses on a set of variables

The Regression Command runs a variety of regression analyses including Simple Multiple, Stepwise, Forward, Backward and Block. Reports from each analysis include summary measures of each regression equation run, an ANOVA table for each regression, and a table of estimated regression coefficients, their standard errors, their t-values, their p-values, and 95% confidence intervals for them for each regression.

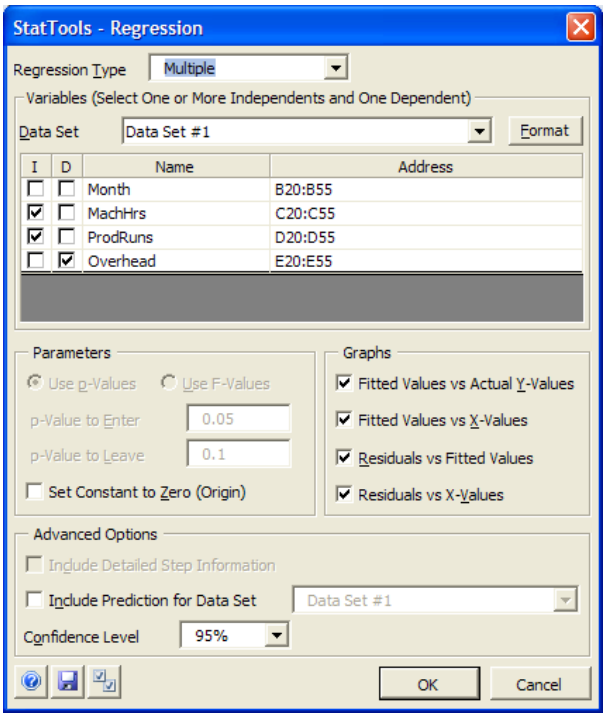
In addition, you have the option of creating two new variables, the fitted values and the residuals, and creating a number of diagnostic scatterplots.

Available Regression Types

The available regression types include **Simple Multiple**, **Stepwise**, **Forward**, **Backward** or **Block**. The Simple Multiple regression procedure builds an equation all at once, using the selected explanatory variables. The other procedures allow variables (or blocks of variables) to enter or leave the equation sequentially. Specifically, the stepwise procedure allows variables to enter one at a time. The next variable to enter is the one most highly correlated with the unexplained part of the response variable. However, the stepwise option also allows variables to leave once they have entered if they no longer contribute significantly. The forward procedure is the same as the stepwise procedure, except that variables are not allowed to leave once they enter. The backward procedure starts with all potential explanatory variables in the equation and then deletes them one at a time if they do not contribute significantly. Finally, the block procedure allows blocks of explanatory variables to enter or not enter *as a block* in a specified order. If one block is not significant and does not enter, then no later blocks are considered for entry.

Regression Dialog Box

These analyses are set up using the **Regression** dialog box:



One dependent variable (D) and one or more independent variables (I) need to be selected for analysis except in the case of a Block regression. With Block regression, one dependent variable (D) and one to seven blocks (B1 to B7) need to be selected. The selected data set must be unstacked data. Variables can be from different data sets.

The options in the **Regression** dialog include:

- **Regression Type**. Selects the type of regression to perform - Simple Multiple, Stepwise, Forward, Backward or Block.

Regression Parameters change with the selected Analysis Type. Options include:

- **Use p-Values** - Affects how variables are added or removed from the regression equation. When selected, you can specify a **p to Enter** and/or a **p to Leave**, depending on the regression method in use. The lower the p-value, the more significant a variable must be to enter or leave the regression equation. The default values that are shown are usually acceptable. Just remember two things. First, the p-value to enter cannot be larger than the p-value to leave. Second, to make it easier for variables to enter (and harder to leave), use *larger* p-values. To make it harder for variables to enter (and easier to leave), use *smaller* p-values. Typical p-values are in the range .01 to .1.
- **Use F-Values** - As with p-Values, affects how variables are added or removed from the regression equation. Allows you to specify a **F to Enter** and/or a **F to Leave**, depending on the regression method in use. Typical values are in the range 2.5 to 4.

StatTools can create several optional scatterplots, as shown in the selected **Graphs** options. These include:

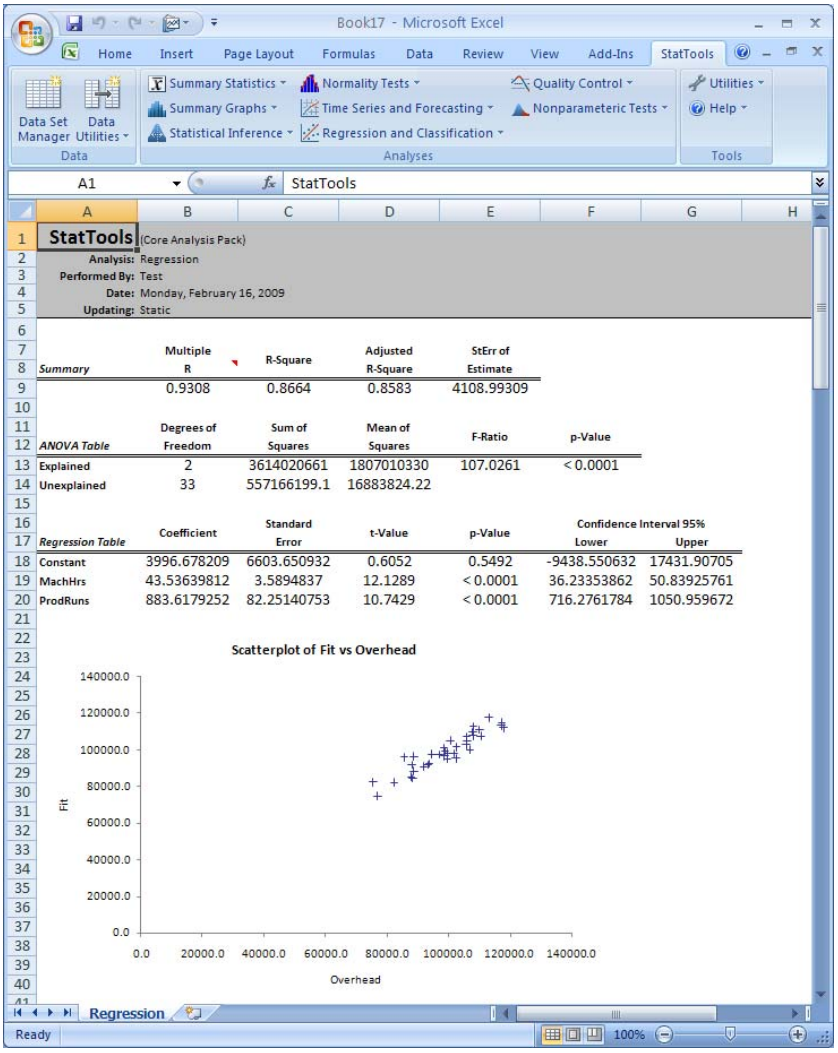
- **Fitted Values vs. Actual Y-Values**
- **Fitted Values vs. X-Values**
- **Residuals vs. Fitted Values**
- **Residuals vs. X-Values**

These plots are typically used in "residual analysis" to check whether the regression assumptions are satisfied. Probably the most useful plot is the one of the residuals (on the vertical axis) versus the fitted (or predicted) values of the response variable.

Advanced Options for a regression analysis include:

- **Include Detailed Step Information**, reporting r-squared and standard error statistics for the independent variable at each intermediate step of the regression.
- **Include Prediction**, where predicted values for the dependent variable are generated for the independent variable values in a second data set. This **prediction data set** must have the same variable names as the original data set that the regression is analyzing. Typically, in the prediction data set, you will have sets of values for independent variables for which you wish to **predict** the value for the dependent variable. The regression equation calculated from the first data set is used to make the predictions. The predicted values for the dependent variable will be entered directly in the prediction data set; filling the column (or row) for the dependent variable with the predicted values. The **Confidence Level** specifies the lower and upper bounds that will be generated for the predicted values.

Stepwise Regression Report



Missing Data and Link to Data

- **Missing Data -** If there are missing values, then any row with missing values for *any* of the selected variables is ignored.
- **Link to Data –** There is no link to the original data. If the data change, you must rerun the analysis.

Logistic Regression Command

Runs a logistic regression on a set of variables

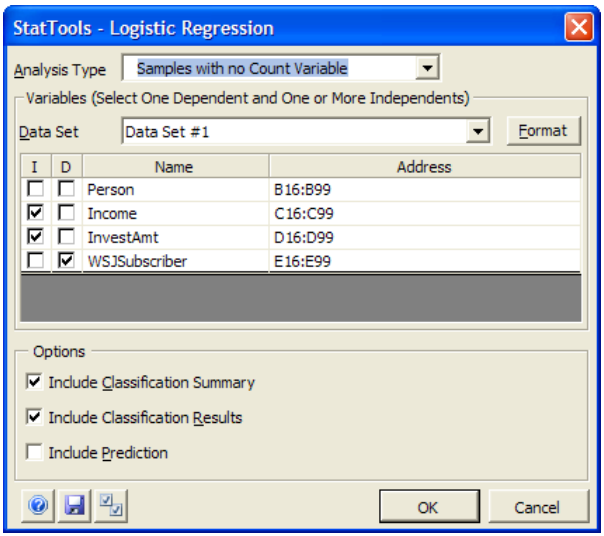
The Logistic Regression command performs a logistic regression analysis on a set of variables. This is essentially a nonlinear type of regression analysis where the response variable is binary: 0 or 1. There should be a 0-1 response variable that specifies whether each observation is a "success" or "failure", plus one or more explanatory variables that can be used to estimate the probability of success.

A second option for logistic regression is to have a "count" variable that specifies the number of "trials" observed at each combination of explanatory variables. Then the response variable should indicate the number of trials resulting in "success." The result of the logistic regression is a regression equation that is similar to a regular multiple regression equation. However, it must be interpreted somewhat differently, as explained below.

StatTools' logistic regression procedure relies on optimization to find the regression equation. This optimization must use a complex nonlinear algorithm, so the procedure can take quite a while, depending on the speed of your PC.

Logistic Regression Dialog Box

This analysis is set up using the **Logistic Regression** dialog box:



One dependent or response variable (D) and one or more independent variables (I) need to be selected for analysis. The data must:

1) Be in "stacked" form, and there must either be a 0-1 response variable that specifies whether each observation is a "success" or "failure". This is referred to here as **Samples with No Count Variable**.

2) Have a "count" variable and an integer response variable. This is referred to here as **Summary of Samples (with Count Variable).**

Variables can be from different data sets. If the Analysis Type is set to **Summary of Samples (with Count Variable)**, an additional count variable needs to be selected.

The options in the **Logistic Regression** dialog include:

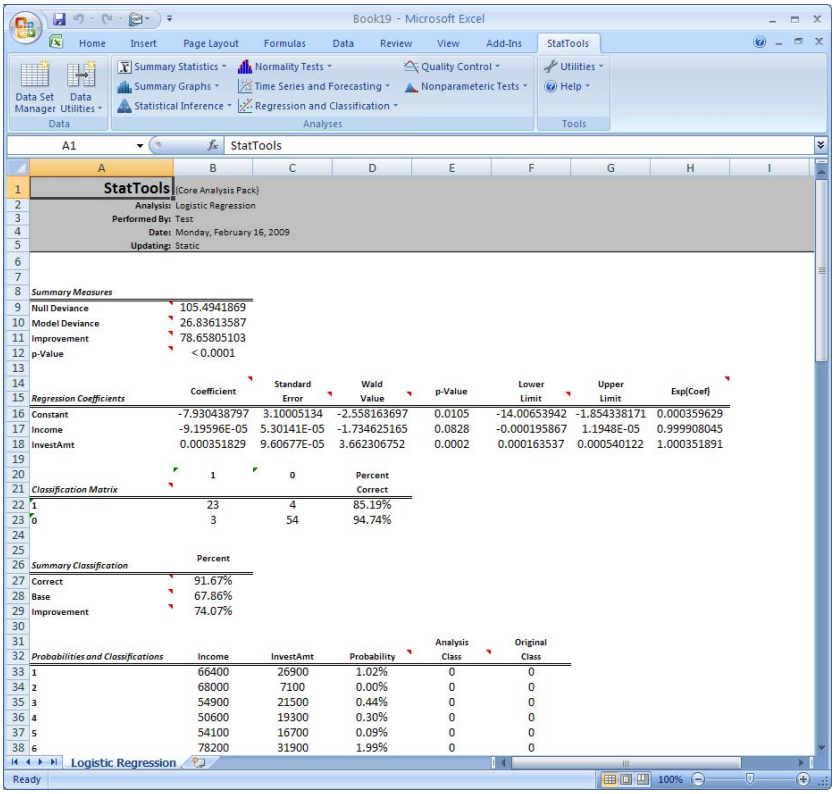
- **Analysis Type**. Selects the type of logistic regression to perform – **Samples with No Count Variable** or **Summary of Samples (with Count Variable)**.

Samples with No Count Variable have a 0-1 response variable that specifies whether each observation is a "success" or "failure", plus one or more explanatory variables that can be used to estimate the probability of success.

Summary of Samples (with Count Variable) have a "count" variable that specifies the number of "trials" observed at each combination of explanatory variables. Then the response variable should indicate the number of trials resulting in "success." For this analysis type a separate column C appears in the Variable Selector, allowing the count variable to be selected.

- **Include Classification Summary** with the regression report.
- **Include Classification Results** with the regression report.
- **Include Prediction**, where predicted values for the dependent variable are generated for the independent variable values in a second data set. This **prediction data set** must have the same variable names as the original data set that the regression is analyzing. Typically, in the prediction data set, you will have sets of values for independent variables for which you wish to **predict** the value for the dependent variable. The regression equation calculated from the first data set is used to make the predictions. The predicted values for the dependent variable will be entered directly in the prediction data set; filling the column (or row) for the dependent variable with the predicted values.

Logistic Regression Report



The report above includes the original data plus data used for classification. The predicted classifications, in column E, are based on whether the estimated probabilities of "success", in column D, are above or below a cutoff value of 0.5, or 50%. The report lists summary statistics for the regression (somewhat similar to R-square for multiple regression), detailed information about the regression equation, and summary results of the classification procedure. (There are plenty of cell comments to help you interpret the results). In this example, we see that 90.5% of the observations are classified correctly. Of course, we are hoping to make this percentage as large as possible. Generally, the only way to improve the results is to use more (or better) explanatory variables. The values in column H **Exp (Coeff)** are generally used to interpret the regression equation. They indicate the estimated change in the odds of "success" when any explanatory variable increases by 1 unit.

Missing Data and Link to Data

- **Missing Data -** If there are missing values, then any row with missing values for *any* of the selected variables is ignored.
- **Link to Data –** There is no link to the original data. If the data change, you must rerun the analysis.

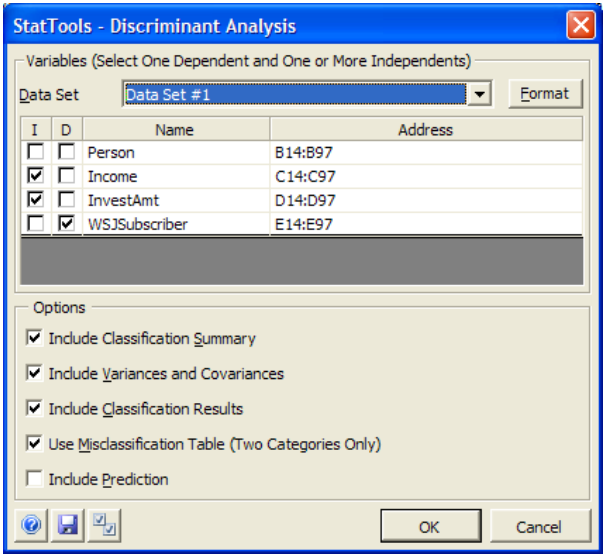
Discriminant Analysis Command

Runs a discriminant analysis on a set of variables

The Discriminant Analysis command performs a discriminant analysis on a data set. In this analysis there is a "category" variable that specifies which of two or more groups an observation is in, plus one or more explanatory variables that can be used to predict group membership. There are two ways to predict group membership. The more general way, valid for any number of groups, is to calculate the "statistical distance" of each observation to the mean of each group and to classify the observation according to the smallest statistical distance. A second method, used for the case of two groups, is to calculate a discriminant function (a linear expression of the explanatory variables) and to classify each observation according to whether its discriminant value is less than or greater than some cutoff value. This second method also allows you to specify prior probabilities of group membership, as well as misclassification costs. Then the classification procedure is equivalent to minimizing the expected cost of misclassification.

Discriminant Analysis Dialog Box

This analysis is set up using the **Discriminant Analysis** dialog box:



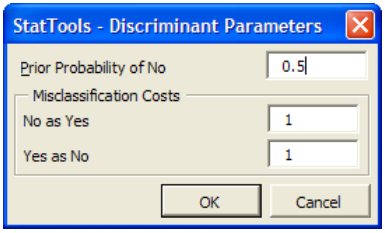
One dependent variable (D) and one or more independent variables (I) need to be selected for analysis. The data must be in "unstacked" form. Variables can be from different data sets.

The options in the **Discriminant Analysis** dialog include:

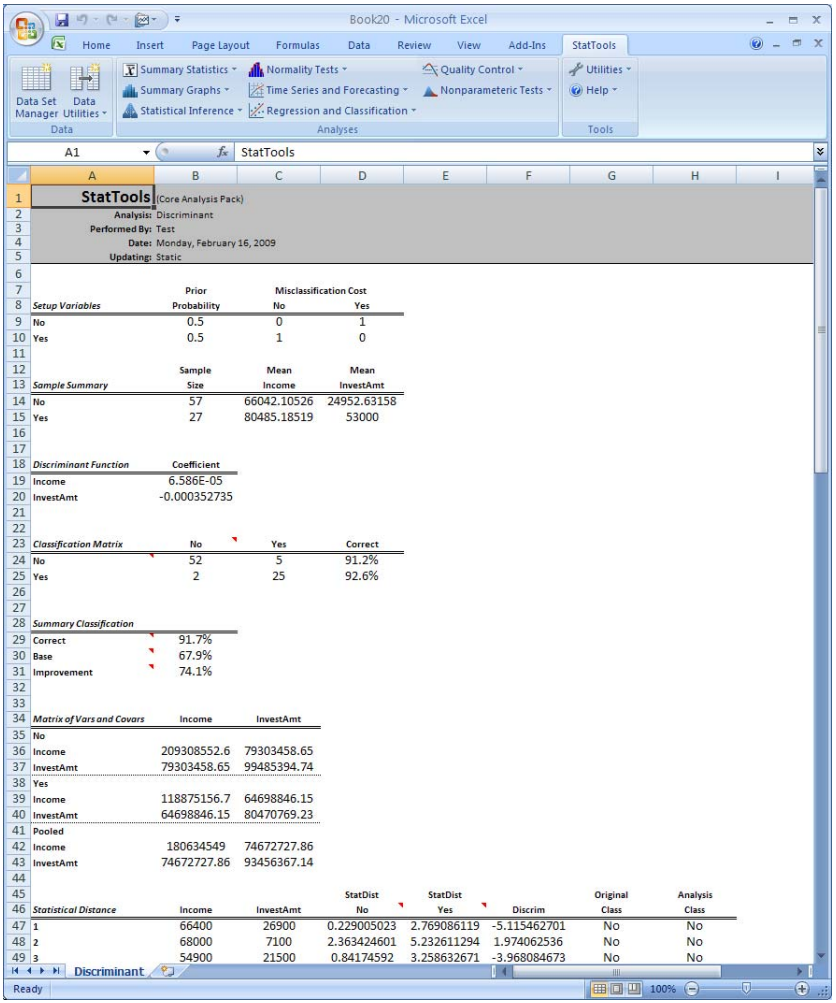
- **Include Classification Summary** with the regression report.
- **Include Variances and Covariances** with the regression report.
- **Include Classification Results** with the regression report.
- **Use Misclassification Table,** selected when you wish to change prior probabilities or misclassification costs.
- **Include Prediction**, where predicted values for the dependent variable are generated for the independent variable values in a second data set. This **prediction data set** must have the same variable names as the original data set that the regression is analyzing. Typically, in the prediction data set, you will have sets of values for independent variables for which you wish to **predict** the value for the dependent variable. The regression equation calculated from the first data set is used to make the predictions. The predicted values for the dependent variable will be entered directly in the prediction data set; filling the column (or row) for the dependent variable with the predicted values.

Misclassification Costs Dialog Box

If there are exactly two groups possible for the dependent category variable (as in this example) and the **Use Misclassification Table** option is selected, a dialog box appears that allows you to specify prior probabilities and/or misclassification costs. The default settings are that each group is equally likely and that the misclassification costs are equal, but you can override these settings.



Discriminant Analysis Report



Book20 - Microsoft Excel

Home | Insert | Page Layout | Formulas | Data | Review | View | Add-Ins | StatTools

Summary Statistics | Normality Tests | Quality Control | Utilities
Summary Graphs | Time Series and Forecasting | Nonparametric Tests | Help
Statistical Inference | Regression and Classification
Data Set Manager | Data Utilities
Data | Analyses | Tools

A1 StatTools

StatTools (Core Analysis Pack)
Analysis: Discriminant
Performed By: Test
Date: Monday, February 16, 2009
Updating: Static

Setup Variables	Prior Probability	Misclassification Cost No	Misclassification Cost Yes
No	0.5	0	1
Yes	0.5	1	0

Sample Summary	Sample Size	Mean Income	Mean InvestAmt
No	57	66042.10526	24952.63158
Yes	27	80485.18519	53000

Discriminant Function	Coefficient
Income	6.586E-05
InvestAmt	-0.000352735

Classification Matrix	No	Yes	Correct
No	52	5	91.2%
Yes	2	25	92.6%

Summary Classification	
Correct	91.7%
Base	67.9%
Improvement	74.1%

Matrix of Vars and Covars	Income	InvestAmt
No		
Income	209308552.6	79303458.65
InvestAmt	79303458.65	99485394.74
Yes		
Income	118875156.7	64698846.15
InvestAmt	64698846.15	80470769.23
Pooled		
Income	180634549	74672727.86
InvestAmt	74672727.86	93456367.14

Statistical Distance	Income	InvestAmt	StatDist No	StatDist Yes	Discrim	Original Class	Analysis Class
1	66400	26900	0.229005023	2.769086119	-5.115462701	No	No
2	68000	7100	2.363424601	5.232611294	1.974062536	No	No
3	54900	21500	0.84174592	3.258632671	-3.968084673	No	No

Discriminant

Ready 100%

The Discriminant Analysis report includes the original data plus data used for classification. The predicted classifications are based on whether the discriminant values shown are below or above a cutoff value. If the prior probabilities and misclassification costs are left at their default values, then this classification procedure is equivalent to basing classification on the smaller of the two statistical distances. In fact, if there were more than two groups, then the discriminant values would not appear, and classification would be based on the smallest of the statistical distances. The report also shows descriptive statistics for the groups and the coefficients of the discriminant function (it does this only when there are two groups), the prior probabilities, misclassification costs, and cutoff value for misclassification (again, only if there are two groups), and the summary results of the classification procedure (with cell comments to help you interpret the

results). In this example, we see that 89% of the observations are classified correctly. Of course, we are hoping to make this percentage as large as possible. Generally, the only way to improve the results is to use more (or better) explanatory variables

Missing Data and Link to Data

- **Missing Data -** If there are missing values, then any row with missing values for *any* of the selected variables is ignored.
- **Link to Data** – There is no link to the original data. If the data change, you must rerun the analysis.

Quality Control Menu

The procedures on the Quality Control menu deal with the analysis of data collected over time, with applications in quality control.

The Pareto chart displays the relative importance of categorized data.

The four types of control charts plot time series data and allow you to see whether a process is in statistical control. We can see whether the data stay within the control limits on the chart, and we can also check for other nonrandom behavior such as long runs above or below the centerline.

Pareto Chart Command

Creates Pareto chart for categorized variable

Pareto charts are useful for determining the most significant items in a group of categorized data, as well as conveying a quick visual representation of their relative importance. Typically Pareto charts are used in the area of Quality Assurance to determine the few factors which have the most significance (Pareto's 80/20 rule).

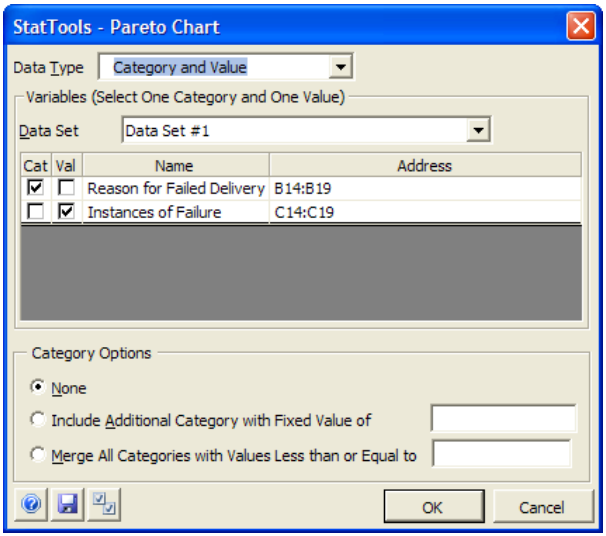
For example, a manufacturer of machine parts has decided to investigate why customers have been rejecting a particular product. When each batch is returned, a reason ("wrong size", "incorrect surface finish", etc.) is entered. After several months of data have been collected, a Pareto chart is plotted. Action is taken to address the largest sources of problems.

StatTools allows you to create Pareto charts based on data in one of two formats – Category Only, or Category and Value. A Category Only variable will typically contain one entry for each reading. In the example above, each cell would correspond to the reason a batch of parts was returned. A cell value might be "incorrect surface finish" and there would likely be many duplicated cells. StatTools will count the number of times each entry appears in the variable and create the corresponding Pareto chart. When Category and Value is selected, the variables you specify are the categories and each corresponding count.

The axes of the Pareto chart are constructed as follows:

- Categories are placed along the horizontal axis
- Frequency (or count) is placed along the left vertical axis
- Cumulative percentage is placed along the right vertical axis

Pareto Chart Dialog Box

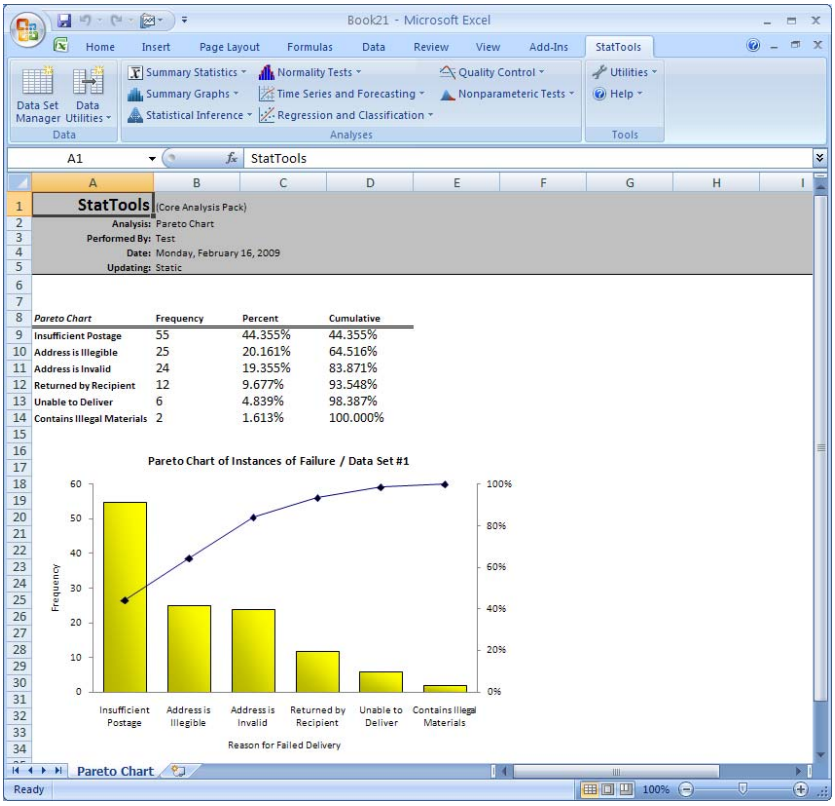


The options in the **Pareto Chart** dialog include:

- **Data Type**. Selects the type of data used to construct the Pareto chart – Category and Value or Category Only.

Category Options in the Pareto Chart dialog include:

- **None** – each distinct category will be represented by a bar in the Pareto chart.
- **Include Additional Category with Fixed Value of** – a bar labeled "Misc" will be added at the extreme right side of the Pareto chart with a frequency equal to the specified value.
- **Merge All Categories with Values Less Than or Equal to** – all categories whose frequency is less than or equal to the specified value will be combined into a category labeled "Misc" and placed at the extreme right side of the Pareto chart.



Missing Data and Link to Data

- **Missing Data -** If there are missing values, then any row with missing values for *any* of the selected variables is ignored.
- **Link to Data –** There is no link to the original data. If the data change, you must rerun the analysis.

X/R Charts Command

Creates X and R control charts for time series variables

This analysis produces X-bar and R charts for time series data. It assumes that data have been collected in small subsamples over time. For example, an operator might collect measurements on the widths of four randomly selected parts every half hour. The subsample size is then 4. If data are collected for 50 half-hour periods, then the data should be arranged in four adjacent columns and 50 adjacent rows, with variable headings such as SubSamp1 through SubSamp4 above the first row of data.

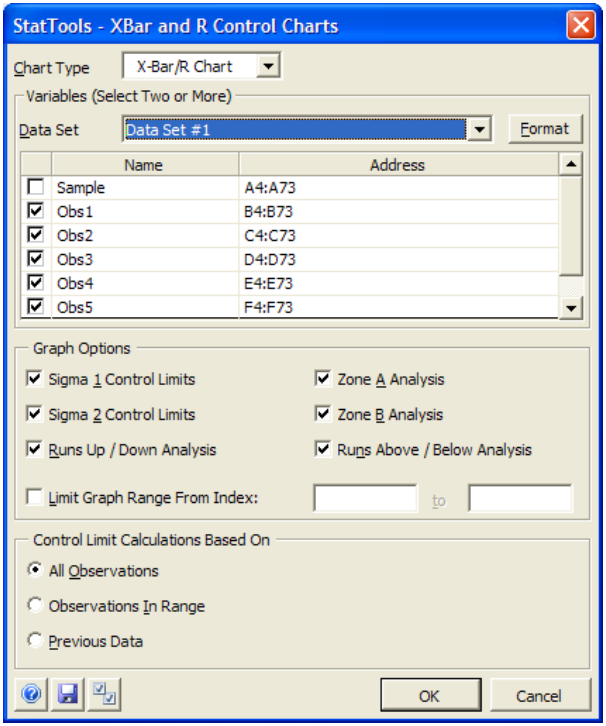
The purpose of the procedure is to check whether the process that is generating the data is in statistical control. To do so, the procedure first calculates an X-bar and an R for each row in the dataset. X-bar is the average of the observations in that row, and R is the range (maximum minus minimum) for the observations in that row.

The X-bars and R's are charted in separate time series plots around centerlines. The centerline for the X-bar chart is the average of the X-bars (sometimes called X-double-bar), and the centerline for the R chart is R-bar, the average of the R's. A simple way to check whether the process is in control is to see whether any of the X-bars or R's fall outside their respective upper and lower control limits (UCL and LCL), which are approximately plus or minus 3 standard deviations from the centerlines. The charts show these control limits, so that it is easy to spot any extreme values.

The procedure also allows you to check for other possible out-of-control behavior, including 8 or more points in a row above or below the centerline, 8 or more points in a row in an uphill or downhill direction, at least 4 of 5 points in a row more than one standard deviation from the centerline, and at least 2 of 3 points in a row more than two standard deviations from the centerline.

XBar and R Control Charts Dialog Box

These graphs are set up using the **XBar and R Control Charts** dialog box:



Two or more variables can be selected for analysis. The selected data set must be unstacked data. Variables can be from different data sets.

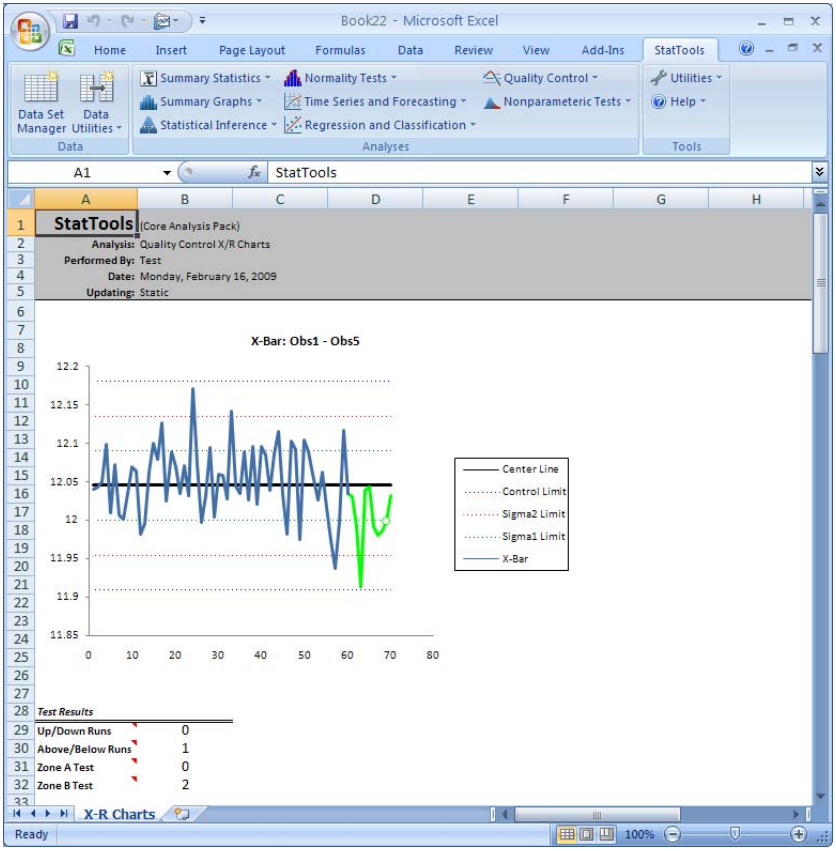
Graph Options in the XBar and R Control Charts dialog include:

- **Sigma 1 and 2 Control Limits** – Adds control limit lines at one and/ or two sigmas from mean line. These extra lines allow you to check for other types of out-of-control behavior (the so-called "zone" rules).
- **Zone A and B Analysis** – number of points beyond Zone A (2 sigma) and Zone B (1 sigma)
- **Runs Up/Down and Runs Above/Below Analyses** – sequential up or down moves of length 8 or greater
- **Limit Graph Range From Index** – Limits the points on the graph to a range of data points for a variable (i.e., range from starting index to ending index)

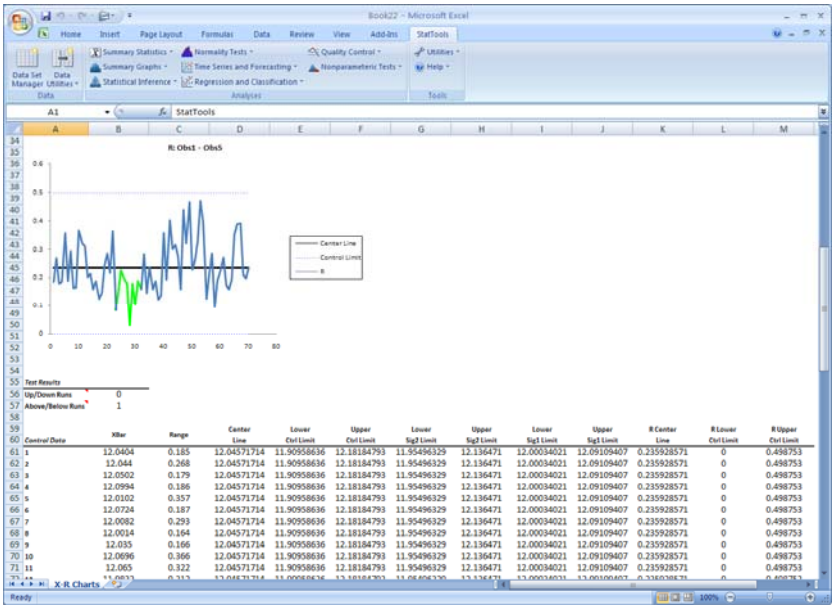
Control Limit Calculations Based On options in the XBar and R Control Charts dialog determine the data on which control limit calculations will be based, including:

- **All Observations** – use all available data in control limit calculations
- **Observations in Range** - use data between **Start Index** and **Stop Index** in control limit calculations
- **Previous Data** – creates control limits from previously observed data. Simply enter the **Subsample Size**, **Average R** and **Average X-Bar** that was calculated from the previous data.

Example X-Bar Chart



Example R Chart



Control Data	Xbar	Range	Center Line	Lower Ctrl Limit	Upper Ctrl Limit	Lower Sig1 Limit	Upper Sig1 Limit	Lower Sig2 Limit	Upper Sig2 Limit	R Center Line	R Lower Ctrl Limit	R Upper Ctrl Limit
1	12.0404	0.185	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
2	12.044	0.268	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
3	12.0502	0.179	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
4	12.0994	0.186	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
5	12.0102	0.357	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
6	12.0724	0.187	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
7	12.0082	0.293	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
8	12.0014	0.164	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
9	12.035	0.166	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
10	12.0696	0.366	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753
11	12.065	0.322	12.04571714	11.90558636	12.18184793	11.95496329	12.136471	12.00034021	12.09109407	0.235928571	0	0.498753

Missing Data and Link to Data

- **Missing Data -** Missing data are not allowed.
- **Link to Data** – Graphs are not linked to data.

P Chart Command

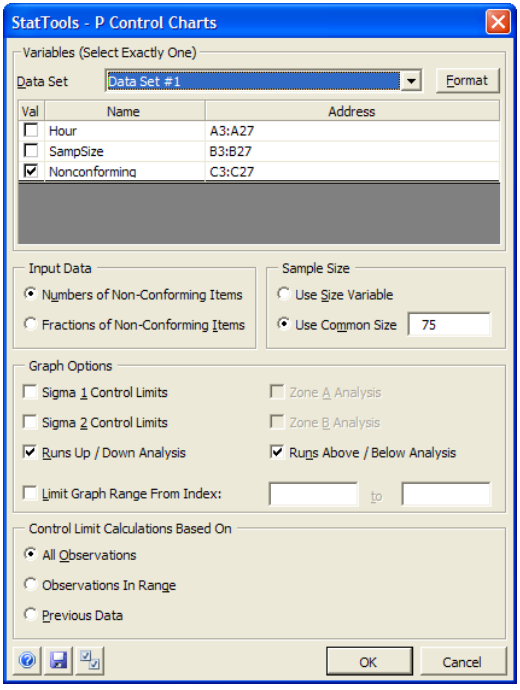
Creates P charts for time series variables

P charts are for "attribute" data. With attribute data each observation indicates the number (or fraction) of items that do not conform to specifications from a sample of items. For example, a process might produce a certain number of items each half hour, some of which are nonconforming. Then a P chart would plot each half hour's fraction of items that are nonconforming. As always, the purpose is to see if the process is in control.

This procedure requires a data set with at least one of the following: a variable that contains the *number* of nonconforming items in each sample or a variable that contains the *fraction* of nonconforming items in each sample. Optionally, there can be a variable that contains the sample sizes. If there is no sample size variable, then you must enter a sample size, which is assumed to be constant across all samples. If there is a sample size variable, however, the sample sizes are not required to be equal.

P Control Charts Dialog Box

This graph is set up using the **P Control Charts** dialog box:



One value variable and optionally, a size variable, are selected for analysis. The selected data set must be unstacked data. Variables can be from different data sets.

Input Data options in the P Control Charts dialog include:

- **Numbers of Non-Conforming Items –** Specifies that the Value variable gives the actual number of non-conforming items out of the total sample
- **Fractions of Non-Conforming Items –** Specifies that the Value variable gives the fraction of non-conforming items in the sample

Sample Size options in the P Control Charts dialog include:

- **Use Size Variable–** Specifies that a size variable is used to give the total size of each sample
- **Use Common Size –** Specifies that no size variable is used as each sample is the entered size.

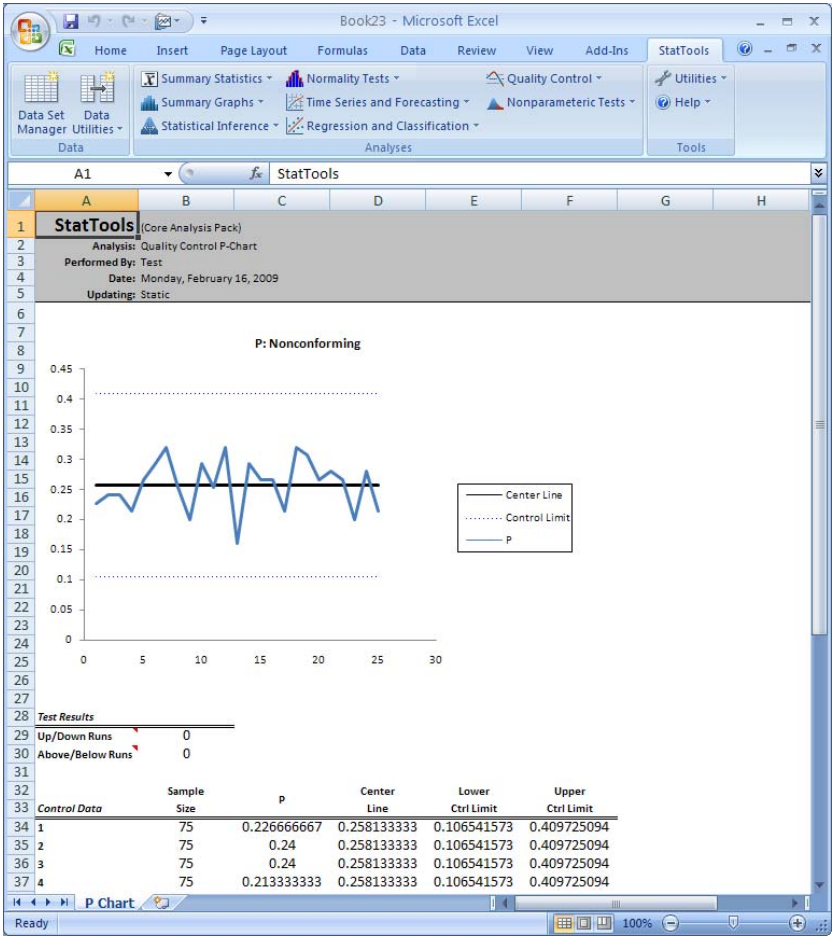
Graph Options in the P Control Charts dialog include:

- **Sigma 1 and 2 Control Limits –** Adds control limit lines at one and/ or two sigmas from mean line. These extra lines allow you to check for other types of out-of-control behavior (the so-called "zone" rules).
- **Zone A and B Analysis –** number of points beyond Zone A (2 sigma) and Zone B (1 sigma)
- **Runs Up/Down and Runs Above/Below Analyses –** sequential up or down moves of length 8 or greater
- **Limit Graph Range From Index –** Limits the points on the graph to a range of data points for a variable (i.e., range from starting index to ending index)

Control Limit Calculations Based On options in the P Control Charts dialog determine the data on which control limit calculations will be based, including:

- **All Observations** – use all available data in control limit calculations
- **Observations in Range** - use data between **Start Index** and **Stop Index** in control limit calculations
- **Previous Data** – creates control limits from previously observed data. Simply enter the **Subsample Size** and **Average P** that was calculated from the previous data.

Example P Chart



Missing Data and Link to Data

- **Missing Data -** Missing data are not allowed.
- **Link to Data** – Graphs are not linked to data.

C Chart Command

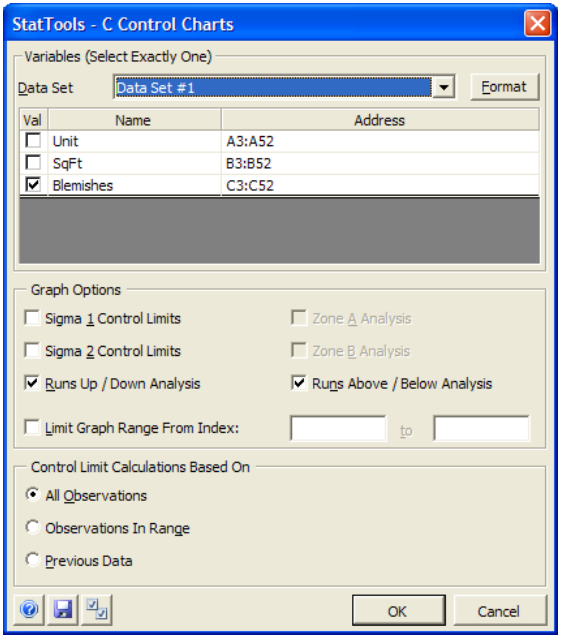
Creates C charts for time series variables

C charts are used to plot the number of defects for items of a constant size. For example, suppose that car doors are produced in batches of 50. In each batch we could count the number of defects (a paint blemish or a rough edge, for example). These counts are then plotted on a chart. As always, the purpose is to check whether the process is in control.

The data set for a C chart must include a variable that contains the count of defects for each item. It is assumed that the item size is equal for each observation. For example, if an "item" is a batch of car doors, then we would assume that each batch has the same number of doors.

C Control Charts Dialog Box

This graph is set up using the **C Control Charts** dialog box:



One or more variables are selected for analysis. The selected data set must be unstacked data. Variables can be from different data sets.

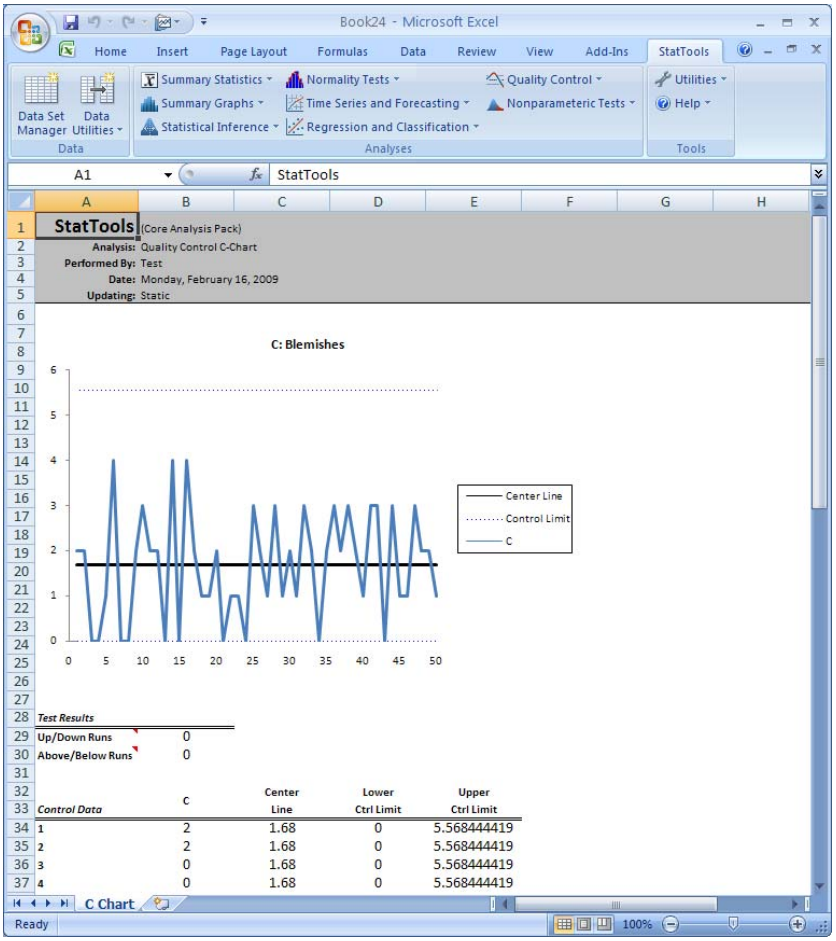
Graph Options in the C Control Charts dialog include:

- **Sigma 1 and 2 Control Limits –** Adds control limit lines at one and/ or two sigmas from mean line. These extra lines allow you to check for other types of out-of-control behavior (the so-called "zone" rules).
- **Zone A and B Analysis –** number of points beyond Zone A (2 sigma) and Zone B (1 sigma)
- **Runs Up/Down and Runs Above/Below Analyses –** sequential up or down moves of length 8 or greater
- **Limit Graph Range From Index –** Limits the points on the graph to a range of data points for a variable (i.e., range from starting index to ending index)

Control Limit Calculations Based On options in the C Control Charts dialog determine the data on which control limit calculations will be based, including:

- **All Observations –** use all available data in control limit calculations
- **Observations in Range** - use data between **Start Index** and **Stop Index** in control limit calculations
- **Previous Data –** creates control limits from previously observed data. Simply enter the **Average C** that was calculated from the previous data.

Example C Chart



Missing Data and Link to Data

- **Missing Data -** Missing data are not allowed.
- **Link to Data** – Graphs are not linked to data.

U Chart Command

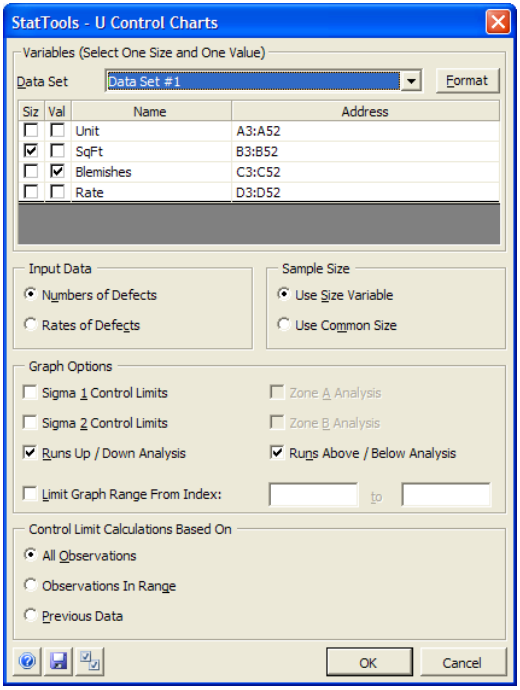
Creates U charts for time series variables

U charts are similar to C charts, but now we plot the rate of defects. Using the car door example, suppose the batch sizes are not necessarily equal, that is, different batches have different numbers of car doors. Then in a U chart, we would plot the rate of defects per car door, that is, the number of defects in a batch divided by the number of doors in the batch. As always, the purpose is to check whether the process is in control.

This procedure requires a data set with at least one of the following: a variable that contains the *number* of defects for each observation or a variable that contains the *rate* of defects in each observation. Optionally, there can be a variable that contains the item sizes. If there is no size variable, then you must enter an item size, which is assumed to be the constant item size for all observations. If there is a size variable, however, the item sizes are not required to be equal.

U Control Charts Dialog Box

This graph is set up using the **U Control Charts** dialog box:



One value variable and optionally, a size variable, are selected for analysis. The selected data set must be unstacked data. Variables can be from different data sets.

Input Data options in the U Control Charts dialog include:

- **Numbers of Defects –** Specifies that the Value variable gives the actual number of non-conforming items out of the total sample.
- **Rates of Defects –** Specifies that the Value variable gives the fraction of non-conforming items in the sample.

Sample Size options in the U Control Charts dialog include:

- **Use Size Variable–** Specifies that a size variable is used to give the total size of each sample
- **Use Common Size –** Specifies that no size variable is used as each sample is the entered size.

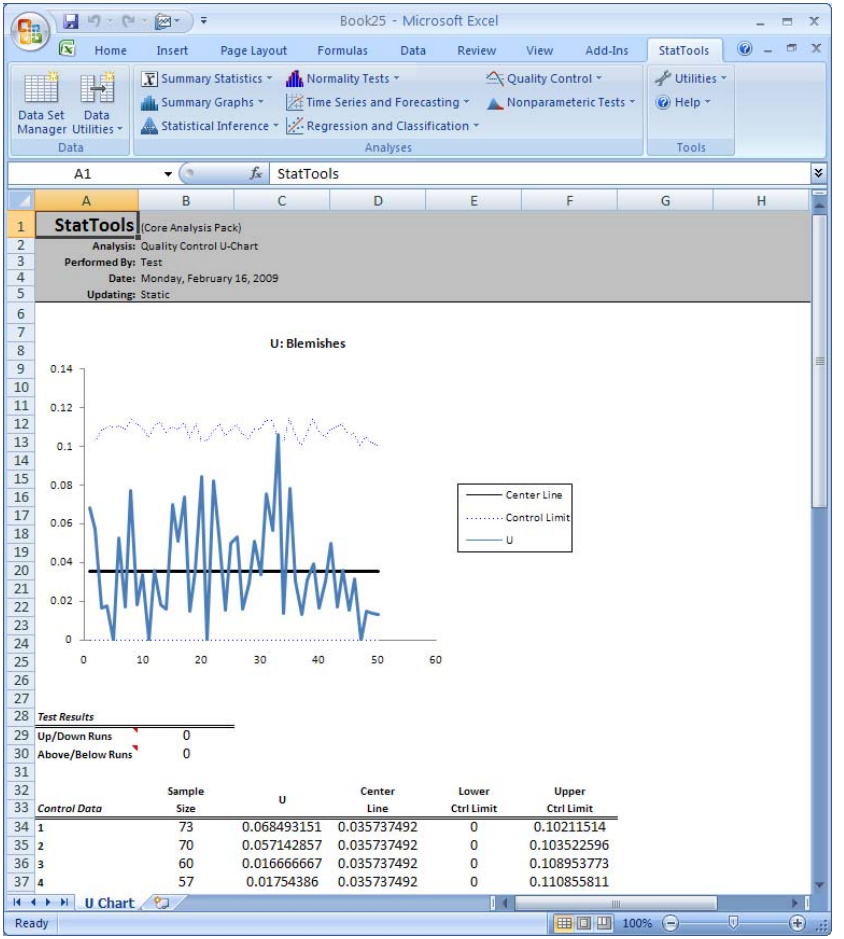
Graph Options in the U Control Charts dialog include:

- **Sigma 1 and 2 Control Limits –** Adds control limit lines at one and/ or two sigmas from mean line. These extra lines allow you to check for other types of out-of-control behavior (the so-called "zone" rules).
- **Zone A and B Analysis –** number of points beyond Zone A (2 sigma) and Zone B (1 sigma)
- **Runs Up/Down and Runs Above/Below Analyses –** sequential up or down moves of length 8 or greater
- **Limit Graph Range From Index –** Limits the points on the graph to a range of data points for a variable (i.e., range from starting index to ending index)

Control Limit Calculations Based On options in the U Control Charts dialog determine the data on which control limit calculations will be based, including:

- **All Observations** – use all available data in control limit calculations
- **Observations in Range** - use data between **Start Index** and **Stop Index** in control limit calculations
- **Previous Data** – creates control limits from previously observed data. Simply enter the **Subsample Size** and **Average U** that was calculated from the previous data.

Example U Chart



Missing Data and Link to Data

- **Missing Data -** Missing data are not allowed.
- **Link to Data** – Graphs are not linked to data.

Nonparametric Tests Menu

"Nonparametric" tests are statistical procedures applied to samples of data to test hypotheses about underlying probability distributions. "Parametric" hypothesis tests are more familiar and widely used; however, the nonparametric alternative offers advantages that make it the more suitable choice in many situations.

Parametric hypothesis tests make assumptions about the type of the underlying distribution (typically, that it is normal), and estimate the parameters of that type of distribution (typically, the mean and standard deviation). In many applications the normality assumption would be incorrect. For example, the numbers of calls per hour to a customer service center and the waiting time at a checkout in a supermarket are not normally distributed. Nonparametric tests do not require any assumptions about the type of the underlying distribution. Some of them make certain general assumptions about the shape of the distribution: in this pack the Wilcoxon Signed-Rank Test assumes that the distribution is symmetric. The other two tests in the pack (the Sign Test and the Mann-Whitney Test) do not assume anything about the distribution shape.

With small sample sizes the nonparametric alternative is often more appropriate. If the sample is large, a normality test can be applied. If the assumption that the distribution is normal turns out to be justified, a parametric test can be used. However, for small sample sizes normality tests have little power to differentiate between the normal and other distributions. Nonparametric tests can provide a way out of the dilemma.

Ordinal Data

For certain types of data, parametric tests cannot be used, while some nonparametric ones can. One such case is **ordinal data** where observations are described in terms of numbers that express places in a ranking; however, the difference between two such numbers is not meaningful. For example, levels of educational attainment can be coded as 0 (less than high school), 1 (some high school), 2 (high school degree), 3 (some college), 4 (college degree), and 5 (post college). When this scale is used, there is no implication that the difference in educational attainment between having a "high school degree" and having "less than high school" is equivalent to the difference between being in the categories "post college" and "some college ", even though in both cases the difference between the ranks is equal 2. This pack includes tests that can be applied to such data - the Sign Test (One-Sample analysis type) and Mann-Whitney Test.

Summary of Uses

In summary, nonparametric tests are applicable in the following situations, in which parametric ones are not:

- when there is little information about the underlying probability distribution,
- when the sample size is too small to reliably test the normality assumption,
- when the data is ordinal.

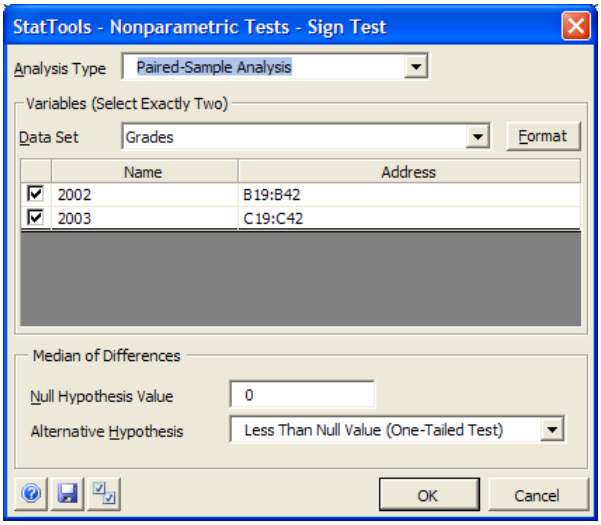
Sign Test Command

Performs sign test on variables

The Sign Test command performs hypothesis tests for the median of a single variable (**One-Sample Analysis**) or for the median of differences for a pair of variables (**Paired-Sample Analysis**). The test does not make any assumptions about the shape of the distribution (and in particular does not assume that it is normal). The One-Sample Analysis can be used with **ordinal data** as described in the Overview.

Sign Test Dialog Box

This analysis is set up using the **Sign Test** dialog box:

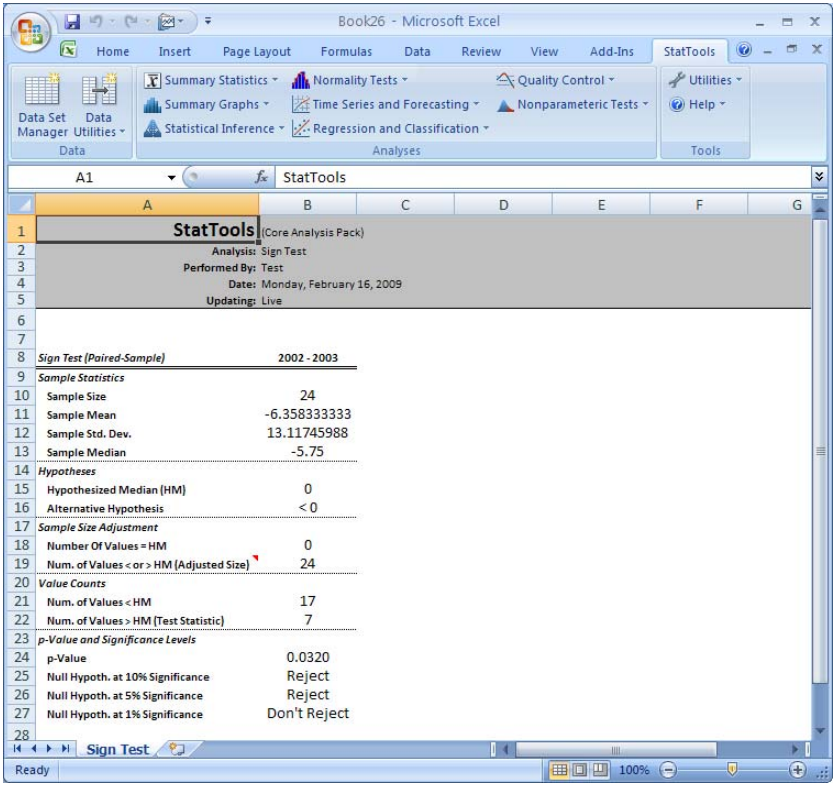


The number of variables selected depends on the Analysis Type used. A One-Sample Analysis requires one or more variables, while a Paired-Sample Analysis requires two variables. For a One-Sample Analysis the selected variables can be stacked or unstacked data; for a Paired-Sample Analysis they have to be unstacked. Variables can be from different data sets.

The options in the **Sign Test** dialog include:

- **Analysis Type.** Selects the type of analysis performed. Options include:
  - **One-Sample Analysis**. Performs hypothesis tests for a single numerical variable.
  - **Paired-Sample Analysis**. This analysis is appropriate when two variables are naturally paired. It is equivalent to a one-sample analysis on the differences between pairs.
- **Median (or Median of Differences)**.
  - **Null Hypothesis Value** or the value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis** or the alternative to the Null Hypothesis Value that will be evaluated during the analysis. The Alternative Hypothesis can be either **"one-tailed"** (that is, greater or less than the null hypothesis) or **"two-tailed"** (that is, not equal to the null hypothesis).

Sign Test Report



Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

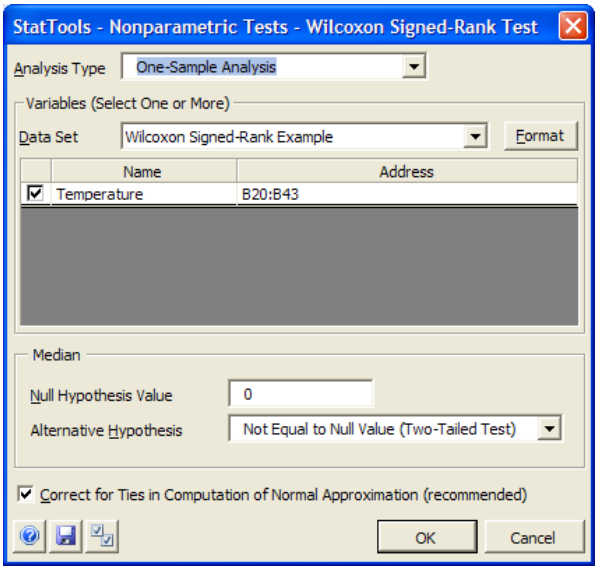
Wilcoxon Signed-Rank Test Command

Performs Wilcoxon Signed-Rank tests on variables

The Wilcoxon Signed-Rank Test command performs hypothesis tests for the median of a single variable (**One Sample Analysis**) or for the median of differences for a pair of variables (**Paired-Sample Analysis**). The test assumes that the probability distribution is symmetric (but it does not assume that it is normal).

This analysis is set up using the **Wilcoxon Signed-Rank Test** dialog box:

Wilcoxon Signed-Rank Test Dialog Box

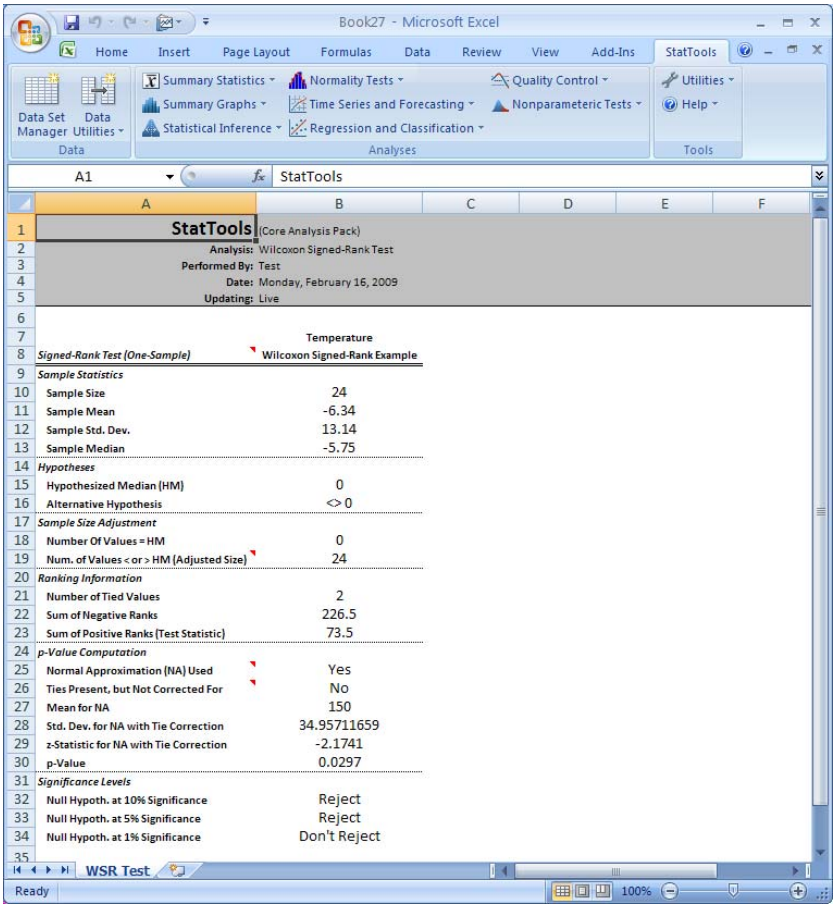


The number of variables selected depends on the Analysis Type used. A One-Sample Analysis requires one or more variables, while a Paired-Sample Analysis requires two variables. For a one-sample analysis the selected variables can be stacked or unstacked data; for a paired-sample analysis they have to be unstacked. Variables can be from different data sets.

The options in the **Wilcoxon Signed-Rank Test** dialog include:

- **Analysis Type.** Selects the type of analysis performed. Options include:
  - **One-Sample Analysis**. Performs hypothesis tests for a single numerical variable.
  - **Paired-Sample Analysis**. This type of analysis is appropriate when two variables are naturally paired. It is equivalent to a one-sample analysis on the differences between pairs.
- **Median (or Median of Differences)**.
  - **Null Hypothesis Value** or the value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis** or the alternative to the Null Hypothesis Value that will be evaluated during the analysis. The Alternative Hypothesis can be either **"one-tailed"** (that is, greater or less than the null hypothesis) or **"two-tailed"** (that is, not equal to the null hypothesis).
- **Tie Correction**. A recommended selection that corrects for tied ranks in the test only when the normal approximation is used. The correction involves counting the numbers of elements in groups of tied ranks and reducing the variance accordingly. The correction for ties will always increase the value of the z statistic, if tied ranks are present. (Note: The tie correction will produce no change in the variance when there are no ties.)

Wilcoxon Signed-Rank Test Report



Book27 - Microsoft Excel

Home | Insert | Page Layout | Formulas | Data | Review | View | Add-Ins | StatTools

Data Set Manager | Data Utilities — Data

Summary Statistics | Summary Graphs | Statistical Inference | Normality Tests | Time Series and Forecasting | Regression and Classification | Quality Control | Nonparametric Tests — Analyses

Utilities | Help — Tools

A1 | StatTools

	A	B
1	**StatTools**	(Core Analysis Pack)
2	**Analysis:**	Wilcoxon Signed-Rank Test
3	**Performed By:**	Test
4	**Date:**	Monday, February 16, 2009
5	**Updating:**	Live
6		
7		Temperature
8	*Signed-Rank Test (One-Sample)*	Wilcoxon Signed-Rank Example
9	*Sample Statistics*	
10	Sample Size	24
11	Sample Mean	-6.34
12	Sample Std. Dev.	13.14
13	Sample Median	-5.75
14	*Hypotheses*	
15	Hypothesized Median (HM)	0
16	Alternative Hypothesis	<> 0
17	*Sample Size Adjustment*	
18	Number Of Values = HM	0
19	Num. of Values < or > HM (Adjusted Size)	24
20	*Ranking Information*	
21	Number of Tied Values	2
22	Sum of Negative Ranks	226.5
23	Sum of Positive Ranks (Test Statistic)	73.5
24	*p-Value Computation*	
25	Normal Approximation (NA) Used	Yes
26	Ties Present, but Not Corrected For	No
27	Mean for NA	150
28	Std. Dev. for NA with Tie Correction	34.95711659
29	z-Statistic for NA with Tie Correction	-2.1741
30	p-Value	0.0297
31	*Significance Levels*	
32	Null Hypoth. at 10% Significance	Reject
33	Null Hypoth. at 5% Significance	Reject
34	Null Hypoth. at 1% Significance	Don't Reject
35		

WSR Test

Ready | 100%

Note: In this report, the p-value is computed using normal approximation when the sample size is greater than 15.

Missing Data and Link to Data

- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

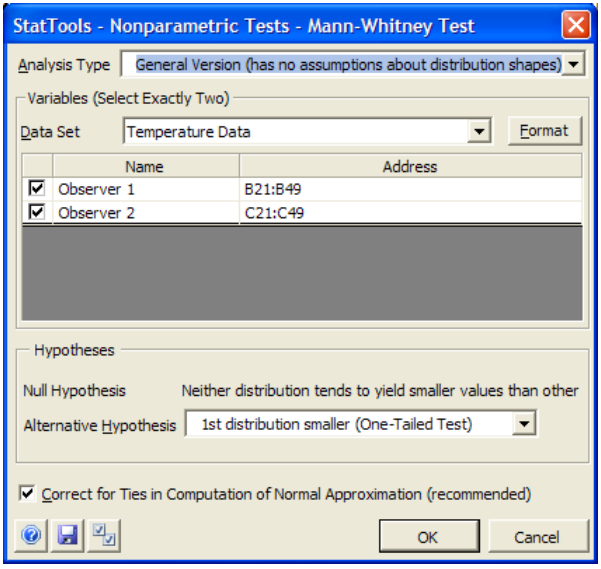
Mann-Whitney Test Command

Performs Mann-Whitney test on variables

The Mann-Whitney Test command performs a hypothesis test on two samples. In one version of the test (the Median Version) the hypothesis states that the medians of the two populations are identical. In this version the probability distributions are assumed to have the same shape. In the other version (the General Version) this assumption is not made, and the hypothesis denies that either probability distribution tends to yield smaller values than the other (more precisely, it states that P[X1>X2] = P[X2>X1], where P[X1>X2] is the probability that an observation from population 1 is greater than an observation from population 2). Note that the Mann-Whitney test can be used to reject the hypothesis that two samples are generated by the same probability distribution. The Mann-Whitney test is often also referred to as the Wilcoxon Rank-Sum test.

Mann-Whitney Test Dialog Box

This analysis is set up using the **Mann-Whitney Test** dialog box:



The analysis requires two variables. They can be stacked or unstacked, and can come from different data sets.

The options in the **Mann-Whitney Test** dialog include:

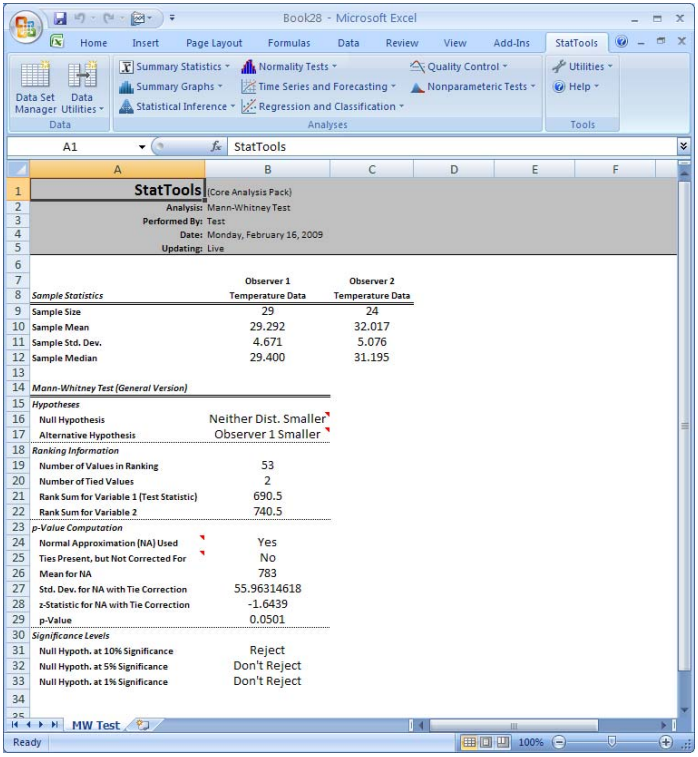
- **Analysis Type**. Selects the formulation of the null hypothesis and the alternatives. Options include:
  - **General Version**. Performs a hypothesis test to see if one probability distribution tends to yield smaller values than the other.
    - **Hypotheses**.
      - **Null Hypothesis**: Denies that either probability distribution tends to yield smaller values that the other. More precisely, it states that P[X1>X2] = P[X2>X1], where P[X1>X2] is the probability that an observation from population 1 is greater than an observation from population 2, and P[X2>X1] has analogous interpretation. For continuous distributions, this is equivalent to saying that both of these probabilities are 0.5 (P[X1>X2] = P[X2>X1] = 0.5).
      - **Alternative Hypothesis**: The Alternative Hypothesis can be either "one-tailed" (that is, one probability is greater or less than the other) or "two-tailed" (that is, the two probabilities are not equal).
  - **Median Version**. Performs a hypothesis test to see if the median of one population is the same or different from the median of the other population. Assumes the two distributions have the same shape.
    - **Hypotheses**.
      - **Null Hypothesis**: States that the two medians are equal.
      - **Alternative Hypothesis**: The Alternative Hypothesis can be either "one-tailed" (that is, the median of the first population is greater or less than that of the second) or "two-tailed" (that is, the medians are not equal).

Note: The calculations performed in the General and Median versions of the test are the same; the versions differ only with regard to the presence of the assumptions of equal distribution shapes, and with regard to the null hypothesis. The point of having these two versions

is to make it clear that the Mann-Whitney test can be applied even if one cannot assume that the distributions have approximately identical shapes, as long as one considers an appropriate null hypothesis. Specifically, if you run the test when the two distributions clearly have different shapes, and the test rejects the null hypothesis, it could be because the medians are different, but it could also be because the variances are different or other reasons.

- **Tie Correction**. A recommended selection that corrects for tied ranks in the test only when the normal approximation is used. The correction involves counting the numbers of elements in groups of tied ranks and reducing the variance accordingly. The correction for ties will always increase the value of the z statistic, if tied ranks are present. (Note: The tie correction will produce no change in the variance when there are no ties.)

Mann-Whitney Test Report



In the report above, the ***p- value*** is computed using normal approximation when the size of one of the two samples is greater than 10 (except when one of the sizes is equal to 11 or 12, while the other is equal to 3 or 4).

Missing Data and Link to Data

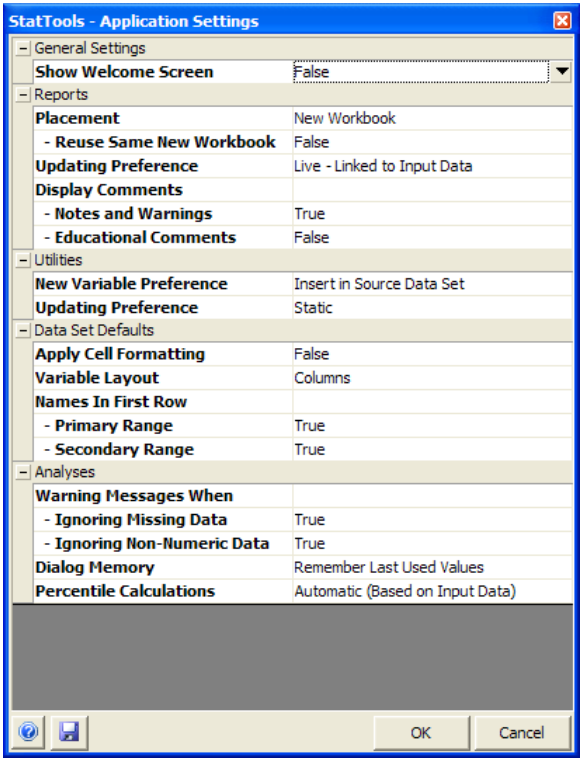
- **Missing Data -** Missing data are allowed. All rows with missing data on the selected variables are ignored.
- **Link to Data –** All of the reports are calculated with formulas that are linked to the data. If the values of the selected variable change, the outputs change automatically.

Utilities Menu

Application Settings Command

Specifies settings for StatTools reports, graphs, utilities, datasets and analyses

The **Application Settings** command allows you to specify general settings for StatTools reports, graphs, utilities, datasets and analyses. These settings apply across analyses and datasets. Other analysis-specific settings are defined in the dialog box for each analysis.



Reports Settings

The **Reports Settings** specify options for reports and graphs created by StatTools analyses. It includes the following:

- **Placement** – selects the location in Excel for new reports and graphs, including:
  - **Active Workbook**, where a new worksheet is created for each report.

- **In New Workbook**, where a new StatTools report workbook is created (if necessary) and each report is placed on a sheet in that workbook.
- **After Last Used Column in Active Sheet**, where each StatTools report is placed on the active sheet to the right of the last used column.
- **Query for Starting Cell,** where, after running an analysis, you have the opportunity to select a cell where the top-left corner of the report or graph will be placed.
- **Reuse Same New Workbook,** where, if a new workbook is created, that same new workbook will be used for all reports

- **Updating Preference –** specifies how results will change when variable data is changed. Options for results updating include:
  - **Live - Values Change With Input Data,** where reports update automatically as input data changes.
  - **Static – Values are Fixed,** or reports do not change with changing input data. Statistics are fixed based on the input data values when the procedure was run.

  Results are made live in StatTools through the use of Excel formulas and custom StatTools functions. For example, the formula:

  =StatMean('Confidence Interval.xls'!Pair)

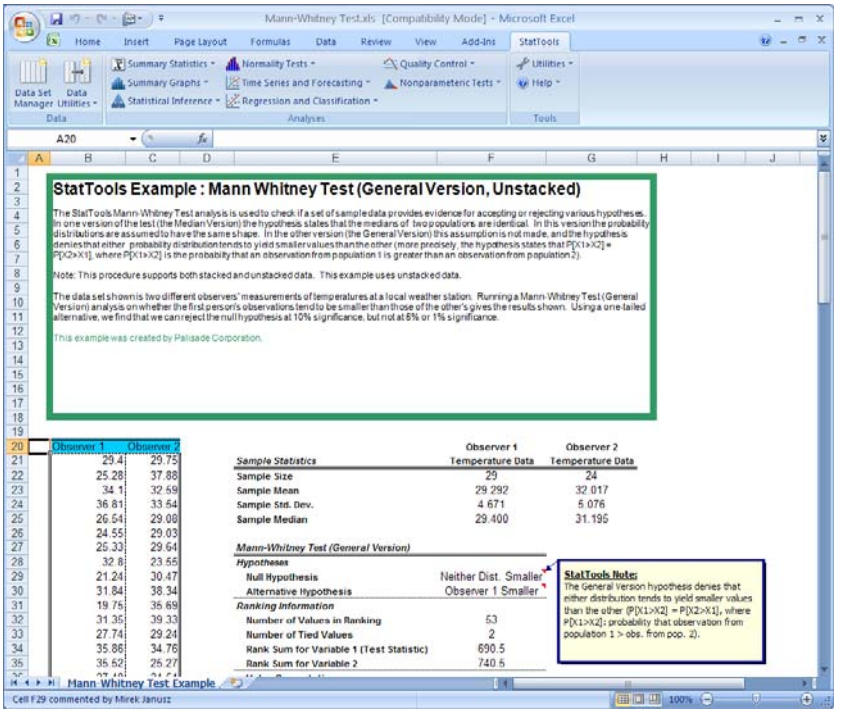
  calculates the mean of the variable *Pair* (which uses data from the Excel range name "Pair" located in the workbook Confidence Interval.xls). As data in the Pair range changes, the value returned by the StatMean function will update.

  Reports and graphs from all StatTools procedures can be updated live, with the following exceptions:

  1) **Regression**
  2) **Discriminant Analysis**
  3) **Logistic Regression**
  4) **Forecast**

  These procedures require lengthy recalculations that would cause Excel to become less responsive during live updating.

- **Display Comments –** specifies which categories of StatTools messages will be included in reports. Notes, warnings and Educational notes may be displayed as pop-up notes on cells in reports, as shown here:



Utilities Settings

The **Utilities Settings** specify options for new variables created by the StatTools Data Utilities. These are found on the Data Utilities menu. They create new variables by transforming, combining or otherwise processing existing variables.

- **New Variable Preference –** Specifies the desired location for new variables that are created by data utilities. The options include:
  - **Insert in Source Data Set**, where each new variable created is inserted at the right-hand side (or bottom) of the data set containing the original variable.
  - **Create New Data Set**, where new variables will be placed in a new data set

  There are, however, cases when the selected New Variable Preference may not be followed. These include **Stack and Unstack** (where new variables will always be placed in a new data set) and when the **original data is from a multi-range data set** (where new variables can only be inserted in the source data set)
- **Updating Preference –** specifies how the values for new variables created by a utility will change when original variable data is changed. Options for updating include:
  - **Live - Values Change With Input Data,** where new variable values update automatically as input data changes.
  - **Static – Values are Fixed,** where new variable values do not change with changing input data. New variable values are fixed based on the input data values when the procedure was run.

  There are, however, cases when the selected Updating Preference will not be followed and new variables will always use the **Static – Values are Fixed** preference. These include **Stack and Unstack** and **Random Samples**. Live updating is not applicable to these utilities.

Data Sets Settings

The **Data Sets Settings** specify options for new data sets created using the Data Set Manager command. These settings are just defaults that appear when you create a new data set. They can be changed if desired in the Data Set Manager dialog box.

- **Apply Cell Formatting –** Specifies if the data set will be formatted by StatTools.
- **Layout –** Selects the default variable layout (row-wise or column-wise)
- **Names in First Column/Row (Primary Range) –** Specifies if names are entered in the first column or row for the first range defined.
- **Names in First Column/Row (Secondary Range) –** Specifies if names are entered in the first column or row of the second and all subsequent ranges for a multi-range data set.

Analyses Settings

The **Analyses** Settings specify the default entries displayed in dialog boxes used to set up analyses. These are just the entries that appear when you first display a dialog box for an analysis. They can be changed if desired in each dialog box.

- **Warning Messages –** Selects whether warning messages will be displayed when running an analysis if StatTools detects missing data in a variable or if StatTools detects non-numeric data.
- **Dialog Memory –** Specifies the default entries in dialog boxes used to set up analyses. Options include:
    - **Remember Last Used Values (by Workbook)**, where a displayed dialog box shows the entries that were made in that dialog the last time it was displayed for the active workbook. If a dialog is displayed for the first time, it shows the saved system default settings for the analysis.
    - **Always Use System Default Values**, where a displayed dialog box shows the saved system default settings for the analysis.

- **Percentile Calculations** – Selects the method to be used for calculating percentiles. Depending on the nature of your data, different methods can give better answers. Available methods (and the type of data they are suited for) include:

    1) Automatic (Based on Input Data)
    2) Interpolated with Asymmetric Endpoints (Continuous)
    3) Interpolated with Symmetric Endpoints (Continuous)
    4) Excel Percentile Function (Continuous)
    5) Closest Observation (Discrete)
    6) Empirical Dist. Function (Discrete)
    7) Empirical Dist. Function with Averaging (Discrete)

Delete Data Sets Command

Deletes StatTools data sets in the active workbook

The **Delete Data Sets** command deletes all defined data sets from the active workbook. The actual data in Excel is not deleted; just the definition of the data sets.

Clear Dialog Memory Command

Clears all memory of entries in analysis dialog boxes

The **Clear Dialog Memory** command clears all "memory" of entries for analysis dialog boxes. Subsequent displayed dialog boxes will initially show the saved system default settings for each analysis.

Unload StatTools Add-in Command

Unloads the StatTools Add-in

The Unload StatTools Add-in Command unloads StatTools, closing all StatTools windows.

Help Menu

StatTools Help

Opens on-line help file for StatTools

The Help menu StatTools Help command opens the main help file for StatTools. All of StatTools's features and commands are described in this file.

Online Manual

Opens online manual for StatTools

The Help menu Online Manual command opens this manual in PDF format. You must have Adobe Acrobat reader installed to view the online manual.

License Activation Command

Displays licensing information for StatTools and allows the licensing of trial versions

The Help menu License Activation command displays the License Activation dialog box, listing the version and licensing information for your copy of StatTools. Using this dialog box you can also convert a trial version of StatTools into an licensed copy.

For more information on licensing your copy of StatTools, see **Chapter 1: Getting Started** in this manual.

About Command

Displays version and copyright information about StatTools

The Help menu About command displays the About dialog box, listing the version and copyright information for your copy of StatTools.

Reference: StatTools Functions

Introduction

Custom worksheet functions are used by StatTools to return calculated statistics to Excel formulas. These functions allow:

1) Statistics calculations to be embedded in worksheet formulas, just as are standard Excel functions.
2) Statistics to be "live", that is, results change when original data changes.

If you look at the formulas in the cells in a StatTools report, you will see StatTools worksheet functions. All StatTools functions begin with the prefix "Stat", such as **StatMean()** or **StatStdDev().** All StatTools functions are displayed in the Excel Insert Function dialog for ease of entry.

StatTools Functions vs. Excel Functions

In some cases StatTools replaces Excel's built-in statistics with its own robust and fast calculations. The accuracy of Excel's built-in statistics calculations has often been questioned, and StatTools uses none of them! Even Excel's worksheet statistics functions – such as STDEV() – are replaced by new, robust StatTools versions – such as StatSTDEV(). StatTools statistics calculations meet the highest tests for accuracy, with performance optimized through the use of C++ .DLLs, not macro calculations.

StatTools functions, as opposed to the built-in Excel functions, support the use of stacked data. The StatTools **StatDestack** function automatically unstacks data from a stacked data set (for a category you specify). It then passes this data to a StatTools statistics function for analysis.

StatTools functions also support the analysis of data that resides on different worksheets. Multi-sheet data sets allow more than 65535 points per variable. They are entered using the **Multiple** button in the Data Set Manager dialog.

Distribution Functions

StatTools includes a set of distribution functions (such as **StatBinomial**) which replace Excel's built-in distribution functions (such as **BinomDist**). Unlike Excel's distribution functions, the StatTools distribution functions can return a number of different values from a probability distribution. The value returned is set by the **statistic** argument (the second to last argument in the function). This argument can be a value 1 to 12 or a string that indicates the statistic you want to get for the entered distribution:

List of Possible Returned Statistics

Entered Value or String	Returned Statistic
1 or "mean"	mean
2 or "stddev"	standard deviation
3 or "variance"	variance
4 or "skewness"	skewness
5 or "kurtosis"	kurtosis
6 or "mode"	mode
7 or "discrete mean"	discrete mean (or the value closest to the true mean that actually could occur)
8 or "x to y"	x to y (the distribution y-value for an entered x-value)
9 or "x to p"	x to p (the distribution p-value for an entered x-value)
10 or "p to x"	p to x (the distribution x-value for an entered p-value)
11 or "x to q"	x to q (the distribution q-value for an entered x-value)
12 or "q to x"	q to x (the distribution x-value for an entered q-value)

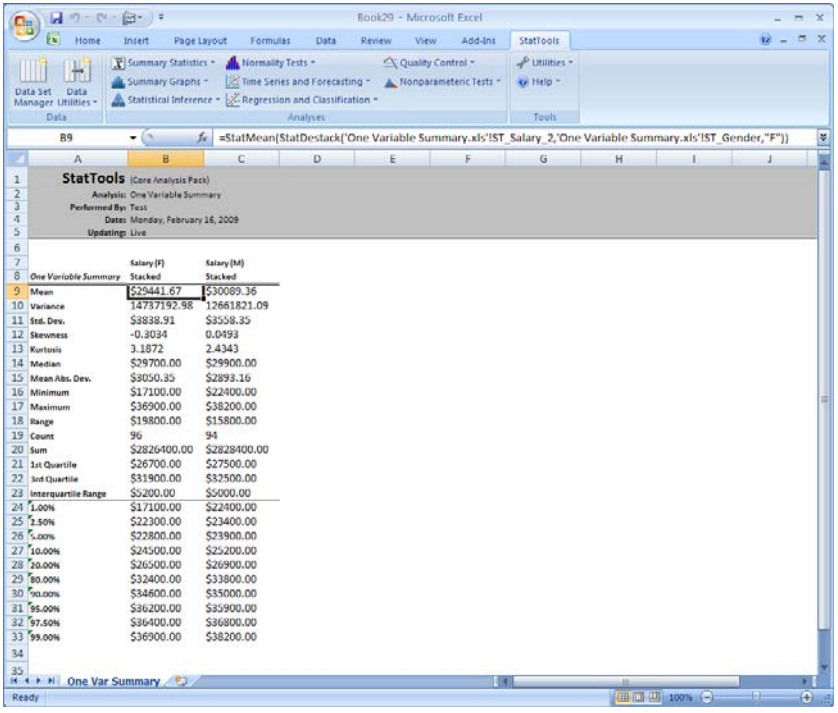
For example, the StatTools distribution function:

StatNormal(10,1,"x to p", 9.5)

Returns the p value associated with the x value of 9.5 in a normal distribution with a mean of 10 and a standard deviation of 1.

"Live" Reports

StatTools uses custom functions to make the results as "live" as possible. That is, whenever it is practical, reports have formulas that link to the original data. For example, suppose you have a variable *Weight* and you want summary measures on *Weight*, such as its mean and standard deviation. The Summary Statistics procedure names the range of weights as Weight, and then it enters formulas in the output cells: **=StatMean(Weight)** and **=StatStdDev(Weight).** StatMean and StatStdDev are built in StatTools functions for calculating mean and standard deviation. These replace the standard built-in Excel functions for the same statistics. Because of these functions, when your data change, the results change automatically, so that you don't have to rerun the procedure.



There are times when it is not practical to link results to data. The prime example is regression. StatTools does not provide the formulas that are used to create regression output; it provides only the numerical results. In such cases, if your data change, you will have to rerun the procedures.

You can use the **Reports Settings command Static option** to not have your reports linked to your data. This is useful if Excel recalculation time becomes an issue as data changes.

Reference: Listing of Statistics Functions

Table of Available Functions

This table lists the custom functions that are added to Excel by StatTools. When used, all functions are preceded by the entry **Stat**.

Function	Returns
AUTOCORRELATION*(data ,numLags)*	Calculates the autocorrelation for the values in a data set
AVEDEV*(Data1,Data2,...DataN)*	Calculates the average absolute deviation of the data from their mean. The arguments can be numbers, arrays, or ranges.
BINOMIAL*(N,P,statistic,value)*	Calculates the *statistic* for the entered binomial distribution
CATEGORYINDICIES*(range, category_name)*	Gets the cell indices for a specified category
CATEGORYNAMES*(range)*	Gets the names of the categories in a range
CATEGORYOCCURRENCECOUNT*(range,category_name)*	Calculates the number of cells in a range in a specified category
CHISQ*(deg_freedom,statistic,value)*	Calculates the *statistic* for the entered one-tailed chi-squared distribution
CORRELATIONCOEFF*(data1, data2)*	Calculates the correlation coefficient between 2 data sets
COUNT*(Data1,Data2,...DataN)*	Calculates the number of elements in its arguments, which can be numbers, arrays, or ranges
COUNTCATEGORIES*(range)*	Counts the number of the categories in a range
COUNTCELLSBYTYPE*(range, type)*	Calculates the number of cells of a specified type in a range
COUNTRANGE*(Range,MinValue,MaxValue ,IncludeMin,IncludeMax)*	Calculates the number of values in the *Range* that fall between *MinValue* and *MaxValue*
COVARIANCE*(data1,data2)*	Calculates the sample covariance between 2 data sets
COVARIANCEP*(data1,data2)*	Calculates the population covariance between 2 data sets. Any missing numbers cause a blank to be returned.

Function	Returns
DESTACK*(data_range, categories_range_1,category_1, categories_range_2,category_2)*	Extracts data in a specified category from stacked data
DURBINWATSON*(data)*	Calculates the Durbin-Watson statistic for the values in a data set
F*(deg_freedom1,deg_freedom2,statistic, value)*	Calculates the *statistic* for the entered F distribution for 2 data sets
GETCELLVALUES*(range)*	Gets the values for all the cells of a specified type in the range
KURTOSIS*(Data1,Data2,... DataN)*	Calculates the sample kurtosis of its arguments, which can be numbers, arrays, or ranges
KURTOSISP*(Data1,Data2,... DataN)*	Calculates the population kurtosis of its arguments, which can be numbers, arrays, or ranges
LN*(x)*	Calculates the natural logarithm of a positive, real number
MAX*(Data1,Data2,...DataN)*	Calculates the maximum of its arguments, which can be numbers, arrays, or ranges
MEAN*(Data1,Data2,...DataN)*	Calculates the arithmetic mean (average) of its arguments, which can be numbers, arrays, or range
MEANABS*(Data1,Data2,...DataN)*	Calculates the arithmetic mean (average) of the absolute values of its arguments, which can be numbers, arrays, or range
MEDIAN*(data,discrete flag)*	Calculates the median of a data set
MIN*(Data1,Data2,...DataN)*	Calculates the minimum of its arguments, which can be numbers, arrays, or ranges
MODE*(Data1,Data2,...DataN)*	Calculates the mode of its arguments, which can be numbers, arrays, or ranges
NORMAL*(mean,std_dev, statistic, value)*	Calculates the *statistic* for the entered normal (Gaussian) distribution
PAIRCOUNT *(Data1,Data2)*	Counts the number of pairs of cells for which each of the cells in the pair is numeric
PAIRMEAN*(Data1,Data2)*	Calculates the mean of the differences between pairs of cells
PAIRMEDIAN*(Data1,Data2)*	Calculates the median of the differences between pairs of cells
PAIRSTDDEV*(Data1,Data2)*	Calculates the sample standard deviation of the differences between pairs of cells
PERCENTILE *(data,p, discreteFlag)*	Calculates the p-th percentile of a data set
PRODUCT*(Data1,Data2,…DataN)*	Calculates the product of its arguments, which can be numbers, arrays, or ranges

Function	Returns
QUARTILE *(data,q,discrete flag)*	Calculates the specified quartile of a data set
RAND*()*	Returns a random number in the range 0 – 1.
RANGE *(Data1,Data2,...DataN)*	Calculates the range (maximum - minimum) of its arguments, which can be numbers, arrays, or ranges
RUNSTEST *(data,cutoff)*	Calculates the runs statistic for the values in a data set
SKEWNESS*(Data1,Data2,...Data N)*	Calculates the sample skewness of its arguments, which can be numbers, arrays, or ranges
SKEWNESSP*(Data1,Data2,...Data N)*	Calculates the population skewness of its arguments, which can be numbers, arrays, or ranges
STANDARDIZE*(x,mean,std_dev)*	Calculates a normalized value from a distribution with the specified mean and standard deviation
STDDEV *(Data1,Data2,...DataN)*	Calculates the sample standard deviation of its arguments, which can be numbers, arrays, or ranges
STDDEVP*(Data1,Data2,...DataN)*	Calculates the population standard deviation of its arguments, which can be numbers, arrays, or ranges
SUM*(Data1,Data2,...DataN)*	Calculates the sum of its arguments, which can be numbers, arrays, or ranges
SUMDEVSQ*(Data1,Data2,...Data N)*	Calculates the sum of the square of the deviation from the mean of its arguments, which can be numbers, arrays, or ranges
SUMSQ*(Data1,Data2,...DataN)*	Calculates the sum of the square of its arguments, which can be numbers, arrays, or ranges
STUDENT*(deg_freedom,statistic, value)*	Calculates the *statistic* for the entered Student's t-distribution
VARIANCE*(Data1,Data2,...Data N)*	Calculates the sample variance of its arguments, which can be numbers, arrays, or ranges
VARIANCEP*(Data1,Data2,...Data N)*	Calculates the population variance of its arguments, which can be numbers, arrays, or ranges

Detailed Function Descriptions

Statistics functions are listed here with their required arguments.

AUTOCORRELATION

Description	**Autocorrelation** (*data,numLags*) calculates the autocorrelation for the values in a data set data using the number of lags *numLags*. *data* is the array or range of data to calculate the autocorrelation for. *numLags* is the number of lags to use.
Examples	**StatAutocorrelation** (*C1:C100,1*) returns the autocorrelation for the data in the range *C1:C100* with 1 *lag*.
Guidelines	*numLags* must greater than or equal to 1.

AVEDEV

Description	**AVEDEV**(*Data1,Data2,...DataN*) calculates the average absolute deviation of *Data1,Data2,...DataN* from their mean.
Examples	**StatAveDev**(*1,2,5*) calculates the average absolute deviation of the 1,2 and 5 from their mean.
Guidelines	*Data1,Data2,...DataN* arguments can be numbers, arrays, or ranges.

BINOMIAL

Description	**BINOMIAL**(*N,P,statistic,value*) calculates the *statistic* for the binomial distribution which has the specified *N* and *P* values.
Examples	**StatBinomial**(*2,.3,"x to p",1*) calculates the p-value for an x-value of 1 in the binomial distribution with N=2 and P =.3
Guidelines	*N* is the number of trials or events and must be an integer >0 *P* is the probability and must be >=0 and <=1. *statistic* is an integer value 1 to 12 or a string indicating the statistic to be returned. For more information see the section Distribution Functions at the start of this chapter. *value* must be >= 0 (when an x value is entered) or >=0 and <=1 (when a p value is entered).

CATEGORYINDICIES

Description	**CATEGORYINDICES**(*range, category_name*) gets the indices for the cells containing the specified *category_name* with the Excel *range*. This is an array function and the returned indices are index numbers (between 1 and # of cells in *range*) giving the positions of the cells containing *category_name* within the *range.*
Examples	**StatCategoryIndices** (C1:C100,"Male") returns the indexes (between 1 and 100) of the cells containing the string Male.
Guidelines	*range* is a valid Excel range *category_name* is a string, value or cell reference specifying the category to find

CATEGORYNAMES

Description	**CATEGORYNAMES**(*range*) gets the names of the categories in the specified Excel *range*. This is an array function and the number of returned names is between 1 and # of cells in *range.*
Examples	**StatCategoryNames**(C1:C100) returns the names of the categories in the range C1:C100.
Guidelines	*range* is a valid Excel range

CATEGORYOCCURRENCECOUNT

Description	**CATEGORYOCCURRENCECOUNT**(*range, category_name*) returns the number of cells containing the specified *category_name* in the Excel *range*.
Examples	**StatCategoryOccurrenceCount**(C1:C100,"Male") returns the number of cells in the range C1:C100 that contain "Male".
Guidelines	*range* is a valid Excel range *category_name* is a string, value or cell reference specifying the category to find

CHISQ

Description	**CHISQ**(*deg_freedom,statistic,value*) calculates the *statistic* for the one-tailed chi-squared distribution using the specified degrees of freedom *deg_freedom*.
Examples	**StatChiDist**(*2,"x to p",5*) calculates the one-tailed chi-squared distribution at the value 5 with 2 degrees of freedom
Guidelines	the number of degrees of freedom *deg_freedom* must be in the range 1-32767 *statistic* is an integer value 1 to 12 or a string indicating the statistic to be returned. For more information see the section Distribution Functions at the start of this chapter. *value* must be >= 0 (when an x value is entered) or >=0 and <=1 (when a p value is entered)

CORRELATION COEFF

Description	**CORRELATIONCOEFF**(*data1,data2*) calculates the correlation coefficient between two data sets *data1* and *data2*.
Examples	**StatCorrelationCoeff**(*A1:A100,B1:B100*) calculates the correlation coefficient between two data sets located in *A1:A100* and *B1:B100*.
Guidelines	*data1* and *data2* must have the same number of elements.

COUNT

Description	**COUNT** (*data1,data2,...dataN*) calculates the number of elements in *data1,data2* through *dataN* , which can be numbers, arrays, or ranges.
Examples	**StatCount**(*A1:A100,B1:B100*) calculates the number of elements in the two data sets located in *A1:A100* and *B1:B100*.
Guidelines	*data1,data2,...dataN* are 1 to 30 arguments, which can be numbers, arrays, or ranges.

COUNTCATEGORIES

Description	**COUNTCATEGORIES**(*range*) returns the number of categories in the specified Excel *range*.
Examples	**StatCountCategories**(C1:C100) returns the number of the categories in the range C1:C100.
Guidelines	*range* is a valid Excel range

COUNTCELLSBYTYPE

Description	**COUNTCELLSBYTYPE** (*range, type*) calculates the number of elements in the entered *range* which are of the specified *type*.
Examples	**StatCountCellByType**(*A1:A100,1*) calculates the number of elements in the data set located in *A1:A100* that are numeric.
Guidelines	*range* is a valid Excel range *type* is 1=numeric, 2=non-empty, 3=non-empty, non-numeric, 4=empty. Note: StatTools considers a cell containing only spaces to be an empty cell.

COUNTRANGE

Description	**COUNTRANGE** (*range,minValue,maxValue, includeMin,includeMax*) calculates number of values in the r*ange* that fall between *minValue* and *maxValue*. Values equaling *minValue* and *MaxValue* can be included by setting *IncludeMin* and/or *IncludeMax* to TRUE.
Examples	**StatCountRange**(*A1:A100,1,10,TRUE,TRUE*) calculates the number of values in the data set located in *A1:A100* that fall in the range >=1 and <=10.
Guidelines	*range* is the range of cells from which you want to count values. *minValue* is the minimum value in the range. *maxValue* is the maximum value in the range. *includeMin* is a boolean value indicating if the MinimumValue should be included in the count. Default is TRUE. *includeMax* is a boolean value indicating if the MaximumValue should be included in the count. Default is TRUE.

COVARIANCE

Description	**COVARIANCE**(*data1,data2*) calculates the sample covariance between the data sets *data1* and *data2*.
Examples	**StatCovariance**(*A1:A100,B1:B100*) calculates the sample covariance between two data sets located in *A1:A100* and *B1:B100*.
Guidelines	*data1* and *data2* can be arrays or ranges. *data1* and *data2* must have the same number of elements.

COVARIANCEP

Description	**COVARIANCEP**(*data1,data2*) calculates the population covariance between 2 data sets *data1* and *data2*.
Examples	**StatCovarianceP**(*A1:A100,B1:B100*) calculates the population covariance between two data sets located in *A1:A100* and *B1:B100*.
Guidelines	*data1* and *data2* can be arrays or ranges.

DESTACK

Description	**DESTACK**(*data_range,categories_range_1,category_1,categories_range_2,category_2*) extracts data in a specified category *category_1* from the stacked data in *data_range*. This function allows other StatTools statistics functions to take stacked data as input. Thus, the StatDestack function will only be seen as "embedded" in other functions, as shown in the example here. It returns an array of data for the specified category, extracted from *data_range*.
Examples	StatMean(**StatDestack**(B1:B100,A1:A100,"Male")) calculates the mean for the values in the range B1:B100 where the corresponding category range A1:A100 has the value "Male".
Guidelines	*data_range* is the range with the stacked data. *categories_range_1* is the range with the first category name. *category_1* is the first category for which to get the data. *categories_range_2 (optional)* is the range with the second category name. *category_2 (optional)* is the second category for which to get the data.

DURBINWATSON

Description	**DURBINWATSON**(*data*) calculates the Durbin-Watson statistic for the values in the data set *data*.
Examples	**StatDurbinWatson**(*A1:A100*) calculates the Durbin-Watson statistic for the range of data located in A1:A100.
Guidelines	*data* can be an array or range of data.

F

Description	**F**(*deg_freedom1,deg_freedom2,statistic,value*) calculates the *statistic* for the F distribution using the numerator degrees of freedom *deg_freedom1* and the denominator degrees of freedom *deg_freedom2.*
Examples	**StatF** (1,1,"x to p",1.5) calculates the F Distribution p value for an x value of 1.5 with a numerator degrees of freedom =1 and a denominator degrees of freedom =1 .
Guidelines	*deg_freedom1* and *deg_freedom2* must be an integer > 0. *statistic* is an integer value 1 to 12 or a string indicating the statistic to be returned. For more information see the section Distribution Functions at the start of this chapter. *value* must be >= 0.

GETCELLVALUES

Description	**GETCELLVALUES** (*range,typeOfCell*) gets the values for all the cells of a specified type in the range
Examples	**StatGetCellValues**(*A1:A100,2*) gets the values for non-empty cells in the range located in *A1:A100*.
Guidelines	*range* is the range of cells from which you want to get the values. *typeOfCell* is 0 for all, 1 for numeric, 2 for non-empty, 3 for non-empty, non-numeric, and 4 for empty cells.

KURTOSIS

Description	**KURTOSIS**(*Data1,Data2,...DataN*) calculates the sample kurtosis of the data specified in *Data1,Data2,...DataN.* Note: StatKurtosis calculated on normally distributed data returns the value 3.
Examples	**StatKurtosis**(*A1:A100,{1;2;3;2.4}*) calculates the sample kurtosis of the data set located in *A1:A100* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

KURTOSISP

Description	**KURTOSISP**(*Data1,Data2,...DataN*) calculates the population kurtosis of the data specified in *Data1,Data2,...DataN.* Note: StatKurtosisP calculated on normally distributed data returns the value 3.
Examples	**StatKurtosisP**(*A1:A100,{1;2;3;2.4}*) calculates the population kurtosis of the data set located in *A1:A100* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

LN

Description	**LN**(*x*) calculates the natural logarithm.
Examples	**StatLN**(*4.5)* calculates natural logarithm of 4.5.
Guidelines	*x* must be a positive, real number.

MAX

Description	**MAX**(*Data1,Data2,...DataN*) calculates the maximum of the data specified in *Data1,Data2,...DataN*.
Examples	**StatMax**(*A1:A100,{1;2;3;2.4}*) calculates the maximum value in the data set located in *A1:A100* and the values *1,2,3* and *2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

MEAN

Description	**MEAN**(*Data1,Data2,...DataN*) calculates the mean of the data specified in *Data1,Data2,...DataN*.
Examples	**StatMean**(*A1:A100,{1;2;3;2.4}*) calculates the mean value in the data set located in *A1:A100* and the values *1,2,3* and *2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

MEANABS

Description	**MEANABS**(*Data1,Data2,...DataN*) calculates the mean of the absolute value of the data specified in *Data1,Data2,...DataN*.
Examples	**StatMeanAbs**(*A1:A100,{1;2;3;2.4}*) calculates the mean of the absolute values of the data set located in *A1:A100* and the values *1,2,3* and *2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

MEDIAN

Description	**MEDIAN**(*data,calcFlag*) calculates the median of the values located in *data*. It may be calculated using any of five alternative methods, as optionally specified by *calcFlag*.
Examples	**StatMedian**(*A1:A100,1*) calculates the median value in the data set located in *A1:A100*. Data is continuous.
Guidelines	*data* is an Excel range. *calcFlag* is an optional argument that can take an integer value in the range -1 to 5. This value corresponds with the desired method for calculating the percentile. -1 or not specified) Automatic (Based on Input Data) 0) Same as Excel's Percentile Function (Continuous) 1) Interpolated with Asymmetric Endpoints (Continuous) 2) Closest Observation (Discrete) 3) Empirical Dist. Function (Discrete) 4) Interpolated with Symmetric Endpoints (Continuous) 5) Empirical Dist. Function with Averaging (Discrete)

MIN

Description	**MIN**(*Data1,Data2,...DataN*) calculates the minimum of the data specified in *Data1,Data2,...DataN*.
Examples	**StatMin**(*A1:A100,{1;2;3;2.4}*) calculates the minimum value in the data set located in *A1:A100* and the values *1,2,3* and *2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges

MODE

Description	**MODE**(*Data,Is_discrete*) calculates the mode of a data set.
Examples	**StatMode**(*A1:A100,FALSE*) calculates the mode of the data set located in A1:A100. Data is continuous.
Guidelines	*Data* is the array or range of data to calculate mode for. Is*_discrete* is an optional argument, and specifies whether the data is to be treated as discrete (true), or continuous (false). If missing, it is automatically determined from the data.

NORMAL

Description	**NORMAL**(*mean,std_dev,statistic,value*) calculates the *statistic* for the normal distribution specified by *mean* and *std_dev.*
Examples	**StatNormal**(*2,1,"x to p",3*) calculates the p value for an x value of 3 in the normal distribution with mean=2 and std dev =1
Guidelines	*mean* is the arithmetic mean of the distribution *std_dev* is the standard deviation of the distribution. It must be > 0. *statistic* is an integer value 1 to 12 or a string indicating the statistic to be returned. For more information see the section Distribution Functions at the start of this chapter. *value* must be >=0 and <=1 when a p value is entered.

PAIRCOUNT

Description	**PAIRCOUNT**(Data1,Data2) counts the number of pairs of cells in Data1 and Data. Only pairs of numeric cells are counted. A missing value in either range will not be counted.
Examples	**StatPairCount** (*A1:A100,B1:B100)* counts the number of pairs of numeric cells in the data sets located in *A1:A100* and *B1:B100.*
Guidelines	*Data1* and *Data2* must be equal sized Excel ranges. Pairs of cells from *Data1* and *Data2* are selected on a row-wise basis, starting from the top left.

PAIRMEAN

Description	**PAIRMEAN**(*Data1,Data2*) calculates the mean of the differences between pairs of cells in *Data1* and *Data2*. Calculations are only made for pairs of numeric cells. A missing value in either range will not be counted.
Examples	**StatPairMean**(*A1:A100,B1:B100)* calculates the mean of the differences between pairs of numeric cells in the data sets located in A1:A100 and B1:B100.
Guidelines	*Data1* and *Data2* must be equal sized Excel ranges. Pairs of cells from *Data1* and *Data2* are selected on a row-wise basis, starting from the top left.

PAIRMEDIAN

Description	**PAIRMEDIAN**(*Data1,Data2,method_Flag*) calculates the median of the differences between pairs of cells in *Data1* and *Data2*. Calculations are only made for pairs of numeric cells. A missing value in either range will not be counted. Median may be calculated using any of five alternative methods, as optionally specified by *method_Flag*.
Examples	**StatPairMedian**(*A1:A100,B1:B100)* calculates the median of the differences between pairs of numeric cells in the data sets located in A1:A100 and B1:B100.
Guidelines	*Data1* and *Data2* must be equal sized Excel ranges. Pairs of cells from *Data1* and *Data2* are selected on a row-wise basis, starting from the top left. *method_Flag* is an optional argument that can take an integer value in the range -1 to 5. This value corresponds with the desired method for calculating the median. -1 or not specified) Automatic (Based on Input Data) 0) Same as Excel's Percentile Function (Continuous) 1) Interpolated with Asymmetric Endpoints (Continuous) 2) Closest Observation (Discrete) 3) Empirical Dist. Function (Discrete) 4) Interpolated with Symmetric Endpoints (Continuous) 5) Empirical Dist. Function with Averaging (Discrete)

PAIRSTDDEV

Description	**PAIRSTDDEV**(*Data1,Data2*) calculates the sample standard deviation of the differences between pairs of cells in *Data1* and *Data2*. Calculations are only made for pairs of numeric cells. A missing value in either range will not be counted.
Examples	**StatPairStdDev**(*A1:A100,B1:B100)* calculates the sample standard deviation of the differences between pairs of numeric cells in the data sets located in A1:A100 and B1:B100.
Guidelines	*Data1* and *Data2* must be equal sized Excel ranges. Pairs of cells from *Data1* and *Data2* are selected on a row-wise basis, starting from the top left.

PERCENTILE

Description	**PERCENTILE** (*data,p,method_Flag*) calculates the *p*-th percentile of *data*. Percentiles may be calculated using any of five alternative methods, as optionally specified by *method_Flag*.
Examples	**StatPercentile** (*A1:A100,.15,0)* calculates the 15th percentile for the data located in *A1:A100*. The data is continuous and uses the weighted average percentile calculation method.
Guidelines	*p* must be 0-1, inclusive *method_Flag* is an optional argument that can take an integer value in the range -1 to 5. This value corresponds with the desired method for calculating the percentile. -1 or not specified) Automatic (Based on Input Data) 0) Same as Excel's Percentile Function (Continuous) 1) Interpolated with Asymmetric Endpoints (Continuous) 2) Closest Observation (Discrete) 3) Empirical Dist. Function (Discrete) 4) Interpolated with Symmetric Endpoints (Continuous) 5) Empirical Dist. Function with Averaging (Discrete)

PRODUCT

Description	**PRODUCT**(*Data1,Data2,...DataN*) calculates the product of the data specified in *Data1,Data2,...DataN.*
Examples	**StatProduct**(*A1:A10,{1;2;3;2.4}*) calculates the product of all value in the data set located in *A1:A10* and the values *1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

QUARTILE

Description	**QUARTILE** (*data,q, method_Flag*) calculates the specified quartile of *data.* Quartiles may be calculated using any of five alternative methods, as optionally specified by *method_Flag.*
Examples	**StatQuartile** (*A1:A100,1,FALSE)* calculates the 1st quartile for the data located in A1:A100. The data is continuous.
Guidelines	*data* must be an Excel range *Q* is the quartile; 0=minimum, 1=1st quartile, 2=2nd quartile (median), 3=3rd quartile, 4=maximum. *method_Flag* is an optional argument that can take an integer value in the range -1 to 5. This value corresponds with the desired method for calculating the percentile. -1 or not specified) Automatic (Based on Input Data) 0) Same as Excel's Percentile Function (Continuous) 1) Interpolated with Asymmetric Endpoints (Continuous) 2) Closest Observation (Discrete) 3) Empirical Dist. Function (Discrete) 4) Interpolated with Symmetric Endpoints (Continuous) 5) Empirical Dist. Function with Averaging (Discrete)

RAND

Description	**RAND**() returns a random number in the range 0 to 1. This function uses the random number generator from Palisade's @RISK product and not Excel's built-in random number generator.
Examples	**StatRand**() returns a random number in the range 0 to 1.

RANGE

Description	**RANGE** (*Data1,Data2,...DataN*) calculates the range (maximum - minimum) of the data specified in *Data1,Data2,...DataN*.
Examples	**StatRange**(*A1:A100,{1;2;3;2.4}*) calculates the range (maximum - minimum) of the data located in *A1:A100* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

RUNSTEST

Description	**RUNSTEST** (*data,cutoff*) calculates the runs statistic for the values in *data* using *cutoff* value.
Examples	**StatRunsTest**(*A1:A100,StatMean(A1:A100)*) calculates the runstest statistic on the data in *A1:A100* using the mean of the data as the cutoff value.
Guidelines	*data* must be an Excel range.

SKEWNESS

Description	**SKEWNESS**(*Data1,Data2,...DataN*) calculates the sample skewness of the data specified in *Data1,Data2,...DataN*.
Examples	**StatSkewness**(*A1:A10,{1;2;3;2.4}*) calculates the sample skewness of all value in the data set located in *A1:A10* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

SKEWNESSP

Description	**SKEWNESSP**(*Data1,Data2,...DataN*) calculates the population skewness of the data specified in *Data1,Data2,...DataN*.
Examples	**StatSkewnessP**(*A1:A10,{1;2;3;2.4}*) calculates the population skewness of all value in the data set located in *A1:A10* and the values *1,2,3* and *2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

STANDARDIZE

Description	**STANDARDIZE**(*x,mean,std_dev*) calculates a normalized value from a distribution with the specified *mean* and standard deviation *std_dev*. *x* is the value to be normalized.
Examples	**StatStandardize**(2,1,3) calculates a normalized value at the value 2 from a distribution with a mean of 1 and a standard deviation of 3.
Guidelines	*x* is the value to be normalized *mean* is the arithmetic mean of the distribution. *std_dev* is the standard deviation of the distribution. It must be > 0.

STDDEV

Description	**STDDEV** (*Data1,Data2,...DataN*) calculates the sample standard deviation of the data specified in *Data1,Data2,...DataN*.
Examples	**StatStdDev**(*A1:A10,{1;2;3;2.4}*) calculates the sample standard deviation of all value in the data set located in *A1:A10* and the values *1,2,3* and *2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

STDDEVP

Description	**STDDEVP** (*Data1,Data2,...DataN*) calculates the population standard deviation of the data specified in *Data1,Data2,...DataN*.
Examples	**StatStdDevP**(*A1:A10,{1;2;3;2.4}*) calculates the population standard deviation of all value in the data set located in *A1:A10* and the values *1,2,3* and *2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

SUM

Description	**SUM**(*Data1,Data2,...DataN*) calculates the sum of the data specified in *Data1,Data2,...DataN*.
Examples	**StatSum**(*A1:A10,{1;2;3;2.4}*) calculates the sum of all value in the data set located in *A1:A10* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

SUMDEVSQ

Description	**SUMDEVSQ**(*Data1,Data2,...DataN*) calculates the sum of the square of the deviation from the mean of its arguments, which can be numbers, arrays, or ranges. Any missing numbers cause a blank to be returned.
Examples	**StatSumDevSq**(*A1:A10,{1;2;3;2.4}*) calculates the sum of the square of the deviation from the mean of all values in the data set located in *A1:A10* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

SUMSQ

Description	**SUMSQ**(*Data1,Data2,...DataN*) calculates the sum of the square of its arguments, which can be numbers, arrays, or ranges. Any missing numbers cause a blank to be returned.
Examples	**StatSumSq**(*A1:A10,{1;2;3;2.4}*) calculates the sum of the square of all values in the data set located in *A1:A10* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

STUDENT

Description	**STUDENT**(*deg_freedom*, *statistic,value*) calculates the *statistic* for the entered Student's t-distribution.
Examples	**StatStudent**(5,*1,"x to p",2*) calculates the p-value from the Student's t-distribution with 5 degrees of freedom at the x-value of 2.
Guidelines	*deg_freedom* is an integer indicating the number of degrees of freedom. It must be in the range 1-32767. *statistic* is an integer value 1 to 12 or a string indicating the statistic to be returned. For more information see the section Distribution Functions at the start of this chapter. *x* is the numeric value at which to evaluate the distribution. It must be >= 0.

VARIANCE

Description	**VARIANCE** (*Data1,Data2,...DataN*) calculates the sample variance of the data specified in *Data1,Data2,...DataN.*
Examples	**StatVariance**(*A1:A10,{1;2;3;2.4}*) calculates the sample variance of all value in the data set located in *A1:A10* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

VARIANCEP

Description	**VARIANCEP** (*Data1,Data2,...DataN*) calculates the population variance of the data specified in *Data1,Data2,...DataN*.
Examples	**StatVarianceP**(*A1:A10,{1;2;3;2.4}*) calculates the sample variance of all value in the data set located in *A1:A10* and *the values 1,2,3 and 2.4*.
Guidelines	*Data1,Data2,...DataN* can be numbers, arrays, or ranges.

Index