Copyright Notice
Copyright © 2008-2015, Palisade Corporation

Trademark Acknowledgments
Microsoft, Excel and Windows are registered trademarks of Microsoft Corporation.
IBM is a registered trademark of International Business Machines, Inc.
Palisade, TopRank, BestFit and RISKview are registered trademarks of Palisade Corporation.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter 1: Welcome to StatTools</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work Where You're Comfortable ...</td>
<td>7</td>
</tr>
<tr>
<td>Robust Statistics inside Excel ...</td>
<td>7</td>
</tr>
<tr>
<td>StatTools Analyses ...</td>
<td>7</td>
</tr>
<tr>
<td>StatTools Data Management ...</td>
<td>8</td>
</tr>
<tr>
<td>StatTools Reporting ...</td>
<td>8</td>
</tr>
<tr>
<td>Data Access and Sharing ...</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 2: Getting Started</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction ...</td>
<td>11</td>
</tr>
<tr>
<td>About This Version ...</td>
<td>11</td>
</tr>
<tr>
<td>Working with your Operating Environment ...</td>
<td>11</td>
</tr>
<tr>
<td>If You Need Help ...</td>
<td>11</td>
</tr>
<tr>
<td>StatTools System Requirements ...</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Installation Instructions</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Installation Instructions ...</td>
<td>14</td>
</tr>
<tr>
<td>The DecisionTools Suite ...</td>
<td>14</td>
</tr>
</tbody>
</table>

| Software Activation | 15 |

<table>
<thead>
<tr>
<th>Instructional Materials</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example Spreadsheets ...</td>
<td>16</td>
</tr>
<tr>
<td>Quick Start Tutorials ...</td>
<td>16</td>
</tr>
<tr>
<td>Guided Tour Videos ...</td>
<td>16</td>
</tr>
<tr>
<td>Examples and Videos for the XDK</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 3: Overview of StatTools</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>StatTools Ribbon ...</td>
<td>17</td>
</tr>
<tr>
<td>Data Sets and the Data Manager ...</td>
<td>17</td>
</tr>
<tr>
<td>StatTools Reports and Charts ...</td>
<td>20</td>
</tr>
</tbody>
</table>
# Table of Contents

## Chapter 4: StatTools Command Reference

- **StatTools Icons** ........................................................................................................... 23
- **Table of StatTools Procedures** .................................................................................. 25

### Data Commands ............................................................................................................. 33
- Data Set Manager Command ......................................................................................... 33
- Data Viewer Command .................................................................................................. 37

### Data Utilities Menu ....................................................................................................... 43
- Stack Command .............................................................................................................. 44
- Unstack Command .......................................................................................................... 46
- Transform Command ...................................................................................................... 47
- Lag Command ................................................................................................................ 49
- Difference Command ..................................................................................................... 50
- Interaction Command .................................................................................................... 52
- Combination Command .................................................................................................. 54
- Dummy Command .......................................................................................................... 55
- Random Sample Command ............................................................................................ 57

### Summary Statistics Menu ............................................................................................ 59
- One-Variable Summary Command .................................................................................. 60
- Correlation and Covariance Command ......................................................................... 62

### Summary Graphs Menu .................................................................................................. 65
- Histogram Command ...................................................................................................... 66
- Scatterplot Command ..................................................................................................... 69
- Box-Whisker Plot Command ........................................................................................... 71
- Bar Chart Command ....................................................................................................... 73

### Statistical Inference Menu .......................................................................................... 77
- Confidence Interval - Mean/ Std. Deviation Command .................................................. 78
- Confidence Interval - Proportion Command ................................................................... 81
- Hypothesis Test - Mean/ Std. Deviation Command ....................................................... 85
- Hypothesis Test - Proportion Command ....................................................................... 88
- Sample Size Selection Command .................................................................................. 93
- One-Way ANOVA Command ......................................................................................... 95
- Two-Way ANOVA Command ......................................................................................... 99
- Chi-Square Independence Test Command ..................................................................... 102
- Chi-square Goodness-of-Fit Test Command ................................................................ 105

### Normality Tests Menu .................................................................................................. 109
Chi-Square Normality Test Command ........................................ 110
Lilliefors Test Command .......................................................... 113
Q-Q Normal Plot Command ....................................................... 115

Time Series and Forecasting Menu ........................................... 117
Time Series Graph Command .................................................... 118
Autocorrelation Command ....................................................... 120
Runs Test for Randomness Command ....................................... 123
Forecasting Command .............................................................. 125

Regression and Classification Menu ........................................ 131
Regression Command .............................................................. 132
Logistic Regression Command .................................................. 142
Discriminant Analysis Command .............................................. 145

Quality Control Menu .............................................................. 149
Pareto Chart Command ............................................................ 150
X/R Charts Command .............................................................. 153
P Chart Command .................................................................... 157
C Chart Command .................................................................... 160
U Chart Command .................................................................... 163

Nonparametric Tests Menu ..................................................... 167
Sign Test Command ................................................................. 169
Wilcoxon Signed-Rank Test Command ..................................... 172
Mann-Whitney Test Command .................................................. 176
Kruskal-Wallis Test Command .................................................. 180

Multivariate Analysis Menu ..................................................... 185
Principal Components Analysis Command ............................. 186
Cluster Analysis Command ........................................................ 190

Utilities Menu .......................................................................... 199
Application Settings Command ................................................ 200
Delete Data Sets Command ....................................................... 205
Clear Dialog Memory Command .............................................. 205
Load DecisionTools Add-In Command ...................................... 205
Unload StatTools Add-In Command .......................................... 205

Help Menu ................................................................................ 207

Chapter 5: StatTools Functions Reference ................................. 209

Introduction .............................................................................. 209
StatTools Functions vs. Excel Functions ................................. 209
<table>
<thead>
<tr>
<th>Distribution Functions</th>
<th>210</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live Reports</td>
<td>211</td>
</tr>
<tr>
<td>Table of Available Functions</td>
<td>213</td>
</tr>
</tbody>
</table>

**Detailed Function Descriptions**

- StatAutocorrelation .................................................. 217
- StatAveDev .......................................................................... 217
- StatBinomial ....................................................................... 217
- StatCategoryIndices ...................................................... 218
- StatCategoryNames .......................................................... 218
- StatCategoryOccurrenceCount ........................................... 218
- StatChiSq ........................................................................... 219
- StatCorrelationCoeff ....................................................... 219
- StatCount .......................................................................... 219
- StatCountCategories ....................................................... 220
- StatCountCellsByType ..................................................... 220
- StatCountRange ................................................................... 220
- StatCovariance .................................................................... 220
- StatCovarianceP ............................................................... 221
- StatDestack ......................................................................... 221
- StatDurbinWatson ............................................................ 221
- StatF .................................................................................... 222
- StatGetCellValue ............................................................... 222
- StatKurtosis ......................................................................... 222
- StatKurtosisP ....................................................................... 223
- StatLN .................................................................................. 223
- Stat ...................................................................................... 223
- Max ...................................................................................... 223
- StatMean .............................................................................. 223
- StatMeanAbs ........................................................................ 223
- Stat ...................................................................................... 223
- Median .................................................................................. 224
- StatMin ................................................................................ 224
- StatMode .............................................................................. 224
- StatNormal .......................................................................... 225
- StatPairCount ................................................................. 225
- Stat ...................................................................................... 225
- PairMean ............................................................................. 225
- StatPairMedian ................................................................. 226
- Stat ...................................................................................... 226
- PairStdDev .......................................................................... 226
- StatPercentile ...................................................................... 227
Chapter 1: Welcome to StatTools

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, enabling 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 will be able to learn StatTools in a matter of minutes. StatTools works just as Excel does, with toolbars, menus, and custom worksheet functions, all inside Excel. Unlike stand-alone statistics software, there is no steep learning curve or upfront training costs with StatTools because you work just as you are used to working in Excel. Your data and variables are in Excel worksheets. 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 statistical calculations has often been questioned, and StatTools uses none of them. Even Excel's worksheet statistical functions, such as Stdev(), are replaced by new, robust StatTools versions, such as StatStdDev(). StatTools statistical 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. A total of over 30 wide-ranging statistical procedures plus eight 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. These provide a comprehensive statistics toolset, right inside of Excel.

StatTools features live, "hot-linked" statistical calculations. If you change a value in Excel, you expect your worksheet to recalculate and generate updated results. The same thing happens in StatTools. If you change a value in your data set, 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 data set and variable manager right in Excel, just as you would expect from a stand-alone statistics package. You can define any number of data sets, 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 data sets and variables can reside in different workbooks and worksheets, allowing you to organize your data as you prefer. Then, you can run statistical analyses that refer to your variables, instead of having to re-select your data multiple times. Because Excel 2007 and higher versions have over one million rows in a single worksheet, the number of cases you can analyze with StatTools Industrial Edition is essentially limited only by the amount of available memory. (StatTools Professional is limited to 10,000 cases.)
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. You can change report titles, number formats, and text just as in any standard Excel worksheet. You can 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 importing your existing data into StatTools is easy. You can use standard Excel capabilities to import data from Microsoft SQL Server, Oracle, Microsoft Access, or any other ODBC compliant database. You can also import data from text files or other applications. In short, if you can import 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.
Chapter 2: Getting Started

Introduction

This introduction describes the contents of the StatTools package and shows how to install StatTools and attach it to your copy of Microsoft Excel.

About This Version

This version of StatTools can be used with Microsoft Excel 2007 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, it assumes that:

- You are familiar with your computer and using the mouse.
- You are familiar with terms such as icons, click, double-click, menu, ribbon, and window.
- You understand basic concepts such as directory (folder) 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 it is available on a per incident charge. To ensure that you are a registered user of StatTools, please register online at


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 at your computer and ready to work.
Before Calling

Before contacting technical support, please review the following checklist:

- Have you consulted the relevant sections of this online manual?
- Have you watched the online Quick Start videos available from the StatTools Welcome screen?
- Have you read the README file? It contains current information on TopRank that might not be included in this manual.
- Can you duplicate the problem consistently? Can you duplicate the problem on a different computer or with a different model?
- Have you consulted our Web site, http://www.palisade.com? This Web site contains the latest FAQ (a searchable database of tech support questions and answers). 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 to:

  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 to:
  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 About StatTools item from the StatTools Help menu in Excel.

**Student Version**

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.
• Go to [http://www.palisade.com](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:
• Microsoft Windows XP or higher.
• Microsoft Excel 2007 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 XP or higher:

1) Double-click the StatTools Setup.exe (or the DTSuite Setup.exe) from your download or installation CD.

2) 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.

If you wish to remove StatTools from your computer, use the Control Panel’s Add/Remove Programs utility and select the entry for StatTools (or the DecisionTools Suite).

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 subfolder of a main “Program Files\Palisade” folder. This is quite similar to how Excel is often installed into a subfolder of a “Microsoft Office” folder.

One subfolder of this Palisade folder will be the StatTools folder (by default called StatTools7). This folder contains the program files, plus example models and other files necessary for StatTools to run. Another subfolder of the Palisade folder is the SYSTEM folder. This contains files required by every add-in in the DecisionTools Suite, including common help files and program libraries.
Software Activation

Activation is a one-time license verification process that is required for your Palisade software to run as a fully licensed product. An **activation ID** is on your printed/emailed invoice and may resemble a dash separated sequence like "DNA-6438907-651282-CDM". If you enter your Activation ID during installation, then your software is activated at the end of the installation process and no further user action is required. If you wish to activate your software after installation, select the **License Manager** command from the StatTools Help menu.

The License Manager can be used to activate, deactivate, and move software licenses. It is also used to manage licenses for network installations. Follow the prompts and dialogs in the License Manager to perform the desired licensing operation.
Instructional Materials

The StatTools package includes a number of instructional materials to help you learn StatTools features and run StatTools procedures.

**Example Spreadsheets**

Evolver includes about 50 example spreadsheets. These use actual data sets to illustrate the various StatTools statistical analyses.

You can find these example spreadsheets from the Evolver Help menu. When you click its **Example Spreadsheets** command, an “example file list” file opens in Excel. This file contains links to all of the example files.

**Quick Start Tutorials**

From the StatTools Welcome screen, which you see when you launch StatTools or you can access at any time from the StatTools Help menu, you can click the **Quick Start** link to see a series of short videos that lead you through the basic features of StatTools. These Quick Start videos are intended for beginners, but they are sufficient to get you started running your own StatTools analyses.

**Guided Tour Videos**

A series of Guided Tour videos is also available from the StatTools Welcome Screen. These are more in-depth videos, and they lead you through practically all of the StatTools procedures.

**Examples and Videos for the XDK**

You might want to learn how to use the macro language of Excel, Visual Basic for Applications (VBA), to automate StatTools procedures. StatTools provides an Excel Developer Kit (XDK) for doing this. You can learn about the XDK from the **Developer Kit (XDK)** command on the StatTools Help menu. There you will see links to example spreadsheet models of StatTools automation, as well as additional instructional videos on using the XDK.
Chapter 3: Overview of StatTools

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 Ribbon

Once you have installed StatTools in Excel 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 you must work with a data set, or a set of statistical variables, usually 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, which you must do before you run most StatTools analyses. You can then use these predefined variables in your statistical analyses, without having to re-select the data on later analyses.
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 can also be laid out by row.

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 in columns.

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.
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 you 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.

If you start with data in unstacked format, the 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.

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 provides 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 can be customized with standard Excel chart commands.

By default, StatTools tries to make the results as "live" as possible. That is, whenever it is practical, reports contain 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 **ST_Weight**, and then it enters formulas in the output cells: **=StatMean(ST_Weight)** and **=StatStdDev(ST_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 provides the option of turning off live updating. This is useful if Excel recalculation time becomes an issue as data changes.

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, you can change this behavior from the Display section of the Advanced tab in the Excel Options dialog.
StatTools Icons

The following icons are shown on the StatTools ribbon in Excel. Note that all but the first two of these have dropdown arrows that lead to a menu.

<table>
<thead>
<tr>
<th>Icon</th>
<th>Function</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Data Set Manager" /></td>
<td>Define a data set and variables, or edit or delete an existing data set and variables</td>
<td>Data group</td>
</tr>
<tr>
<td><img src="image" alt="Data Viewer" /></td>
<td>Create a quick graphical view of variables</td>
<td>Data group</td>
</tr>
<tr>
<td><img src="image" alt="Data Utilities" /></td>
<td>Create one or more new variables with various utilities</td>
<td>Data group</td>
</tr>
<tr>
<td><img src="image" alt="Summary Statistics" /></td>
<td>Calculate summary statistics for variables</td>
<td>Analyses group</td>
</tr>
<tr>
<td><img src="image" alt="Summary Graphs" /></td>
<td>Create summary graphs for variables</td>
<td>Analyses group</td>
</tr>
<tr>
<td><img src="image" alt="Statistical Inference" /></td>
<td>Calculate confidence intervals and run standard hypothesis tests</td>
<td>Analyses group</td>
</tr>
<tr>
<td>Icon</td>
<td>Description</td>
<td>Group</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>Normality Tests</td>
<td>Check normality of variables</td>
<td>Analyses group</td>
</tr>
<tr>
<td>Time Series and Forecasting</td>
<td>Run standard time series and forecasting procedures</td>
<td>Analyses group</td>
</tr>
<tr>
<td>Regression and Classification</td>
<td>Estimate a variety of regression equations and perform classification procedures</td>
<td>Analyses group</td>
</tr>
<tr>
<td>Quality Control</td>
<td>Create standard statistical process control charts</td>
<td>Analyses group</td>
</tr>
<tr>
<td>Nonparametric Tests</td>
<td>Run several well-known nonparametric procedures</td>
<td>Analyses group</td>
</tr>
<tr>
<td>Multivariate Analysis</td>
<td>Run principal components and cluster analysis procedures</td>
<td>Analyses group</td>
</tr>
<tr>
<td>Utilities</td>
<td>Run several useful StatTools utilities</td>
<td>Help group</td>
</tr>
<tr>
<td>Help</td>
<td>Find help and a variety of instructional examples and videos</td>
<td>Help group</td>
</tr>
</tbody>
</table>
# Table of StatTools Procedures

The StatTools procedures come in natural groups. This table provides a brief description of each procedure in each group. More detailed information on these procedures is provided in the rest of this chapter.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Description</th>
<th>Missing Data?</th>
<th>Live vs. Static Reports</th>
<th>Data Requirement</th>
<th>Invalid Data</th>
<th># of Vars.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary Statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-Variable Summary Command</td>
<td>Creates a table of summary statistics, including average, median, standard deviation, quartiles, percentiles, and others, for selected variables</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-100</td>
</tr>
<tr>
<td>Correlation and Covariance Command</td>
<td>Creates a table of correlations and/or covariances for selected variables</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>Ignored</td>
<td>1-250</td>
</tr>
<tr>
<td><strong>Summary Graphs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histogram Command</td>
<td>Creates a histogram for selected variables, and gives you the option of defining the histogram's categories or &quot;bins&quot;</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Partially live – data changes update graph when data is within the graph's X-axis range</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-100</td>
</tr>
<tr>
<td>Scatterplot Command</td>
<td>Creates a scatterplot for each pair of selected variables</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>Not allowed</td>
<td>1-10</td>
</tr>
<tr>
<td>Box-Whisker Plot Command</td>
<td>Creates a single boxplot (if a single variable is selected) or side-by-side boxplots (if multiple variables are selected)</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-12</td>
</tr>
<tr>
<td>Bar Chart</td>
<td>Creates bar charts for variables, all on one chart or in separate charts</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-100</td>
</tr>
<tr>
<td>Procedure</td>
<td>Description</td>
<td>Missing Data?</td>
<td>Live vs. Static Reports</td>
<td>Data Requirement</td>
<td>Invalid Data</td>
<td># of Vars.</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Confidence Interval - Mean/ Std. Deviation Command</strong></td>
<td>Calculates a confidence interval for the mean and/or standard deviation of a single variable (one-sample analysis), or the difference between means for a pair of variables (two-sample analysis or paired-sample analysis)</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired-sample analysis)</td>
</tr>
<tr>
<td><strong>Confidence Interval - Proportion Command</strong></td>
<td>Calculates a confidence interval for the proportion of items in a sample that belong to a given category (one-sample analysis), or the difference between proportions of items in a given category (two-sample analysis)</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-250 (one-sample analysis) Exactly 2 (two-sample analysis)</td>
</tr>
<tr>
<td><strong>Hypothesis Test - Mean/ Std. Deviation</strong></td>
<td>Performs hypothesis tests for the mean and/or standard deviation of single variables (one-sample analysis), or the difference between means for a pair of variables (two-sample analysis or paired-sample analysis)</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-250 (one-sample analysis) Exactly 2 (two-sample analysis; paired-sample analysis)</td>
</tr>
<tr>
<td><strong>Hypothesis Test - Proportion</strong></td>
<td>Performs a hypothesis test for the proportion of items in a sample that belong to a given category (one-sample analysis), or for the difference between proportions of items in a given category (two-sample analysis)</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-250 (one-sample analysis) Exactly 2 (two-sample analysis)</td>
</tr>
<tr>
<td>Procedure</td>
<td>Description</td>
<td>Missing Data?</td>
<td>Live vs. Static Reports</td>
<td>Data Requirement</td>
<td>Invalid Data</td>
<td># of Vars.</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------</td>
<td>-------------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>Sample Size</td>
<td>Determines the sample size(s) required to obtain a confidence interval, for a mean, a proportion, the difference between means, or the difference between proportions, with a prescribed half-length</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>One-Way ANOVA Command</td>
<td>An extension of the two-sample analysis for comparing two population means, it tests whether two or more means are all equal</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>2-50</td>
</tr>
<tr>
<td>Two-Way ANOVA Command</td>
<td>Performs a two-way analysis of variance, usually done in the context of an experimental design where there are two &quot;factors&quot; that are each set at several &quot;treatment levels</td>
<td>Not allowed</td>
<td>Live</td>
<td>Stacked data</td>
<td>Not allowed</td>
<td>2 categories, 1 value variable</td>
</tr>
<tr>
<td>Chi-Square Independence Test Command</td>
<td>Uses a chi-square test to check whether row and column attributes in a contingency table are statistically independent</td>
<td>No</td>
<td>Live (as long as table size does not change)</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Chi-Square Goodness-of-Fit Test</td>
<td>Performs a hypothesis test for goodness of fit of the frequency distribution of a categorical variable</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>Ignored</td>
<td>1</td>
</tr>
<tr>
<td>Procedure</td>
<td>Description</td>
<td>Missing Data?</td>
<td>Live vs. Static Reports</td>
<td>Data Requirement</td>
<td>Invalid Data</td>
<td># of Vars.</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>-------------------------</td>
<td>-------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Normality Tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-Square Normality Test Command</td>
<td>Runs a chi-square test of normality for a selected variable</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Partially Live (bin positioning will not change but occupation and graphs will)</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1</td>
</tr>
<tr>
<td>Lilliefors Test Command</td>
<td>Provides a more powerful test for normality (more likely to detect non-normality if it exists) than the chi-square goodness-of-fit test</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1-10</td>
</tr>
<tr>
<td>Q-Q Normal Plot Command</td>
<td>Creates a quantile-quantile (Q-Q) plot for a selected variable as an informal test of normality</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>1</td>
</tr>
<tr>
<td><strong>Time Series &amp; Forecasting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Series Graph Command</td>
<td>Creates a time series plot of one or more time series variable(s), all on the same chart</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Unstacked data</td>
<td>Not allowed</td>
<td>1-100</td>
</tr>
<tr>
<td>Procedure</td>
<td>Description</td>
<td>Missing Data?</td>
<td>Live vs. Static Reports</td>
<td>Data Requirement</td>
<td>Invalid Data</td>
<td># of Vars.</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>-------------------------</td>
<td>------------------------</td>
<td>------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Autocorrelation Command</td>
<td>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</td>
<td>Allowed at beginning or end of data</td>
<td>Live</td>
<td>Unstacked data</td>
<td>Up to 32,000 cases allowed</td>
<td>Not allowed</td>
</tr>
<tr>
<td>Runs Test for Randomness Command</td>
<td>Performs a runs test to check whether a variable (usually a time series variable) is random</td>
<td>Allowed at beginning or end of data</td>
<td>Live</td>
<td>Unstacked data</td>
<td>Not allowed</td>
<td>1 or more</td>
</tr>
<tr>
<td>Forecasting Command</td>
<td>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</td>
<td>Allowed at beginning of data only</td>
<td>Live</td>
<td>Unstacked data</td>
<td>Up to 32,000 cases allowed</td>
<td>Not allowed</td>
</tr>
</tbody>
</table>

**Regression & Classification**

<table>
<thead>
<tr>
<th>Regression Command</th>
<th>Runs a variety of regression analyses, including Multiple, Stepwise, Forward, Backward, and Block</th>
<th>Allowed at beginning, middle, and end of data</th>
<th>Static</th>
<th>Unstacked data</th>
<th>Not allowed</th>
<th>1 dependent; 1-1000 independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>Performs a logistic regression analysis on a data set, essentially a nonlinear type of regression analysis where the dependent variable is binary (0 or 1)</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Static</td>
<td>Stacked data</td>
<td>Not allowed</td>
<td>1 dependent; 1-250 independent</td>
</tr>
<tr>
<td>Discriminant Analysis Command</td>
<td>Performs a discriminant analysis, where there should be a &quot;category&quot; 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</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Static</td>
<td>Unstacked data</td>
<td>Not allowed except in dependent variables</td>
<td>1 dependent; 1-250 independent</td>
</tr>
<tr>
<td>Procedure</td>
<td>Description</td>
<td>Missing Data?</td>
<td>Live vs. Static Reports</td>
<td>Data Requirement</td>
<td>Invalid Data</td>
<td># of Vars.</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>-------------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Quality Control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pareto Chart Command</td>
<td>Creates a Pareto chart that indicates the relative importance of the categories in a categorical variable</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Static</td>
<td>Unstacked data</td>
<td>Ignored</td>
<td>1 category, or 1 value and 1 category</td>
</tr>
<tr>
<td>X/R Charts Command</td>
<td>Creates X-bar and R charts that indicate whether a process is in statistical control</td>
<td>Not allowed</td>
<td>Static</td>
<td>Unstacked data</td>
<td>Not allowed</td>
<td>2-25</td>
</tr>
<tr>
<td>P Chart Command</td>
<td>Creates P charts that indicate whether a process is in statistical control</td>
<td>Not allowed</td>
<td>Static</td>
<td>Unstacked data</td>
<td>Not allowed</td>
<td>1 variable 1 size variable</td>
</tr>
<tr>
<td>C Chart Command</td>
<td>Creates C charts that indicate whether a process is in statistical control</td>
<td>Not allowed</td>
<td>Static</td>
<td>Unstacked data</td>
<td>Not allowed</td>
<td>1</td>
</tr>
<tr>
<td>U Chart Command</td>
<td>Creates U charts that indicate whether a process is in statistical control</td>
<td>Not allowed</td>
<td>Static</td>
<td>Unstacked data</td>
<td>Not allowed</td>
<td>1</td>
</tr>
<tr>
<td><strong>Non-Parametric Tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sign Test Command</td>
<td>Performs hypothesis tests for the median of a single variable or for the median of differences for a pair of variables</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>See next section for details</td>
</tr>
<tr>
<td>Procedure</td>
<td>Description</td>
<td>Missing Data?</td>
<td>Live vs. Static Reports</td>
<td>Data Requirement</td>
<td>Invalid Data</td>
<td># of Vars.</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------</td>
<td>-------------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>Wilcoxon Signed-Rank Test</td>
<td>Similar to the Sign Test, but assumes that the probability distribution is symmetric</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>See next section for details</td>
</tr>
<tr>
<td>Mann-Whitney Test</td>
<td>Performs a hypothesis test for comparing two populations</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>2</td>
</tr>
<tr>
<td>Kruskal-Wallis Test</td>
<td>Performs a hypothesis test for comparing three or more populations</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Live</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>3-25</td>
</tr>
<tr>
<td>Multivariate Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principal Components Analysis</td>
<td>Procedure for reducing the dimension (number of variables) of multivariate data</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Static</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>2-1000</td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>Searches for patterns in a data set in order to classify observations into groups of similar items</td>
<td>Allowed at beginning, middle, and end of data</td>
<td>Static</td>
<td>Stacked and unstacked data</td>
<td>Ignored</td>
<td>2-1000</td>
</tr>
<tr>
<td>Data Utilities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stack Command</td>
<td>Takes a data set with separate variables for each group in separate columns, and “stacks” them into two columns: a “category” column, and a “value” column</td>
<td>Yes – anywhere in variable</td>
<td>Static</td>
<td>Unstacked data only Up to 65535 cases allowed</td>
<td>n/a</td>
<td>1-100</td>
</tr>
<tr>
<td>Unstack Command</td>
<td>Does the exact opposite of the Stack procedure</td>
<td>Yes – anywhere in variable</td>
<td>n/a</td>
<td>Stacked data only</td>
<td>n/a</td>
<td>1-32</td>
</tr>
<tr>
<td>Dummy Command</td>
<td>Creates dummy (0-1) variables based on existing variables</td>
<td>Yes – anywhere in variable</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>n/a</td>
<td>1</td>
</tr>
<tr>
<td>Interaction</td>
<td>Creates one or more new “product” variables from a pair of numeric variables, a categorical and a numeric variable, or two categorical variables</td>
<td>Yes – anywhere in variable</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>Not allowed</td>
<td>2-32 from same dataset</td>
</tr>
<tr>
<td>Procedure</td>
<td>Description</td>
<td>Missing Data?</td>
<td>Live vs. Static Reports</td>
<td>Data Requirement</td>
<td>Invalid Data</td>
<td># of Vars.</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------</td>
<td>------------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>Combination</td>
<td>Creates a new variable by taking the product, sum, average, min, max or min-max range from one or more variables</td>
<td>Yes – anywhere in variable</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>n/a</td>
<td>2 value vars., or 1 value and 1 category var, or 2 category var</td>
</tr>
<tr>
<td>Lag Command</td>
<td>Creates one or more new lagged variables based on an existing variable</td>
<td>Yes – anywhere in variable</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>Ignored</td>
<td>1</td>
</tr>
<tr>
<td>Transform Command</td>
<td>Applies any of four nonlinear transformations (natural logarithm, square, square root, or reciprocal) to selected variables to create new variables</td>
<td>Yes – anywhere in variable</td>
<td>Live or static</td>
<td>Unstacked data only</td>
<td>Ignored</td>
<td>1-100</td>
</tr>
<tr>
<td>Difference Command</td>
<td>Creates any number of difference variables from an original variable</td>
<td>Yes – anywhere in variable</td>
<td>Live</td>
<td>Unstacked data only</td>
<td>n/a</td>
<td>1</td>
</tr>
<tr>
<td>Random Sample Command</td>
<td>Generates any number of random samples from a data set, where sampling is with or without replacement</td>
<td>Yes – anywhere in variable</td>
<td>Static</td>
<td>Stacked data only</td>
<td>Ignored</td>
<td>1-32</td>
</tr>
</tbody>
</table>
Data Commands

The two commands in this section let you define StatTools data sets and then view their basic properties.

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 can be analyzed in StatTools procedures. The Data Set Manager dialog 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.

StatTools is analogous to most stand-alone statistical software packages in that it is structured around variables. For most analyses, you must 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 having to re-select the data you wish to analyze.

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 can also be laid out by Rows.

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 in columns.

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.
The **Data Set** options in the Data Set Manager dialog 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.
- **Apply Cell Formatting.** Adds a grid and colors that identifies the data set.
The **Variables** options in the Data Set Manager dialog 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 contains the data for a variable.
  
  ![Example of Columns layout](image)

  - **Rows.** With this layout, each row of the data set contains the data for a variable.

  ![Example of Rows layout](image)

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

  ![Example of Names in First Column](image)

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

- **Excel Range Name.** This 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 meaningful range names helps make formulas more readable.

- **Output Format.** Specifies the format for values shown for a variable in reports from StatTools analyses. The **Automatic** 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:
Data Commands

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).

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.

Actual data capacities might be less than this, depending on the system configuration, memory limitations, and your version of Excel. Also, specific StatTools analyses may have different limitations.

Data Set and Variable Capacities

Note: the Data Set Manager dialog 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 Viewer Command

Allows you to easily review statistics and graphs for any data set defined in StatTools.

Clicking the Data Viewer icon on the StatTools ribbon brings up the Data Viewer Options dialog, where you can choose a data set (previously defined in the StatTools Data Set Manager) and select variables from that data set. Both stacked and unstacked data formats are supported.

The following options are available:

- **Ignore Entire Row/Column if Any Cell Has a Missing or Non-Numeric Value.** Setting this option will discard an entire row or column if any of the values in that row or column are missing or non-numeric.

- **Correlation Calculations.** The data viewer displays correlation coefficients between your variables. This option controls how these coefficients are calculated.

If your data set contains more than one variable, then after you specify the Data Viewer options and click “OK,” you will be presented with the Multivariate Data Viewer window. This window has several different “views,” each giving a different perspective of your data set. You can change the view by clicking the view buttons at the bottom of the window.
**Summary View**

The summary view gives a quick overview of your data set in tabular format. Each row of the table corresponds to one variable, and shows the variables name, a thumbnail graph, and some statistical summary information.

You can control which columns are displayed in the table by clicking the **Configure Columns** button at the bottom of the window.

You can click and drag an individual thumbnail off the window to get a full-sized graph.

You can export this table to Excel by clicking the **Export** button at the bottom on the window.

**Correlation View**

The correlation view gives you visual and numeric information about the correlation patterns between variables using a two-dimensional array of graphs. The diagonals contain histograms of each variable, while the off-diagonal graphs show scatterplots of each possible pair of variables. (Note: the graphs below the diagonals are always reflections of the graphs above the diagonal.) The correlation coefficients associated with each pair are shown at the top of each scatterplot.
You can click and drag an individual graph off the window to get a full-sized graph.

You can export the correlation coefficient to Excel by clicking the **Export** button at the bottom on the window.

Note: the Correlation View is not available if the data set is in stacked format, because in this case the data values cannot necessarily be paired.
Overlay View

The overlay view shows all the variables simultaneously in a single graph.

You can control which variables are displayed (up to a maximum of 10) in the overlay graph by clicking the Graph Options button at the bottom of the window and choosing the Select Variables to Graph menu item. By default, StatTools will only show variables with the same formatting as the first variable in the data set. For example, if you have a data set with some numeric values, some date values, and some percentage values, StatTools by default will show only the numeric values together on the overlay graph.

You can export this graph to Excel by clicking the Export button at the bottom of the window.

You can change the formatting of the graph by clicking the Graph Options button at the bottom of the window.
The Trend and Boxplot views show summary statistics of all the variables in a compact graphical format. (These are relevant only for certain data sets, such as in the following screenshots, where there is a variable for each of 10 exams.)

You can control which inputs are displayed in the graphs (up to a maximum of 500) by clicking the **Graph Options** button at the bottom of the window, and choosing the **Select Variables to Graph** menu item. By default, StatTools will only show variables with the same formatting as the first variable in the data set. For example, if you have a data set with some numeric values, some date values, and some percentage values, StatTools by default will show only the numeric values together on the graph.

You can export the graphs to Excel by clicking the **Export** button at the bottom of the window.

You can change the formatting of the graphs by clicking the **Graph Options** button at the bottom of the window.
Data Utilities Menu

The commands in the Data Utilities menu are useful for transforming your data in various ways. You usually do this as a preliminary step before running StatTools procedures.
Stack Command

Converts a set of variables to stacked format from unstacked format

The Stack command converts data from "unstacked" format, where a data set includes at least two value variables, to a "stacked" format, in which the data set includes a category variable and a value variable. For example, if you are comparing mean household incomes in several different neighborhoods, an unstacked format would include 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 format 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 more convenient to work with than the unstacked version.

<table>
<thead>
<tr>
<th>Stacked format</th>
<th>Unstacked format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Gender</td>
</tr>
<tr>
<td>1</td>
<td>Female</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
</tr>
<tr>
<td>7</td>
<td>Female</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
</tr>
<tr>
<td>10</td>
<td>Female</td>
</tr>
</tbody>
</table>
Variables are stacked using the **Stacking Utility** dialog:

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 you click OK, 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 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 categorical variable. In stacked format, you typically select 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 Exam Score, this command unstacks them into separate Exam Score (Female) and Exam Score (Male) columns.

Variables are unstacked using the Unstacking Utility dialog:

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 you click OK, the variables are unstacked and a new data set is 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 performs a nonlinear transformation of any variable using one of four possible transformations: natural logarithm, square, square root, or reciprocal. In addition, you can 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.

Variables are transformed using the Transformation Utility dialog:

The data set is always treated as unstacked data. Only variables from a single data set at a time can be transformed.
The options in the Transformation Utility dialog include:

- **Transformation Function.** 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. You can also enter a custom **Formula** to calculate a new variable value based on a mathematical expression, such as:

\[(\text{Variable} \times 1.5)^2\]

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

<table>
<thead>
<tr>
<th>Family</th>
<th>Salary</th>
<th>Culture</th>
<th>Sports</th>
<th>Dining</th>
<th>Log(Salary)</th>
<th>Log(Culture)</th>
<th>Log(Sports)</th>
<th>Log(Dining)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$54,000</td>
<td>$1,020</td>
<td>$990</td>
<td>$1,510</td>
<td>10.90778916</td>
<td>6.927567906</td>
<td>6.897704943</td>
<td>7.31986493</td>
</tr>
<tr>
<td>2</td>
<td>$57,500</td>
<td>$1,100</td>
<td>$460</td>
<td>$1,180</td>
<td>10.95964023</td>
<td>7.003065459</td>
<td>6.131226489</td>
<td>7.073209717</td>
</tr>
<tr>
<td>3</td>
<td>$53,300</td>
<td>$700</td>
<td>$780</td>
<td>$1,390</td>
<td>10.88369161</td>
<td>6.802394763</td>
<td>6.65929392</td>
<td>7.371489295</td>
</tr>
<tr>
<td>4</td>
<td>$43,500</td>
<td>$570</td>
<td>$860</td>
<td>$1,750</td>
<td>10.68061022</td>
<td>6.345636361</td>
<td>6.766932289</td>
<td>7.467371067</td>
</tr>
<tr>
<td>5</td>
<td>$57,200</td>
<td>$900</td>
<td>$1,390</td>
<td>$2,120</td>
<td>10.95439918</td>
<td>6.802394763</td>
<td>7.237069026</td>
<td>7.659171368</td>
</tr>
<tr>
<td>6</td>
<td>$63,400</td>
<td>$820</td>
<td>$1,880</td>
<td>$3,090</td>
<td>11.05729194</td>
<td>6.70930434</td>
<td>7.539027056</td>
<td>8.03592637</td>
</tr>
<tr>
<td>7</td>
<td>$58,500</td>
<td>$1,340</td>
<td>$710</td>
<td>$1,540</td>
<td>10.97678203</td>
<td>7.200424893</td>
<td>6.66526497</td>
<td>7.339537696</td>
</tr>
<tr>
<td>8</td>
<td>$55,600</td>
<td>$1,250</td>
<td>$680</td>
<td>$1,800</td>
<td>10.92598384</td>
<td>7.13089883</td>
<td>6.522092798</td>
<td>7.495541944</td>
</tr>
<tr>
<td>9</td>
<td>$61,300</td>
<td>$1,190</td>
<td>$1,220</td>
<td>$2,330</td>
<td>11.02365312</td>
<td>7.081708666</td>
<td>7.106606138</td>
<td>7.753623547</td>
</tr>
<tr>
<td>10</td>
<td>$61,100</td>
<td>$640</td>
<td>$1,480</td>
<td>$2,670</td>
<td>11.02026716</td>
<td>6.461468176</td>
<td>7.298797367</td>
<td>7.888933751</td>
</tr>
<tr>
<td>11</td>
<td>$77,200</td>
<td>$900</td>
<td>$820</td>
<td>$2,850</td>
<td>11.25415474</td>
<td>6.802394763</td>
<td>6.70930434</td>
<td>7.965074273</td>
</tr>
<tr>
<td>12</td>
<td>$58,800</td>
<td>$710</td>
<td>$1,080</td>
<td>$2,200</td>
<td>10.98109713</td>
<td>6.56526497</td>
<td>6.98471632</td>
<td>7.696212639</td>
</tr>
<tr>
<td>13</td>
<td>$62,900</td>
<td>$1,240</td>
<td>$1,230</td>
<td>$2,430</td>
<td>11.04930144</td>
<td>7.122866659</td>
<td>7.114769448</td>
<td>7.795646536</td>
</tr>
<tr>
<td>14</td>
<td>$61,900</td>
<td>$1,270</td>
<td>$1,000</td>
<td>$2,110</td>
<td>11.03327546</td>
<td>7.146772179</td>
<td>6.907756279</td>
<td>7.664443228</td>
</tr>
<tr>
<td>15</td>
<td>$70,500</td>
<td>$1,180</td>
<td>$590</td>
<td>$1,820</td>
<td>11.24504302</td>
<td>7.073209717</td>
<td>6.53691598</td>
<td>7.50659178</td>
</tr>
</tbody>
</table>

Data Utilities Menu
Lag Command

Creates one or more new lagged variables based on an existing variable

The Lag command creates one or more new lagged variables 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 is sales three months earlier, in August.

Variables are lagged using the Lag Utility dialog:

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

The options in the Variable Lag Utility dialog include:

- **Number of Lags.** The number of time periods to lag values when creating the new variable(s). A new variable is created for each of the lags up to the entered Number of Lags.
Difference Command

Creates one or more difference variables from an existing variable

The Difference command creates one or more difference variables from an existing variable. It is used primarily for time series variables. You select a variable to be differenced and 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 difference value is the March value minus the February 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 variables are created using the Differences Utility dialog:

![Difference Utility Dialog](image)

The data set is always treated as unstacked data. Only one variable at a time can be used for creating difference variables.
The options in the Differences Utility dialog include:

- **Number of Differences.** The number of difference variables to create.

<table>
<thead>
<tr>
<th>Month</th>
<th>Sales</th>
<th>Difference1(Sales)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-12</td>
<td>226</td>
<td></td>
</tr>
<tr>
<td>Feb-12</td>
<td>254</td>
<td>28</td>
</tr>
<tr>
<td>Mar-12</td>
<td>204</td>
<td>-50</td>
</tr>
<tr>
<td>Apr-12</td>
<td>193</td>
<td>-11</td>
</tr>
<tr>
<td>May-12</td>
<td>191</td>
<td>-2</td>
</tr>
<tr>
<td>Jun-12</td>
<td>166</td>
<td>-25</td>
</tr>
<tr>
<td>Jul-12</td>
<td>175</td>
<td>9</td>
</tr>
<tr>
<td>Aug-12</td>
<td>217</td>
<td>42</td>
</tr>
<tr>
<td>Sep-12</td>
<td>167</td>
<td>-50</td>
</tr>
<tr>
<td>Oct-12</td>
<td>192</td>
<td>25</td>
</tr>
<tr>
<td>Nov-12</td>
<td>127</td>
<td>-65</td>
</tr>
<tr>
<td>Dec-12</td>
<td>148</td>
<td>21</td>
</tr>
</tbody>
</table>

*Data Set with Difference Variable*
Interaction Command

Creates one or more interaction variables from existing variables

The Interaction command creates one or more interaction variables from existing variables. Interaction variables can be created from two numeric variables, one numeric and one categorical variable, or two categorical variables.

If both variables are numeric (non-categorical), it creates their product. If one variable is numeric and the other is categorical, it creates the products of the numeric 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 variables are created using the Interaction Utility dialog:

The data set is always treated as unstacked data. Exactly two variables must be selected when creating interaction variables.

The options in the Interaction Utility dialog 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.
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 numeric "measurement" variables. Then the interaction variable is their product. Second, one variable can be a numeric "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. (This option is illustrated in the screenshots shown here.) 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, StatTools will create 2x5=10 interaction variables.

<table>
<thead>
<tr>
<th>Gender</th>
<th>EduLevel</th>
<th>YrsExp</th>
<th>Salary</th>
<th>Interaction(YrsExp, Gender = Female)</th>
<th>Interaction(YrsExp, Gender = Male)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3</td>
<td>3</td>
<td>32000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>14</td>
<td>39200</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>12</td>
<td>33200</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>8</td>
<td>30600</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>3</td>
<td>3</td>
<td>29000</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>3</td>
<td>3</td>
<td>30500</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>3</td>
<td>4</td>
<td>30000</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>3</td>
<td>8</td>
<td>27000</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>4</td>
<td>34000</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>3</td>
<td>9</td>
<td>29500</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>3</td>
<td>9</td>
<td>26800</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>8</td>
<td>31300</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>
Combination Command

Creates a combination variable from several existing variables

The Combination command allows you to create a combination variable from several existing variables. The existing variables are combined from one of the following operations: product, sum, average, min, max, or min-max range.

Combination Utility Dialog

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

The data set is always treated as unstacked data. One or more variables can be used for creating a combination variable.

The options in the Combination Utility dialog include:

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

Data Set with Combination Variable

<table>
<thead>
<tr>
<th>Student</th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
<th>Test4 (Average)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62</td>
<td>68</td>
<td>79</td>
<td>76</td>
</tr>
<tr>
<td>2</td>
<td>85</td>
<td>75</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>75</td>
<td>83</td>
<td>91</td>
</tr>
<tr>
<td>4</td>
<td>66</td>
<td>62</td>
<td>64</td>
<td>71</td>
</tr>
<tr>
<td>5</td>
<td>77</td>
<td>75</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>60</td>
<td>65</td>
<td>67</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>83</td>
<td>83</td>
<td>77</td>
</tr>
<tr>
<td>8</td>
<td>73</td>
<td>81</td>
<td>72</td>
<td>65</td>
</tr>
<tr>
<td>9</td>
<td>64</td>
<td>58</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>10</td>
<td>46</td>
<td>65</td>
<td>56</td>
<td>63</td>
</tr>
</tbody>
</table>
Dummy Command

Creates one or more dummy (0-1) variables based on an existing variable

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

- 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.

- 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 with value 0 when Weight > 160 and value 1 when Weight <= 160.

Dummy variables are created using the Dummy Utility dialog:

![Dummy Utility Dialog]

The data set is always treated as unstacked data. Exactly one variable should be selected for creating dummy variables.
The options in the Dummy Utility dialog box include:

- **Options.** The method for creating the dummy (or dummies): a dummy variable for each category of a categorical variable, or a single dummy variable from a numeric variable. The **Cutoff value** specifies the cutoff to use when basing a 0-1 dummy variable on a numeric variable.

### Data Set with Dummy Variable

<table>
<thead>
<tr>
<th>Employee</th>
<th>EduLevel</th>
<th>Gender</th>
<th>Salary</th>
<th>Gender = Female</th>
<th>Gender = Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>Male</td>
<td>$32,000</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Female</td>
<td>$39,100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Female</td>
<td>$33,200</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>Female</td>
<td>$30,600</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>Male</td>
<td>$29,000</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>Female</td>
<td>$30,500</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>Female</td>
<td>$30,000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>Male</td>
<td>$27,000</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>Female</td>
<td>$34,000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>Female</td>
<td>$29,500</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Random Sample Command

Generates any number of random samples from selected variables

The Random Samples command generates 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 Samples are generated using the Random Sample Utility dialog:

![Random Sample Utility Dialog]

The data set is always treated as unstacked data. One or more variables can be selected for generating random samples.
The options in the Random Sample Utility dialog 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 members 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

<table>
<thead>
<tr>
<th>Employee(1)</th>
<th>EduLevel(1)</th>
<th>Gender(1)</th>
<th>Salary(1)</th>
<th>Employee(2)</th>
<th>EduLevel(2)</th>
<th>Gender(2)</th>
<th>Salary(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>3</td>
<td>Female</td>
<td>34900</td>
<td>78</td>
<td>3</td>
<td>Male</td>
<td>33000</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>Female</td>
<td>28500</td>
<td>107</td>
<td>3</td>
<td>Female</td>
<td>36200</td>
</tr>
<tr>
<td>197</td>
<td>3</td>
<td>Male</td>
<td>55000</td>
<td>192</td>
<td>3</td>
<td>Male</td>
<td>55000</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>Female</td>
<td>26800</td>
<td>147</td>
<td>3</td>
<td>Male</td>
<td>41000</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>Female</td>
<td>27000</td>
<td>45</td>
<td>1</td>
<td>Female</td>
<td>45300</td>
</tr>
<tr>
<td>70</td>
<td>3</td>
<td>Male</td>
<td>33500</td>
<td>189</td>
<td>3</td>
<td>Female</td>
<td>43000</td>
</tr>
<tr>
<td>66</td>
<td>2</td>
<td>Male</td>
<td>29600</td>
<td>88</td>
<td>3</td>
<td>Male</td>
<td>37000</td>
</tr>
<tr>
<td>66</td>
<td>3</td>
<td>Female</td>
<td>29800</td>
<td>10</td>
<td>2</td>
<td>Female</td>
<td>33000</td>
</tr>
<tr>
<td>83</td>
<td>1</td>
<td>Female</td>
<td>35000</td>
<td>101</td>
<td>1</td>
<td>Female</td>
<td>35300</td>
</tr>
<tr>
<td>185</td>
<td>3</td>
<td>Male</td>
<td>46500</td>
<td>137</td>
<td>3</td>
<td>Male</td>
<td>42000</td>
</tr>
</tbody>
</table>
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.
One-Variable Summary Command

Calculates summary statistics for selected 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.

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

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.

Note that if you choose the stacked format and then select a categorical variable such as Gender and a numeric variable such as Score, StatTools will create separate Score summary measures for males and females. This option is illustrated in the screenshots you see here.
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 `StatMean` and `StatStdDev` to allow hot-linking to data. The report is placed in the location specified using the Settings command.

  ![StatTools Report](image)

- **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 numeric values in the range.)

- **Link to Data.** All of the summary measures are calculated by formulas that are linked to the data. If any of the data change, the summary measures change automatically.
Correlation and Covariance Command

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

The Correlation and Covariance command produces a table of correlations and/or a table of covariances between any set of selected numeric 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 show (1) only the correlations (or covariances) below the diagonal, (2) only those above the diagonal, or (3) those below and above the diagonal.

This analysis is set up using the Correlation and Covariance dialog:

Two or more variables must be selected for analysis. The data set is always treated as unstacked data. Variables can be from different data sets.
The options in the Correlation 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.

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

### StatTools Report

<table>
<thead>
<tr>
<th>Linear Correlation Table</th>
<th>AXP</th>
<th>FDX</th>
<th>GM</th>
<th>IBM</th>
<th>MCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AXP</td>
<td>1.00</td>
<td>0.400</td>
<td>0.483</td>
<td>0.358</td>
<td>0.354</td>
</tr>
<tr>
<td>FDX</td>
<td>1.00</td>
<td>0.260</td>
<td>0.292</td>
<td>0.265</td>
<td></td>
</tr>
<tr>
<td>GM</td>
<td>1.00</td>
<td>0.259</td>
<td>0.318</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>1.00</td>
<td>0.397</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCD</td>
<td></td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Covariance Table</th>
<th>AXP</th>
<th>FDX</th>
<th>GM</th>
<th>IBM</th>
<th>MCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AXP</td>
<td>0.00558</td>
<td>0.00369</td>
<td>0.00319</td>
<td>0.00279</td>
<td>0.00205</td>
</tr>
<tr>
<td>FDX</td>
<td>0.01295</td>
<td>0.00231</td>
<td>0.00309</td>
<td>0.00216</td>
<td></td>
</tr>
<tr>
<td>GM</td>
<td>0.00661</td>
<td>0.00196</td>
<td>0.00185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>0.00666</td>
<td>0.00254</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCD</td>
<td></td>
<td>0.00512</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• **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. If any of the data change, these summary measures update automatically.
Summary Graphs Menu

The commands on the Summary Graphs menu create useful charts for statistical analyses. The resulting charts are Excel charts, so they can be modified just like regular Excel charts.
Histogram Command
Creates histograms for selected 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.

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

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.

Note that if you choose the stacked format and then select a categorical variable such as Gender and a numeric variable such as Score, StatTools will create separate Score histograms for males and females. In this case, it is a good idea to specify the bins so that the *same* bins are used for both genders.
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.
**Missing Data and Link to Data**

- **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable 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.
Scatterplot Command

Creates scatterplots between pairs of selected variables

The Scatterplot command creates a scatterplot for each pair of variables you select. StatTools allows you to choose which of the variables will be placed on the horizontal (X) axis and which will be on the vertical (Y) axis. Each scatterplot shows the correlation between the two variables (if you request it).

This graph type is set up using the Scatterplot dialog:

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 scatterplots are created. (Note that if you select the same variables for X and for Y, as in the above dialog, you get a matrix of scatterplots.) The data set is always treated as unstacked data. Variables can be from different data sets.
The options in the **Scatterplot** dialog include:

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

- **Chart Type.** Selects the type of graph to be generated. Options include:
  - **Simple XY Chart.** Generates typical Excel-like scatterplot.
  - **Break Down by Category Variable.** Points in each scatterplot are colored based on categories, as specified by an additional categorical variable.

**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 selected 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 displays the summary statistics (quartiles, interquartile range, etc.) that are used to form the box-whisker plot(s).

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

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.

Note that if you choose a stacked format and then select a categorical variable such as Gender and a numeric variable such as Score, StatTools will create “side-by-side” box-whisker plots. This is probably the most common use for box-whisker plots: to compare two or more populations graphically.
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. This is primarily for learning purposes for users not familiar with box-whisker plots.

- **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable 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 plots. However, the scale of the horizontal axis might need to be updated manually if the scale of the data changes significantly.
-bar chart command

creates bar charts for selected variables

the bar chart command creates multiple bar charts representing each selected variable, or a single bar chart for all the selected variables. in each case you can select the categories to be included in the chart.

it also creates a frequency table for each variable selected. bar charts are particularly useful for categorical variables, even if the categories are coded numerically.

this graph type is set up using the bar chart dialog:

one or more variables can be selected. the data can be in the stacked or unstacked format. variables can be selected from different data sets.
The options in the Bar Chart dialog include:

- **Analysis Type.** Specifies how many charts to create. Options include:
  - **One Chart per Variable.** Creates a single bar chart for each selected variable.
  - **All Variables in Single Chart.** Creates a single bar chart for all selected variables.

- **Chart Type.** Specifies the type of bar chart to be created. Options include:
  - **Simple.** Compares categories of a single variable. When the "Simple" type is selected and all variables are to be placed in a single chart, the resulting chart consists of multiple simple charts placed horizontally next to each other.
  - **Clustered.** Compares categories of several variables. Categories are displayed as a series of vertical bars.
  - **Stacked.** Compares categories of several variables. Categories are stacked in a single bar.

Clustered and Stacked bar charts should be used when selected variables have categories in common.

- **Categories to Analyze.** Allows you to select which categories of the selected variables will be shown in the chart(s).

- **Options.** The available options change with the selected combination of Analysis Type and Data Type. Options include:
  - **X-Axis.** This option is available for Clustered and Stacked bar charts. It lets you decide whether Categories or Variables should show on the X-Axis. By changing this option, you can flip the chart. For example, the original chart might answer questions like "What percent of responders answered Yes to the question whether they use e-mail?", and "What percent of responders answered Yes to the question whether they use WiFi?", etc. The flipped chart can answer what percent of Yes answers were obtained in answer to the question about e-mail, what percent in answer to the question about WiFi, etc.
  - **Y-Axis.** Shows either Frequency or Relative Frequency on the Y-Axis.
– **Calculate Frequencies Relative To.** This option affects the Relative Frequency values. If All Categories is selected, the Relative Frequency is computed by reference to all categories in the data set, even though they were not selected during the analysis. If Selected Categories is selected, the Relative Frequency is computed by reference to just the categories selected in the analysis.

<table>
<thead>
<tr>
<th></th>
<th>Mobile Data?</th>
<th>Social Networks?</th>
<th>WiFi?</th>
<th>e-mail?</th>
<th>GPS?</th>
<th>Multimedia?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
</tr>
<tr>
<td>No</td>
<td>350</td>
<td>371</td>
<td>162</td>
<td>393</td>
<td>575</td>
<td>403</td>
</tr>
<tr>
<td>Yes</td>
<td>642</td>
<td>620</td>
<td>828</td>
<td>558</td>
<td>411</td>
<td>551</td>
</tr>
<tr>
<td>Freq, Relative to All Categories</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mobile Data?</th>
<th>Social Networks?</th>
<th>WiFi?</th>
<th>e-mail?</th>
<th>GPS?</th>
<th>Multimedia?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel. Freq</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
<td>Survey Data</td>
</tr>
<tr>
<td>No</td>
<td>0.35</td>
<td>0.37</td>
<td>0.16</td>
<td>0.39</td>
<td>0.58</td>
<td>0.40</td>
</tr>
<tr>
<td>Yes</td>
<td>0.64</td>
<td>0.62</td>
<td>0.83</td>
<td>0.60</td>
<td>0.41</td>
<td>0.59</td>
</tr>
<tr>
<td>Freq, Relative to All Categories</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

**Bar Chart Report**

**Missing Data and Link to Data**

- **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable are ignored.

- **Link to Data.** Bar Charts are partially linked to data. If new categories are included in the data, you will need to rerun the Bar Chart analysis.
Statistical Inference Menu

The commands on the Statistical Inference menu perform the most common statistical inference analyses: creating confidence intervals and performing standard hypothesis tests.
Confidence Interval - Mean/ Std. Deviation Command

Calculates confidence intervals for mean and/or standard deviation of selected variables

The Confidence Interval for Mean/ Std. Deviation command calculates a confidence interval for the mean and/or 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.

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

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 each 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.

- **Input Type.** Selects the type of information available for running the analysis. Options include:
  - **Data Set.** Select this option if you have the data set.
  - **Summary Statistics.** Select this option if you don’t have the data set but you know some statistics from it.

- **Known Population Standard Deviation.** Check this option if the standard deviation of the population is known.

- **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.
**Missing Data and Link to Data**

- **Missing Data.** Missing data are allowed. They are ignored in a pairwise manner when a Paired-Sample Analysis is used. Otherwise, for each variable, the missing data for that variable 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 creates a confidence interval for the proportion of items in a sample that belong to a given category (One-Sample Analysis), or a confidence interval for the difference between the proportions of items in a given category from two populations (Two-Sample Analysis). There are three data formats supported by this procedure: Population Sample (one or more columns listing information on each sampled member), Summary Table with Counts (a contingency table), and Summary Table with Proportions (a contingency tables showing counts as percentages of totals).

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

This particular dialog requests a one-sample analysis for the proportion of voters who favor the Democrat, where the data set lists the voter preference for each sampled member. Here are the results:
Alternatively, the following dialog requests the difference between two proportions: the proportion of females who favor the Democrat minus the proportion of males who favor the Democrat, where the data set now lists the gender and voter preference of each sampled member.
The results in this case are the following:

A **One-Sample** analysis requires one or more variables for analysis. The first dialog above shows the setup when the data set format is **Population Sample**. When the data set format is **Summary Table with Counts**, you must select a “Cat” variable for category and a “Ct” variable for the count. Similarly, when the data set format is **Summary Table with Proportions**, you must select a “Cat” variable for category and a “%” variable for the percentage.

The **Two-Sample** analysis requires two variables. In the typical case where the data set format is **Population Sample** and the data are stacked, you must select a “C1” variable for “comparison” variable (in this case Gender) and a “C2” variable for category. This is shown in the second dialog above. When the data set format is **Summary Table with Counts**, you must select a “Cat” variable for category and a “Ct” variable for each count (male and female in this case). Similarly, when the data set format is **Summary Table with Proportions**, you must select a “Cat” variable for category and a “%” variable for each percentage (male and female).
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 the difference between 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, specifies the size of the first sample and (for two-sample analysis only) the size of the second sample.

- **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable 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/or standard deviation of selected variables

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

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

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 one or more numeric variables.
  - **Two-Sample Analysis.** Performs hypothesis tests for the difference between means from two independent populations.
  - **Paired-Sample Analysis.** This is similar to 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.

- **Input Type.** Selects the type of information available for running the analysis. Options include:
  - **Data Set.** Select this option if you have the data set.
  - **Summary Statistics.** Select this option if you don’t have the data set but you know some statistics from it.

- **Known Population Standard Deviation.** Check this option if the standard deviation of the population is known.

- **Hypothesis Tests to Perform.** Specifies the hypothesis tests that will be performed on the selected variables. Options change based on the analysis type selected. For each hypothesis test selected, options include:
  - **Null Hypothesis Value**, or the value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis Type.** Can be either "one-tailed" (that is, greater than or less than the null hypothesis value) or "two-tailed" (that is, not equal to the null hypothesis value).
Missing Data and Link to Data

- **Missing Data.** Missing data are allowed. They are ignored in a pairwise manner when a Paired-Sample Analysis is used. Otherwise, for each variable, the missing data for that variable 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.

This analysis is set up using the Hypothesis Test for Proportion dialog:
This particular dialog requests a one-sample analysis for the proportion of voters who favor the Democrat, where the data set lists the voter preference for each sampled member. The alternative hypothesis is that the proportion of voters who favor the Democrat is at least 0.5. Here are the results:

Alternatively, the following dialog tests the difference between two proportions: the proportion of females who favor the Democrat minus the proportion of males who favor the Democrat, where the data set now lists the gender and voter preference of each sampled member. Now the alternative hypothesis is that these two proportions are not equal.
Hypothesis Test for Proportion Dialog (Two-Sample)
The results in this case are the following:

A **One-Sample** analysis requires one or more variables for analysis. The first dialog above shows the setup when the data set format is **Population Sample**. When the data set format is **Summary Table with Counts**, you must select a “Cat” variable for category and a “Ct” variable for the count. Similarly, when the data set format is **Summary Table with Proportions**, you must select a “Cat” variable for category and a “%” variable for the percentage.

The **Two-Sample** analysis requires two variables. In the typical case where the data set format is **Population Sample** and the data are stacked, you must select a “C1” variable for “comparison” variable (in this case Gender) and a “C2” variable for category. This is shown in the second dialog above. When the data set format is **Summary Table with Counts**, you must select a “Cat” variable for category and a “Ct” variable for each count (male and female in this case). Similarly, when
the data set format is **Summary Table with Proportions**, you must select a “Cat” variable for category and a “%” variable for each percentage (male and female).

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 the differences between two proportions 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(s). Options change based on the analysis type selected. Options include:
  - **Null Hypothesis Value**, the value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis Type**, the alternative to the Null Hypothesis Value that will be evaluated during the analysis. The Alternative Hypothesis Type can be either "one-tailed" (greater than or less than the null hypothesis) or "two-tailed" (not equal to the null hypothesis).

- **First Sample Size and Second Sample Size.** For the **Summary Table with Proportions Data Type**, specifies the size of the first sample and (for two-sample analysis only) the size of the second sample.

- **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable 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(s) required to calculate confidence intervals

The Sample Size Selection command determines the sample size(s) 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, because sample sizes are typically determined before you collect data. You need to specify the confidence level, the desired half-length, and any other parameters necessary for determining sample size.

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

The options in the Sample Size Selection dialog include:

- **Parameter to Estimate.** Selects the type of parameter that will be estimated from the sample(s). Options include **Mean**, **Proportion**, **Differences of Means**, and **Differences of Proportions**.

- **Confidence Level.** The confidence level (usually 90%, 95%, or 99%, but other values are allowed).

- **Half-Length of Interval.** The desired “plus or minus” part for the confidence interval.

- In the last box (or last two boxes), enter the estimated standard deviation when Parameter to Estimate is **Mean** or **Difference of Means**; enter the estimated proportion when Parameter to Estimate is **Proportion**; and enter estimates of both proportions when Parameter to Estimate is **Difference of Proportions**.
### Sample Size Estimate Report

<table>
<thead>
<tr>
<th>StatTools Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis: Sample Size Selection</td>
</tr>
<tr>
<td>Performed By: Palisade</td>
</tr>
<tr>
<td>Date: Thursday, April 9, 2015</td>
</tr>
<tr>
<td>Updating: Live</td>
</tr>
</tbody>
</table>

#### Sample Size for Mean

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence Level</td>
<td>95.00%</td>
</tr>
<tr>
<td>Half-length of Interval</td>
<td>0.1</td>
</tr>
<tr>
<td>Std Dev (estimate)</td>
<td>2</td>
</tr>
<tr>
<td>Sample Size</td>
<td>1537</td>
</tr>
</tbody>
</table>

### Missing Data and Link to Data

- **Missing Data.** - Not relevant.
- **Link to Data.** - Not relevant.
One-Way ANOVA Command

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.

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

The selected data sets can be stacked or unstacked data. With stacked data, you select one categorical variable and one numeric variable. With unstacked data, you select several numeric variables, such as a Salary variable for each of several neighborhoods. 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 differences: **No Correction, 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 to 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.
In this One-Way ANOVA report, summary statistics for each population (in this case, each brand) 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 distances from the five brands are not all equal. To see which means are different from which others, you look at the confidence intervals at the bottom of the report. Those pairs with values in bold have significantly different means (because this intervals don’t include 0).
• **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 selected 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, and observations on this variable would be collected for a number of balls of each Brand / Temp combination. The purpose of the study is to see whether there are significant mean differences across the various treatment level combinations.

The data for two-way ANOVA must be in stacked form. That is, there must be two categorical variables (corresponding to Brand and Temp in the above example), and there must be a numeric variable (corresponding to Yards 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).

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

![Two-Way ANOVA Dialog](image)

Two categorical variables (C1 and C2) and one numeric variable (Val) need to be selected for analysis. The selected data set needs to be in stacked format.
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.)

The purpose of interactions is to estimate whether the effect of one factor depends on the level of the other factor. For example, one brand might go farther in cold temperatures and another might go farther in warm temperatures. This would be an interaction. StatTools provides graphs (not shown here) for detecting possible interactions.
• **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. 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.

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

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** (headers in the leftmost column and topmost row of the table); **Columns Title** (the title you want to use to represent the columns in the report); **Rows Title** (the title you want to use to represent the rows in the report).
This report shows the main result of the test, a p-value. If this p-value is small (as here), you can conclude that the row and column attributes are not independent. You can then analyze the tables 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. If the counts in the original contingency table change, the outputs in the report change automatically.
Chi-square Goodness-of-Fit Test Command

Performs a hypothesis test for goodness of fit of the frequency distribution of a categorical variable

The Chi-square Goodness-of-Fit Test is applied to see if the frequency distribution of a categorical variable fits a specific pattern. Two values are involved, an observed value, which is the frequency of a category from a sample, and the expected value which is calculated based upon the claimed distribution.

The hypothesis tests can be performed assuming a uniform distribution (equal proportions for each category) or a specific distribution.

This analysis is set up using the Chi-square Goodness-of-Fit Test dialog:

You should select one variable for testing. The selected data set is always treated as unstacked data.
The options in the Chi-square Goodness-of-Fit Test dialog include:

- **Analysis Type.** Selects the type of analysis performed. Options include:

- **Equal Proportions.** The null hypothesis states that the frequency distribution of the variable is uniform which means that each category have equal proportions.

- **Specific Proportions.** In this case, the null hypothesis states that the frequency distribution of the variable is different to the uniform distribution, so expected values should be specified by the users.

- **Input Type.** Selects the type of information available for running the analysis. Options include:
  - **Data Set.** Select this option if you have the data set.
  - **Summary Statistics.** Select this option if you don’t have the data set but you know some statistics from it.
  - **Unknown Data Set.** Check this option if the data set is not available but the frequency table is available instead.
This report shows the main result of the test, a p-value. If this p-value is small null hypothesis is rejected. A bar chart of the observed and expected values is also shown.
**Missing Data and Link to Data**

- **Missing Data.** Missing data are allowed. All cells with missing data in the selected variables are ignored.

- **Link to Data.** The report is generated with formulas that are linked to the data. If the values of the selected variables change, the reported values change automatically. However, if new categories are included in the data, you will need to rerun the analysis.
Normality Tests Menu

Because many statistical procedures assume that a variable 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 whether 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, you 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.

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

This analysis is set up using the Chi-Square Normality Test dialog:

![Chi-Square Normality Test Dialog]

You should select one variable for testing. The selected data set is always treated as unstacked data.
The options in the **Chi-Square 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.** Specifies that the minimum and maximum of your data set will be used to calculate the minimum and maximum of equal interval bins. The 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 extend 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 extend 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.

The following report shows the results of the test. The p-value of 0.4776 is good evidence that the amounts are normally distributed. More evidence to this effect appears in the histograms 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 non-normality. 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 non-normality. 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. For each variable, the missing data for that variable are ignored.

- **Link to Data.** The histogram and all formulas for the test are linked to the original data. If the data change, the histogram and the test results change automatically.
Lilliefors Test Command

Tests whether 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 a given 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, there is have evidence that the data do not come from a normal distribution.

This analysis is set up using the Lilliefors Test dialog:

One or more variables can be selected for testing. The selected data set is always treated as unstacked data. Variables can be from different data sets.

The following report shows the results of the test. There is no p-value (as in most hypothesis tests), but you can see from the test results 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, you might conclude that these data are "close enough" to being normally distributed for practical purposes.
• **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable are ignored.

• **Link to Data.** The CDFs and all formulas for the test are linked to the original data. If the data change, the graph and the test results change automatically.
Q-Q Normal Plot Command

Tests informally whether 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).

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

You should select one variable for plotting. The selected data set is always treated as unstacked data.

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

- **Plot Using Standardized Q-Values.** Specifies that a standardized Q-Value, instead of Q-Q data, should be used on the Y-axis of the graph. This makes comparisons of the Y-axis values between Q-Q Normal plots clearer.
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, you look for obvious curvature in the plot, and arguably none is apparent in this plot.

**Missing Data and Link to Data**

- **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable are ignored.
- **Link to Data.** The plots and all formulas are linked to the original data. If the data change, the plot changes automatically.
The procedures on the Time Series and Forecasting menu are used to analyze 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 trend and seasonality).
Time Series Graph Command

Creates time series graphs for selected variables

The Time Series Graph command plots one or more time series variables. 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. If you want to use it to label the horizontal axis of the chart, you must select it as the "label" variable.

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

One or more variables can be selected for graphing. The selected data set is always treated as unstacked data. Variables can be from different data sets. The Label variable (in the Lbl column) appears on the X-axis.

The options in the Time Series Graph dialog include:

- **Plot All Variables on a Single Graph.** Plots all variables in a single graph.

- **Use Two Y-Axes.** Displays 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.
**Missing Data and Link to Data**

- **Missing Data.** Missing data are allowed. For each variable, the missing data for that variable are ignored.
- **Link to Data.** Graphs are linked to data. If the data changes, the graph updates automatically.
**Autocorrelation Command**

Calculates the autocorrelations for selected variables

The Autocorrelation Command calculates the autocorrelations for any selected numeric variables. Typically, these variables will be time series variables, although StatTools will perform the calculations for any variables. 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.

This analysis is set up using the **Autocorrelation** dialog:

![Autocorrelation Dialog](image)

One or more variables can be selected for analysis. The selected data set is always treated as 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.
In the following Autocorrelation report, the corresponding autocorrelation appears for each lag, along with an *approximate* standard error.

### Autocorrelation Report

#### StatTools Report

- **Analysis:** Autocorrelation
- **Performed By:** Palisade
- **Date:** Thursday, April 9, 2015
- **Updating:** Live

#### Autocorrelation of Sales / Sales Data

![Autocorrelation Chart]

#### Autocorrelation Table

<table>
<thead>
<tr>
<th>Number of Values</th>
<th>Sales Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>0.1443</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.3492</td>
</tr>
<tr>
<td>Lag #1</td>
<td>0.0772</td>
</tr>
<tr>
<td>Lag #2</td>
<td>0.0814</td>
</tr>
<tr>
<td>Lag #3</td>
<td>-0.0095</td>
</tr>
<tr>
<td>Lag #4</td>
<td>-0.1353</td>
</tr>
<tr>
<td>Lag #5</td>
<td>0.0206</td>
</tr>
<tr>
<td>Lag #6</td>
<td>-0.1494</td>
</tr>
<tr>
<td>Lag #7</td>
<td>-0.1492</td>
</tr>
<tr>
<td>Lag #8</td>
<td>-0.2626</td>
</tr>
<tr>
<td>Lag #9</td>
<td>-0.1792</td>
</tr>
<tr>
<td>Lag #10</td>
<td>0.0121</td>
</tr>
<tr>
<td>Lag #12</td>
<td>-0.0516</td>
</tr>
</tbody>
</table>
• **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. If the data change, the autocorrelations (and the correlogram) update automatically.
Runs Test for Randomness Command

Tests whether selected variables are random

The Runs Test for Randomness command allows you to check the "randomness" of a sequence of values in one or more variables, usually time series variables. It reports the number of "runs" 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, you would expect neither too few nor too many runs. The runs test reports a p-value for the test. If this p-value is small, you can conclude that sequence is probably not random, with either too many or too few runs.

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

One or more variables can be selected for analysis. The selected data set is always treated as 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.

The following report shows the number of runs, 20, and the expected number of runs under randomness, 24.83. This provides some evidence of non-randomness (less zigzagging than expected), but the difference isn’t significant at the usual levels of significance.

<table>
<thead>
<tr>
<th>Sales</th>
<th>Sales Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Runs Test for Randomness</strong></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>48</td>
</tr>
<tr>
<td>Below Mean</td>
<td>22</td>
</tr>
<tr>
<td>Above Mean</td>
<td>26</td>
</tr>
<tr>
<td>Number of Runs</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>182.67</td>
</tr>
<tr>
<td>E(R)</td>
<td>24.8333</td>
</tr>
<tr>
<td>StdDev(R)</td>
<td>3.4027</td>
</tr>
<tr>
<td>Z-Value</td>
<td>-1.4204</td>
</tr>
<tr>
<td>P-Value (two-tailed)</td>
<td>0.1555</td>
</tr>
</tbody>
</table>

**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. If the data change, the reports updates automatically.
Forecasting Command

Generates forecasts for selected time series variables

The Forecasting command provides 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 trend and 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 that 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 also 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.
The analysis is set up using the **Forecasting** dialog (with three tabs):
One variable can be selected for analysis. The selected data set is always treated as unstacked data.

The options in the **Forecasting Settings** tab of 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 (zero 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 constants (between 0 and 1) that minimize the RMSE (for the non-holdout period).

- **Method.** Specifies 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.** (Moving Average method only) The number of consecutive observations used in each moving average.
  
  - **Level.** (All Exponential Smoothing methods) A smoothing parameter between 0 and 1.
  
  - **Trend.** (Holt's and Winter's Exponential Smoothing methods) A second smoothing parameter between 0 and 1.
  
  - **Seasonality.** (Winter's Exponential Smoothing method only) A third smoothing parameter between 0 and 1.

  **Note:** If Optimize Parameters is selected, the Level, Trend, and Seasonality smoothing constants are found through the optimization procedure.

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

- **Deseasonalize.** Specifies that the data will be deseasonalized before forecasting. For 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.

- **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 time series graphs.

- **Starting Label.** Specifies the entry for the first time scale label on the graphs.
The options on the **Graphs to Display** tab specify the forecast graphs that will be generated. Available graphs include:

- **Forecast Overlay**. The time series graph of the data values generated by the forecast.
- **Original Series**. The time series graph of the observed data.
- **Forecast Errors**. The error between the forecast and observed data.
- **Deseasonalized Forecast Overlay**. The time series graph of the data values generated by the forecast after the observed data have been deseasonalized.
- **Deseasonalized Original Series**. The time series graph of the observed data after it has been deseasonalized.
- **Deseasonalized Forecast Errors**. The error between the forecast and observed after the observed data has been deseasonalized.
• **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 Multiple, Stepwise, Forward, Backward, and Block. The classification methods include logistic regression and discriminant analysis.
Regression Command

Runs regression analyses on a set of variables

The Regression analysis supports different regression models, including Dummy Variable Regression (regression with categorical variables), Polynomial Regression, and Regression with Interactions between Variables. The optional Regression Wizard helps you set up these models if you decide to use it.

With each of the models, different methods for the selection of variables to include in the equation are available. These are referred to as "Regression Types" and include Multiple, Stepwise, Forward, Backward, and Block methods. Reports from each regression include summary measures, an ANOVA table, and a table of estimated regression coefficients, their standard errors, their t-values, their p-values, and 95% confidence intervals.

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

Regression Wizard

The Regression Wizard offers an easy way to create regression models, like Dummy Variable Regression or Polynomial Regression. These models require the inclusion of variable transformations ("derived variables") in the analysis; the Wizard makes it easy to include the transformations.

The Wizard guides you through four steps:

Step 1: Model selection. Here you specify that type of model that you want to create. Options include:

- **Standard Regression Model.** Multiple Regression model.
- **Dummy Variable Regression Model.** Multiple Regression model which includes category independent variables. The 0/1 coding for the categories is performed automatically.
- **Polynomial Regression Model.** Multiple Regression model including powers of independent variables. You can also include all the possible interactions between the original variables.
- **Regression Model with Interactions between Variables.** Multiple Regression model including all the possible interactions between independent variables.
• **Regression Model with Nonlinear Variable Transformations.** Regression model where some nonlinear transformations are applied to the variables. Options include:

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$\log(y) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p$</td>
</tr>
<tr>
<td>Reciprocal</td>
<td>$1/y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p$</td>
</tr>
<tr>
<td>Logarithmic</td>
<td>$y = b_0 + b_1\log(x_1) + b_2\log(x_2) + \ldots + b_p\log(x_p)$</td>
</tr>
<tr>
<td>Power</td>
<td>$\log(y) = b_0 + b_1\log(x_1) + b_2\log(x_2) + \ldots + b_p\log(x_p)$</td>
</tr>
</tbody>
</table>

**Step 2: Selection of independent variables.** Here you select one or more independent variables and any desired variable transformations.

**Step 3: Selection of the dependent variable.** Here you select the dependent variable. With some model types you also select a transformation of the dependent variable.

**Step 4: Choice of selection method.** Here you choose the "Regression Type" - Multiple, Stepwise, Forward, Backward, or Block.

When the Wizard completes, the settings can be adjusted using the Regression dialog.

The available regression types (methods of selecting variables) include **Multiple, Stepwise, Forward, Backward, and Block**.

The Multiple regression procedure builds an equation all at once, using the selected independent variables. (It includes the procedure usually called “simple” regression, where there is a single independent variable.) 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 dependent 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 independent 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 independent 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.

These analysis are set up using the Regression dialog:
The options in the Regression dialog include:

- **Setup Using Wizard.** If you select this option, a sequence of dialogs will guide you through the process of creating different regression models.

- **Regression Type.** Here you choose the method of selecting variables: Multiple, Stepwise, Forward, Backward, or Block.

- **Fix Independent Variables.** If you select this option, an additional F column will appear in the variable selection grid. The F column allows you to select "fixed" variables, that is, variables that will never be excluded from the equation during the variable selection process.

One dependent variable (D) and one or more independent variables (I) must 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) of independent variables must be selected.
Because categorical variables must be transformed before they can be used in a regression analysis, selected variables need to be designated as either "Category" variables or "Numeric" variables. The selected data set must be in the unstacked format. Variables can be from different data sets. However, the derived variables option is not available with multiple data sets.

If the **Include Derived Variables** check box is selected, transformations of variables such as the square of a variable can be added to the analysis. When the **Add** button is clicked, the **Add Derived Variables** dialog shows.

The Add Derived Variables dialog box offers a quick and easy way to create new variables which are transformations of the original variables. Options include:

- **Power.** Select this option to apply power transformations to the variables.
- **Logarithm.** Select this option to apply logarithmic transformations to the variables.
- **Interaction without Category Variable.** Select this option to include an interaction term which is the product of two or more numeric variables.
- **Interaction with Category Variable.** Select this option to include an interaction term based on one or more numeric variables and one category variable.
The Parameters 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), you should use larger p-values. To make it harder for variables to enter (and easier to leave), you should use smaller p-values. Typical p-values are in the range 0.01 to 0.1.

- **Use F-Values.** As with p-Values, this setting affects how variables are added or removed from the regression equation. It allows you to specify an F to Enter and/or an F to Leave, depending on the regression method in use. Typical values are in the range 2.5 to 4.
- **Set Constant to Zero.** Forces the regression to go through the origin.

- **Standardize Numeric Independent Variables.** Subtracts the mean from each variable and divides it by the standard deviation. This is sometimes useful if the independent variables have very different magnitudes.

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**
- **Residual vs. Order Index**
- **Histogram of Residuals**
- **Identify Outliers in Graphs.** Highlights outliers of regression in each scatter plot.

These plots are typically used in "residual analysis" to check whether the regression assumptions are satisfied.
Advanced Options for a regression analysis include:

- **Display Regression Equation.** This spells out the regression equation(s) in the report. It is useful when one or more categorical variables are included. Then a separate equation for each combination of categories is shown.

- **Check Multicollinearity.** This includes the Variance Inflation Factor (VIF) of each independent variable in the report. Large VIF values indicate multicollinearity. The correlation matrix can also be included.

- **Include Detailed Step Information.** This reports R-squared and standard error statistics at each intermediate step of the regression (when any procedure other than Multiple is used).
• **Include Prediction for Data Set.** This generates predicted values for the dependent variable for 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 Prediction intervals and confidence intervals (for the means) are also included.

• **Identify Outliers in Data Set.** Inserts an extra column in the data set to indicate if an observation is or not an outlier of regression.

**Regression Report**

![Regression Report Image]
If the **Include Prediction for Data Set** option is selected, predictions will be included to the right of the selected data set. A column for predictions, two columns for the prediction interval, and two columns for the confidence interval are included in the report.

<table>
<thead>
<tr>
<th>Month</th>
<th>Machine Hours</th>
<th>Production Runs</th>
<th>Overhead Prediction</th>
<th>Lower Limit</th>
<th>Prediction</th>
<th>Upper Limit</th>
<th>Confidence Lower Limit</th>
<th>Confidence Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>1374</td>
<td>24</td>
<td>85001.5225</td>
<td>79015.8492</td>
<td>93987.1957</td>
<td>82065.4556</td>
<td>87957.0883</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>1510</td>
<td>35</td>
<td>100645.1033</td>
<td>91957.7338</td>
<td>109332.4728</td>
<td>93847.4465</td>
<td>102942.3420</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>1215</td>
<td>21</td>
<td>75565.3451</td>
<td>66086.2771</td>
<td>84644.4131</td>
<td>75042.1451</td>
<td>89008.0452</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1629</td>
<td>27</td>
<td>92778.4150</td>
<td>89988.4748</td>
<td>107597.0553</td>
<td>96321.2543</td>
<td>101027.6158</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>1858</td>
<td>28</td>
<td>109707.0164</td>
<td>99558.2390</td>
<td>119857.7938</td>
<td>104157.2959</td>
<td>115256.7369</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>1763</td>
<td>40</td>
<td>114135.8803</td>
<td>106874.7952</td>
<td>125396.9053</td>
<td>112457.7269</td>
<td>119842.0336</td>
<td></td>
</tr>
</tbody>
</table>

**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 a nonlinear type of regression analysis where the dependent variable is binary (0 or 1). There should be a 0-1 dependent variable that specifies whether the category of each observation, plus one or more independent variables that can be used to estimate the probability of category 1.

A second data set format for logistic regression is to have a "count" variable that specifies the number of "trials" observed at each combination of independent variables. Then the dependent variable should indicate the number of trials resulting in value 1.

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.

The StatTools logistic regression procedure relies on optimization to find the regression equation. This optimization uses a complex nonlinear algorithm, so the procedure can take a while, depending on the number of variables and observations, and the speed of your PC.

This analysis is set up using the Logistic Regression dialog:
One dependent variable (D) and one or more independent variables (I) need to be selected for analysis. The data must:

- Be numeric. If you have categorical variable such as Gender, with values Male and Female, for example, you must convert it to a dummy variable for the analysis.
- Be in "stacked" form (the Samples with No Count Variable option), where there is a 0-1 dependent variable, or (the Summary of Samples (with Count Variable) option) have a "count" variable and an integer dependent 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).
- **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 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.
This 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). A few of the classifications are listed at the bottom. For each observation, the estimated probability of the dependent variable being 1 is calculated from the regression. If this probability is greater than 0.5, the observation is classified as a 1 (a lasagna trier); otherwise, it is classified as a 0.

- **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 independent 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 independent 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.

This analysis is set up using the Discriminant Analysis dialog:

![Discriminant Analysis Dialog](image)

One dependent variable (D) and one or more independent variables (I) need to be selected for analysis. The data set is always treated as unstacked data. 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 want 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 being analyzed. 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 discriminant function estimated 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.

If there are exactly two groups possible for the dependent category variable (as in this example) and the **Use Misclassification Table** option is selected, the following dialog 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.

*Misclassification Costs Dialog*
The Discriminant Analysis report 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, 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). A few of the classifications are shown at the bottom. 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. If there were more than two groups, the discriminant values would not appear, and classification would be based on the smallest of the statistical distances.

- **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 to quality control.

The Pareto chart displays the relative importance of various categories in categorical data.

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

Creates Pareto chart for categorical variable

Pareto charts are useful for determining the most significant items in a group of categorical 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 with the most significance (Pareto’s 80/20 rule).

For example, suppose 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.

The chart is set up using the Pareto Chart dialog:

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 counts the number of times each entry appears in the variable and creates the
The corresponding Pareto chart. When Category and Value is selected, the variables you specify are the categories and counts. 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.

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.

- **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 with frequency 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 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-Bar 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 five randomly selected parts every half hour. The subsample size is then 5. If data are collected for 70 half-hour periods, then the data should be arranged in five adjacent columns and 50 adjacent rows, with variable headings such as SubSamp1 through SubSamp5 above the first row of data.

The purpose of the procedure is to check whether the process 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-bar and R values are charted in separate time series plots around centerlines. The centerline for the X-bar chart is the average of the X-bar values, and the centerline for the R chart is the average of the R values. A simple way to check whether the process is in control is to see whether any of the X-bar values or R values 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.
These graphs are set up using the **XBar and R Control Charts** dialog:

Two or more variables can be selected for analysis. The selected data set is always treated as unstacked data. Variables can be from different data sets.

Graph options in the dialog include:

- **Sigma 1 and 2 Control Limits**. Adds control limit lines at one and/or two standard deviations (sigmas) from the 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 Analysis** and **Runs Above/Below Analysis**. Sequential up or down moves of length 8 or greater

- **Limit Graph Range From Index**. Limits the points on the graph to a range from a starting index to an ending index
The options in the **Control Limit Calculations Based On** section of the dialog determine the data on which control limit calculations are 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, where you enter the **Subsample Size**, **Average R** and **Average X-Bar** that was calculated from the previous data.

**Example X-Bar Chart**

![Example X-Bar Chart](chart.png)

**Test Results**

- Up/Down Runs: 0
- Above/Below Runs: 1
- Zone A Test: 0
- Zone B Test: 2
**Example R Chart**

**Missing Data and Link to Data**

- **Missing Data.** Missing data are not allowed.
- **Link to Data.** Graphs are not linked to the data.
P Chart Command

Creates P charts for time series variables

P charts are created for "attribute" data, where 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 non-conforming. Then a P chart would plot each half hour's fraction of items that are non-conforming. Again, 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 non-conforming items in each sample or a variable that contains the fraction of non-conforming 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.

This graph is set up using the P Control Charts dialog:
You must select one value variable and optionally, a size variable, for analysis. The selected data set is always treated as unstacked data. Variables can be from different data sets.

Input Data options in the dialog include:

- **Numbers of Non-Conforming Items.** Specifies that the Value variable lists the number of non-conforming items in the sample.

- **Fractions of Non-Conforming Items.** Specifies that the Value variable lists the fraction of non-conforming items in the sample.

Sample Size options in the 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 and that instead, each sample is the entered size.

Graph options in the dialog include:

- **Sigma 1 and 2 Control Limits.** Adds control limit lines at one and/or two standard deviations (sigmas) from the 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 Analysis** and **Runs Above/Below Analysis.** Sequential up or down moves of length 8 or greater

- **Limit Graph Range From Index.** Limits the points on the graph to a range from a starting index to an ending index)
**Control Limit Calculations Based On** options in the P Control Charts dialog determine the data on which control limit calculations are 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, using 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 the 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. On each door you 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 car door, then each door should have the same square footage.

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

You should select exactly one variable analysis. The selected data set is always treated as unstacked data.
Graph options in the dialog include:

- **Sigma 1 and 2 Control Limits.** Adds control limit lines at one and/or two standard deviations (sigmas) from the 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 Analysis** and **Runs Above/Below Analysis.** Sequential up or down moves of length 8 or greater.

- **Limit Graph Range From Index.** Limits the points on the graph to a range from a starting index to an ending index.

The options in the **Control Limit Calculations Based On** section of the dialog determine the data on which control limit calculations are 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, using the **Average C** that was calculated from the previous data.
Example C Chart

Test Results

<table>
<thead>
<tr>
<th>Test</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up/Down Runs</td>
<td>0</td>
</tr>
<tr>
<td>Above/Below Runs</td>
<td>0</td>
</tr>
</tbody>
</table>

Missing Data and Link to Data

- **Missing Data.** Missing data are not allowed.
- **Link to Data.** Graphs are not linked to the data.
**U Chart Command**

Creates U charts for time series variables

U charts are similar to C charts, but they plot the *rate* of defects. Using the car door example, suppose the car door square footages are not necessarily equal. Then in a U chart, you would plot the rate of defects per car door. As always, the purpose is to check whether the process is in control.

This procedure requires a data set with 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.

This graph is set up using the **U Control Charts** dialog:
You should select one value variable and optionally, a size variable for analysis. The selected data set is always treated as unstacked data. Variables can be from different data sets.

Input Data options in the dialog include:

- **Numbers of Defects.** Specifies that the Value variable lists the number of defects.
- **Rates of Defects.** Specifies that the Value variable lists the rate of defects (for example, defects per square foot).

Sample Size options in the 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 and instead each sample is the entered size.

Graph options in the dialog include:

- **Sigma 1 and 2 Control Limits.** Adds control limit lines at one and/or two standard deviations (sigmas) from the 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 Analysis** and **Runs Above/Below Analysis.** Sequential up or down moves of length 8 or greater.
- **Limit Graph Range From Index.** Limits the points on the graph to a range from a starting index to an ending index.
The options in the **Control Limit Calculations Based On** section of the dialog determine the data on which control limit calculations are 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, by using the **Subsample Size** and **Average U** that was calculated from the previous data.

**Missing Data and Link to Data**

- **Missing Data.** Missing data are not allowed.
- **Link to Data.** Graphs are not linked to the data.
Nonparametric tests are statistical procedures for testing hypotheses about underlying probability distributions. Parametric hypothesis tests are more familiar and widely used, but the nonparametric alternatives offer advantages that make them more suitable 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 is not valid. 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. For example, the Wilcoxon signed-rank test assumes that the distribution is symmetric. In contrast, two other nonparametric tests, the sign test and the Mann-Whitney test, do not assume anything about the distribution shape.

With small sample sizes, a 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 and is usually preferable. 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.
For certain types of data, parametric tests cannot be used, while some nonparametric ones can. One such case is with 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 2. StatTools includes tests that can be applied to such data: the sign test (one-sample analysis type) and the Mann-Whitney test.

In summary, nonparametric tests are applicable in the following situations, in which parametric tests 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 are ordinal.
Sign Test Command

Performs sign test on selected variables

The Sign Test command performs a hypothesis test 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.

This analysis is set up using the Sign Test dialog:

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 (shown here) requires exactly 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.** 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.** The value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis.** The alternative to the Null Hypothesis Value that will be evaluated during the analysis. The Alternative Hypothesis can be either one-tailed (greater than or less than the null hypothesis value) or two-tailed (not equal to the null hypothesis value).

### Sign Test Report

<table>
<thead>
<tr>
<th>StatTools Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis: Sign Test</td>
</tr>
<tr>
<td>Performed By: Palisade</td>
</tr>
<tr>
<td>Date: Thursday, April 9, 2015</td>
</tr>
<tr>
<td>Updates: Live</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Statistics</th>
<th>Grade 1 - Grade 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>24</td>
</tr>
<tr>
<td>Sample Mean</td>
<td>-6.358333333</td>
</tr>
<tr>
<td>Sample Std. Dev.</td>
<td>13.11745988</td>
</tr>
<tr>
<td>Sample Median</td>
<td>-5.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Grade 1 - Grade 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesized Median (HV)</td>
<td>0</td>
</tr>
<tr>
<td>Alternative Hypothesis</td>
<td>&lt; 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Size Adjustment</th>
<th>Grade 1 - Grade 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Values = HM</td>
<td>0</td>
</tr>
<tr>
<td>Num. of Values &lt; HV (Adjusted Size)</td>
<td>24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Counts</th>
<th>Grade 1 - Grade 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of Values &lt; HV</td>
<td>17</td>
</tr>
<tr>
<td>Num. of Values &gt; HV</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p-Value and Significance Levels</th>
<th>Grade 1 - Grade 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-Value</td>
<td>0.0320</td>
</tr>
<tr>
<td>Null Hypoth. at 10% Significance</td>
<td>Reject</td>
</tr>
<tr>
<td>Null Hypoth. at 5% Significance</td>
<td>Reject</td>
</tr>
<tr>
<td>Null Hypoth. at 1% Significance</td>
<td>Don't Reject</td>
</tr>
</tbody>
</table>
• **Missing Data.** Missing data are allowed. They are ignored in a pairwise manner when a Paired-Sample Analysis is used. Otherwise, for each variable, the missing data for that variable 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 update 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 difference 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:

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 exactly two variables. For a one-sample analysis, the data set can be stacked or unstacked data; for a paired-sample analysis, it has 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.** The value of the population parameter under the null hypothesis.
  - **Alternative Hypothesis.** The alternative to the Null Hypothesis Value that will be evaluated during the analysis. The Alternative Hypothesis can be either **one-tailed** (greater than or less than the null hypothesis value) or **two-tailed** (not equal to the null hypothesis value).

- **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 test 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

**Analysis:** Wilcoxon Signed-Rank Test  
**Performed By:** Palisade  
**Date:** Friday, April 10, 2015  
**Updating:** Live

### Temperature Data

<table>
<thead>
<tr>
<th>Sample Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>24</td>
</tr>
<tr>
<td>Sample Mean</td>
<td>-6.34</td>
</tr>
<tr>
<td>Sample Std. Dev.</td>
<td>13.14</td>
</tr>
<tr>
<td>Sample Median</td>
<td>-5.75</td>
</tr>
</tbody>
</table>

### Hypotheses

- Hypothesized Median (HM): 0  
- Alternative Hypothesis: $\neq 0$

### Sample Size Adjustment

- Number Of Values = HM: 0  
- Num. of Values $<$ or $>$ HM (Adjusted Size): 24

### Ranking Information

- Number of Tied Values: 2  
- Sum of Negative Ranks: 226.5  
- Sum of Positive Ranks (Test Statistic): 73.5

### p-Value Computation

- Normal Approximation (NA) Used: Yes  
- Ties Present, but Not Corrected For: No  
- Mean for NA: 150  
- Std. Dev. for NA with Tie Correction: 34.95711659  
- z-Statistic for NA with Tie Correction: -2.1741  
- p-Value: 0.0297

### Significance Levels

- Null Hypoth. at 10% Significance: Reject  
- Null Hypoth. at 5% Significance: Reject  
- Null Hypoth. at 1% Significance: Don't Reject

---

**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. They are ignored in a pairwise manner when a Paired-Sample Analysis is used. Otherwise, for each variable, the missing data for that variable 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 \( P[X_1>X_2] = P[X_2>X_1] \), where \( P[X_1>X_2] \) 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.

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

The analysis requires two variables. The data set can be stacked or unstacked, and the variables 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. The null hypothesis denies that either probability distribution tends to yield smaller values that the other. More precisely, it states that \( P[X_1 > X_2] = P[X_2 > X_1] \), where \( P[X_1 > X_2] \) is the probability that an observation from population 1 is greater than an observation from population 2, and \( P[X_2 > X_1] \) has an analogous interpretation. For continuous distributions, this is equivalent to saying that both of these probabilities are 0.5 (\( P[X_1 > X_2] = P[X_2 > X_1] = 0.5 \)). The Alternative Hypothesis can be either one-tailed (one probability is greater or less than the other) or two-tailed (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. The null hypothesis states that the two medians are equal. The Alternative Hypothesis can be either one-tailed (the median of the first population is greater or less than that of the second) or two-tailed (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 you cannot assume that the distributions have approximately identical shapes, as long as you consider 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 test statistic, if tied ranks are present. (Note: The tie correction will produce no change in the variance when there are no ties.)

In this report, 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. For each variable, the missing data for that variable 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.
Kruskal-Wallis Test Command

Performs Kruskal-Wallis test on variables

The Kruskal-Wallis Test command performs a hypothesis test on three or more samples. It extends the Mann-Whitney test, which applies to pairs of variables. In the general version of the test, the hypothesis states that all samples come from the same distribution. In this case, continuous distributions are assumed. On the other hand, if it is possible to assume that all population distributions have the same shape (normal or not), the hypothesis states that the medians of the populations are identical.

This analysis is set up using the Kruskal-Wallis dialog:

Three or more variables can be selected. The data set can be in stacked or unstacked form. Variables can be from different data sets.
The options in the Kruskal-Wallis Test dialog include:

**Analysis Type.** Selects the type of analysis performed. Options include:

- **General Version.** It assumes that the value variable has a continuous distribution in each population and the Null Hypothesis states that all samples come from the same distribution.

- **Median Version.** It assumes that all population distributions have the same shape and the Null Hypothesis states that the medians of the populations are identical.

**Mann-Whitney U Test.** The Kruskal-Wallis test, like a one-way ANOVA, indicates whether there are significant differences across categories, but it does not tell you specifically which categories differ, so StatTools provides the option to use the Mann-Whitney U Test with the Bonferroni correction. The Bonferroni correction is an adjustment made to p-values when several statistical tests are being performed simultaneously on the same data.

**Correct for Ties.** Because the variance of the sampling distribution of the KW test statistic is influenced by ties, you can correct for ties in the calculation of this test statistics.
In this report, the p-value is computed using the Chi-Square approximation. StatTools has three different methods to compute the p-value:

- **Chi-square Approximation.** The p-value is computed assuming that sampling distribution of the KW test statistic is closely approximated by the chi-square distribution. The approximation improves as both the number of variables and the number of observations within each variable increase.
• **Exact.** The computation of the p-value is based on the true distribution of the KW test statistic. These computations are very intensive because it is necessary to rearrange the ranks of the sample data in all possible ways (the value of the test statistic needs to be computed with each possible scenario), so it is used for small samples only.

• **Monte Carlo.** This provides an unbiased estimate of the exact p-value and displays a confidence interval using the Monte Carlo sampling method. It is used when data sets are too large to use the exact method or when the assumptions necessary for the chi-square approximation method are not met.

Note that StatTools uses the exact method if the number of variables \( k \leq 4 \) and the number of observations within each variable \( n_i \leq 4 \). The Monte Carlo method is used if the exact method cannot be used, and one of the following is true: there are more than three variables \( (k > 3) \) with at least one of the variables containing a small number of observations \( (n_i \leq 5) \), or there are exactly three variables with the total number of observation in the three variables \( N \leq 30 \). The Monte Carlo method runs 10000 iterations (that is, it selects 10000 samples) and can support simulations with the total number of observations in all the variables \( N \leq 25000 \). If the exact or Monte Carlo methods cannot be used, the chi-square approximation is used.

**Missing Data and Link to Data**

• **Missing Data.** Missing data are allowed. All cells with missing data in the selected variables are ignored.

• **Link to Data.** The live reports are calculated with formulas that are linked to the data. If the values of the selected variables change, the reported values change automatically; the p-value calculation method also changes automatically, as needed.
Multivariate Analysis Menu

The two procedures in this section fall in the general area of multivariate analysis. They work on a number of variables simultaneously. The first tries to find correlated variables that can be combined, so that the dimension of the data set can be reduced. The second tries to find subsets of cases that are “alike.”
Principal Components Analysis Command

Performs Principal Components Analysis on selected variables

The Principal Component Analysis is a procedure for reducing the dimensionality of a multivariate data set. Specifically, it transforms a set of correlated variables into a smaller number of uncorrelated variables.

This analysis is often used as input to other analyses:

- In linear regression it is used if there are multicollinearity problems, or if the number of independent variables is relatively large relative to the number of observations.
- It is useful to construct a scatterplot of the first two principal components to check for outliers and the presence of clusters.

This analysis is set up using the Principal Components Analysis dialog:

Two or more variables can be selected. The selected data set must be in the unstacked format. Variables can be from different data sets.
The options in the Principal Components Analysis dialog include:

- **Matrix Type.** Selects the matrix from which principal components will be extracted. Options include:
  - **Covariance Matrix.** Computes the principal components from a covariance matrix. Use this option if the scales of measurement used for the different variables are commensurate (for example, centimeters and inches are commensurate, while inches and degrees of Fahrenheit are not.)
  - **Correlation Matrix.** Computes the principal components from a correlation matrix. Use this option if the scales of measurement used for the different variables are not commensurate.

The options in the Number of Principal Components to Compute section include:

- **Fixed Number of Components.** Specifies the number of principal components to compute. Use this option if you already know the number of components you wish to compute.

- **Components which Account for Specified Percentage of Total Variance.** Retains sufficient number of components to account for a specified percentage of the total variance.

- **Components with Greater than Average Component Variance.** Retains a component if its variance is greater than the average of the variances of all the components.

There are several additional options available:

- **Scree Plot.** Plots the percentage of total variance explained by each component.

- **Score Plot for First Two Components.** Plots the first principal component versus the second principal component.

- **Break Down by Category Variable.** Use this option to create a scatter plot for the first two principal components colored by a category variable.

- **Include Ellipsis for Each Category.** This option creates an ellipsis to cluster points of the same category.

- **Place Components in.** Specifies whether variables representing the principal components will be generated. If included, you can choose where to put those variables: in the
original data set or a new one. When numerical variables are selected from different data sets, the only option available is to place the components in a new data set. If the "Place Components in" option is selected, columns with the values of the components are placed to the right of the selected data set.

Scree Plot

<table>
<thead>
<tr>
<th>Component</th>
<th>Variance</th>
<th>Total Variance Explained of Total Variance</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51.630</td>
<td>51.630</td>
<td>51.630</td>
</tr>
<tr>
<td>2</td>
<td>30.753</td>
<td>82.383</td>
<td>82.383</td>
</tr>
<tr>
<td>3</td>
<td>28.125</td>
<td>99.508</td>
<td>99.508</td>
</tr>
<tr>
<td>4</td>
<td>24.983</td>
<td>124.491</td>
<td>124.491</td>
</tr>
<tr>
<td>5</td>
<td>21.908</td>
<td>146.403</td>
<td>146.403</td>
</tr>
<tr>
<td>6</td>
<td>19.827</td>
<td>166.229</td>
<td>166.229</td>
</tr>
<tr>
<td>7</td>
<td>17.509</td>
<td>183.738</td>
<td>183.738</td>
</tr>
<tr>
<td>8</td>
<td>15.075</td>
<td>198.813</td>
<td>198.813</td>
</tr>
<tr>
<td>9</td>
<td>11.064</td>
<td>210.877</td>
<td>210.877</td>
</tr>
<tr>
<td>10</td>
<td>10.259</td>
<td>221.136</td>
<td>221.136</td>
</tr>
</tbody>
</table>

Data: [StatTools Report]

Score Plot

<table>
<thead>
<tr>
<th>Component</th>
<th>Exam 1</th>
<th>Exam 2</th>
<th>Exam 3</th>
<th>Exam 4</th>
<th>Exam 5</th>
<th>Exam 6</th>
<th>Exam 7</th>
<th>Exam 8</th>
<th>Exam 9</th>
<th>Exam 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.259</td>
<td>0.254</td>
<td>0.235</td>
<td>0.242</td>
<td>0.261</td>
<td>0.257</td>
<td>0.263</td>
<td>0.262</td>
<td>0.258</td>
<td>0.270</td>
</tr>
<tr>
<td>2</td>
<td>-0.344</td>
<td>-0.368</td>
<td>-0.339</td>
<td>-0.413</td>
<td>-0.562</td>
<td>-0.374</td>
<td>-0.259</td>
<td>-0.252</td>
<td>-0.362</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Score Plot for First Two Principal Components (94.53 % of Total Variance Explained)
Data Set with Principal Components

Missing Data and Link to Data

- **Missing Data** - Missing data are allowed. All rows with missing data in the selected variables are ignored. Missing categories will be renamed as “N/A” when breaking down by a category variable.

- **Link to Data** – There is no link to the original data. If the data changes, you must rerun the analysis.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Exam 1</th>
<th>Exam 2</th>
<th>Exam 3</th>
<th>Exam 4</th>
<th>Exam 5</th>
<th>Exam 6</th>
<th>Exam 7</th>
<th>Exam 8</th>
<th>Exam 9</th>
<th>Exam 10</th>
<th>PC1</th>
<th>PC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>75</td>
<td>79</td>
<td>73</td>
<td>68</td>
<td>74</td>
<td>80</td>
<td>86</td>
<td>90</td>
<td>83</td>
<td>84</td>
<td>249.2990</td>
<td>-25.6536</td>
</tr>
<tr>
<td>F</td>
<td>97</td>
<td>77</td>
<td>77</td>
<td>73</td>
<td>72</td>
<td>77</td>
<td>78</td>
<td>68</td>
<td>75</td>
<td>72</td>
<td>228.5998</td>
<td>-38.9350</td>
</tr>
<tr>
<td>M</td>
<td>73</td>
<td>67</td>
<td>70</td>
<td>66</td>
<td>63</td>
<td>73</td>
<td>65</td>
<td>61</td>
<td>71</td>
<td>62</td>
<td>207.9636</td>
<td>-38.2430</td>
</tr>
<tr>
<td>F</td>
<td>83</td>
<td>81</td>
<td>72</td>
<td>81</td>
<td>79</td>
<td>70</td>
<td>73</td>
<td>81</td>
<td>72</td>
<td>81</td>
<td>239.7710</td>
<td>-49.3879</td>
</tr>
<tr>
<td>F</td>
<td>77</td>
<td>83</td>
<td>82</td>
<td>90</td>
<td>76</td>
<td>64</td>
<td>69</td>
<td>68</td>
<td>59</td>
<td>63</td>
<td>222.4346</td>
<td>-67.0375</td>
</tr>
<tr>
<td>M</td>
<td>75</td>
<td>76</td>
<td>69</td>
<td>83</td>
<td>81</td>
<td>70</td>
<td>79</td>
<td>85</td>
<td>77</td>
<td>70</td>
<td>236.4790</td>
<td>-45.7031</td>
</tr>
<tr>
<td>F</td>
<td>60</td>
<td>61</td>
<td>56</td>
<td>62</td>
<td>57</td>
<td>73</td>
<td>65</td>
<td>70</td>
<td>81</td>
<td>77</td>
<td>209.5833</td>
<td>-14.6635</td>
</tr>
<tr>
<td>M</td>
<td>73</td>
<td>78</td>
<td>76</td>
<td>72</td>
<td>75</td>
<td>84</td>
<td>97</td>
<td>94</td>
<td>85</td>
<td>96</td>
<td>259.1273</td>
<td>-23.1083</td>
</tr>
<tr>
<td>F</td>
<td>94</td>
<td>91</td>
<td>82</td>
<td>92</td>
<td>88</td>
<td>80</td>
<td>77</td>
<td>74</td>
<td>73</td>
<td>83</td>
<td>256.8007</td>
<td>-56.0052</td>
</tr>
<tr>
<td>F</td>
<td>59</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>52</td>
<td>74</td>
<td>74</td>
<td>68</td>
<td>78</td>
<td>66</td>
<td>200.7141</td>
<td>-6.1006</td>
</tr>
<tr>
<td>F</td>
<td>71</td>
<td>72</td>
<td>76</td>
<td>84</td>
<td>77</td>
<td>79</td>
<td>67</td>
<td>73</td>
<td>74</td>
<td>73</td>
<td>230.8406</td>
<td>-45.8205</td>
</tr>
<tr>
<td>F</td>
<td>66</td>
<td>70</td>
<td>67</td>
<td>76</td>
<td>72</td>
<td>80</td>
<td>73</td>
<td>75</td>
<td>79</td>
<td>82</td>
<td>231.2206</td>
<td>-29.8359</td>
</tr>
<tr>
<td>M</td>
<td>72</td>
<td>70</td>
<td>71</td>
<td>69</td>
<td>66</td>
<td>63</td>
<td>56</td>
<td>57</td>
<td>55</td>
<td>55</td>
<td>193.4346</td>
<td>-55.2790</td>
</tr>
<tr>
<td>F</td>
<td>70</td>
<td>62</td>
<td>79</td>
<td>63</td>
<td>76</td>
<td>74</td>
<td>73</td>
<td>77</td>
<td>77</td>
<td>74</td>
<td>226.7279</td>
<td>-31.6146</td>
</tr>
<tr>
<td>M</td>
<td>96</td>
<td>62</td>
<td>56</td>
<td>74</td>
<td>61</td>
<td>68</td>
<td>69</td>
<td>61</td>
<td>62</td>
<td>71</td>
<td>202.7457</td>
<td>-32.8389</td>
</tr>
</tbody>
</table>
Cluster Analysis Command

Performs Cluster Analysis on a data set

The Cluster Analysis command searches for patterns in a data set in order to classify observations or variables into groups of similar items. This procedure uses several "agglomerative hierarchical" methods. Each of these methods begins with each observation in a separate cluster and then, in a series of steps, combine clusters to end with one single cluster containing all observations. By selecting one of the steps in the procedure before you get to a single cluster, you obtain a division of the data set into groups of similar items.

The analysis supports a variety of agglomerative hierarchical methods and distance measures. You can also select among three options to determine the final number of cluster to be generated. The clustering process is summarized by a graph known as a "dendrogram" and by a table showing the steps of the process.
This analysis is set up using the Cluster Analysis dialog:

Two or more variables can be selected. The selected data set must be in the unstacked format. Variables can be from different data sets.

On top of the Cluster Analysis dialog the user can select:

- **Analysis Type.** Selects the type of analysis to perform:
  Cluster Observations to classify observations into groups, and Cluster Variables to classify variables into groups.
The first tab in this dialog (Clustering Settings) allows you to specify:

**Agglomerative Method.** Hierarchical Clustering is a sequential process. In each step of the agglomerative hierarchical approach, an observation or a cluster of observations is merged into another cluster. To decide which pair of clusters should be merged at each step, these methods compute distances between observations, use these to compute distances between clusters, and then merge the two clusters with smallest distance. Different Agglomerative Clustering Methods are different ways of defining the distance between clusters:

- **Single Linkage (Nearest Neighbor).** Here, the distance between two clusters $A$ and $B$ is defined as the minimum distance between a point in $A$ and a point in $B$:
  \[ d(A,B) = \min \{d(x_i,y_j), \text{for } x_i \text{ in } A \text{ and } y_j \text{ in } B \} \]
  where $d(x_i,y_j)$ is the distance between vectors $x_i$ and $y_j$.

- **Complete Linkage (Farthest Neighbor).** Here, the distance between two clusters $A$ and $B$ is defined as the maximum distance between a point in $A$ and a point in $B$:
  \[ d(A,B) = \max \{d(x_i,y_j), \text{for } x_i \text{ in } A \text{ and } y_j \text{ in } B \} \]
  where $d(x_i,y_j)$ is the distance between vectors $x_i$ and $y_j$.

- **Average Linkage.** Here, the distance between two clusters $A$ and $B$ is defined as the average of the $n_A*n_B$ distances between the $n_A$ points in $A$ and the $n_B$ points in $B$:
  \[ d(A,B) = \frac{\sum\sum d(x_i,y_j)}{(n_A*n_B)} \]
  where the sum is over all $x_i$ in $A$ and all $y_j$ in $B$.

- **Centroid.** Here, the distance between two clusters $A$ and $B$ is defined as the distance between the mean vectors (often called centroids) of the two clusters:
  \[ d(A,B) = d(\bar{x}_A, \bar{y}_B) \]
  where $\bar{x}_A$ and $\bar{y}_B$ are the mean vectors for the observation vectors in $A$ and the observation vectors in $B$, respectively.
- **Median.** Similar to centroid, except that the centroid for the new cluster is computed as the median (mid-point) between old group centroids:

\[ \text{centroid}(A,B) = \left( \bar{x}_A + \bar{y}_B \right) / 2 \]

where \( \bar{x}_A \) and \( \bar{y}_B \) are the “centroids” for cluster A and cluster B, respectively. (Note that this median is not the ordinary median used in statistics.) The distance between clusters is the distance between cluster centroids computed in this way.

- **Ward’s Method.** It is more precisely known as Ward's Minimum Variance method because it uses the minimum variance criterion for choosing the pair of clusters to merge at each step.

**Distance Measure.** Selects the distance measure to be used with the selected agglomerative method. Options include:

- **Euclidean**
- **Squared Euclidean**
- **Mahalanobis**
- **Manhattan** (also referred to as "city block" distance)
- **Correlation**, if you are clustering variables
- **Absolute Correlation**, if you are clustering variables
- **Standardize Variables.** Subtracts the mean from each variable and divides it by the standard deviation before the distance matrix is calculated. Use this option if the scales of measurement are in different units and you wish to minimize the effect of scale differences.

**Final Number of Clusters.** Specifies the final number of clusters to be generated. Options include:

- **Fixed Number**
- **Using Similarity Level.** The similarity level between two clusters A and B is defined as:

\[ s(A,B) = 100(1-d(A,B))/d_{max} \]

where \( d(A,B) \) is the distance between clusters A and B and \( d_{max} \) is the maximum value in the original distance matrix D.

- **Automatic (Using Mojena Method).** The number of clusters selected correspond to the first stage in the dendrogram
satisfying the condition:
\[ a_j > \bar{a} + k s a \text{ for } j=1,2,...,n \]

where \( a_1, a_2, \ldots, a_n \) are the distance values for stages with \( n, n-1, \ldots, 1 \) clusters, \( \bar{a} \) and \( s a \) are, respectively, the mean and standard deviation of the individual \( a \) values, and \( k \) is a constant. Milligan and Cooper recommended \( k = 1.25 \), based on a simulation study ("An examination of procedures for determining the number of clusters in a data set", Psychometrika, June 1985, Volume 50, Issue 2, pp. 159-179).
Options in this dialog specify settings for the graphs, tables and variables that will be generated. Options include:

- **Scree Plot.** Plots the number of clusters on the x-axis (starting with the one-cluster solution at the very left) against the distance at which objects or clusters are combined on the y-axis.

- **Dendrogram.** Plots a tree-like diagram which displays both the cluster-subcluster relationships and the order in which the clusters were merged. Some settings are available for this chart:
  - **Orientation.** Horizontal or Vertical
  - **Y-Axis.** Selects the measure to be shown on the y-axis.
  - **Identify Clusters by Color.** The maximum number of colors available is 15.

- **Profile Plot.** This is a series of colored lines where each line represents a cluster. For each cluster, the mean of each variable is plotted, and the points are connected by a colored line representing that cluster. Lines that are close to each other indicate clusters that are similar and lines that are far apart indicate clusters that are not similar.

- **Distance Matrix.** Shows the initial distance matrix.

- **Agglomeration Steps.** Shows details of the agglomeration steps.

- **Use Observation Labels from Variable.** Selects a variable containing labels for observations. This variable should not have repeated values. It is used in the dendrogram and agglomeration steps table.

- **Place Cluster Membership in.** Inserts a variable in order to identify which cluster each observation belongs to. If included, you can choose where to put those variables: in the original data set or a new one.
Cluster Analysis
Report

StatTools Report
- Analysis: Cluster Analysis
- Performed By: Phukado
- Date: Wednesday, January 20, 2010
- Updating: Data

Cluster Analysis
- Analysis Type: Cluster Observations
- Agglomerative Method: Ward
- Distance Measure: Euclidean
- Standardized Variables
- Final Number of Clusters: Fixed Number
- Number of Cluster: 3
- Similarity Level: 97.41%

Score Plot

Dendrogram
If the "Place Cluster Membership in" option is selected, a column indicating group membership is placed to the right of the selected data set.

### Data Set with Cluster Membership

| Alcohol | Malic Acid | Ash | Alkalinity | Magnesium |总磷 | 非极性磷 | Polyphenols | 颜色 | Hue | Umarell | Proteins | Cluster |
|---------|------------|-----|------------|-----------|-----|---------|------------|------|-----|---------|----------|---------|---------|
| 14.30   | 1.71       | 2.43| 15.6       | 127       | 2.8 | 0.26    | 2.29       | 5.84 | 1.04| 3.02    | 1985     | 2       |
| 11.24   | 2.14       | 2.61| 11.7       | 100       | 2.03| 2.79    | 0.25       | 1.18 | 4.83| 1.95    | 4.4      | 195     |
| 13.15   | 2.36       | 2.67| 18.4       | 185       | 2.8 | 3.24    | 0.3        | 2.81 | 5.68| 1.93    | 3.17     | 1195    |
| 14.17   | 1.95       | 2.5 | 16.8       | 113       | 3.85| 3.49    | 0.24       | 2.18 | 7.6  | 0.16    | 3.4      | 1480    |
| 13.04   | 2.59       | 4.07| 26.8       | 138       | 2.8 | 2.95    | 0.35       | 1.82 | 4.33| 1.84    | 2.95     | 715     |
| 12.36   | 2.95       | 11.5| 132       | 1.97      | 3.37| 3.48    | 0.34       | 1.97 | 6.75| 1.95    | 2.95     | 1400    |
| 14.93   | 2.67       | 2.45| 14.0       | 96        | 2.5 | 2.02    | 0.3        | 1.96 | 5.25| 1.92    | 3.56     | 1290    |
| 14.06   | 2.15       | 3.63| 17.6       | 131       | 2.6 | 2.57    | 0.31       | 1.25 | 3.05| 1.96    | 3.56     | 1395    |
| 14.03   | 1.94       | 2.17| 14.0       | 97        | 2.8 | 2.94    | 0.25       | 1.96 | 3.2  | 1.96    | 2.95     | 1080    |
| 13.88   | 2.77       | 16  | 98        | 198       | 2.98| 3.15    | 0.32       | 1.85 | 7.22| 1.51    | 3.55     | 1685    |
| 14.25   | 2.59       | 2.3 | 159       | 105       | 2.99| 3.42    | 0.22       | 2.80 | 5.75| 1.25    | 3.17     | 1510    |

### Missing Data and Link to Data

- **Missing Data.** Missing data are allowed. All rows with missing data in the selected variables are ignored.

- **Link to Data.** There is no link to the original data. If the data change, you must rerun the analysis.
Utilities Menu

The Utilities menu includes the Application Settings command, along with several other handy commands discussed here.
Application Settings Command

Specifies settings for StatTools reports, graphs, utilities, data sets, and analyses

The Application Settings command allows you to specify general settings for StatTools reports, graphs, utilities, data sets, and analyses. These settings apply across analyses and data sets. Other analysis-specific settings are defined in the dialog for each analysis.
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.** A new worksheet is created for each report.
  - **In New Workbook.** 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.** Each StatTools report is placed on the active sheet to the right of the last used column.
  - **Query for Starting Cell.** 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.** If a new workbook is created, that same new workbook will be used for all reports.
  - **Overwrite Existing Reports.** Specifies whether you want existing reports (for a given procedure) to be overwritten by new ones.

- **Updating Preference.** Specifies how results will change when variable data is changed. Options for results updating include:
  - **Live - Values Change With Input Data.** Reports update automatically as input data changes.
  - **Static - Values are Fixed.** 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:

\[=\text{StatMean(ST_Pair)}\]

calculates the mean of the variable with range name *ST_Pair*. As data in the ST_Pair range changes, the value returned by the StatMean function updates automatically.
Reports and graphs from all StatTools procedures can be updated live, with the following exceptions: Regression, Logistic Regression, Discriminant Analysis, and Forecasting. 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:

```
StatTools Report

Hypothesis Test (One-Sample)

- Sample Size
- Sample Mean
- Sample Std Dev
- Hypothesized Mean
- Alternative Hypothesis
- Standard Error of Mean
- Degrees of Freedom
- t-Test Statistic
- p-Value
- Null Hypoth. at 10% Significance
- Null Hypoth. at 5% Significance
- Null Hypoth. at 1% Significance

StatTools Note: These results are based on the assumption that the variable(s) are approximately normally distributed. If this is not the case, then these results might not be valid, especially if the sample size is small.
```
Utilities Settings

The Utilities settings specify options for new variables created by the StatTools from the Data Utilities menu.

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

Note, however, that the Stack and Unstack procedures always place new variables in a new worksheet.

- **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.** New variable values update automatically as input data changes.
  - **Static – Values are Fixed.** 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 is not followed and new variables always use the **Static – Values are Fixed** preference. These include Stack, 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.

- **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 (or Row).** Specifies if names are entered in the first column (or row) of the data set.
The **Analyses** settings specify the default entries displayed in dialogs used to set up analyses. These are the entries that appear when you first display a dialog for an analysis. They can be changed if desired in each dialog.

- **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 dialogs used to set up analyses. Options include:
  - **Remember Last Used Values (by Workbook).** A displayed dialog 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.** A displayed dialog 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 provide “better” values. Available methods (and the type of data they are suited for) include:
  - Automatic (Based on Input Data)
  - Interpolated with Asymmetric Endpoints (Continuous)
  - Interpolated with Symmetric Endpoints (Continuous)
  - Excel Percentile Function (Continuous)
  - Closest Observation (Discrete)
  - Empirical Dist. Function (Discrete)
  - 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. This deletes only the definitions of the data sets, not the actual Excel data.

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 dialogs. Subsequent displayed dialogs will initially show the saved system default settings for each analysis.

Load DecisionTools Add-In Command

Loads another selected DecisionTools Add-in

The Load DecisionTools Add-in Command provides a handy way to load any of the other Palisade add-ins, such as @RISK or PrecisionTree, while keeping StatTools open.

Unload StatTools Add-In Command

Unloads the StatTools Add-in

Help Menu

The Help menu provides a number of self-explanatory commands for learning more about StatTools. You can open the document you are reading now, you can show the Welcome screen (with links to the Quick Start and Guided Tour videos), and you can open an “example file list” file with links to a variety of StatTools example files.

In addition, the License Activation command displays the License Activation dialog, listing the version and licensing information for your copy of StatTools. From here, you can also convert a trial version of StatTools into a licensed copy.

Finally, the About StatTools command displays the version and copyright information for your copy of StatTools.
Introduction

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

- Statistics calculations to be embedded in worksheet formulas, just as with standard Excel functions.
- 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 functions. All StatTools functions begin with the prefix "Stat", such as StatMean() and 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 statistical 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.
Distribution Functions

StatTools includes a set of distribution functions (such as \textit{StatBinomial}) which replace Excel's built-in distribution functions (such as \textit{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 \texttt{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 for the entered distribution:

<table>
<thead>
<tr>
<th>Entered Value or String</th>
<th>Returned Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or &quot;mean&quot;</td>
<td>mean</td>
</tr>
<tr>
<td>2 or &quot;stddev&quot;</td>
<td>standard deviation</td>
</tr>
<tr>
<td>3 or &quot;variance&quot;</td>
<td>variance</td>
</tr>
<tr>
<td>4 or &quot;skewness&quot;</td>
<td>skewness</td>
</tr>
<tr>
<td>5 or &quot;kurtosis&quot;</td>
<td>kurtosis</td>
</tr>
<tr>
<td>6 or &quot;mode&quot;</td>
<td>mode</td>
</tr>
<tr>
<td>7 or &quot;discrete mean&quot;</td>
<td>discrete mean (or the value closest to the true mean that actually could occur)</td>
</tr>
<tr>
<td>8 or &quot;x to y&quot;</td>
<td>x to y (the distribution y-value for an entered x-value)</td>
</tr>
<tr>
<td>9 or &quot;x to p&quot;</td>
<td>x to p (the distribution p-value for an entered x-value)</td>
</tr>
<tr>
<td>10 or &quot;p to x&quot;</td>
<td>p to x (the distribution x-value for an entered p-value)</td>
</tr>
<tr>
<td>11 or &quot;x to q&quot;</td>
<td>x to q (the distribution q-value for an entered x-value)</td>
</tr>
<tr>
<td>12 or &quot;q to x&quot;</td>
<td>q to x (the distribution x-value for an entered q-value)</td>
</tr>
</tbody>
</table>

For example, the StatTools distribution function:

=\text{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 mean 10 and standard deviation 1.
Live Reports

StatTools uses custom functions to make the results "live" whenever 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 this variable, such as its mean and standard deviation. The One-Variable Summary procedure enters formulas in the output cells: =StatMean(ST_Weight) and =StatStdDev(ST_Weight), where ST_Weight is the StatTools range name for the variable 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, the results change automatically when your data change, 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 (in the Application Settings dialog) to not link your reports to your data. This is useful if Excel recalculation time becomes an issue as data changes.
### Table of Available Functions

This table lists the custom functions that are added to Excel by StatTools. Note that these functions are preceded by **Stat**.

<table>
<thead>
<tr>
<th>Function</th>
<th>Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>StatAutocorrelation</strong>(<em>data</em>, <em>numLags</em>)</td>
<td>Calculates the autocorrelation for a set of values, usually a time series</td>
</tr>
<tr>
<td><strong>StatAveDev</strong>(<em>Data1</em>,<em>Data2</em>,...<em>Data N</em>)</td>
<td>Calculates the average absolute deviation of the data from their mean. The arguments can be numbers, arrays, or ranges.</td>
</tr>
<tr>
<td><strong>StatBinomial</strong>(<em>N</em>,<em>P</em>,<em>statistic</em>,<em>value</em>)</td>
<td>Calculates the <em>statistic</em> for the specified binomial distribution</td>
</tr>
<tr>
<td><strong>StatCategoryIndices</strong>(<em>range</em>, <em>category_name</em>)</td>
<td>Returns the cell indices for a specified category</td>
</tr>
<tr>
<td><strong>StatCategoryNames</strong>(<em>range</em>)</td>
<td>Returns the names of the categories in a range</td>
</tr>
<tr>
<td><strong>StatCategoryOccurrenceCount</strong>(<em>range</em>, <em>category_name</em>)</td>
<td>Calculates the number of cells in a range in a specified category</td>
</tr>
<tr>
<td><strong>StatChiSq</strong>(<em>deg_freedom</em>,<em>statistic</em>,<em>value</em>)</td>
<td>Calculates the <em>statistic</em> for the specified one-tailed chi-square distribution</td>
</tr>
<tr>
<td><strong>StatCorrelationCoeff</strong>(<em>data1</em>, <em>data2</em>,<em>flag</em>)</td>
<td>Calculates the correlation coefficient between two sets of values. Last argument specifies whether linear (Pearson) or rank-order (Spearman) correlation is calculated.</td>
</tr>
<tr>
<td><strong>StatCount</strong>(<em>Data1</em>,<em>Data2</em>,...<em>DataN</em>)</td>
<td>Calculates the number of elements in its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatCountCategories</strong>(<em>range</em>)</td>
<td>Counts the number of the categories in a range</td>
</tr>
<tr>
<td><strong>StatCountCellsByType</strong>(<em>range</em>, <em>type</em>)</td>
<td>Calculates the number of cells of a specified type in a range</td>
</tr>
<tr>
<td><strong>StatCountRange</strong>(<em>Range</em>,<em>MinValue</em>,<em>MaxValue</em>,<em>IncludeMin</em>,<em>IncludeMax</em>)</td>
<td>Calculates the number of values in a range that fall between a minimum and maximum value</td>
</tr>
<tr>
<td><strong>StatCovariance</strong>(<em>data1</em>,<em>data2</em>)</td>
<td>Calculates the sample covariance between two data sets</td>
</tr>
<tr>
<td><strong>StatCovarianceP</strong>(<em>data1</em>,<em>data2</em>)</td>
<td>Calculates the population covariance between two sets of values. Any missing numbers cause a blank to be returned.</td>
</tr>
<tr>
<td>Function</td>
<td>Returns</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>StatDestack(\text{data_range, }) (\text{categories_range_1,category_1, categories_range_2,category_2})</td>
<td>Extracts data in a specified category from stacked data</td>
</tr>
<tr>
<td>StatDurbinWatson(\text{data})</td>
<td>Calculates the Durbin-Watson statistic for a set of values, usually a time series</td>
</tr>
<tr>
<td>StatF(\text{deg_freedom1,deg_freedom2,statistic, value})</td>
<td>Calculates the statistic for the specified F distribution</td>
</tr>
<tr>
<td>StatGetCellValues(\text{range})</td>
<td>Returns the values for all the cells of a specified type in the range</td>
</tr>
<tr>
<td>StatKurtosis(\text{Data1,Data2,...DataN})</td>
<td>Calculates the sample kurtosis of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td>StatKurtosisP(\text{Data1,Data2,...DataN})</td>
<td>Calculates the population kurtosis of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td>StatLN(\text{x})</td>
<td>Calculates the natural logarithm, defined for positive numbers only</td>
</tr>
<tr>
<td>StatMax(\text{Data1,Data2,...DataN})</td>
<td>Calculates the maximum of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td>StatMean(\text{Data1,Data2,...DataN})</td>
<td>Calculates the arithmetic mean (average) of its arguments, which can be numbers, arrays, or range</td>
</tr>
<tr>
<td>StatMeanAbs(\text{Data1,Data2,...DataN})</td>
<td>Calculates the arithmetic mean (average) of the absolute values of its arguments, which can be numbers, arrays, or range</td>
</tr>
<tr>
<td>StatMedian(\text{data, discrete flag})</td>
<td>Calculates the median of a set of values</td>
</tr>
<tr>
<td>StatMin(\text{Data1,Data2,...DataN})</td>
<td>Calculates the minimum of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td>StatMode(\text{Data1,Data2,...DataN})</td>
<td>Calculates the mode of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td>StatNormal(\text{mean, std_dev, statistic, value})</td>
<td>Calculates the statistic for the specified normal distribution</td>
</tr>
<tr>
<td>StatPairCount(\text{Data1,Data2})</td>
<td>Counts the number of pairs of cells for which each of the cells in the pair is numeric</td>
</tr>
<tr>
<td>StatPairMean(\text{Data1,Data2})</td>
<td>Calculates the mean of the differences between pairs of cells</td>
</tr>
<tr>
<td>StatPairMedian(\text{Data1,Data2})</td>
<td>Calculates the median of the differences between pairs of cells</td>
</tr>
<tr>
<td>StatPairStdDev(\text{Data1,Data2})</td>
<td>Calculates the sample standard deviation of the differences between pairs of cells</td>
</tr>
<tr>
<td>StatPercentile(\text{data,p, discreteFlag})</td>
<td>Calculates the p-th percentile of a data set</td>
</tr>
<tr>
<td>StatProduct(\text{Data1,Data2,...DataN})</td>
<td>Calculates the product of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td>Function</td>
<td>Returns</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>StatQuartile</strong> <em>(data,q,discrete flag)</em></td>
<td>Calculates the specified quartile of a set of values</td>
</tr>
<tr>
<td><strong>StatRand()</strong></td>
<td>Returns a random number between 0 and 1</td>
</tr>
<tr>
<td><strong>StatRange</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the range (maximum minus minimum) of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatRunsTest</strong>(data,cutoff)</td>
<td>Calculates the runs statistic for the values in a set of values, usually a time series</td>
</tr>
<tr>
<td><strong>StatSkewness</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the sample skewness of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatSkewnessP</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the population skewness of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatStandardize</strong>(x,mean,std_dev)</td>
<td>Calculates a standardized value from a distribution with the specified mean and standard deviation</td>
</tr>
<tr>
<td><strong>StatStdDev</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the sample standard deviation of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatStdDevP</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the population standard deviation of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatSum</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the sum of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatSumDevSq</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the sum of the squares of the deviations from the mean of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatSumSq</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the sum of the squares of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatStudent</strong>(deg_freedom,statistic,value)</td>
<td>Calculates the statistic for the specified Student's t-distribution</td>
</tr>
<tr>
<td><strong>StatVariance</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the sample variance of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
<tr>
<td><strong>StatVarianceP</strong>(Data1,Data2,...DataN)</td>
<td>Calculates the population variance of its arguments, which can be numbers, arrays, or ranges</td>
</tr>
</tbody>
</table>
Detailed Function Descriptions

Statistics functions are listed here with their required arguments.

### StatAutocorrelation

<table>
<thead>
<tr>
<th>Description</th>
<th>StatAutocorrelation(data, numLags) calculates the autocorrelation for a set of values, usually a time series, for a specified number of lags. data is the array or range of data values. numLags is the number of lags to use.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatAutocorrelation(C1:C100,1) returns the autocorrelation for the data in the range C1:C100 with 1 lag.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>numLags must greater than or equal to 1.</td>
</tr>
</tbody>
</table>

### StatAveDev

<table>
<thead>
<tr>
<th>Description</th>
<th>StatAveDev(Data1, Data2,...DataN) calculates the average absolute deviation of Data1, Data2,...DataN from their mean.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatAveDev(1,2,5) calculates the average absolute deviation of the 1, 2, and 5 from their mean.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>Data1, Data2,...DataN arguments can be numbers, arrays, or ranges.</td>
</tr>
</tbody>
</table>

### StatBinomial

<table>
<thead>
<tr>
<th>Description</th>
<th>StatBinomial(N,P,statistic,value) calculates the statistic for the binomial distribution with the specified values of N and P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatBinomial(2,.3,&quot;x to p&quot;,1) calculates the p-value for an x-value of 1 in the binomial distribution with N=2 and P=.3.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>N is the number of trials and must be a positive integer. P is the probability of “success” and must be from 0 to 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 nonnegative (when an x value is entered) or from 0 to 1 (when a p value is entered).</td>
</tr>
</tbody>
</table>
### StatCategoryIndices

**Description**  
*StatCategoryIndices*(range, category_name) returns the indices for the cells containing the specified `category_name` in the Excel *range*. This is an array function and the returned indices are index numbers (between 1 and # of cells in *range*). That is, they are the positions of the cells containing `category_name` within the *range*.

**Examples**  
*StatCategoryIndices* (C1:C100,"Male") returns the indices (between 1 and 100) of the cells containing the text *Male*.

**Guidelines**  
*range* is an Excel range.  
*category_name* is a string, value, or cell reference specifying the category to find.

### StatCategoryNames

**Description**  
*StatCategoryNames*(range) returns the names of the categories in the specified Excel *range*. This is an array function. 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 an Excel range.

### StatCategoryOccurrenceCount

**Description**  
*StatCategoryOccurrenceCount*(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 the text *Male*.

**Guidelines**  
*range* is an Excel range.  
*category_name* is a string, value, or cell reference specifying the category to find.
**StatChiSq**

<table>
<thead>
<tr>
<th>Description</th>
<th>StatChiDist(deg_freedom,statistic,value) calculates the statistic for the one-tailed chi-squared distribution using the specified degrees of freedom deg_freedom.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatChiDist(2,&quot;x to p&quot;,5) calculates the one-tailed chi-squared distribution at the value 5 with 2 degrees of freedom</td>
</tr>
<tr>
<td>Guidelines</td>
<td>the number of degrees of freedom deg_freedom must be an integer from 1 to 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 nonnegative (when an x value is entered) or from 0 to 1 (when a p value is entered)</td>
</tr>
</tbody>
</table>

**StatCorrelationCoeff**

<table>
<thead>
<tr>
<th>Description</th>
<th>StatCorrelationCoeff(data1,data2,flag) calculates the correlation coefficient between two paired sets of values, data1 and data2. Linear (Pearson) or rank-order (Spearman) coefficient can be calculated depending on flag value.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatCorrelationCoeff(A1:A100,B1:B100) calculates the correlation coefficient between the paired sets of values in the ranges A1:A100 and B1:B100.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>data1 and data2 must have the same number of elements. If flag is omitted or equal zero, then the linear (Pearson) correlation coefficient is calculated; if flag is not equal zero, then the rank-order (Spearman) correlation coefficient is calculated.</td>
</tr>
</tbody>
</table>

**StatCount**

<table>
<thead>
<tr>
<th>Description</th>
<th>StatCount(data1,data2,...dataN) calculates the number of numeric values in data1, data2, through dataN, which can be numbers, arrays, or ranges.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatCount(A1:A100,B1:B100) calculates the number of numeric values in the ranges A1:A100 and B1:B100.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>data1,data2,...dataN are 1 to 30 arguments, which can be numbers, arrays, or ranges.</td>
</tr>
</tbody>
</table>
### StatCountCategories

**Description**

StatCountCategories(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 an Excel range

### StatCountCellsByType

**Description**

StatCountCellByType(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 values in the range A1:A100.

**Guidelines**

range is an Excel range.

type is one of: 1=numeric, 2=non-empty, 3=non-empty, non-numeric; 4=empty. StatTools considers a cell containing only spaces to be an empty cell.

### StatCountRange

**Description**

StatCountRange(range,minValue,maxValue,includeMin,includeMax) calculates the number of values in the range 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 range A1:A100 that are between 1 and 10, inclusive.

**Guidelines**

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.

### StatCovariance

**Description**

StatCovariance(data1,data2) calculates the sample covariance between two sets of paired values, data1 and data2.

**Examples**

StatCovariance(A1:A100,B1:B100) calculates the sample covariance between the paired values in the ranges A1:A100 and B1:B100.

**Guidelines**

data1 and data2 must have the same number of values.
<table>
<thead>
<tr>
<th>StatCovarianceP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Examples</strong></td>
</tr>
<tr>
<td><strong>Guidelines</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>StatDestack</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Examples</strong></td>
</tr>
<tr>
<td><strong>Guidelines</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>StatDurbinWatson</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Examples</strong></td>
</tr>
<tr>
<td><strong>Guidelines</strong></td>
</tr>
</tbody>
</table>
### StatF

**Description**

StatF\((\text{deg\_freedom}_1,\text{deg\_freedom}_2,\text{statistic},\text{value})\) calculates the statistic for the F distribution using the numerator degrees of freedom \(\text{deg\_freedom}_1\) and the denominator degrees of freedom \(\text{deg\_freedom}_2\).

**Examples**

StatF\((1,1,"x \to p",1.5)\) calculates the F Distribution p value for an x value of 1.5 with numerator degrees of freedom 1 and denominator degrees of freedom 1.

**Guidelines**

\(\text{deg\_freedom}_1\) and \(\text{deg\_freedom}_2\) must be positive integers.

\(\text{statistic}\) 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.

\(\text{value}\) must be nonnegative.

### StatGetCellValues

**Description**

StatGetCellValues\((\text{range},\text{typeOfCell})\) returns 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**

\(\text{typeOfCell}\) is 0 for all, 1 for numeric, 2 for non-empty, 3 for non-empty, non-numeric, and 4 for empty cells.

### StatKurtosis

**Description**

StatKurtosis\((\text{Data}_1,\text{Data}_2,...\text{Data}_N)\) calculates the sample kurtosis of the data specified in \(\text{Data}_1,\text{Data}_2,...\text{Data}_N\). (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 in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**

\(\text{Data}_1,\text{Data}_2,...\text{Data}_N\) can be numbers, arrays, or ranges.
### StatKurtosisP

**Description**  
StatKurtosisP is a function that 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 in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**  
Data1, Data2, ..., DataN can be numbers, arrays, or ranges.

### StatLN

**Description**  
StatLN is a function that calculates the natural logarithm.

**Examples**  
StatLN(4.5) calculates the natural logarithm of 4.5.

**Guidelines**  
x must be positive.

### StatMax

**Description**  
StatMax is a function that 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 in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**  
Data1, Data2, ..., DataN can be numbers, arrays, or ranges.

### StatMean

**Description**  
StatMean is a function that calculates the mean of the data specified in Data1, Data2, ..., DataN.

**Examples**  
StatMean(A1:A100, {1;2;3;2.4}) calculates the mean value of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**  
Data1, Data2, ..., DataN can be numbers, arrays, or ranges.

### StatMeanAbs

**Description**  
StatMeanAbs is a function that 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 in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**  
Data1, Data2, ..., DataN can be numbers, arrays, or ranges.
### StatMedian

**Description**
StatMedian\((data,calcFlag)\) calculates the median of the values located in \(data\). It can 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 in the range A1:A100.

**Guidelines**
- \(calcFlag\) 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)

### StatMin

**Description**
StatMin\((Data1,Data2,...DataN)\) 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 in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**
\(Data1,Data2,...DataN\) can be numbers, arrays, or ranges.

### StatMode

**Description**
StatMode\((Data,Is\_discrete)\) calculates the mode of a data set.

**Examples**
StatMode\((A1:A100,FALSE)\) calculates the mode of the data in the range A1:A100.

**Guidelines**
- \(Is\_discrete\) is an optional argument and specifies whether the data are to be treated as discrete (TRUE), or continuous (FALSE). If missing, it is automatically determined from the data.
### StatNormal

**Description**

StatNormal(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 standard deviation 1

**Guidelines**

- `std_dev` must be positive.
- `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 from 0 to 1 when a p value is entered.

### StatPairCount

**Description**

StatPairCount(Data1,Data2) counts the number of pairs of cells in Data1 and Data2. 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 ranges 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.

### StatPairMean

**Description**

StatPairMean(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 ranges 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.
**StatPairMedian**

<table>
<thead>
<tr>
<th>Description</th>
<th>StatPairMedian(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. The median can be calculated using any of five alternative methods, as optionally specified by method_Flag.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatPairMedian(A1:A100, B1:B100) calculates the median of the differences between pairs of numeric cells in the ranges A1:A100 and B1:B100.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>-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)</td>
</tr>
</tbody>
</table>

**StatPairStdDev**

<table>
<thead>
<tr>
<th>Description</th>
<th>StatPairStdDev(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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>StatPairStdDev(A1:A100, B1:B100) calculates the sample standard deviation of the differences between pairs of numeric cells in the ranges A1:A100 and B1:B100.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>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.</td>
</tr>
</tbody>
</table>
### StatPercentile

<table>
<thead>
<tr>
<th>Description</th>
<th>( \text{StatPercentile}(\text{data}, p, \text{method_Flag}) ) calculates the ( p )-th percentile of ( \text{data} ). Percentiles can be calculated using any of five alternative methods, as optionally specified by ( \text{method_Flag} ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>( \text{StatPercentile}(A1:A100, .15, 0) ) calculates the 15th percentile for the data in the range A1:A100. It uses the weighted average percentile calculation method.</td>
</tr>
</tbody>
</table>
| Guidelines | \( p \) must be from 0 to 1, inclusive.  
\( \text{method\_Flag} \) 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) |

### StatProduct

<table>
<thead>
<tr>
<th>Description</th>
<th>( \text{StatProduct}(\text{Data}_1, \text{Data}_2, \ldots, \text{Data}_N) ) calculates the product of the data specified in ( \text{Data}_1, \text{Data}_2, \ldots, \text{Data}_N ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>( \text{StatProduct}(A1:A100, {1;2;3;2.4}) ) calculates the product of all value in the in the range A1:A100 and the values 1, 2, 3, and 2.4.</td>
</tr>
<tr>
<td>Guidelines</td>
<td>( \text{Data}_1, \text{Data}_2, \ldots, \text{Data}_N ) can be numbers, arrays, or ranges.</td>
</tr>
</tbody>
</table>
### StatQuartile

**Description**

StatQuartile(data,q, method_Flag) calculates the specified quartile of data. Quartiles can be calculated using any of five alternative methods, as optionally specified by method_Flag.

**Examples**

StatQuartile (A1:A100,1) calculates the first quartile for the data in the range A1:A100.

**Guidelines**

- 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)

### StatRand

**Description**

StatRand() 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 from 0 to 1.

### StatRange

**Description**

StatRange(Data1,Data2,...DataN) calculates the range (maximum minus 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 in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**

Data1,Data2,...DataN can be numbers, arrays, or ranges.
### StatRunsTest

**Description**
StatRunsTest(data,cutoff) calculates the runs statistic for the values in data using cutoff value.

**Examples**
StatRunsTest(A1:A100,StatMean(A1:A100)) calculates the runs test statistic on the data in the range A1:A100 using the mean of the data as the cutoff value.

**Guidelines**
data must be an Excel range.

### StatSkewness

**Description**
StatSkewness(Data1,Data2,...DataN) calculates the sample skewness of the data specified in Data1,Data2,...DataN.

**Examples**
StatSkewness(A1:A100,{1;2;3;2.4}) calculates the sample skewness of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**
Data1,Data2,...DataN can be numbers, arrays, or ranges.

### StatSkewnessP

**Description**
StatSkewnessP(Data1,Data2,...DataN) calculates the population skewness of the data specified in Data1,Data2,...DataN.

**Examples**
StatSkewnessP(A1:A100,{1;2;3;2.4}) calculates the population skewness of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**
Data1,Data2,...DataN can be numbers, arrays, or ranges.

### StatStandardize

**Description**
StatStandardize(x,mean,std_dev) calculates a standardized value from a distribution with the specified mean and standard deviation std_dev. x is the value to be standardized.

**Examples**
StatStandardize(2,1,3) calculates a standardized value at the value 2 from a distribution with mean 1 and standard deviation 3.

**Guidelines**
std_dev must be positive.
<table>
<thead>
<tr>
<th><strong>StatStdDev</strong></th>
<th>Description</th>
<th><strong>StatStdDev</strong> <em>(Data1, Data2, ..., DataN)</em> calculates the sample standard deviation of the data specified in <em>Data1, Data2, ..., DataN</em>.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examples</strong></td>
<td><strong>StatStdDev</strong> <em>(A1:A100, {1; 2; 3; 2.4})</em> calculates the sample standard deviation of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.</td>
<td></td>
</tr>
<tr>
<td><strong>Guidelines</strong></td>
<td><em>Data1, Data2, ..., DataN</em> can be numbers, arrays, or ranges.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>StatStdDevP</strong></th>
<th>Description</th>
<th><strong>StatStdDevP</strong> <em>(Data1, Data2, ..., DataN)</em> calculates the population standard deviation of the data specified in <em>Data1, Data2, ..., DataN</em>.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examples</strong></td>
<td><strong>StatStdDevP</strong> <em>(A1:A100, {1; 2; 3; 2.4})</em> calculates the population standard deviation of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.</td>
<td></td>
</tr>
<tr>
<td><strong>Guidelines</strong></td>
<td><em>Data1, Data2, ..., DataN</em> can be numbers, arrays, or ranges.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>StatSum</strong></th>
<th>Description</th>
<th><strong>StatSum</strong> <em>(Data1, Data2, ..., DataN)</em> calculates the sum of the data specified in <em>Data1, Data2, ..., DataN</em>.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examples</strong></td>
<td><strong>StatSum</strong> <em>(A1:A100, {1; 2; 3; 2.4})</em> calculates the sum of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.</td>
<td></td>
</tr>
<tr>
<td><strong>Guidelines</strong></td>
<td><em>Data1, Data2, ..., DataN</em> can be numbers, arrays, or ranges.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>StatSumDevSq</strong></th>
<th>Description</th>
<th><strong>StatSumDevSq</strong> <em>(Data1, Data2, ..., DataN)</em> calculates the sum of the squares of deviations from the mean of its arguments, which can be numbers, arrays, or ranges. Any missing numbers cause a blank to be returned.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examples</strong></td>
<td><strong>StatSumDevSq</strong> <em>(A1:A100, {1; 2; 3; 2.4})</em> calculates the sum of the squares of deviations from the mean of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.</td>
<td></td>
</tr>
<tr>
<td><strong>Guidelines</strong></td>
<td><em>Data1, Data2, ..., DataN</em> can be numbers, arrays, or ranges.</td>
<td></td>
</tr>
</tbody>
</table>
### StatSumSq

**Description**
StatSumSq(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:A100,{1;2;3;2.4}) calculates the sum of the squares of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**
Data1,Data2,...DataN can be numbers, arrays, or ranges.

### StatStudent

**Description**
StatStudent(deg_freedom,statistic,value) calculates the statistic for the specified 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 2.

**Guidelines**
deg_freedom must be an integer from 1 to 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 must be nonnegative.

### StatVariance

**Description**
StatVariance(Data1,Data2,...DataN) calculates the sample variance of the data specified in Data1,Data2,...DataN.

**Examples**
StatVariance(A1:A100,{1;2;3;2.4}) calculates the sample variance of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**
Data1,Data2,...DataN can be numbers, arrays, or ranges.

### StatVarianceP

**Description**
StatVarianceP(Data1,Data2,...DataN) calculates the population variance of the data specified in Data1,Data2,...DataN.

**Examples**
StatVarianceP(A1:A100,{1;2;3;2.4}) calculates the sample variance of the data in the range A1:A100 and the values 1, 2, 3, and 2.4.

**Guidelines**
Data1,Data2,...DataN can be numbers, arrays, or ranges.
Palisade Corporation
  Contacting, 12
  DecisionTools Suite, 14
  Getting Help, 11

StatTools
  Cell Comments, 21
  Data Set, 17, 33
  Data Set Limits, 36
  Data Set Manager, 17, 33
  Data Viewer, 37
  Developer Kit (XDK), 16
  Example Spreadsheets, 16
  Guided Tour Videos, 16
  Icons, 23
  Installation, 14
  Instructional Materials, 16
  License Activation, 207
  License Manager, 15
  Missing Values, 19
  Quick Start Tutorials, 16
  Report Placement, 20
  Ribbon, 17
  Software Activation, 15
  Stacked Data, 19
  Student Version, 13
  System Requirements, 13
  Uninstalling, 14
  Unstacked Data, 19
  Variable, 33
  Variable Range Names, 20

StatTools Commands
  Application Settings, 200
  Autocorrelation, 29, 120
  Bar Chart, 73
  Box-Whisker Plot, 25, 71
  C Chart, 30, 160
  Chi-Square Independence Test, 27, 102
  Chi-Square Normality Test, 28, 110
  Clear Dialog Memory, 205
  Cluster Analysis, 190
  Combination, 54
  Confidence Interval - Mean/Std. Deviation, 26, 78
  Confidence Interval - Proportion, 26, 81
  Correlation and Covariance, 25, 62
  Data Set Manager, 33
  Data Viewer, 37
  Delete Data Sets, 205
  Difference, 32, 50
  Discriminant Analysis, 29, 145
  Forecasting, 29, 125
  Histogram, 25, 66
  Hypothesis Test - Mean/Std. Deviation, 85
  Hypothesis Test - Proportion, 88
  Interaction, 52
  Kruskal-Wallis Test, 180
  Lag, 32, 49
  Lilliefors Test, 28, 113
  Load DecisionTools Add-In, 205
  Logistic Regression, 142
  Mann-Whitney Test, 176
  One-Variable Summary, 25, 60
  One-Way ANOVA, 27, 95
  P Chart, 30, 157
  Pareto Chart, 30, 150
  Principal Components Analysis, 186
  Q-Q Normal Plot, 28, 115
  Random Sample, 32, 57
  Regression, 29, 132
  Runs Test for Randomness, 29, 123
  Sample Size Selection, 93
  Scatterplot, 25, 69
  Sign Test, 30, 169
<table>
<thead>
<tr>
<th>Term</th>
<th>Page 1</th>
<th>Page 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stack</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Time Series Graph</td>
<td>28, 118</td>
<td></td>
</tr>
<tr>
<td>Transform</td>
<td>32, 47</td>
<td></td>
</tr>
<tr>
<td>Two-Way ANOVA</td>
<td>27, 99</td>
<td></td>
</tr>
<tr>
<td>U Chart</td>
<td>30, 163</td>
<td></td>
</tr>
<tr>
<td>Unload StatTools Add-In</td>
<td>205</td>
<td></td>
</tr>
<tr>
<td>Unstack</td>
<td>31, 46</td>
<td></td>
</tr>
<tr>
<td>Wilcoxon Signed-Rank Test</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>X/R Charts</td>
<td>30, 153</td>
<td></td>
</tr>
<tr>
<td>StatTools Functions</td>
<td>213</td>
<td></td>
</tr>
<tr>
<td>StatTools Menus</td>
<td></td>
<td>213</td>
</tr>
<tr>
<td>Data Utilities</td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>Help</td>
<td>207</td>
<td></td>
</tr>
<tr>
<td>Multivariate Analysis</td>
<td>185</td>
<td></td>
</tr>
<tr>
<td>Nonparametric Test</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>Normality Tests</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>Quality Control</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>Regression and Classification</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>Statistical Inference</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>Summary Graphs</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Summary Statistics</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Time Series and Forecasting</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>Utilities</td>
<td>199</td>
<td></td>
</tr>
</tbody>
</table>