



ADDITIONS AND ENHANCEMENTS



This document describes new features added to Version 18 of Statgraphics Centurion. It includes 30 new procedures and significant enhancements to many existing procedures. For detailed information about each feature, consult the PDF references indicated (accessible under *Help – Procedure Documentation* on the main menu.)

Table of Contents

Data Management and User Interface.....	4
Big Data	4
Create Data and Code Columns	9
Data Types	10
Graphics Profile Designer.....	11
Operators	12
R - Installation and Configuration	13
Repeat Analysis For	14
Replace Censored Values	15
System Preferences	16
Changes to Existing Analyses.....	19
Bivariate Density Statlet	19
Box-and-Whisker Plots	20
Capability Analysis for Variable Data	21

Design of Experiments Wizard: Definitive Screening Designs.....	22
Forecasting and Automatic Forecasting.....	23
General Linear Models	24
Monte Carlo Simulation – ARIMA Model Simulation	25
Monte Carlo Simulation – General Simulation Models	26
Multivariate Capability Analysis	27
Multiple Sample Comparison	28
Oneway ANOVA.....	29
Piechart/Donut Chart	30
Simple Regression, Polynomial Regression, Box-Cox Transformations, and Calibration Models.....	31
Subset Analysis	32
Surface and Contour Plots	33
Tabulation	34
X-bar and R Charts	35
New Statistical Analyses.....	36
Attribute Capability Analysis Statlet	36
Capability Control Chart Design Statlet	37
Capability Control Charts for Attributes.....	38
Capability Control Charts for Variables	39
Classification and Regression Trees	40
Demographic Map Visualizer (Locations).....	41
Diamond Plot	42
Distribution Fitting (Arbitrarily Censored Data)	43
Equivalence and Noninferiority Tests (Comparing Mean to Target)	44
Equivalence and Noninferiority Tests (Comparing Paired Samples)	45
Equivalence and Noninferiority Tests (Comparing Two Means).....	46
Equivalence and Noninferiority Tests (2x2 Crossover Study)	47
Heat Map.....	48
Likert Plot	49
Multidimensional Scaling	50
Multiple Diamond Plot	51
Multiple Violin Plot Statlet.....	52
Multivariate Normal Random Numbers	53

Multivariate Normality Test	54
Multivariate Tolerance Limits	55
Orthogonal Regression.....	56
Population Pyramid Statlet	57
Sunflower Plot	58
Text Mining	59
Time Series Baseline Plot	60
Tornado and Butterfly Plots.....	61
Trivariate Density Statlet	62
Violin Plot Statlet.....	63
Wind Rose Statlet	64
X-13ARIMA-SEATS Seasonal Adjustment.....	65

Data Management and User Interface

Big Data

To handle big data, a special file type called a *Statgraphics Big Data file* has been developed. These files have the extension *.sgb* rather than *.sgd*. They differ in 2 important ways from standard Statgraphics data files:

1. They store numeric data in binary format rather than as text. This avoids the step of converting each data value to a number when it is read into the program.
2. Data is stored column-by-column rather than row-by-row. This dramatically reduces execution time when individual columns are read into memory.

Using SGB files, Statgraphics Centurion is capable of analyzing data sets consisting of many millions of records and a large number of columns.

Creating a Big Data SGB File

In order to analyze big data, the data must be placed in an SGB file. This may be done in either of 2 ways:

1. If the data is not too big, it may be possible to read it into the Statgraphics DataBook using the usual methods. One may then select *File – Save As – Save Data File As* from the main menu and select *SG Centurion Big Data File (.sgb)*.
2. Convert a text file directly to a Statgraphics SGB file by selecting *File – Big Data - Create Big Data SGB File* from the main menu.

To convert the text to an SGB file, information about each field must be entered:

Create Big Data SGB File

File preview

```

Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,Actua
2008,1,3,4,2003,1955,2211,2225,WN,335,N712SW,128,150,116,-14,8,IAD,TPA,810,4,8,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,754,735,1002,1000,WN,3231,N772SW,128,145,113,2,19,IAD,TPA,810,5,10,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,628,620,804,750,WN,448,N428WN,96,90,76,14,8,IND,BWI,515,3,17,0,,0,NA,NA,NA,NA,NA
2008,1,3,4,926,930,1054,1100,WN,1746,N612SW,88,90,78,-6,-4,IND,BWI,515,3,7,0,,0,NA,NA,NA,NA,NA
  
```

File type

☐ Fixed width

☐ Tab delimited

☒ Other Delimiter:

File header

☐ None

☒ Field name

☐ Field description

Global missing value indicators

☒ NA ☐ . (period) ☐ 999 ☐ em-dash

☐ N/A ☐ - (hyphen) ☐ -999 ☐ all negatives

☐ NAN ☐ (null) ☐ --- ☐ other:

Number of fields:

Maximum rows:

Extraction rows:

	Name	Type	Width	M.V.	Description
1	Year	Integer	4	NA	1987-2008
2	Month	Integer	2	NA	Month (1-12)
3	DayofMonth	Integer	2	NA	Day (1-31)
4	DayOfWeek	Integer	1	NA	1 (Monday) - 7 (Sunday)
5	DepTime	Integer	4	NA	actual departure time (local, hhmm)
6	CRSDepTime	Integer	4	NA	scheduled departure time (local, hhmm)
7	ArrTime	Integer	4	NA	actual arrival time (local, hhmm)
8	CRSArrTime	Integer	4	NA	scheduled arrival time (local, hhmm)
9	UniqueCarrier	Character	2	NA	unique carrier code
10	FlightNum	Integer	4	NA	flight number
11	TailNum	Character	6	NA	plane tail number
12	ActualElapsedTime	Integer	4	NA	in minutes
13	CRSElapsedTime	Integer	4	NA	in minutes
14	AirTime	Integer	4	NA	in minutes
15	ArrDelay	Integer	4	NA	arrival delay, in minutes
16	DepDelay	Integer	4	NA	departure delay, in minutes

OK Cancel Help

Field information

Back Next

The field information may be extracted from the text file or copies from a datasheet.

Modifying SGB Files

Statgraphics SGB files cannot be modified using the data editor. However, changes can be made to the existing data by selecting *File – Big Data - Modify Big Data SGB File*.

Modify Big Data SGB File

	Name	Type	Width	Description	
1	Year	Integer	4	1987-2008	Recode
2	Month	Integer	2	Month (1-12)	Recode
3	DayofMonth	Integer	2	Day (1-31)	Recode
4	DayOfWeek	Integer	1	1 (Monday) - 7 (Sunday)	Recode
5	DepTime	Integer	4	actual departure time (local, hhmm)	Recode
6	CRSDepTime	Integer	4	scheduled departure time (local, hhmm)	Recode
7	ArrTime	Integer	4	actual arrival time (local, hhmm)	Recode
8	CRSArrTime	Integer	4	scheduled arrival time (local, hhmm)	Recode
9	UniqueCarrier	Character	2	unique carrier code	Recode
10	FlightNum	Integer	4	flight number	Recode
11	TailNum	Character	6	plane tail number	Recode
12	ActualElapsedTime	Integer	4	in minutes	Recode
13	CRSElapsedTime	Integer	4	in minutes	Recode
14	AirTime	Integer	4	in minutes	Recode
15	ArrDelay	Integer	4	arrival delay, in minutes	Recode
16	DepDelay	Integer	4	departure delay, in minutes	Recode

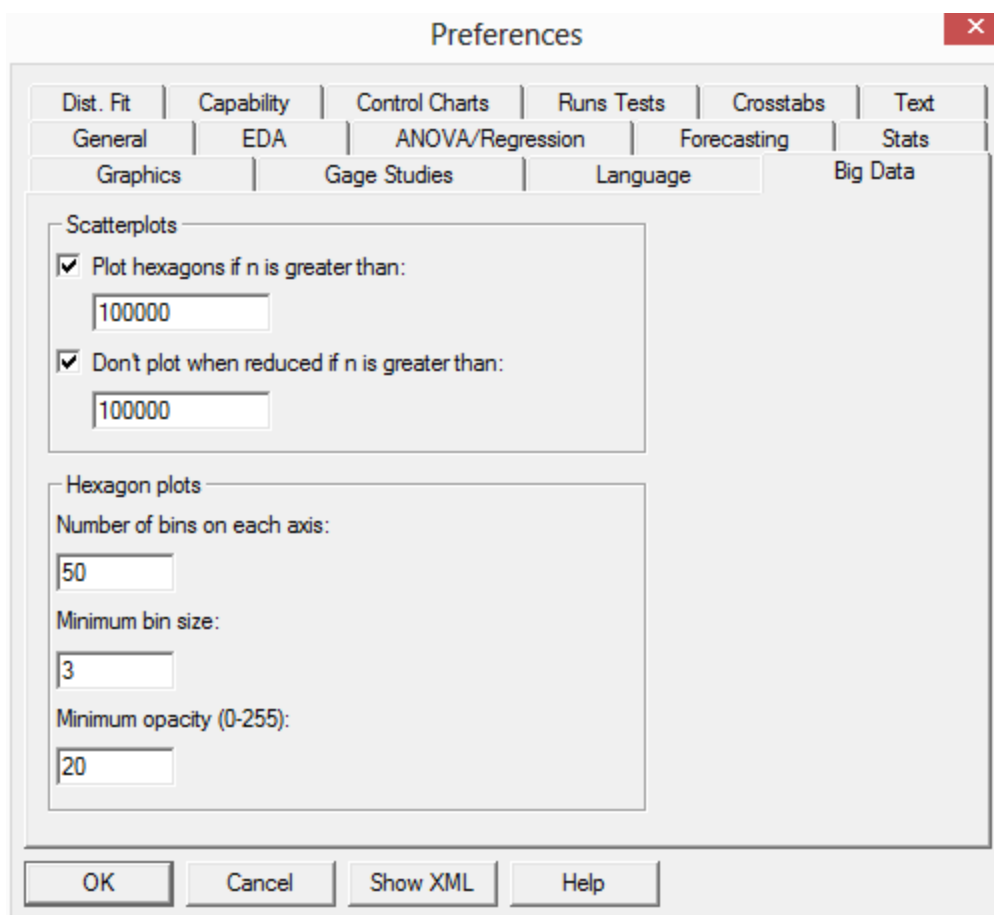
Apply Cancel Help Back Next

Combining Big Data Files

2 big data SGB files with identical variable structure may be combined into a single file by selecting *File – Big Data – Combine Big Data SGB Files* from the main Statgraphics Centurion menu.

Calculations and Graphs

There is a new page on the *Preferences* dialog box that controls several issues pertinent the analysis of Big Data:

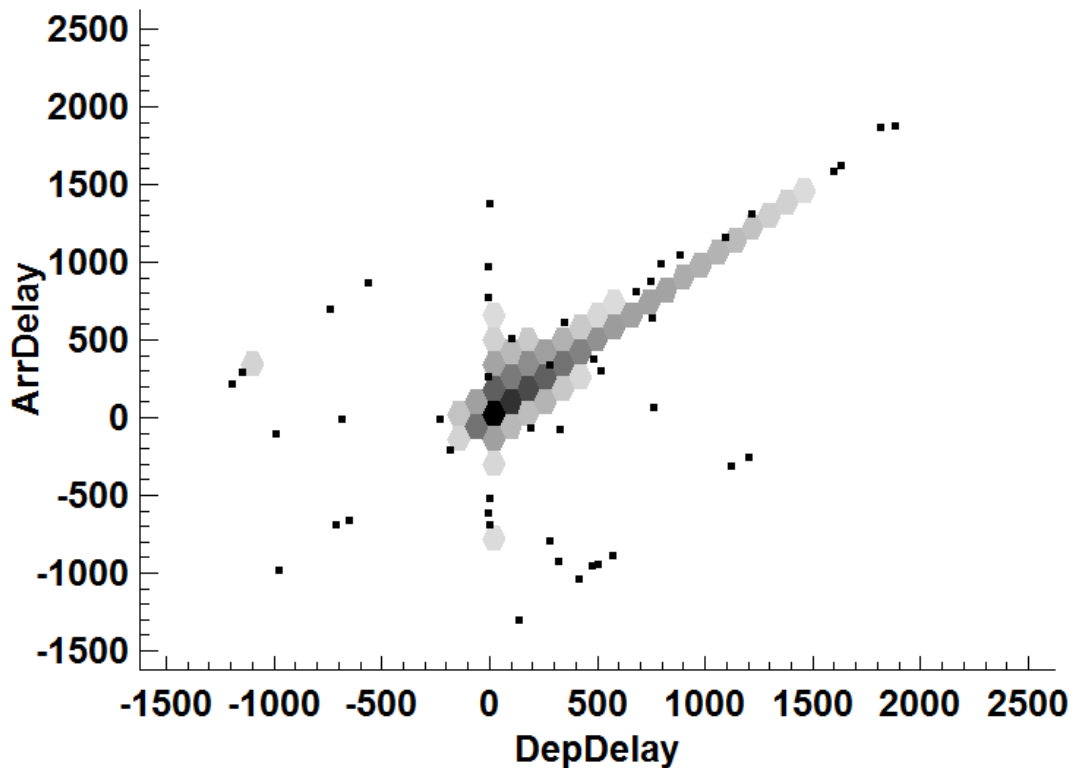


The main issue is that plots that consist of many point symbols take a long time to produce (as much as 7 seconds per million points). Consequently, the system is set up by default to same time in 2 ways:

1. By not plotting points symbols when the panes containing the graphs are not maximized when the number of points is very large.
2. By displaying *Hexagon Plots* instead of standard scatterplots for very large data sets.

A typical hexagon plot is shown below, which shows a plot calculated from nearly 7 million points:

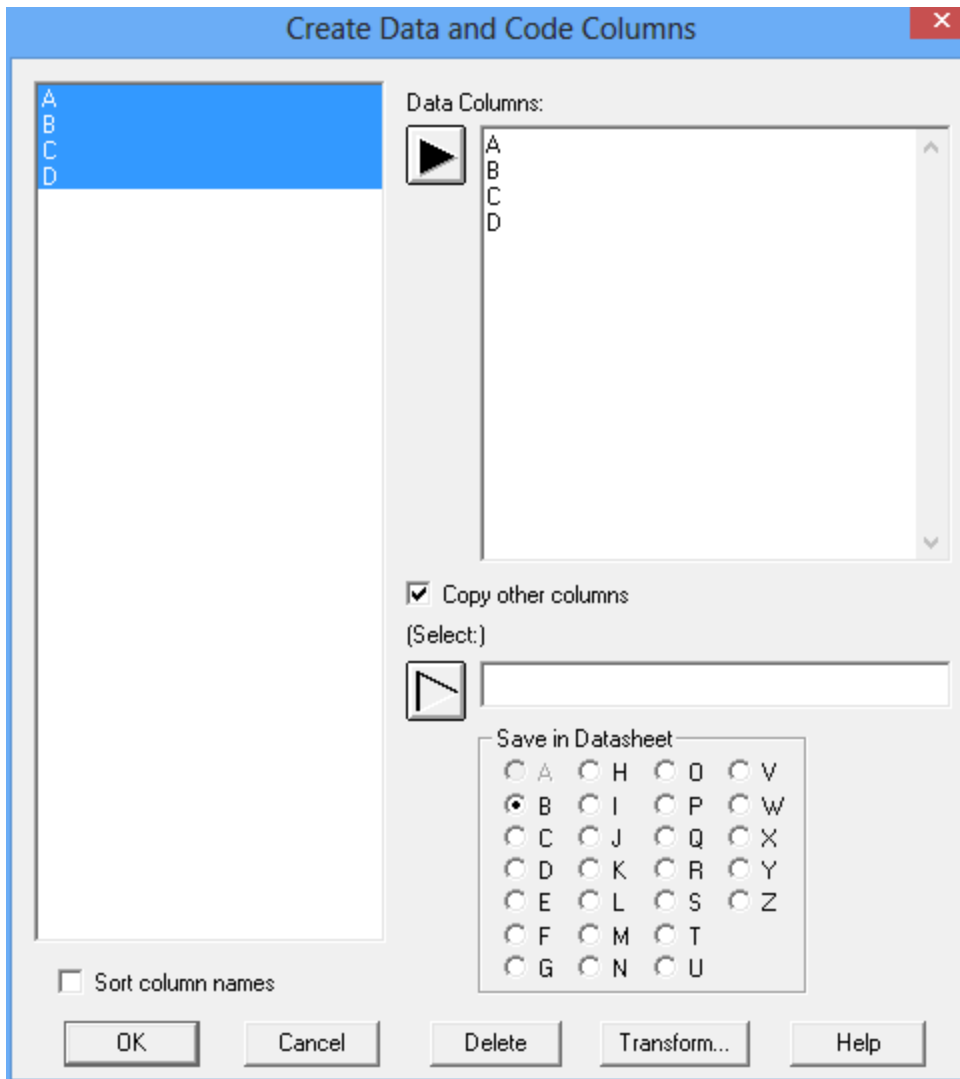
Plot of ArrDelay vs DepDelay



In this display, the X-Y region is divided into a grid of hexagonal bins. The number of observations in each bin is then calculated. If a bin contains more than a specified number of points, a shaded hexagon is plotted instead of the point symbols. The opacity of the hexagon is related to the relative number of points in the associated hexagonal region.

Create Data and Code Columns

This new selection under *Edit* on the main menu converts multiple data columns into a pair of data and code columns for use in procedures such as *Multifactor ANOVA*. For example, suppose you had 4 columns named A, B, C and D with 10 rows in each. This selection will combine the data into a single column with 40 rows and create a corresponding code column with either “A”, “B”, “C”, or “D”. The main dialog box for the procedure is shown below.

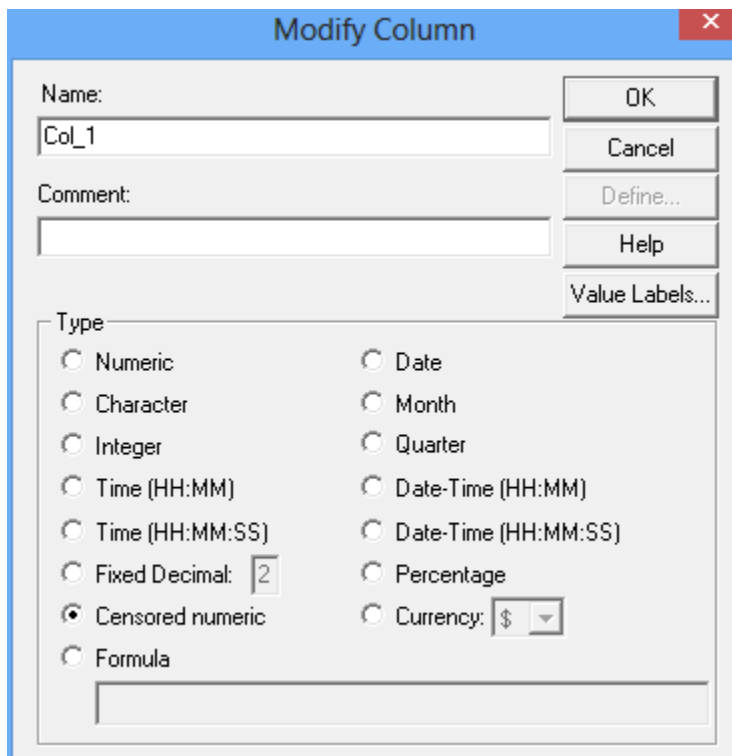


Reference: *Edit Menu*

Data Types

Two new data types have been added to the program:

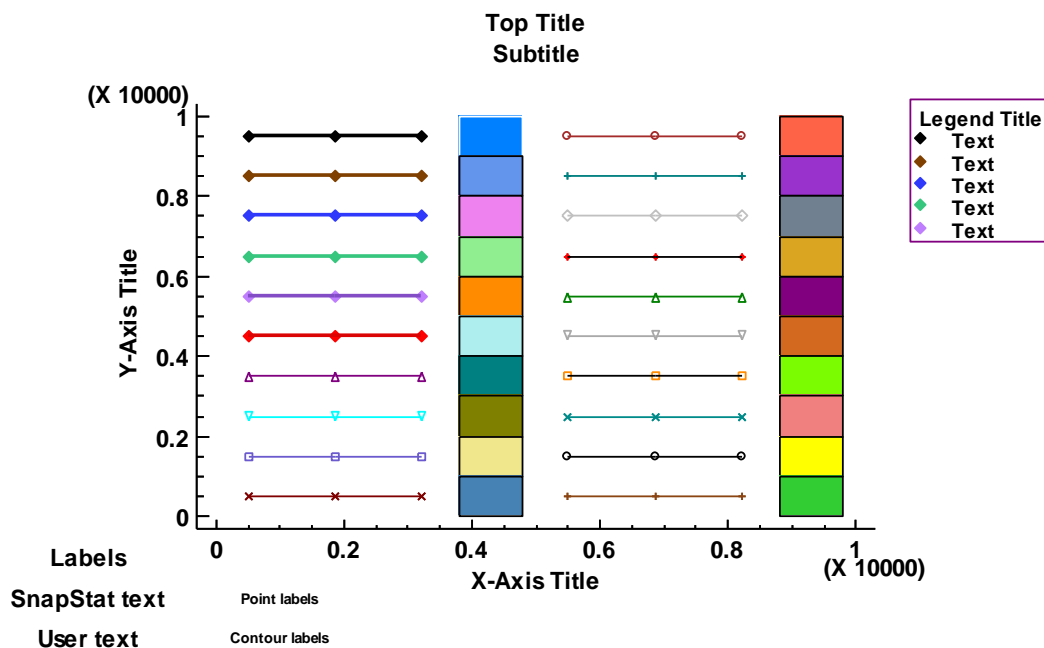
1. **Censored numeric** – allows for the entry of censored numeric data. Valid entries are strings such as “<0.1” for left-censoring, “>10” for right-censoring, and “[0.1,10]” for interval censoring.
2. **Currency** – this is a special data type such similar to percentage. Data may be entered as monetary amounts such as “\$9.99” or as a number such as “9.99” without the currency symbol. If added without the symbol, it will be added automatically. When defining the column, the symbol may be selected from a choice of a preceding “\$”, a preceding “€”, a following “€”, a preceding “£”, or a preceding “¥”.



Reference: *Edit Menu*

Graphics Profile Designer

Added 2 new text size for identifying points on a graph and for labeling contour levels.



Reference: *Graphics Options*

Operators

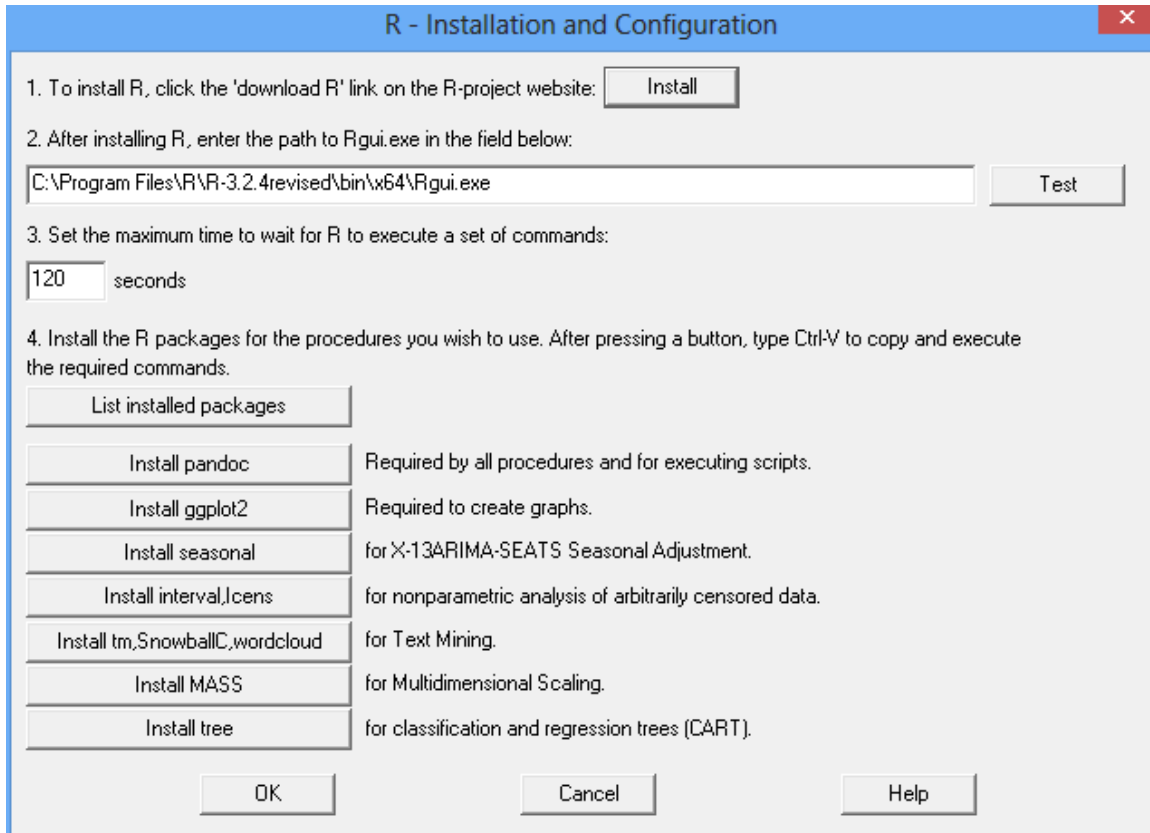
Several new operators have been added to the program for use in Statgraphics expressions:

1. DAY – extracts the day of the month from a date variable.
2. DAYOFYEAR– calculates number of days in year up to date specified.
3. DAYOFMONTH – calculates number of days in month up to date specified.
4. DAYOFWEEK– calculates number of days in week up to date specified.
5. GAMMA – returns the value of the gamma function.
6. HOUR – extracts hour from a date-time variable.
7. ISMISSING – returns 1 for each missing value and 0 for each non-missing value.
8. ISNOTMISSING– returns 0 for each missing value and 1 for each non-missing value.
9. LNGAMMA– returns the value of the loggamma function.
10. MINUTE– extracts hour from a date-time variable.
11. MOD – finds remainder after dividing one number by another.
12. MONTH – extracts month from a date variable.
13. POWER – raises numeric values to a power.
14. QUARTER – extracts quarter from a date variable.
15. SECOND– extracts second from a date-time variable.
16. SIGN – returns -1, 0 or 1 depending on sign of numeric value.
17. YEAR – extracts year from a date variable.

Reference: *STATGRAPHICS Operators*.

R - Installation and Configuration

This new menu item assists users in setting up R and the libraries required to use the Statgraphics/R interface. It also sets the maximum number of seconds that Statgraphics will wait for R to respond to a request.



R - Installation and Configuration

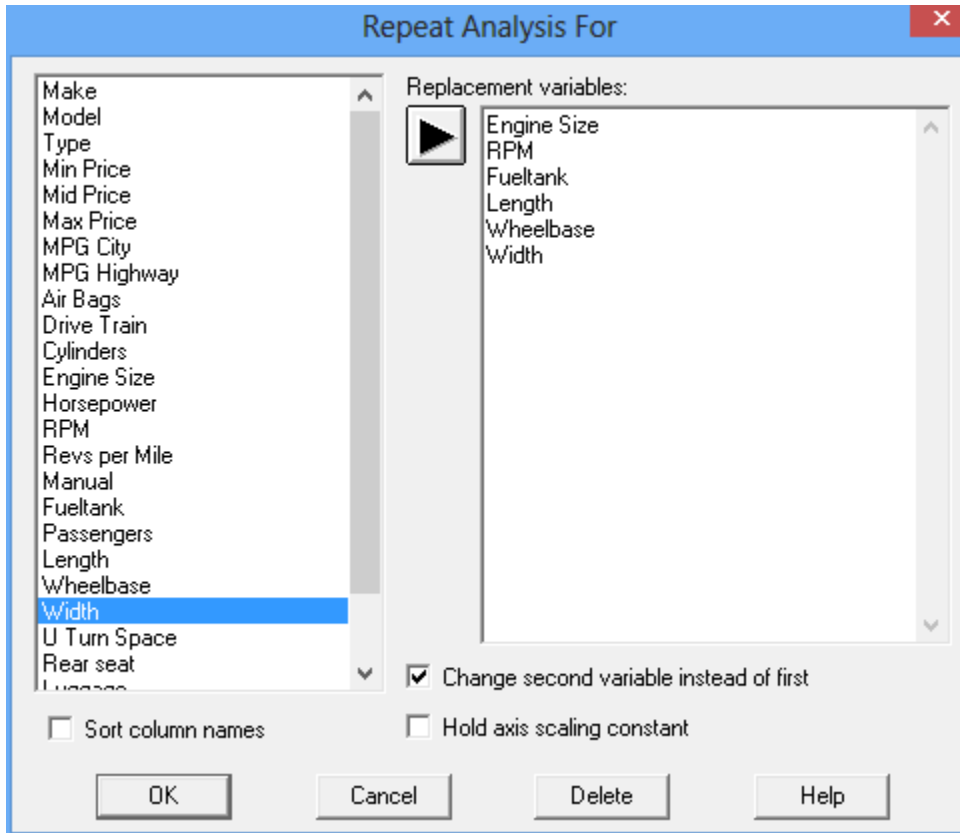
1. To install R, click the 'download R' link on the R-project website:
2. After installing R, enter the path to Rgui.exe in the field below:
3. Set the maximum time to wait for R to execute a set of commands:
 seconds
4. Install the R packages for the procedures you wish to use. After pressing a button, type Ctrl-V to copy and execute the required commands.

<input type="button" value="List installed packages"/>	
<input type="button" value="Install pandoc"/>	Required by all procedures and for executing scripts.
<input type="button" value="Install ggplot2"/>	Required to create graphs.
<input type="button" value="Install seasonal"/>	for X-13ARIMA-SEATS Seasonal Adjustment.
<input type="button" value="Install interval,lcens"/>	for nonparametric analysis of arbitrarily censored data.
<input type="button" value="Install tm,SnowballC,wordcloud"/>	for Text Mining.
<input type="button" value="Install MASS"/>	for Multidimensional Scaling.
<input type="button" value="Install tree"/>	for classification and regression trees (CART).

Reference: *R – Installation and Configuration*

Repeat Analysis For

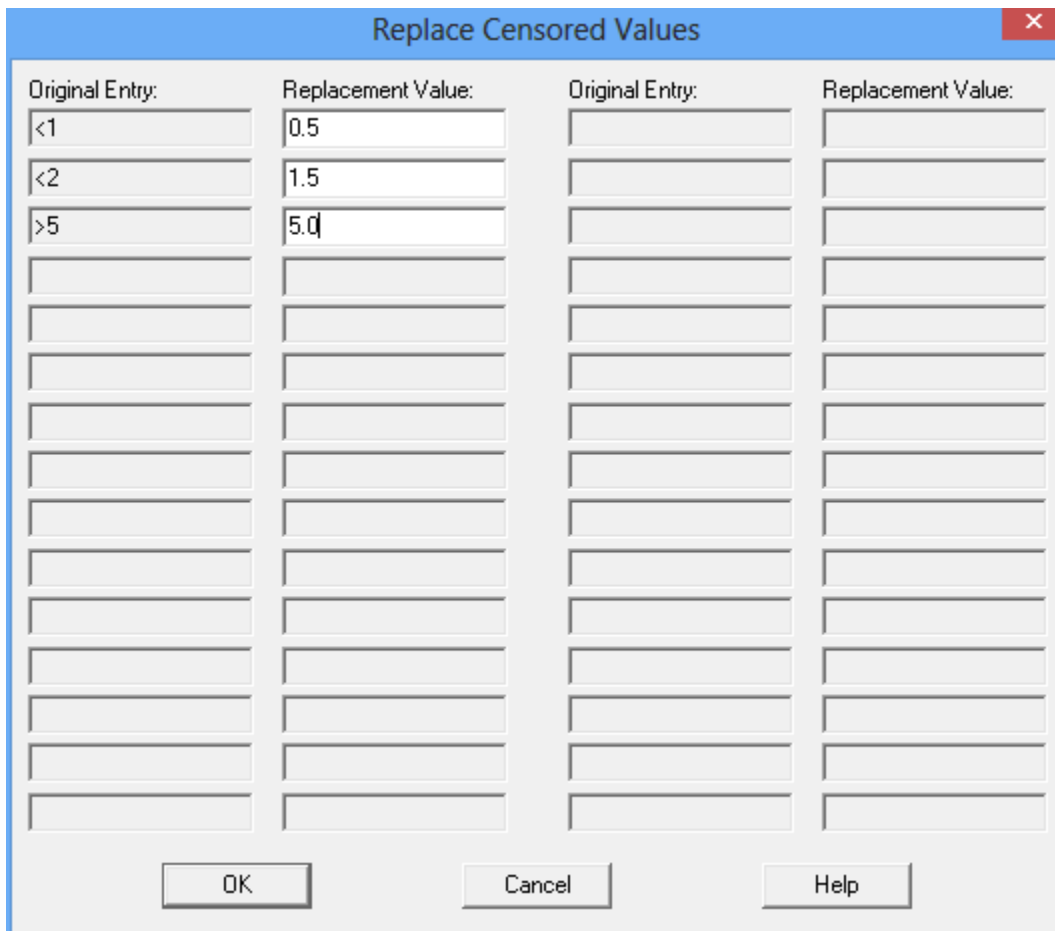
This new selection under *Edit* on the main menu lets you run an existing analysis on a selected set of other variables. A new window is generated for each variable. The main dialog box for the procedure is shown below.



Reference: *Edit Menu*

Replace Censored Values

This new selection under *Edit* on the main menu lets you replace any censored values in a column defined as type *Censored numeric* with specific data values. For example, the dialog box below specifies replacement values for 3 entries: “<1”, “<2”, and “>5”. Replacement of censored values is necessary in order to use analyses that are not designed to handle censored data.



The dialog box titled "Replace Censored Values" contains two columns of input fields. The first column has three rows with "Original Entry" values "<1", "<2", and ">5", and corresponding "Replacement Value" values "0.5", "1.5", and "5.0". The second column is empty. At the bottom are "OK", "Cancel", and "Help" buttons.

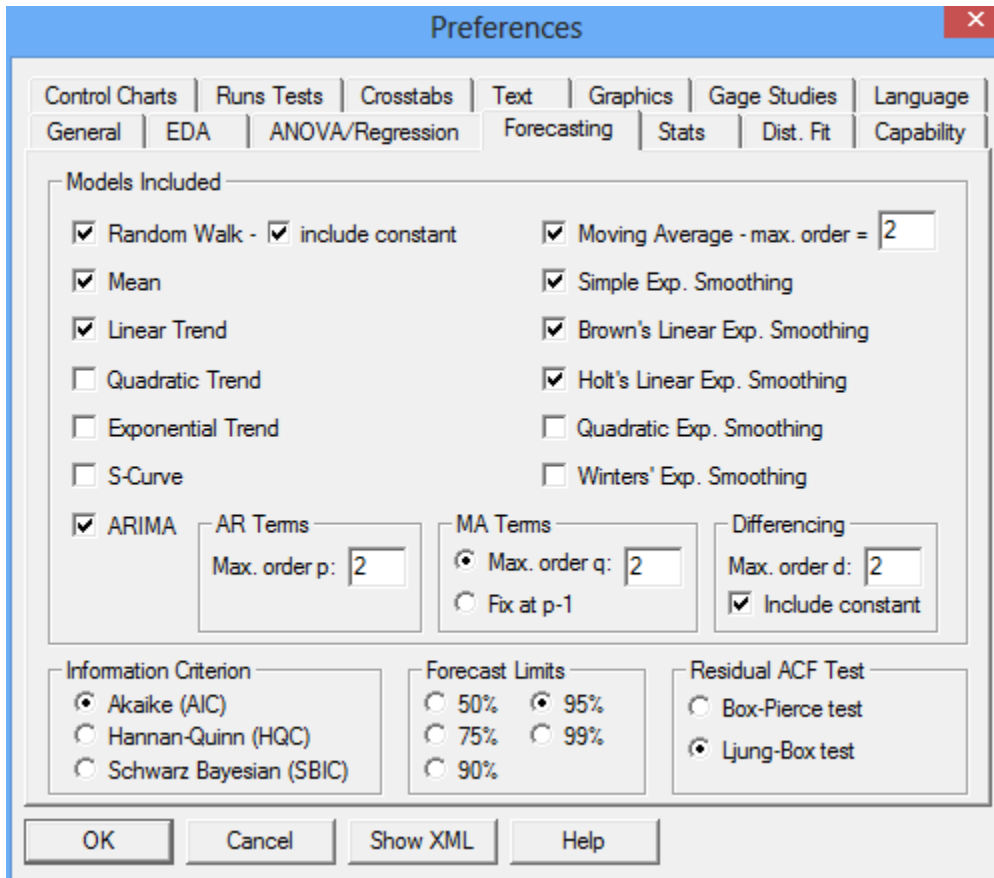
Original Entry:	Replacement Value:	Original Entry:	Replacement Value:
<1	0.5		
<2	1.5		
>5	5.0		

OK Cancel Help

Reference: *Edit Menu*

System Preferences

A change has been made to the *Forecasting* tab to allow a choice between the “Box-Pierce” test and the “Ljung-Box” test when testing for autocorrelations in the residuals:



The screenshot shows the 'Preferences' dialog box with the 'Forecasting' tab selected. The 'Residual ACF Test' section at the bottom right now includes the 'Ljung-Box test' option, which is selected (indicated by a radio button). The 'Box-Pierce test' is also present but unselected.

Control Charts	Runs Tests	Crosstabs	Text	Graphics	Gage Studies	Language
General	EDA	ANOVA/Regression	Forecasting	Stats	Dist. Fit	Capability

Models Included

- ☒ Random Walk - ☒ include constant
- ☒ Mean
- ☒ Linear Trend
- ☐ Quadratic Trend
- ☐ Exponential Trend
- ☐ S-Curve
- ☒ ARIMA
 - AR Terms: Max. order p:
 - MA Terms:
 - ☒ Max. order q:
 - ☐ Fix at p-1
 - Differencing:
 - Max. order d:
 - ☒ Include constant
- ☒ Moving Average - max. order =
- ☒ Simple Exp. Smoothing
- ☒ Brown's Linear Exp. Smoothing
- ☒ Holt's Linear Exp. Smoothing
- ☐ Quadratic Exp. Smoothing
- ☐ Winters' Exp. Smoothing

Information Criterion

- ☒ Akaike (AIC)
- ☐ Hannan-Quinn (HQC)
- ☐ Schwarz Bayesian (SBIC)

Forecast Limits

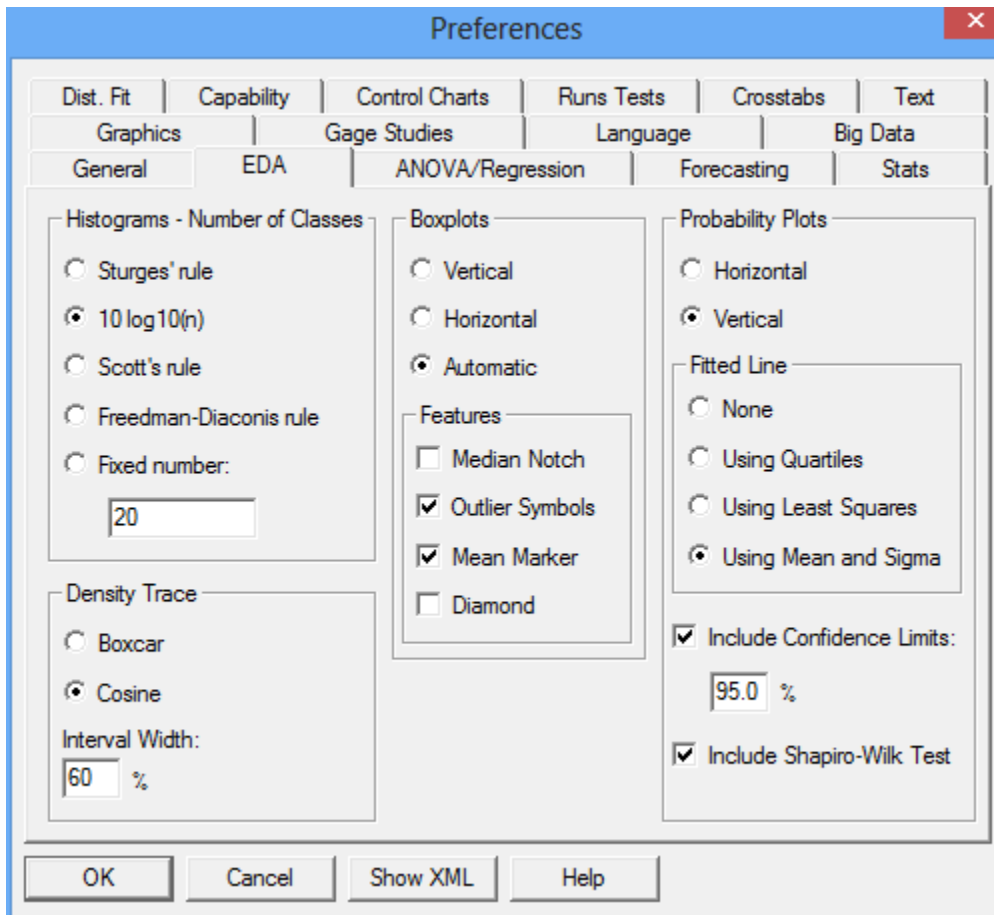
- ☐ 50%
- ☐ 75%
- ☐ 90%
- ☒ 95%
- ☐ 99%

Residual ACF Test

- ☐ Box-Pierce test
- ☒ Ljung-Box test

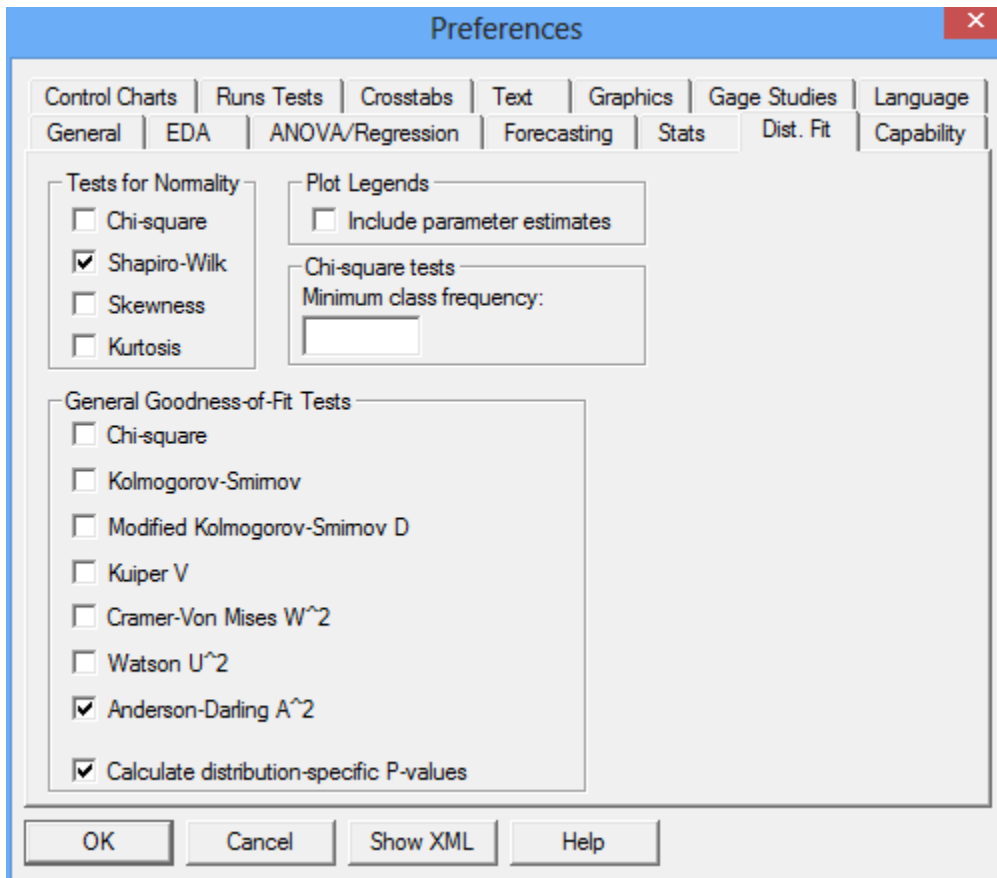
OK Cancel Show XML Help

A change has also been made to the EDA tab to allow for inclusion of a “diamond” on *Boxplots* to show confidence intervals for the mean:



In addition, a new “Automatic” preference for *Boxplots* allow the program to pick what it considers to be the best orientation on a case-by-case basis.

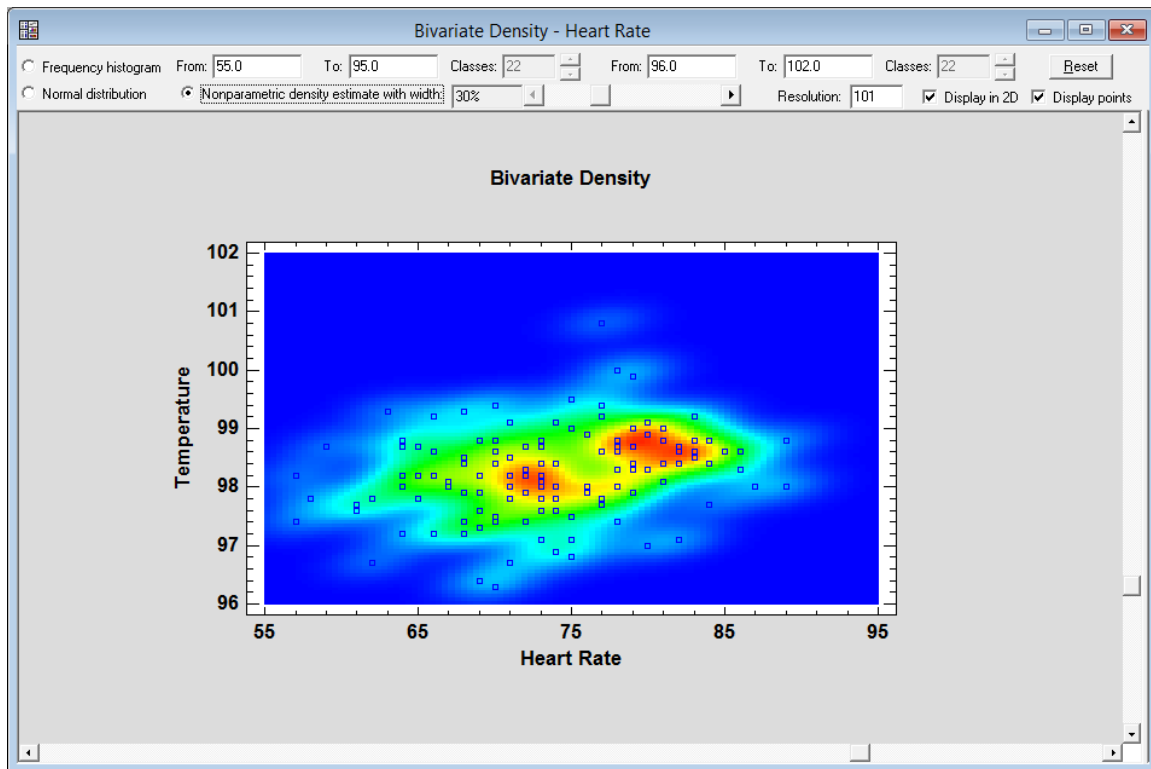
Finally, a change has been made to the *Dist. Fit* tab to specify the minimum frequency for a class when performing a chi-square test below which it will be combined with adjacent classes. Valid entries are integers between 1 and 10.



Reference: *Preferences*

Changes to Existing Analyses

Bivariate Density Statlet

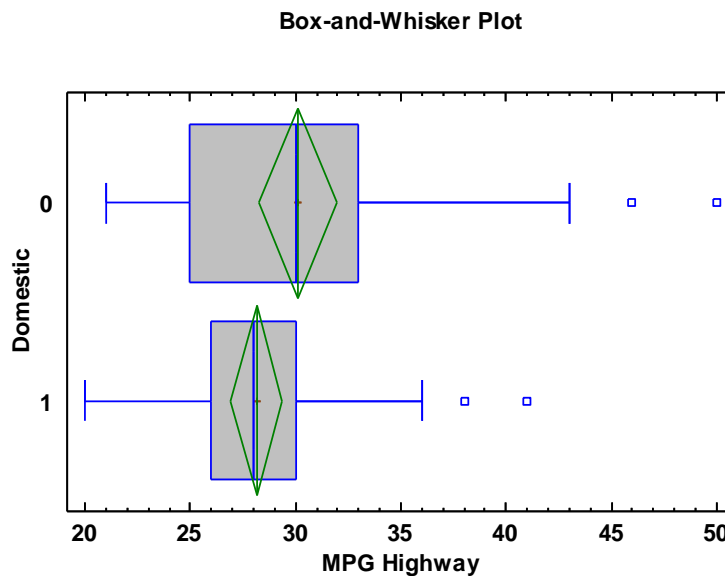


Reference: *Bivariate Density Statlet*

Box-and-Whisker Plots

A new feature has been added to all box-and-whisker plots. It adds a diamond to the plot covering the range of a $100(1-\alpha)\%$ confidence interval for the mean. The change has been made in the following procedures:

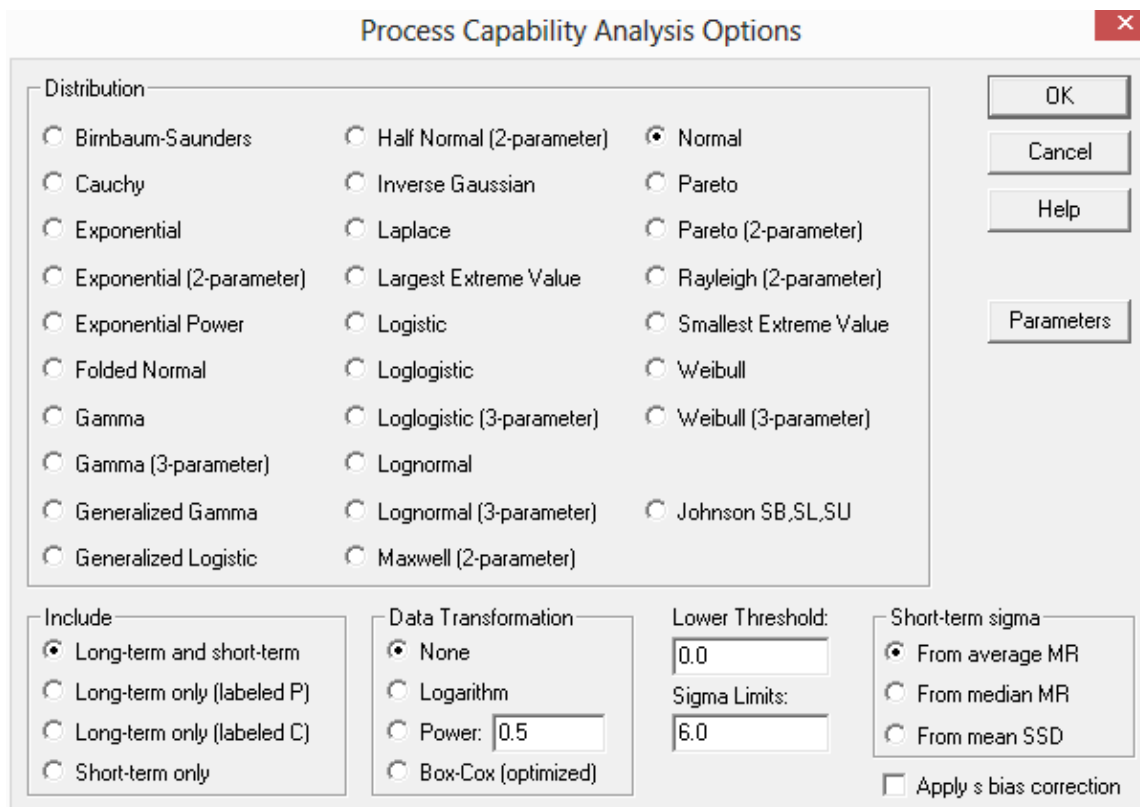
- a. Box-and-Whisker Plots (One Sample)
- b. Box-and-Whisker Plots (Multiple Samples)
- c. Multiple Sample Comparison
- d. Multiple Variable Analysis
- e. One Variable Analysis
- f. Oneway ANOVA
- g. Two Sample Comparison
- h. Paired Sample Comparison
- i. Subset Analysis
- j. Rowwise Statistics
- k. Matrix Plot
- l. Outlier Identification
- m. Monte Carlo Simulation (General Simulation Models)
- n. Monte Carlo Simulation (Random Number Generation)
- o. Item Reliability Analysis
- p. Gage Studies (Average and Range Method)
- q. Gage Studies (ANOVA Method)
- r. Gage Studies (Range Method)



Reference: *Box-and-Whisker Plot*

Capability Analysis for Variable Data

The Analysis Options dialog box has been changed to allow the selection of the method for analyzing the short-term sigma:



The dialog box is titled "Process Capability Analysis Options" and features a close button (X) in the top right corner. It is divided into several sections:

- Distribution:** A grid of radio buttons for selecting a distribution. The "Normal" distribution is selected. Other options include Birnbaum-Saunders, Cauchy, Exponential, Exponential (2-parameter), Exponential Power, Folded Normal, Gamma, Gamma (3-parameter), Generalized Gamma, Generalized Logistic, Half Normal (2-parameter), Inverse Gaussian, Laplace, Largest Extreme Value, Logistic, Loglogistic, Loglogistic (3-parameter), Lognormal, Lognormal (3-parameter), Maxwell (2-parameter), Pareto, Pareto (2-parameter), Rayleigh (2-parameter), Smallest Extreme Value, Weibull, and Weibull (3-parameter). Johnson SB,SL,SU is also listed.
- Include:** Radio buttons for "Long-term and short-term" (selected), "Long-term only (labeled P)", "Long-term only (labeled C)", and "Short-term only".
- Data Transformation:** Radio buttons for "None" (selected), "Logarithm", "Power: 0.5", and "Box-Cox (optimized)".
- Lower Threshold:** A text box containing "0.0".
- Sigma Limits:** A text box containing "6.0".
- Short-term sigma:** Radio buttons for "From average MR" (selected), "From median MR", and "From mean SSD".
- Apply s bias correction:** An unchecked checkbox.

On the right side of the dialog, there are buttons for "OK", "Cancel", "Help", and "Parameters".

The default value is determined from the system preferences.

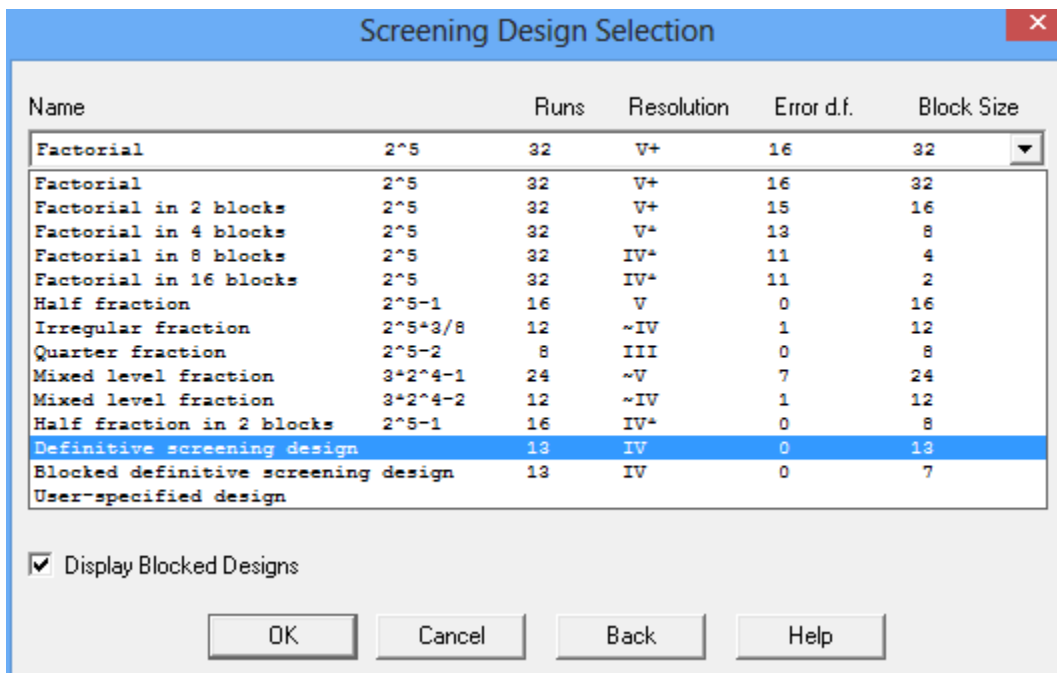
An option has also been added to find the best-fitting Johnson curve. Johnson curves are available with all possible combinations of skewness and kurtosis.

Reference: *Capability Analysis (Variable Data)*

Design of Experiments Wizard: Definitive Screening Designs

A new type of experimental design has been added to the Design of Experiments Wizard. Called *Definitive Screening Designs*, these designs are small designs capable of estimating models involving both linear and quadratic effects, although second-order interactions are partially confounded with themselves and with quadratic effects. In addition, designs for 6 or more factors collapse into designs which can estimate the full second-order model (including interactions) for any 3 factors.

DSDs may be constructed for any combination of continuous and 2-level categorical factors where the total number of factors is between 4 and 16. Both blocked and unblocked designs are available. They appear on the list of screening designs during the *Select Design* step.



Reference: *DOE Wizard – Definitive Screening Designs*

Forecasting and Automatic Forecasting

The Ljung-Box test replaces the Box-Pierce test when testing for autocorrelation remaining in the residuals. The Ljung-Box test has been shown to have superior performance.

Model Comparison

Key:

RMSE = Root Mean Squared Error

RUNS = Test for excessive runs up and down

RUNM = Test for excessive runs above and below median

AUTO = Ljung-Box test for excessive autocorrelation

MEAN = Test for difference in mean 1st half to 2nd half

VAR = Test for difference in variance 1st half to 2nd half

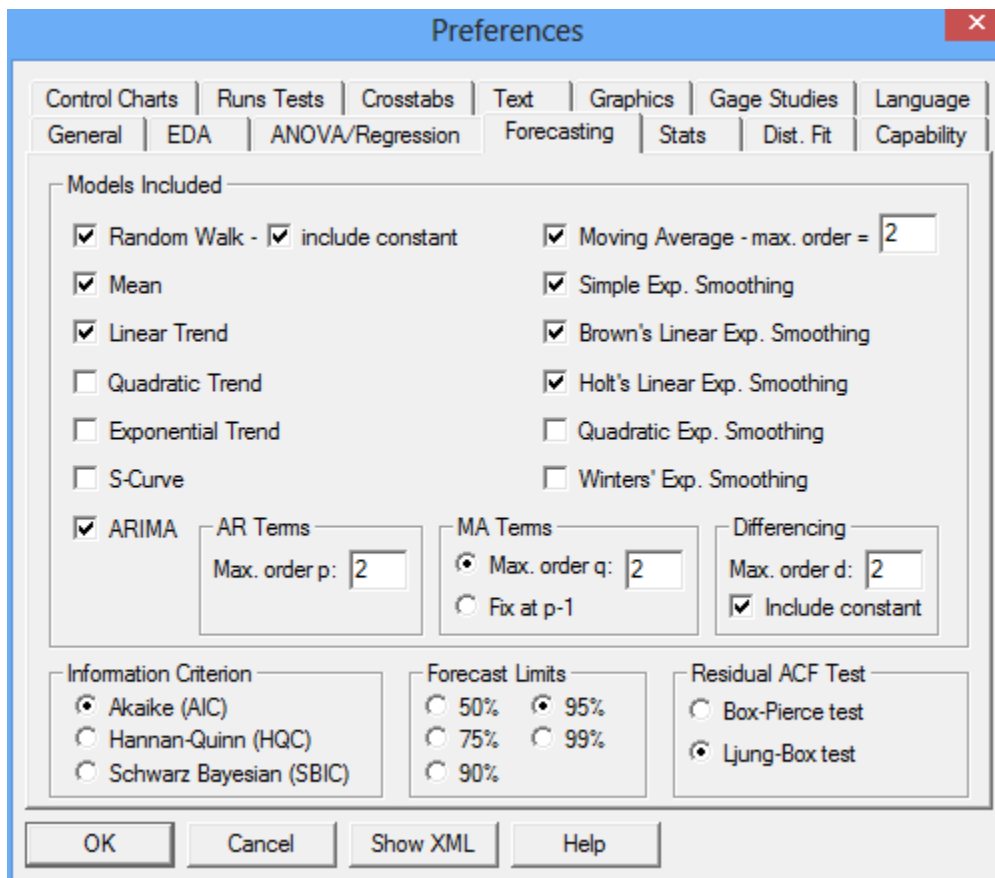
OK = not significant ($p \geq 0.05$)

* = marginally significant ($0.01 < p \leq 0.05$)

** = significant ($0.001 < p \leq 0.01$)

*** = highly significant ($p \leq 0.001$)

If desired, users can switch back to the Box-Pierce test using the *Forecasting* tab on the *Preferences* dialog box.



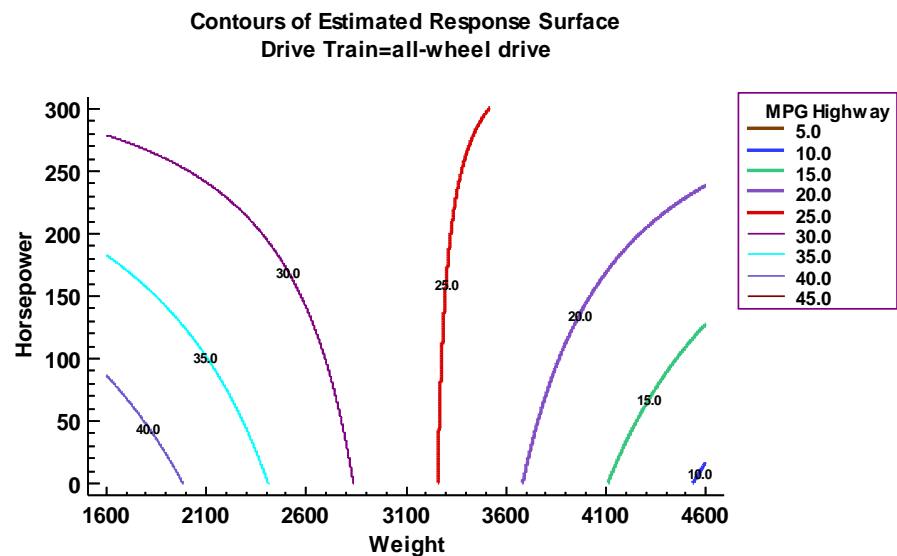
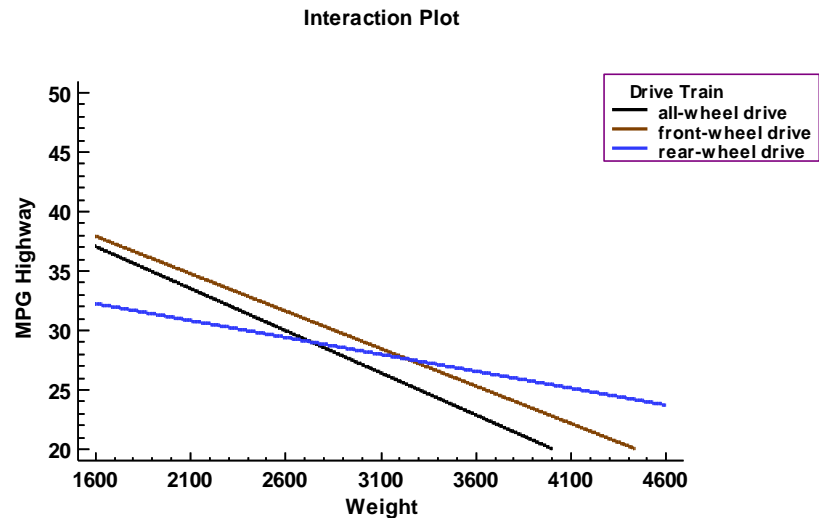
The screenshot shows the 'Preferences' dialog box with the 'Forecasting' tab selected. The 'Models Included' section contains several checked options: Random Walk (with 'include constant' checked), Mean, Linear Trend, Moving Average (with 'max. order' set to 2), Simple Exp. Smoothing, Brown's Linear Exp. Smoothing, Holt's Linear Exp. Smoothing, and ARIMA (with 'AR Terms' set to 'Max. order p: 2', 'MA Terms' set to 'Max. order q: 2' and 'Fix at p-1' unselected, and 'Differencing' set to 'Max. order d: 2' and 'Include constant' checked). Under 'Information Criterion', 'Akaike (AIC)' is selected. Under 'Forecast Limits', '95%' is selected. Under 'Residual ACF Test', 'Ljung-Box test' is selected. The 'OK', 'Cancel', 'Show XML', and 'Help' buttons are at the bottom.

Reference: *Forecasting*

General Linear Models

The following changes have been made:

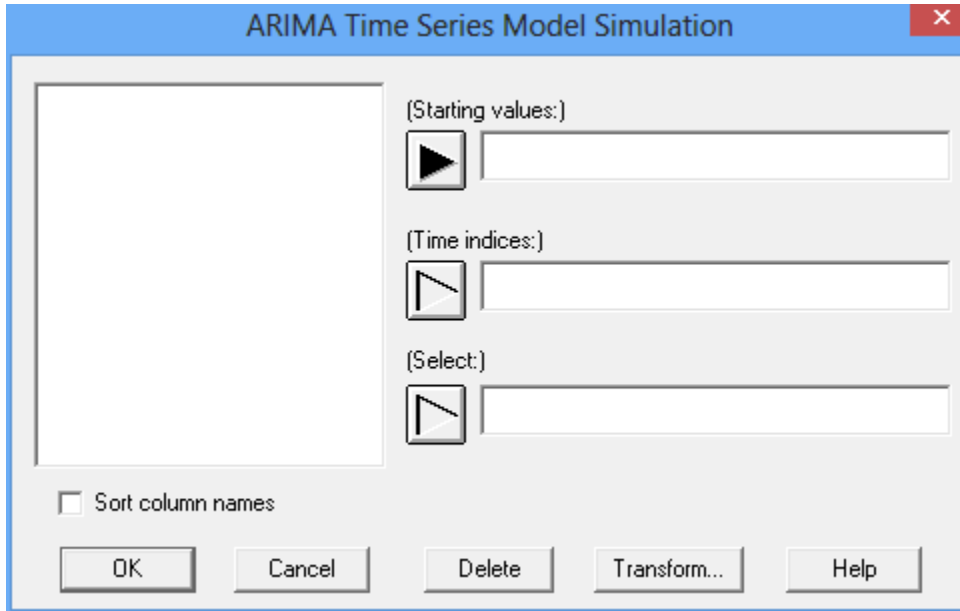
1. The interaction plot now displays interaction between one categorical factor and one quantitative factor.
2. Contour plots label each contour.



Reference: *General Linear Models*

Monte Carlo Simulation – ARIMA Model Simulation

A new data input dialog box has been added that may be used to set the starting values for the time series:



Starting values: optional data to be used to set the starting values. The simulated data is assumed to begin immediately after the end of this data.

Time indices: optional values indicating the time at which each of the starting values was recorded. If supplied, these values will be used to scale the plot of the simulated data.

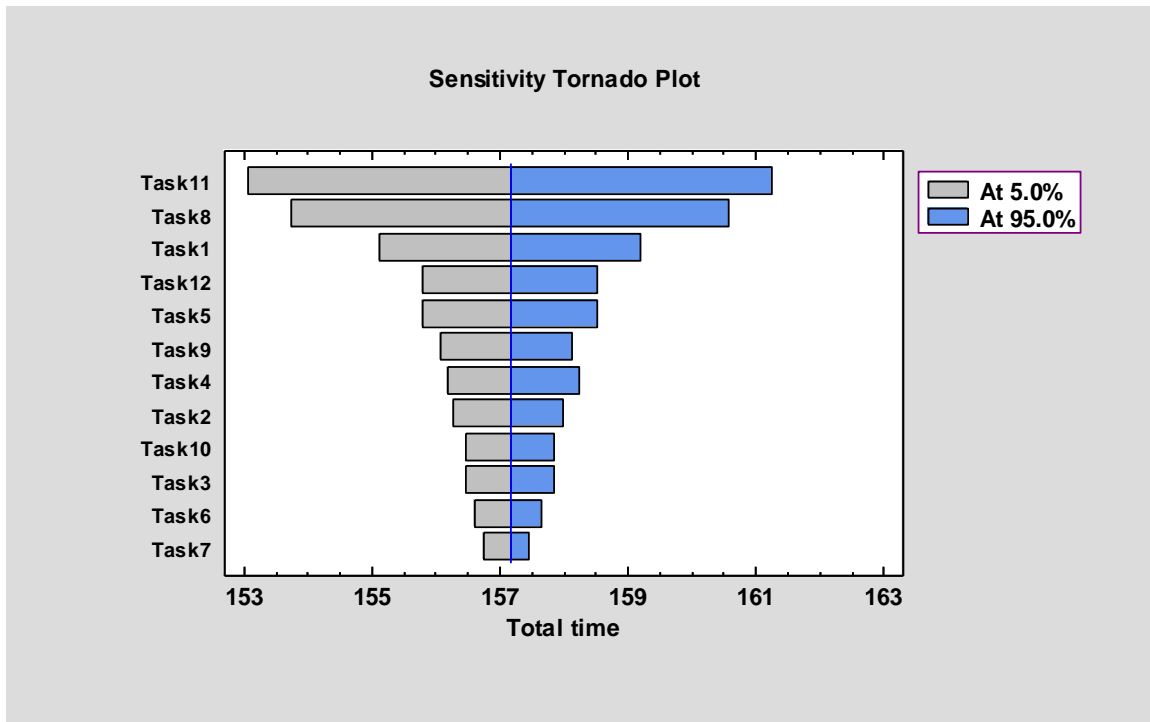
Select: optional subset selection.

If no starting values are supplied, the simulation will generate random starting values by simulating twice as much data as requested and discarding the first half.

The previous data input dialog box has been moved to Analysis Options.

Monte Carlo Simulation – General Simulation Models

A new graph has been added, called a “Sensitivity Tornado Plot”:

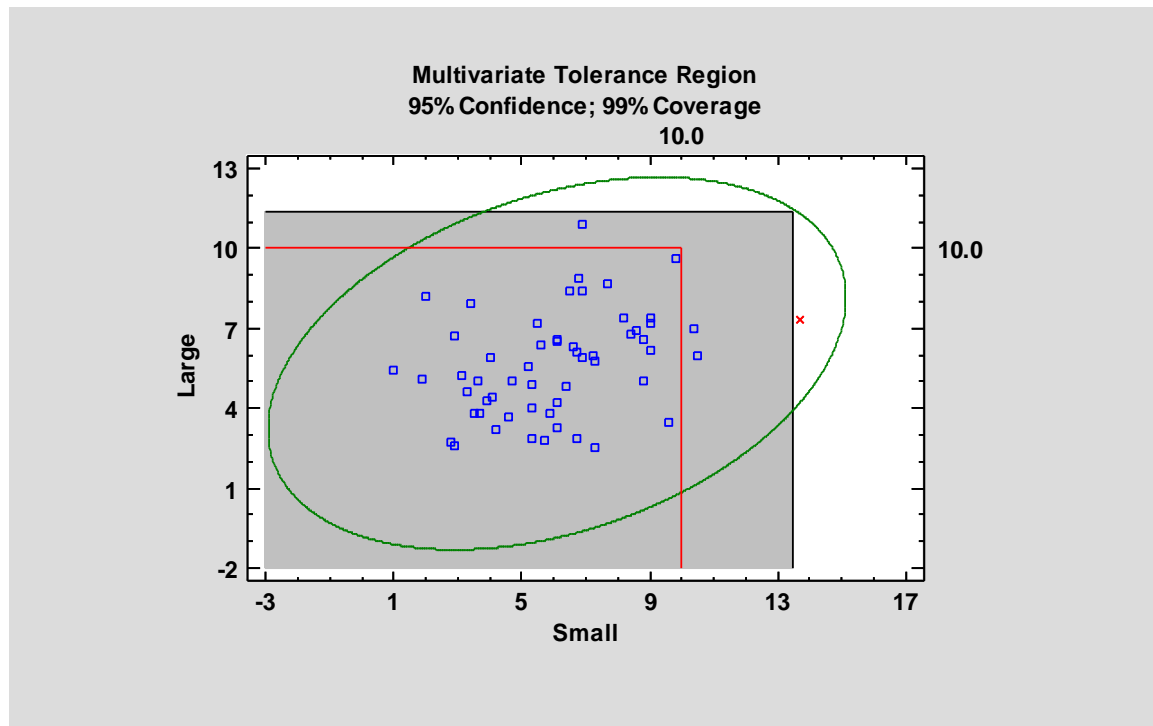


It shows the effect of each input variable on a response when it is changed over a specified percentage of its probability distribution, with all other variables held at their median values. The variables are sorted from top to bottom in order of their overall effect.

Multivariate Capability Analysis

Several new features were added to this procedure:

1. Confidence intervals may now be calculated for multivariate capability indices using bootstrapping.
2. Multivariate tolerance limits may be calculated using either of 2 approaches:
 - a. Simultaneous limits for each variable using a Bonferroni approach.
 - b. Exact elliptical tolerance regions based on Monte Carlo methods.
3. A table lists the normalized squared distance of each multivariate observation from the centroid of the variables.

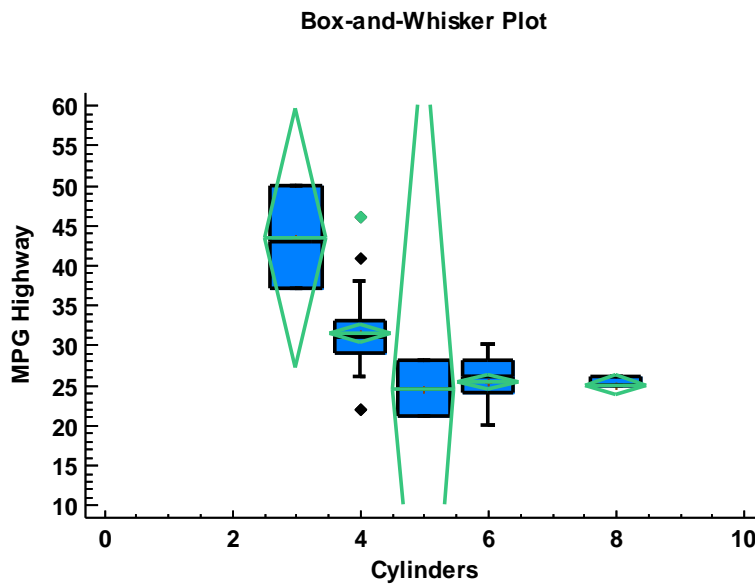


Reference: *Multivariate Capability Analysis*

Multiple Sample Comparison

The following changes have been made:

1. The *Analysis Options* dialog box now gives the option to treat factor levels as numeric. This improves the display of numeric factors in tables and graphs.
2. A diamond may be added to the multiple box-and-whisker plot showing confidence intervals for each level mean.

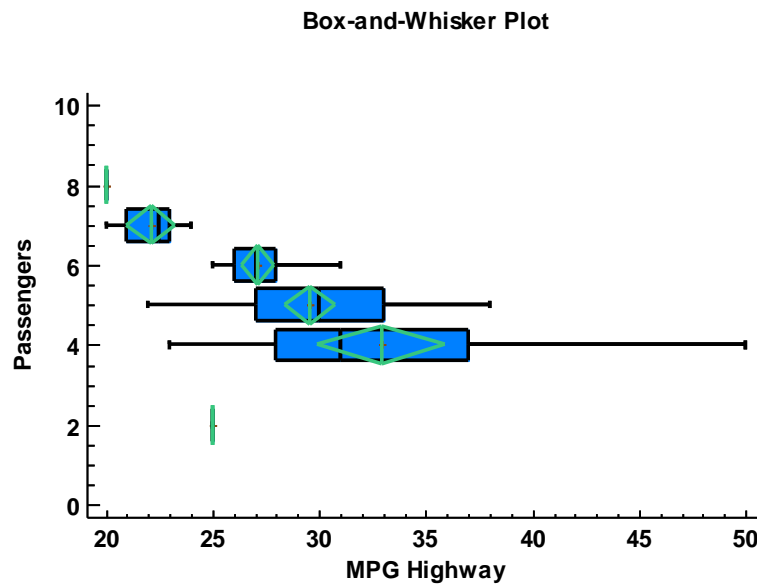


Reference: *Multiple Sample Comparison*

Oneway ANOVA

The following changes have been made:

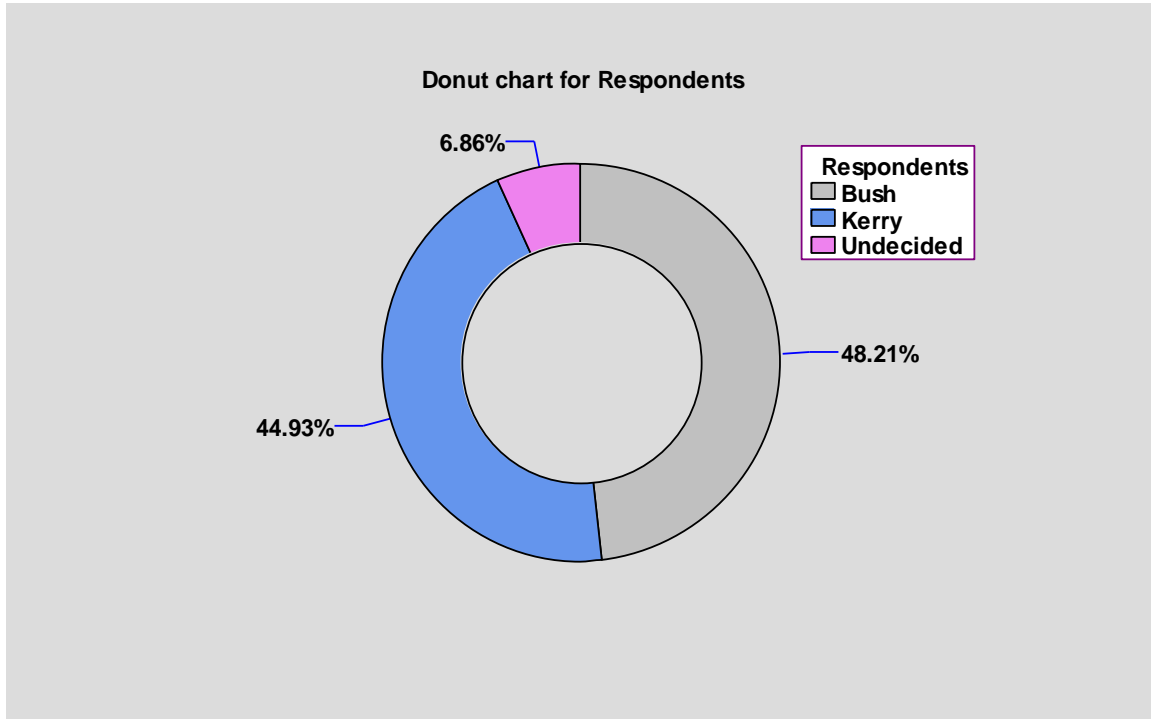
3. The *Analysis Options* dialog box now gives the option to treat factor levels as numeric. This improves the display of numeric factors in tables and graphs.
4. A diamond may be added to the multiple box-and-whisker plot showing confidence intervals for each level mean.



Reference: *One-Way ANOVA*

Piechart/Donut Chart

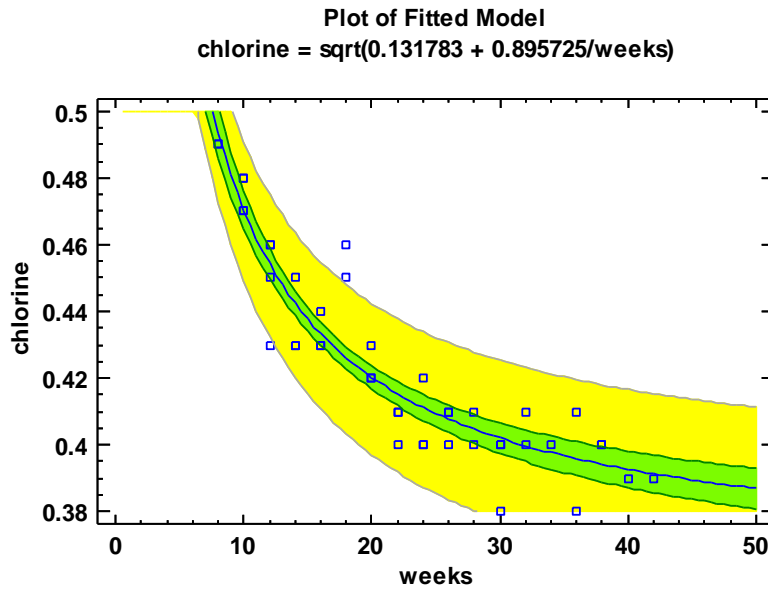
This menu item was renamed and a donut chart added as an alternative to the piechart. The donut chart is similar to the piechart except that the center is removed.



Reference: *Piechart*

Simple Regression, Polynomial Regression, Box-Cox Transformations, and Calibration Models

A new feature has been added to the Plot of Fitted Model that shades the area between the prediction and confidence limits.

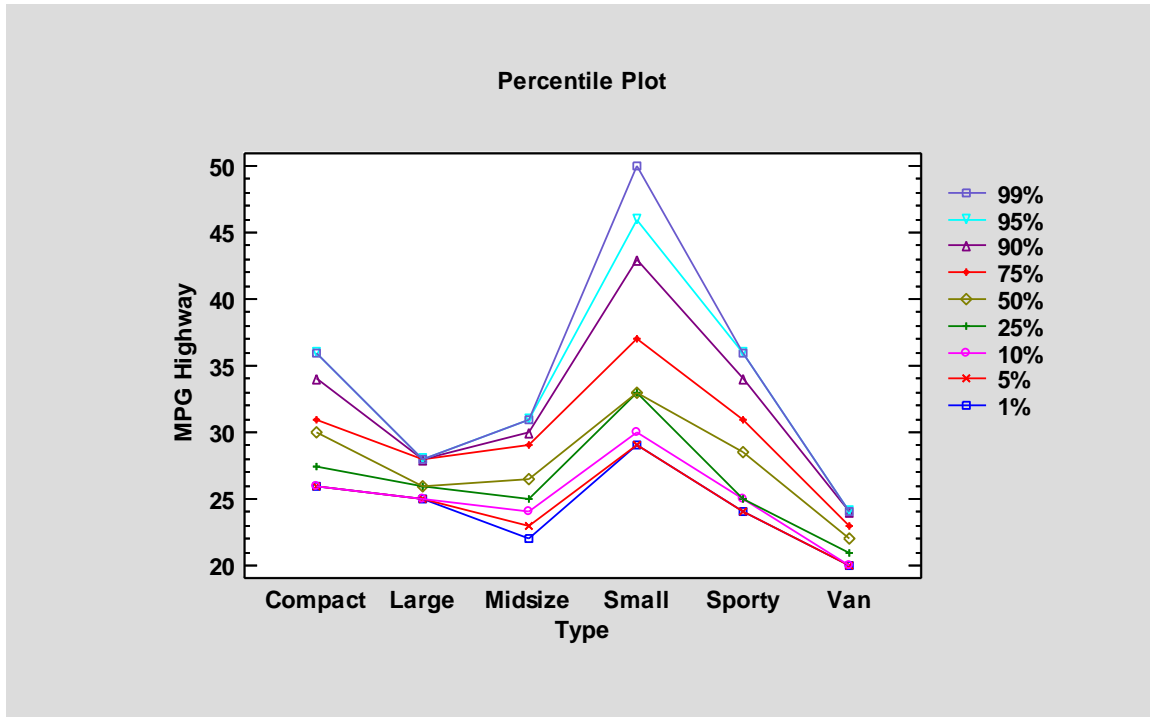


Reference: *Simple Regression*

Subset Analysis

Several new features were added to this procedure:

1. A table of percentiles may now be created for each level of the code variable.
2. A plot of the percentiles by code level may be created.
3. A diamond may be added to the multiple box-and-whisker plot showing confidence intervals for each level mean.
4. Axis scaling has been improved for numeric codes.

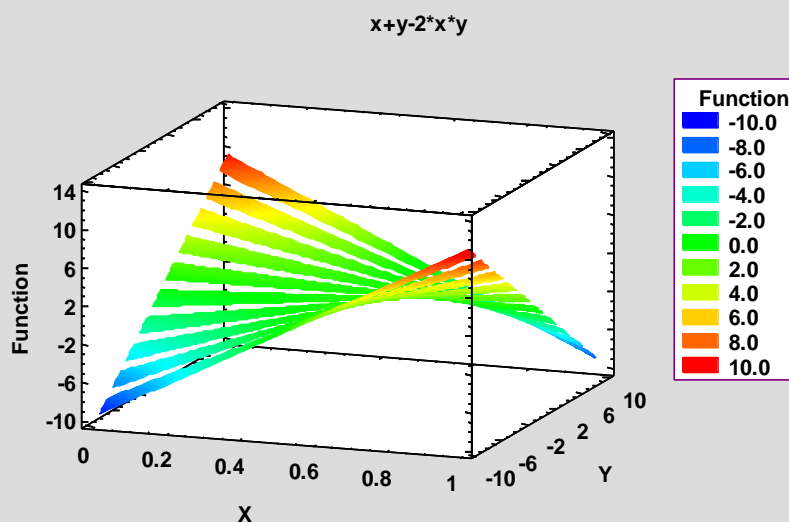
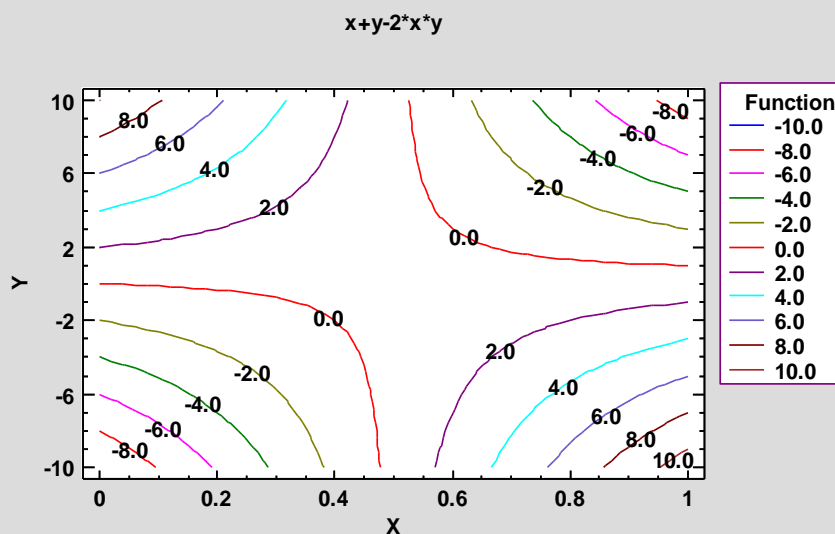


Reference: *Subset Analysis*

Surface and Contour Plots

The following features have been added:

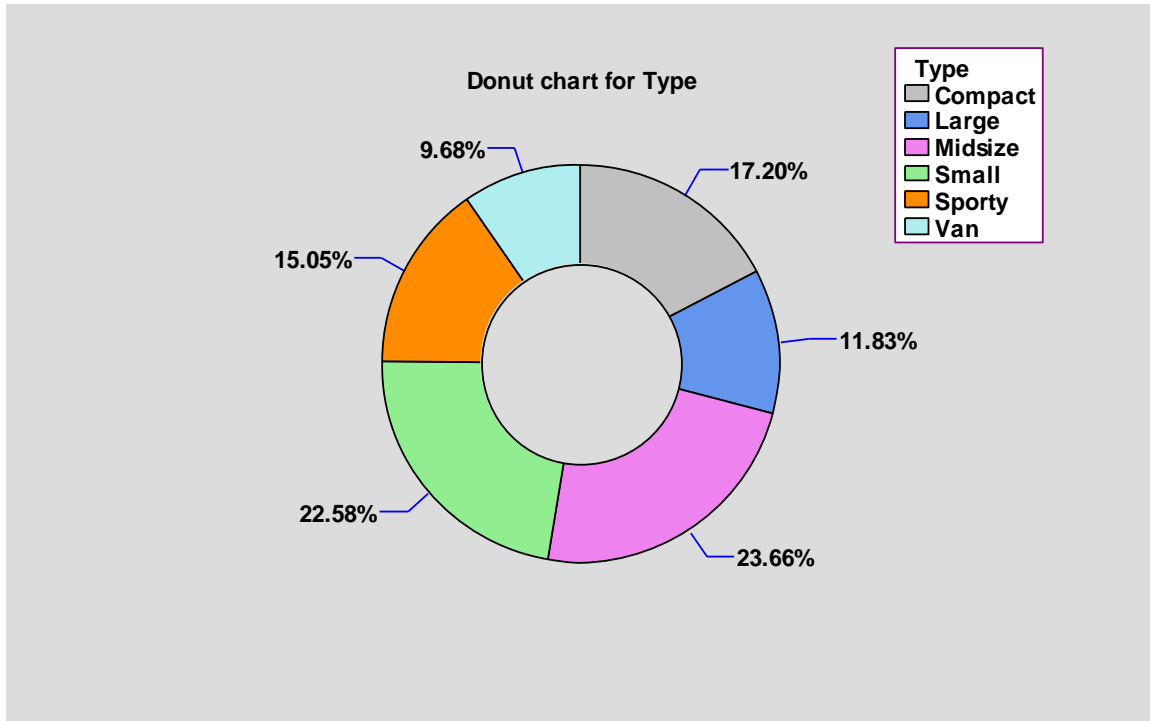
1. Levels are now labeled on contour plots.
2. A ribbon plot has been added.



Reference: *Surface and Contour Plots*

Tabulation

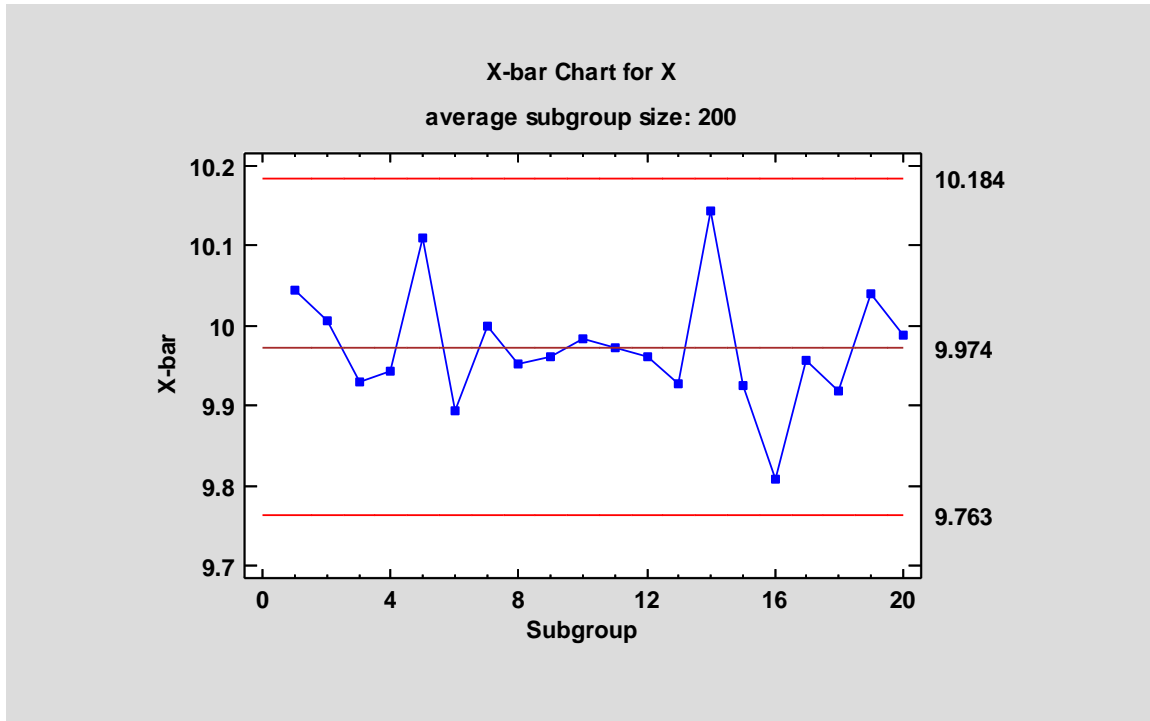
An option has been added to create a donut chart. It is similar to a piechart except that the center is missing.



Reference: *Tabulation*

X-bar and R Charts

In earlier versions, these charts were limited to a maximum subgroup size of $n = 100$. That limitation has been removed.



Reference: *X-bar and R charts*

New Statistical Analyses

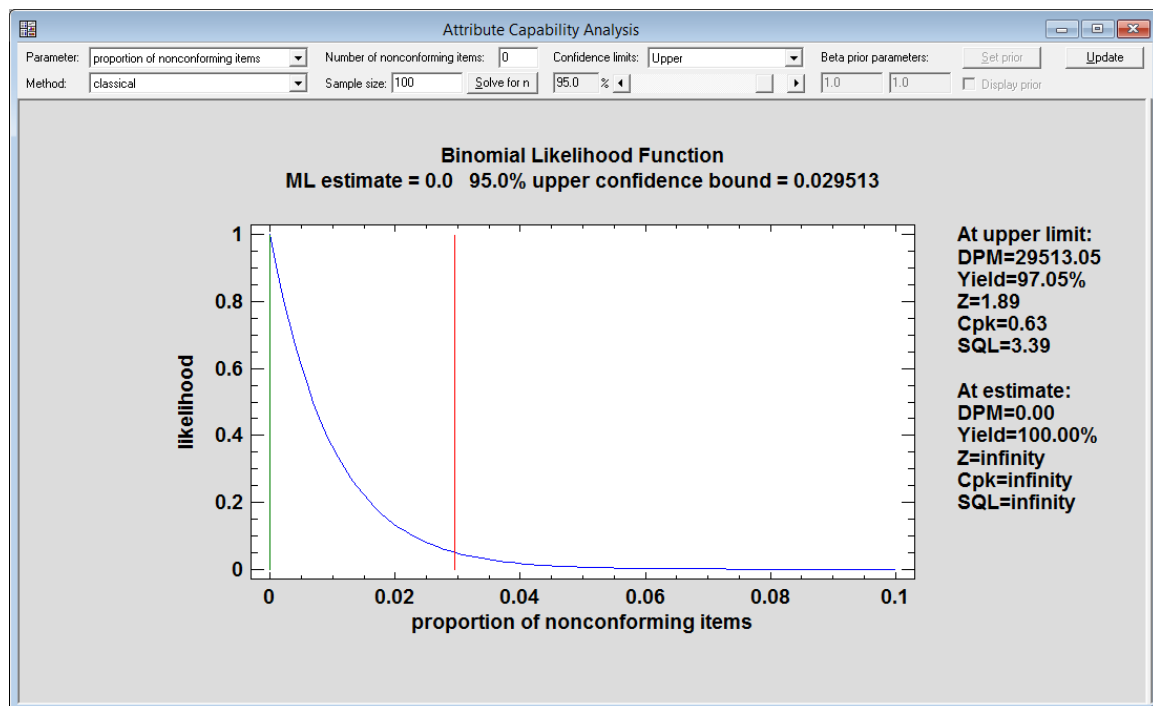
Attribute Capability Analysis Statlet

This Statlet performs a capability analysis based on attribute data. The data may consist of either the number of nonconforming items in a sample or the total number of nonconformities if an item can have more than one nonconformity. The analysis is based on either the binomial or the Poisson distribution.

The Statlet will calculate:

1. Parameter estimates and confidence limits or upper confidence bounds.
2. Capability indices (at both the best estimate and the upper bound).
3. DPM (defects per million).

The analysis may be based on either a classical or Bayesian approach.



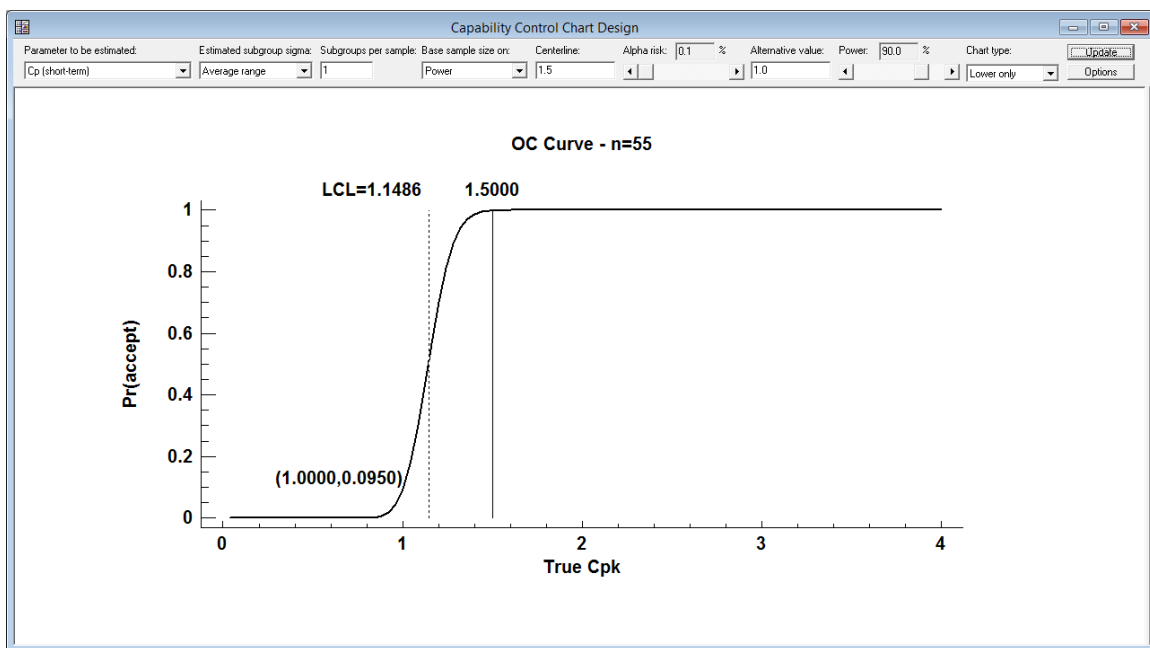
Reference: *Attribute Capability Analysis Statlet*

Capability Control Chart Design Statlet

This new Statlet assists analysts in determining how large samples should be when constructing capability control charts. *Capability control charts* monitor processes which have been shown to be stable and capable of producing results that yield small numbers of nonconformities.

Capability control charts may be constructed for::

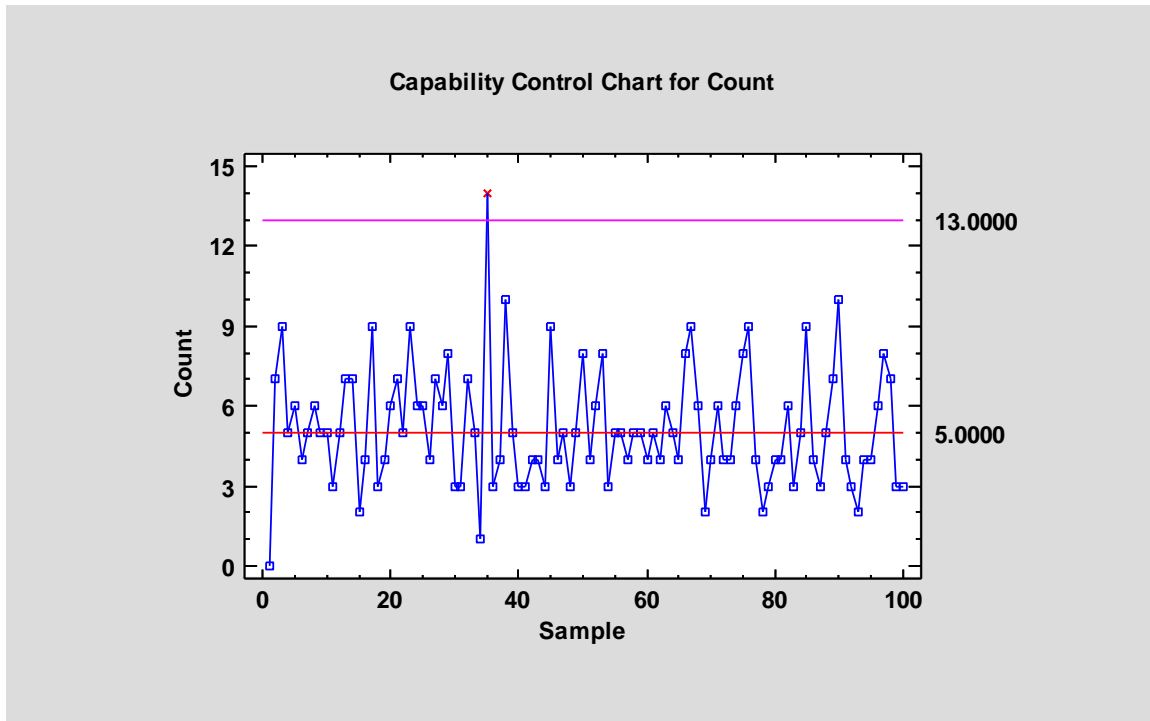
1. The short-term capability index C_p .
2. The long-term capability index P_p .
3. The short-term capability index C_{pk} .
4. The long-term capability index P_{pk} .
5. The proportion of nonconforming items.
6. The rate of nonconformities.



Reference: *Capability Control Chart Design Statlet*

Capability Control Charts for Attributes

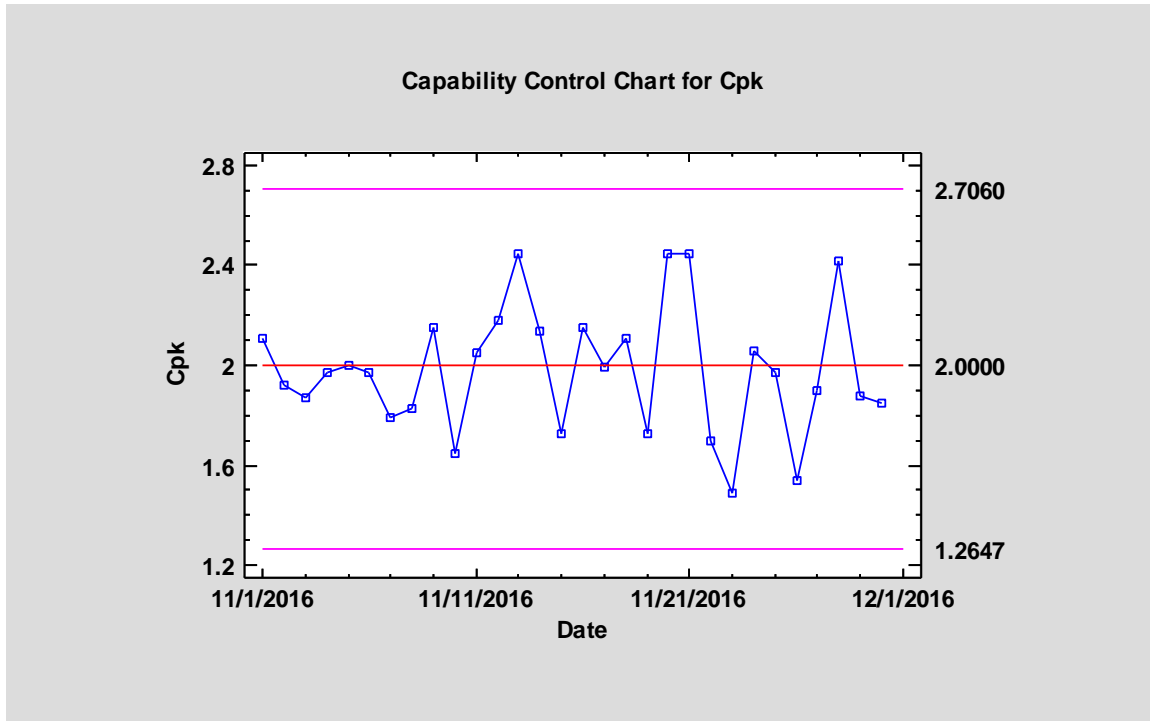
This procedure constructs Phase II statistical process control charts for monitoring either the proportion of nonconforming items or the rate of nonconformities in a process. Given a process that is deemed to be capable of satisfying stated requirements based on the analysis of attribute data, these charts monitor continued compliance with those requirements.



Reference: *Capability Control Charts for Attributes*

Capability Control Charts for Variables

This procedure constructs Phase II statistical process control charts for monitoring capability indices such as C_p and C_{pk} . Given a process that is deemed to be capable of satisfying stated requirements based on the analysis of variable data, these charts monitor continued compliance with those requirements.



Reference: *Capability Control Charts for Variables*

Classification and Regression Trees

The *Classification and Regression Trees* procedure implements a machine-learning process to predict observations from data. It creates models of 2 forms:

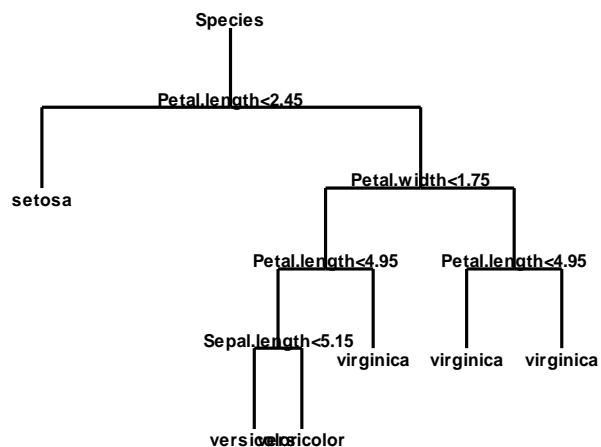
1. *Classification models* that divide observations into groups based on their observed characteristics.
2. *Regression models* that predict the value of a dependent variable.

The models are constructed by creating a tree, each node of which corresponds to a binary decision. Given a particular observations, one travels down the branches of the tree until a terminating leaf is found. Each leaf of the tree is associated with a predicted class or value.

Observations are typically divided into three sets:

1. A *training* set which is used to construct the tree.
2. A *validation* set for which the actual classification or value is known, which can be used to validate the model.
3. A *prediction* set for which the actual classification or value is not known but for which predictions are desired.

The dependent variable may be either categorical or quantitative, as may the predictor variables. The calculations are performed by the “tree” package in R.



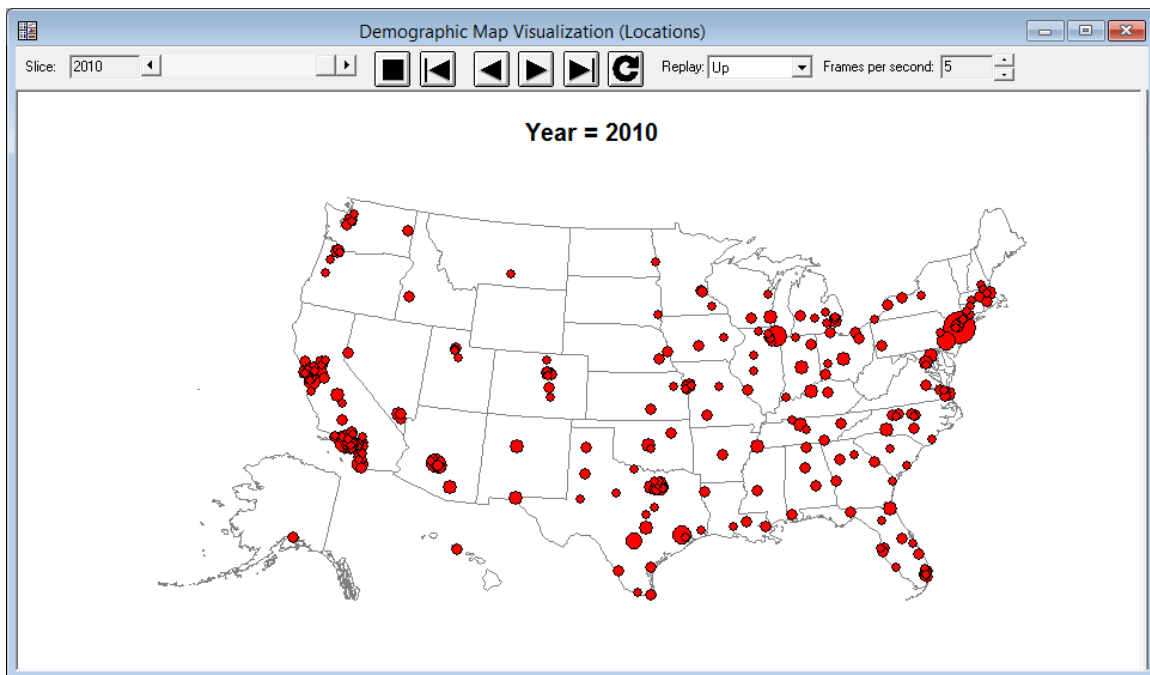
Reference: *Classification and Regression Trees*

Demographic Map Visualizer (Locations)

This new Statlet is designed to illustrate changes in location statistics over time. Given data at each of k locations during p time periods, the program generates a dynamic display that illustrates how the data have changed at each location. Typical applications include plotting:

1. Population and other demographic measurements.
2. Unemployment indices, housing starts, and other economic indices.
3. Environmental statistics.

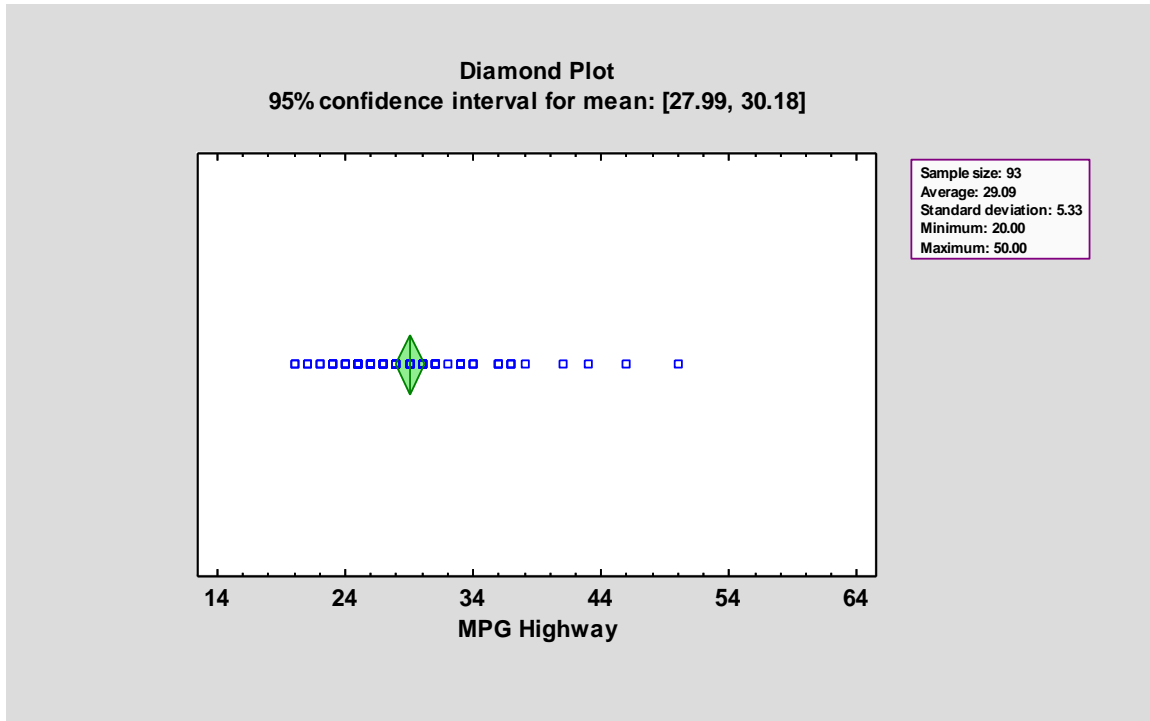
The data at each location is drawn using a bubble whose size is proportional to the observed data value. As time changes, the analyst can follow changes in the data at each location. Various options are offered for smoothing the data and for dealing with missing values.



Reference: *Demographic Map Visualizer (Locations)*

Diamond Plot

The **Diamond Plot** procedure creates a plot for a single quantitative variable showing the n sample observations together with a confidence interval for the population mean.



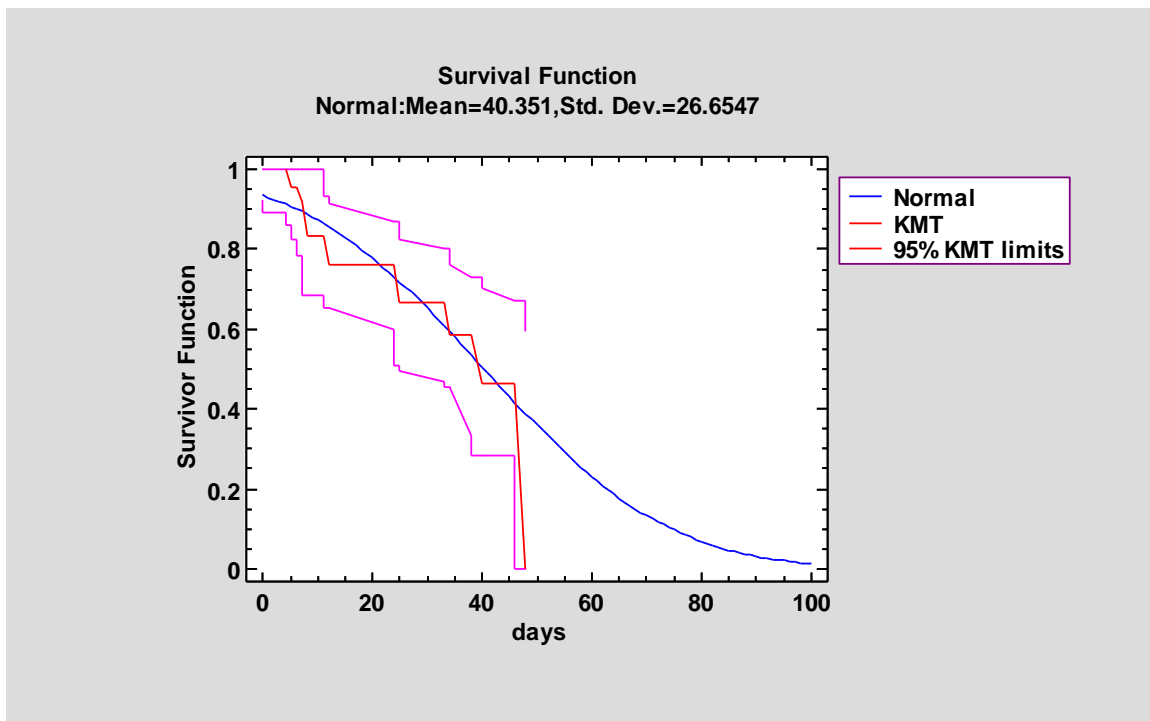
Reference: *Diamond Plot*

Distribution Fitting (Arbitrarily Censored Data)

The **Distribution Fitting (Arbitrarily Censored Data)** procedure analyzes data in which one or more observations are not known exactly. In particular, observations may be:

1. **Left-censored:** known to be less than a stated value.
2. **Right-censored:** known to be greater than a stated value.
3. **Interval censored:** known to fall within a stated interval.

The procedure calculates summary statistics, fits distributions, creates graphs, and calculates a nonparametric estimate of the survival function.

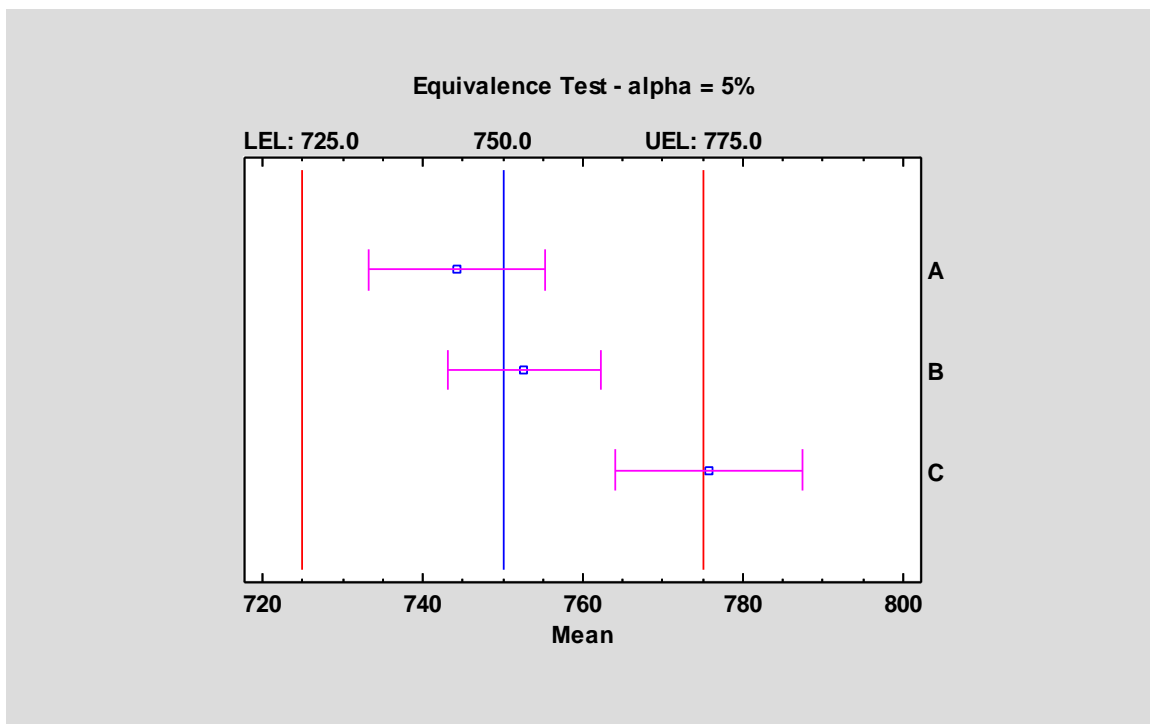


Reference: *Distribution Fitting (Arbitrarily Censored Data)*

Equivalence and Noninferiority Tests (Comparing Mean to Target)

This procedure tests whether the mean of a sample obtained from a single population may be considered to be equivalent to a target value. A mean is considered to be “equivalent” if it falls within a specified interval surrounding that target value. Unlike standard hypothesis tests which are designed to prove that a mean is significantly different than a specified value, equivalence tests are designed to prove that the mean is essentially equivalent to the target.

The procedure may also be used to demonstrate noninferiority. A sample is considered to be “noninferior” if the difference between the mean and the target value is no greater than (or no less than) a specified value. This situation corresponds to a one-sided test of equivalence.

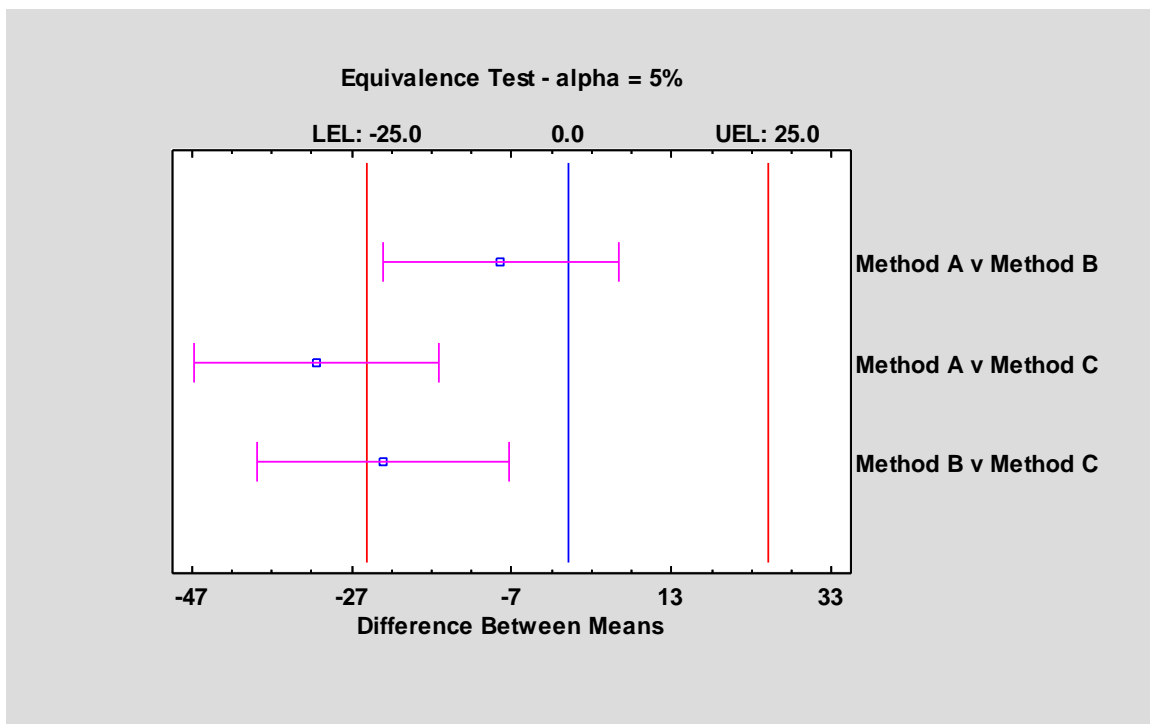


Reference: *Equivalence and Noninferiority Tests (Comparing Mean to Target)*

Equivalence and Noninferiority Tests (Comparing Paired Samples)

This procedure tests whether the means of 2 samples may be considered equivalent, assuming that the data in the 2 samples consist of matched pairs. Two samples are considered to be “equivalent” if the difference between their respective means falls within some specified interval surrounding 0. Unlike standard hypothesis tests which are designed to prove superiority of one method over another, equivalence tests are designed to prove that two methods have essentially the same mean.

The procedure may also be used to demonstrate noninferiority. Two samples are considered to be “noninferior” if the difference between their respective means is no greater than (or no less than) a specified value. This situation corresponds to a one-sided test of equivalence.

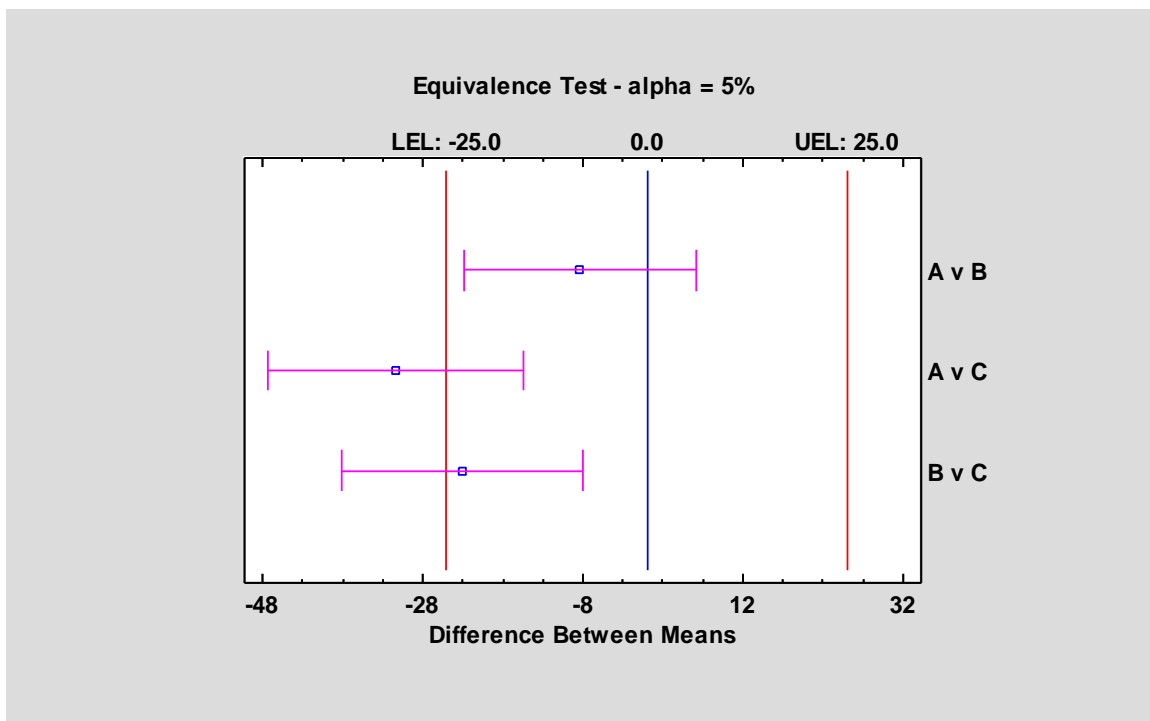


Reference: *Equivalence and Noninferiority Tests (Comparing Paired Samples)*

Equivalence and Noninferiority Tests (Comparing Two Means)

This procedure tests whether the means of 2 samples may be considered equivalent. Two samples are considered to be “equivalent” if the difference between their respective means falls within some specified interval surrounding 0. Unlike standard hypothesis tests which are designed to prove superiority of one method over another, equivalence tests are designed to prove that two methods have essentially the same mean.

The procedure may also be used to demonstrate noninferiority. Two samples are considered to be “noninferior” if the difference between their respective means is no greater than (or no less than) a specified value. This situation corresponds to a one-sided test of equivalence.

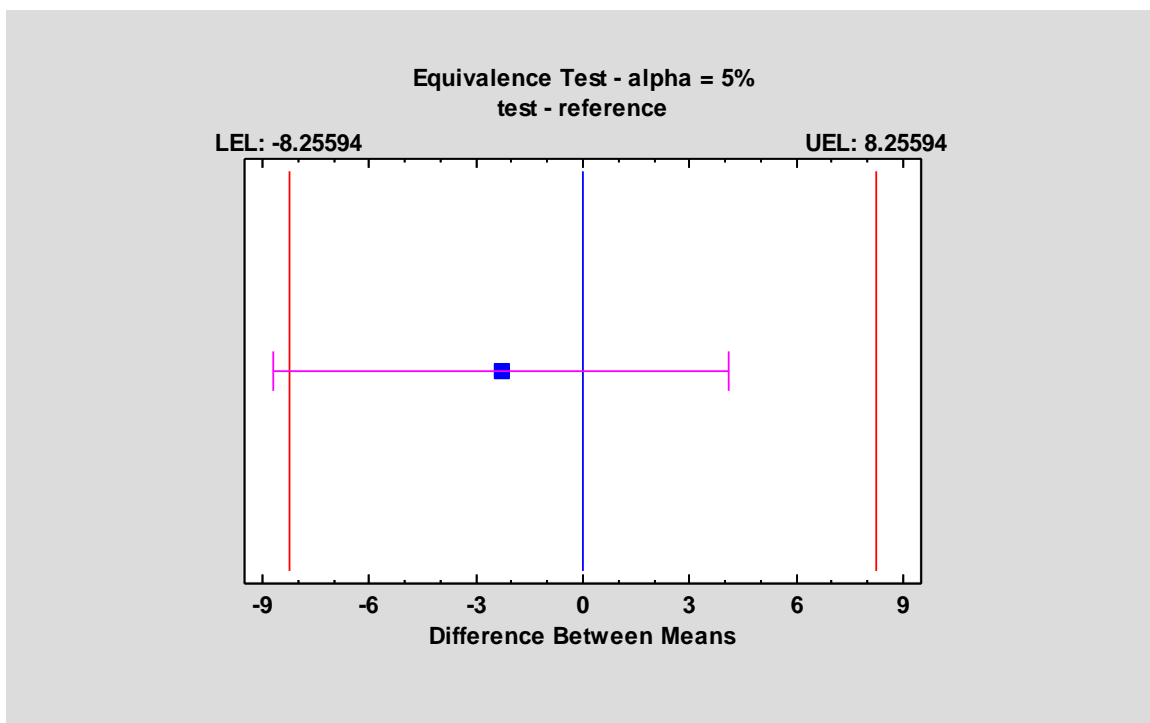


Reference: *Equivalence and Noninferiority Tests (Comparing Two Means)*

Equivalence and Noninferiority Tests (2x2 Crossover Study)

This procedure is used to demonstrate the equivalence of 2 treatments based on a 2x2 crossover study. In such a study, subjects are randomly assigned to 2 sequences. In one sequence, treatment #1 is applied first, followed by treatment #2. In the other sequence, treatment #2 is applied first followed by treatment #1. We wish to demonstrate equivalence between the means of the 2 treatments.

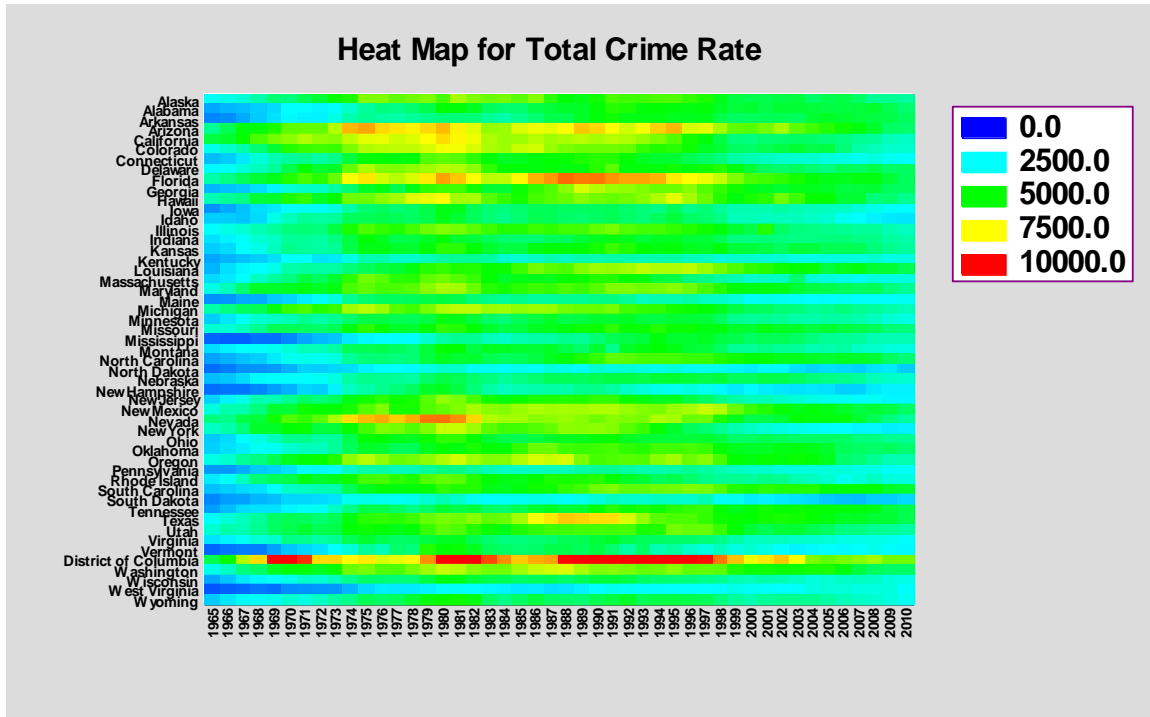
The procedure may also be used to demonstrate noninferiority. Two samples are considered to be “noninferior” if the difference between their respective means is no greater than (or no less than) a specified value. This situation corresponds to a one-sided test of equivalence.



Reference: *Equivalence and Noninferiority Tests (2x2 Crossover Study)*

Heat Map

The **Heat Map** procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient color scale is used to represent values of the quantitative variable.



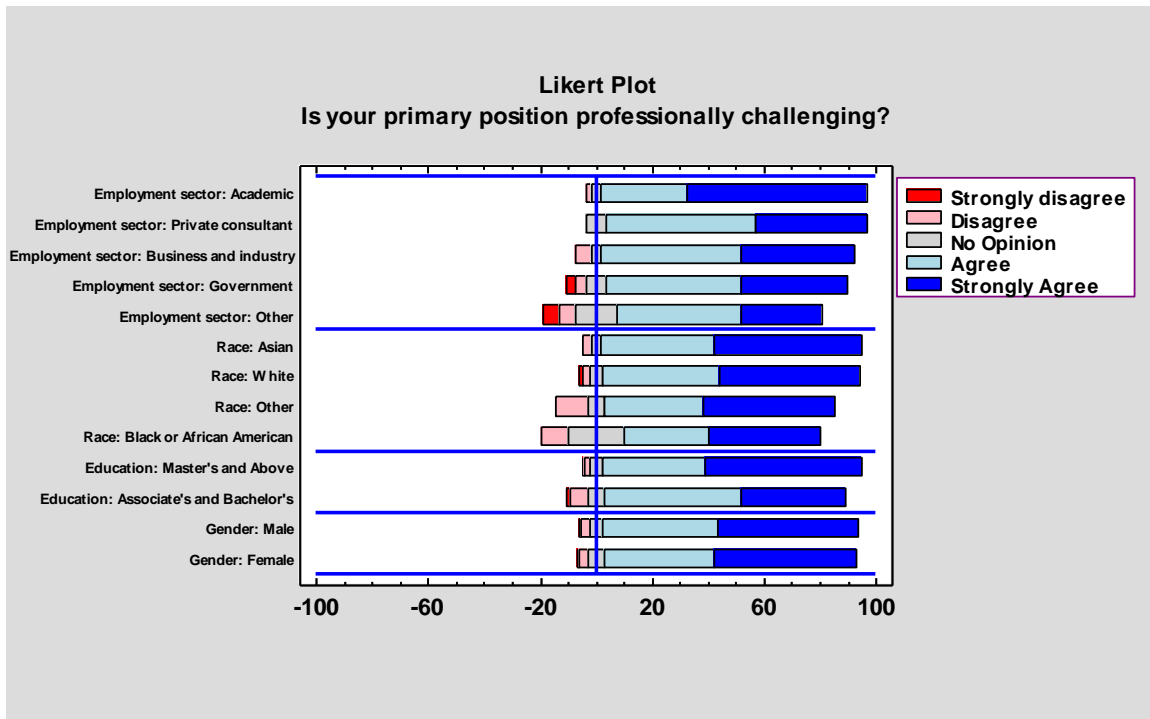
Reference: *Heat Map*

Likert Plot

The **Likert Plot** procedure analyzes data recorded on a Likert scale. Likert scales are commonly used in survey research to record user responses to a statement. A typical 5-level Likert scale might code user responses according to:

- 1 = Strongly disagree
- 2 = Disagree
- 3 = No opinion
- 4 = Agree
- 5 = Strongly agree

This analysis calculates summary statistics and displays the results using a diverging stacked barchart.



Reference: *Likert Plot*

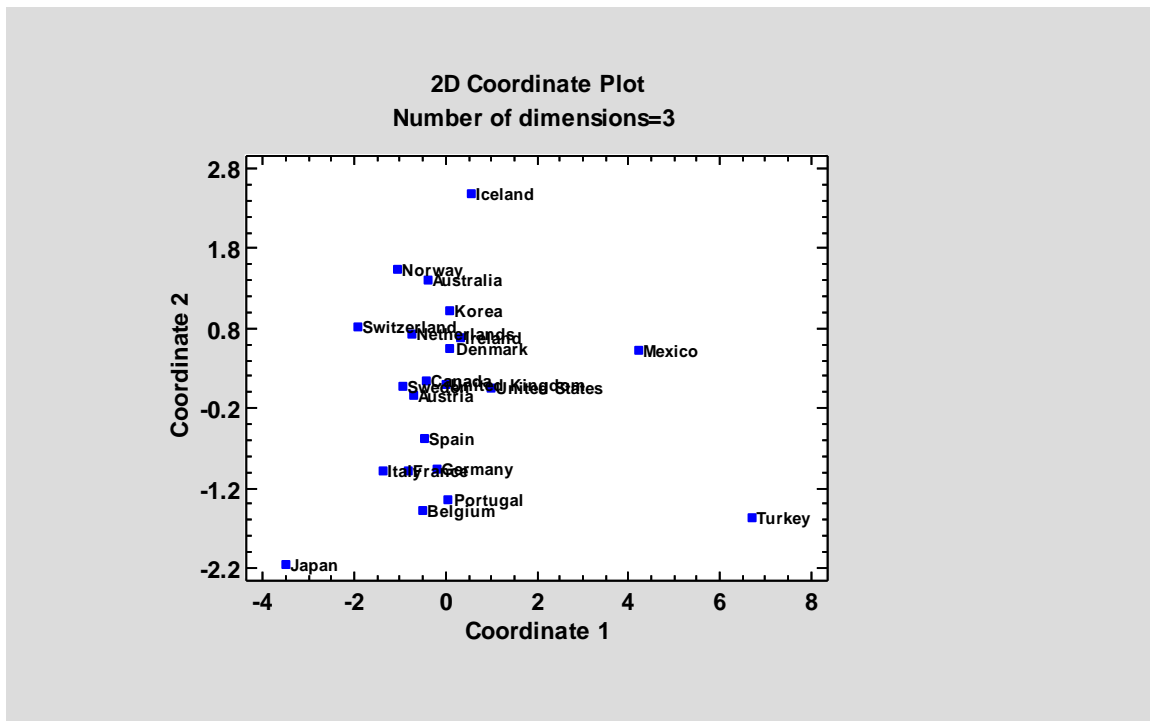
Multidimensional Scaling

The *Multidimensional Scaling* procedure is designed to display multivariate data in a low-dimensional space. Given an n by n matrix of distances between each pair of n multivariate observations, the procedure searches for a low-dimensional representation of those observations that preserves the distances between them as well as possible. The primary output is a map of the points in that low-dimensional space (usually 2 or 3 dimensions).

Input to the procedure may be either:

1. An n by n matrix of distances or “dissimilarities”.
2. An n by p matrix of observations for p variables, from which a distance matrix may be constructed.

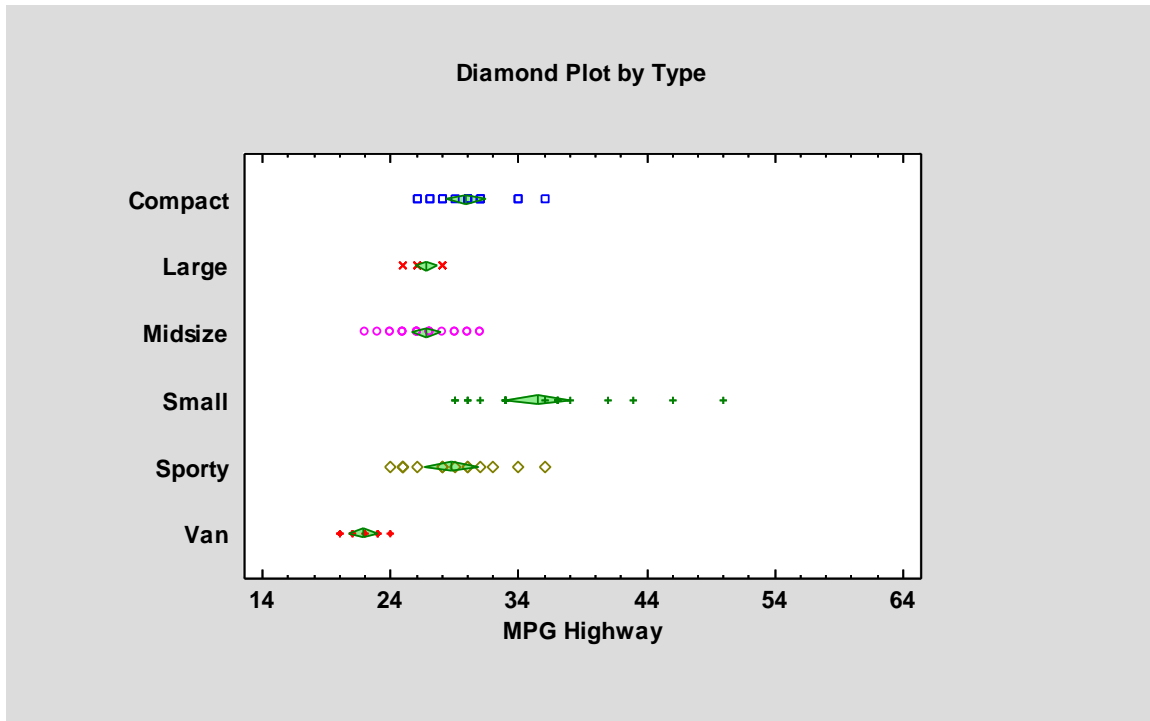
The calculations are performed by R using the “cmdscale” and “isoMDS” functions.



Reference: *Multidimensional Scaling*

Multiple Diamond Plot

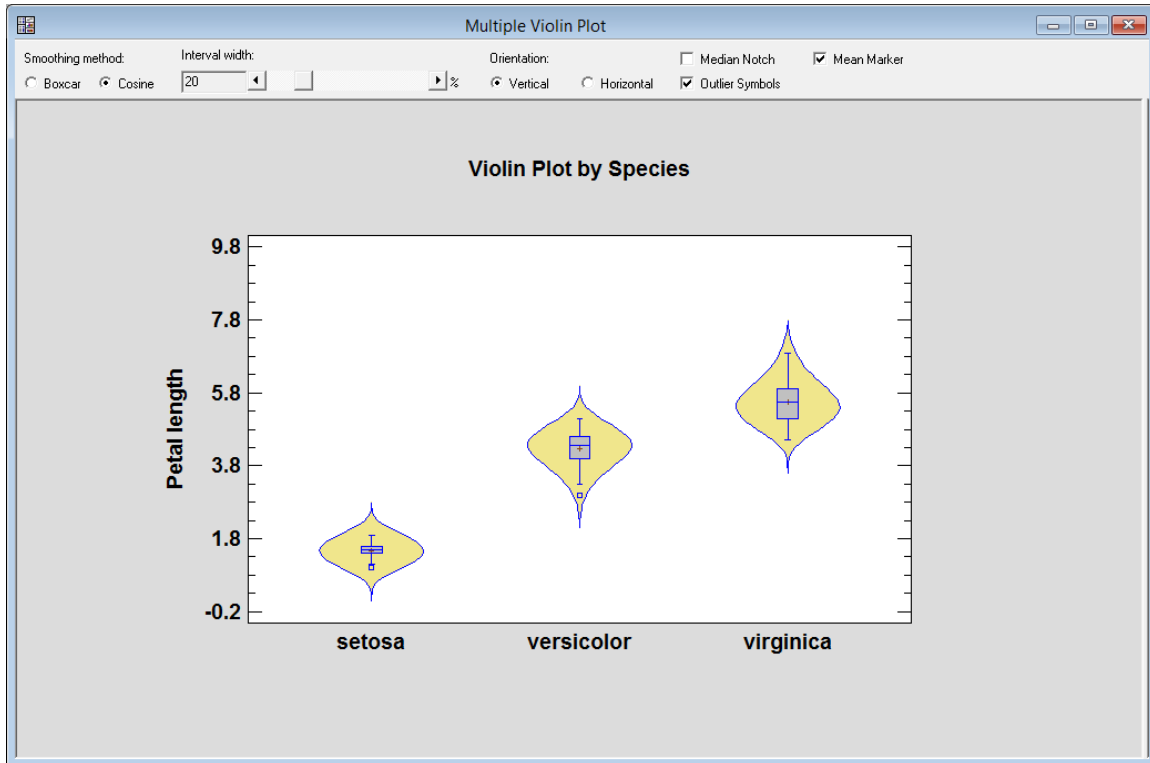
The **Diamond Plot** procedure creates a plot for two or more samples showing the sample observations together with confidence intervals for their respective population means.



Reference: *Multiple Diamond Plot*

Multiple Violin Plot Statlet

The *Multiple Violin Plot Statlet* displays data for 2 or more quantitative samples using a combination of a box-and-whisker plot and a nonparametric density estimator. It is very useful for visualizing the shape of the probability density function for the populations from which the data came.



Reference: *Multiple Violin Plot Statlet*

Multivariate Normal Random Numbers

This procedure generates random numbers from a multivariate normal distribution involving up to 12 variables. The user inputs the variable means, standard deviations, and the correlation matrix. Random samples are generated which may be saved to the Statgraphics databook.

Multivariate Normal Random Numbers

<i>Variable</i>	<i>Mean</i>	<i>Standard Deviation</i>	<i>Sample Mean</i>	<i>Sample Standard Deviation</i>
X1	2.0	0.02	1.99958	0.0208044
X2	250.0	10.0	249.3	10.4658

Correlations

	X1	X2
X1	1.0	0.9
X2	0.9	1.0

Sample Correlations

	X1	X2
X1	1.0	0.914191
X2	0.914191	1.0

Sample size: 200

Seed for random number generator: 11344

Reference: *Multivariate Normal Random Numbers*

Multivariate Normality Test

This procedure tests whether a set of random variables could reasonably have come from a multivariate normal distribution. It includes Royston's H test and tests based on a chi-square plot of the squared distances of each observation from the sample centroid.

Multivariate Normality Test

Data variables:

stiffness (psi)

bending strength (psi)

	Mean	Standard deviation
stiffness	1860.5	352.214
bending strength	8354.13	1867.17

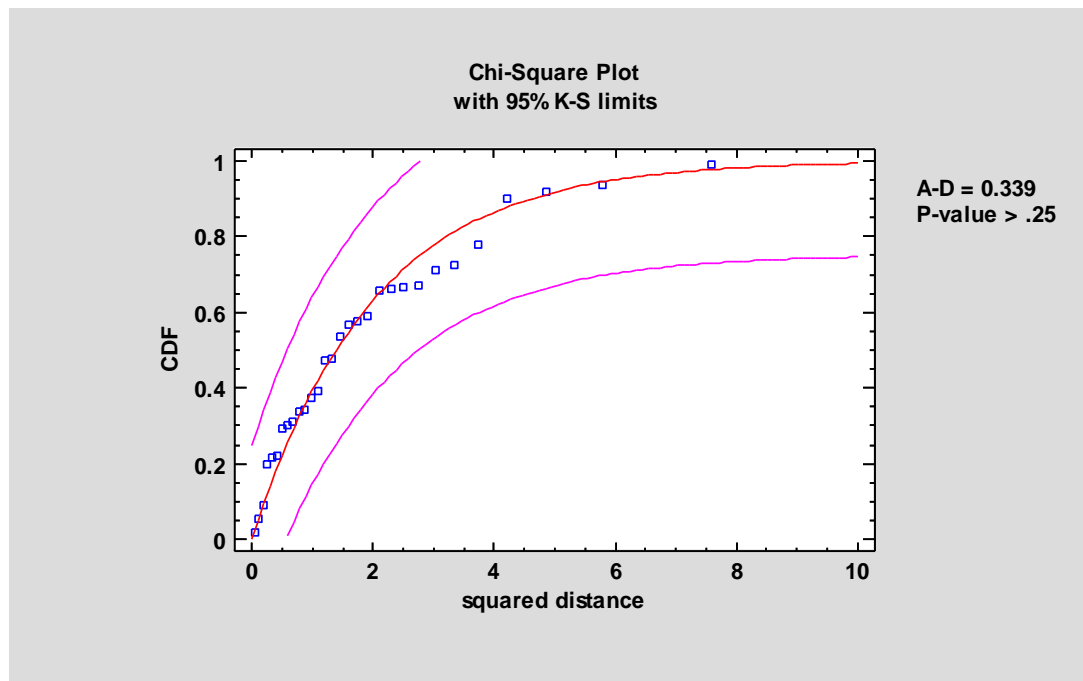
Sample Correlations

	stiffness	bending strength
stiffness	1.0	0.549872
bending strength	0.549872	1.0

Number of observations = 30

Goodness-of-Fit Test

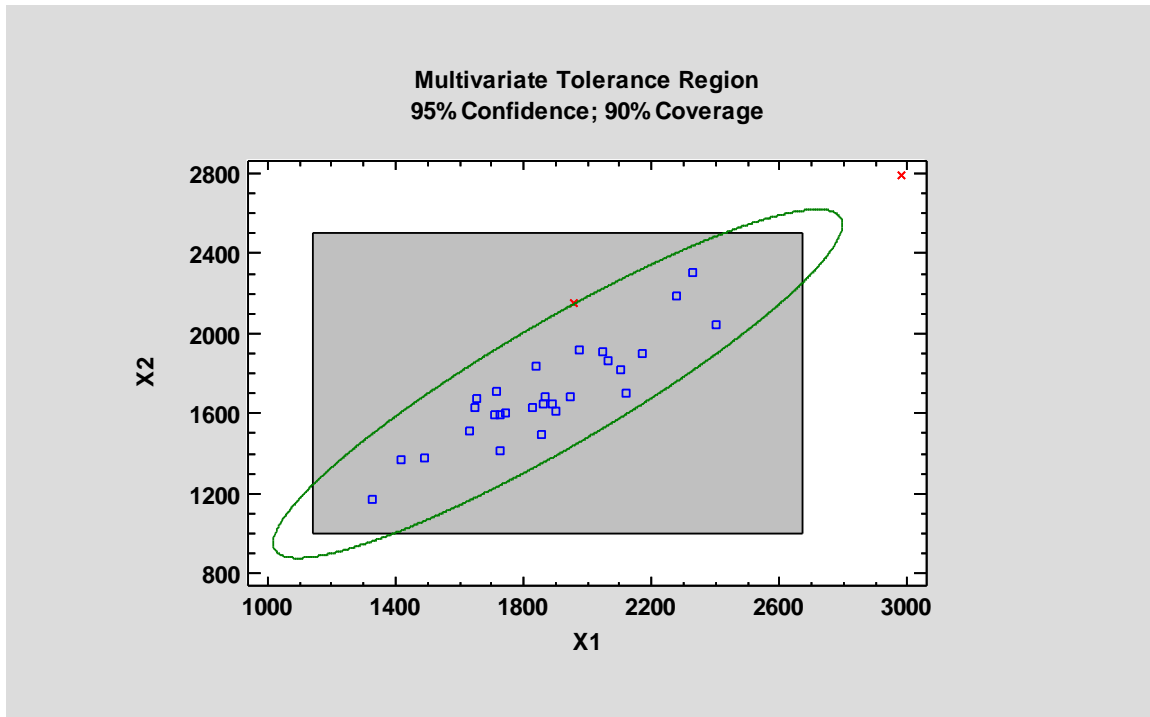
Test	Statistic	P-Value
Shapiro-Wilk W - stiffness	0.975	0.6798
Shapiro-Wilk W - bending strength	0.976	0.6980
Royston's H	0.325	0.8545



Reference: *Multivariate Normality Test*

Multivariate Tolerance Limits

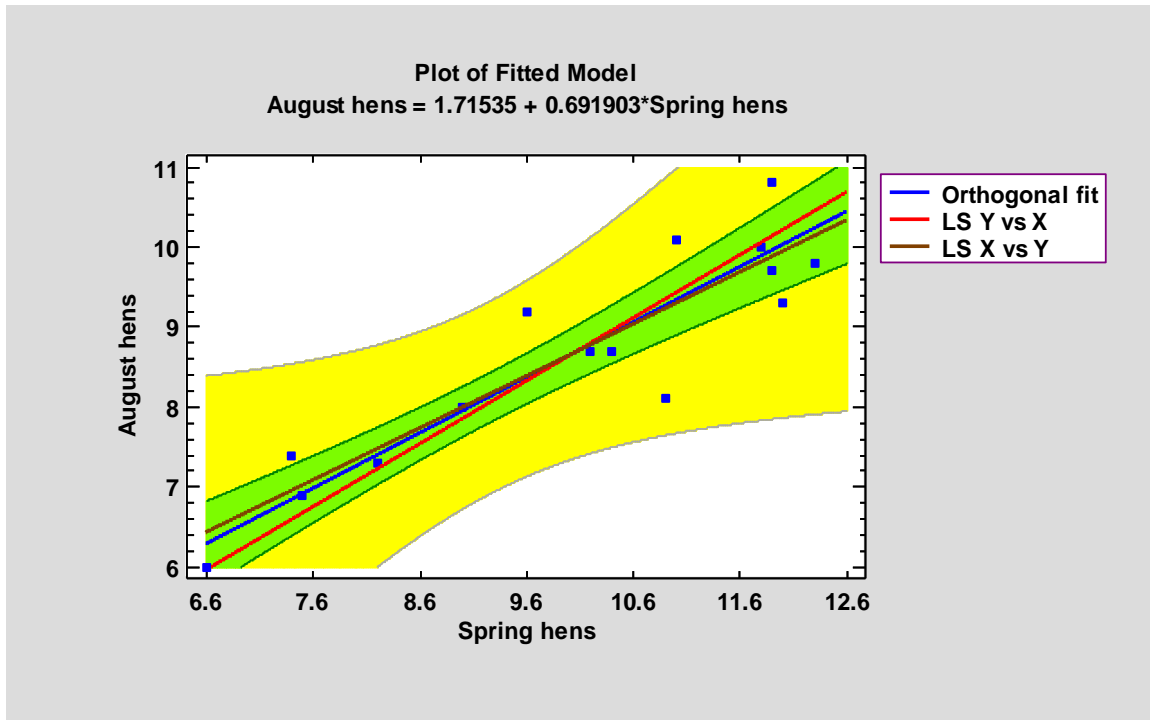
The *Multivariate Tolerance Limits* procedure creates statistical tolerance limits for data consisting of more than one variable. It includes a tolerance region that bounds a selected $p\%$ of the population with $100(1-\alpha)\%$ confidence. It also includes joint simultaneous tolerance limits for each of the variables using a Bonferroni approach. The data are assumed to be a random sample from a multivariate normal distribution. Multivariate tolerance limits are often compared to specifications for multiple variables to determine whether or not most of the population is within spec.



Reference: *Multivariate Tolerance Limits*

Orthogonal Regression

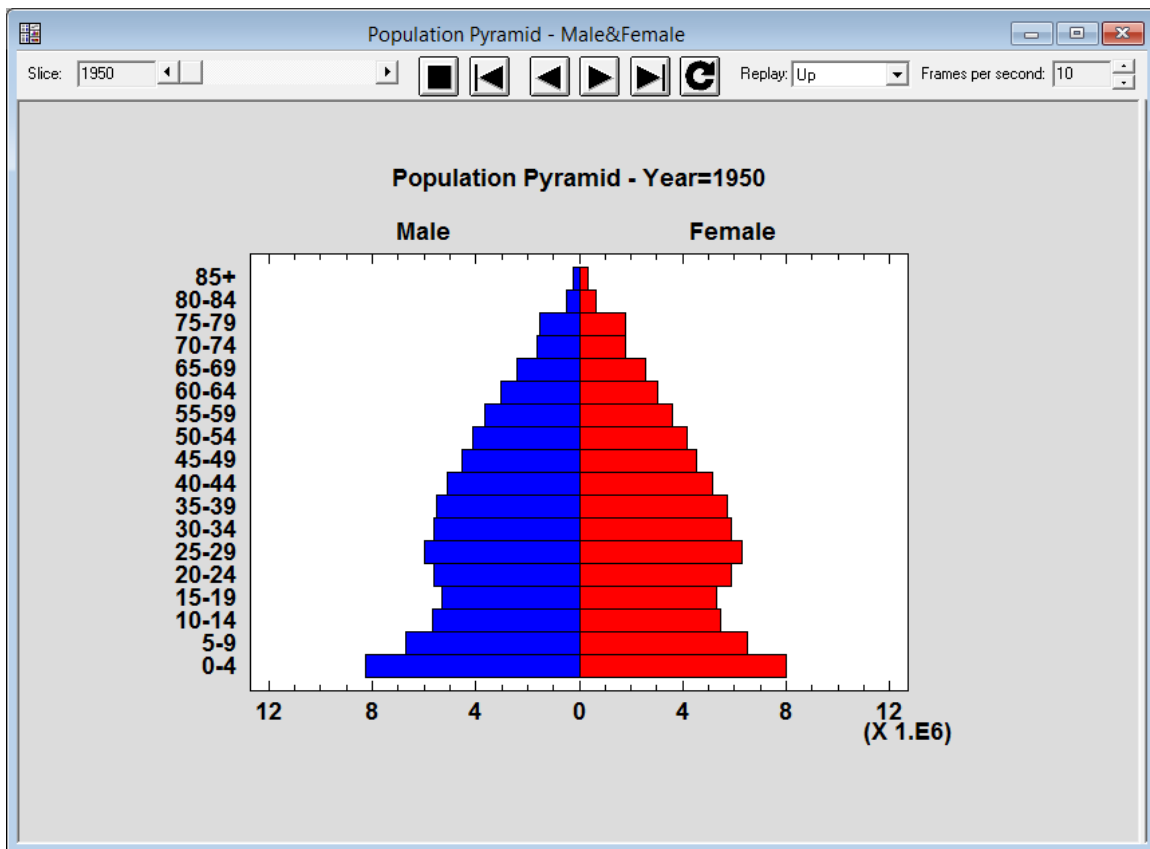
The **Orthogonal Regression** procedure is designed to construct a statistical model describing the impact of a single quantitative factor X on a dependent variable Y, when both X and Y are observed with error. Any of 27 linear and nonlinear models may be fit. Tests are run to determine the statistical significance of the model. The fitted model may be plotted with confidence limits and/or prediction limits. Residuals may also be plotted and unusual observations identified.



Reference: *Orthogonal Regression*

Population Pyramid Statlet

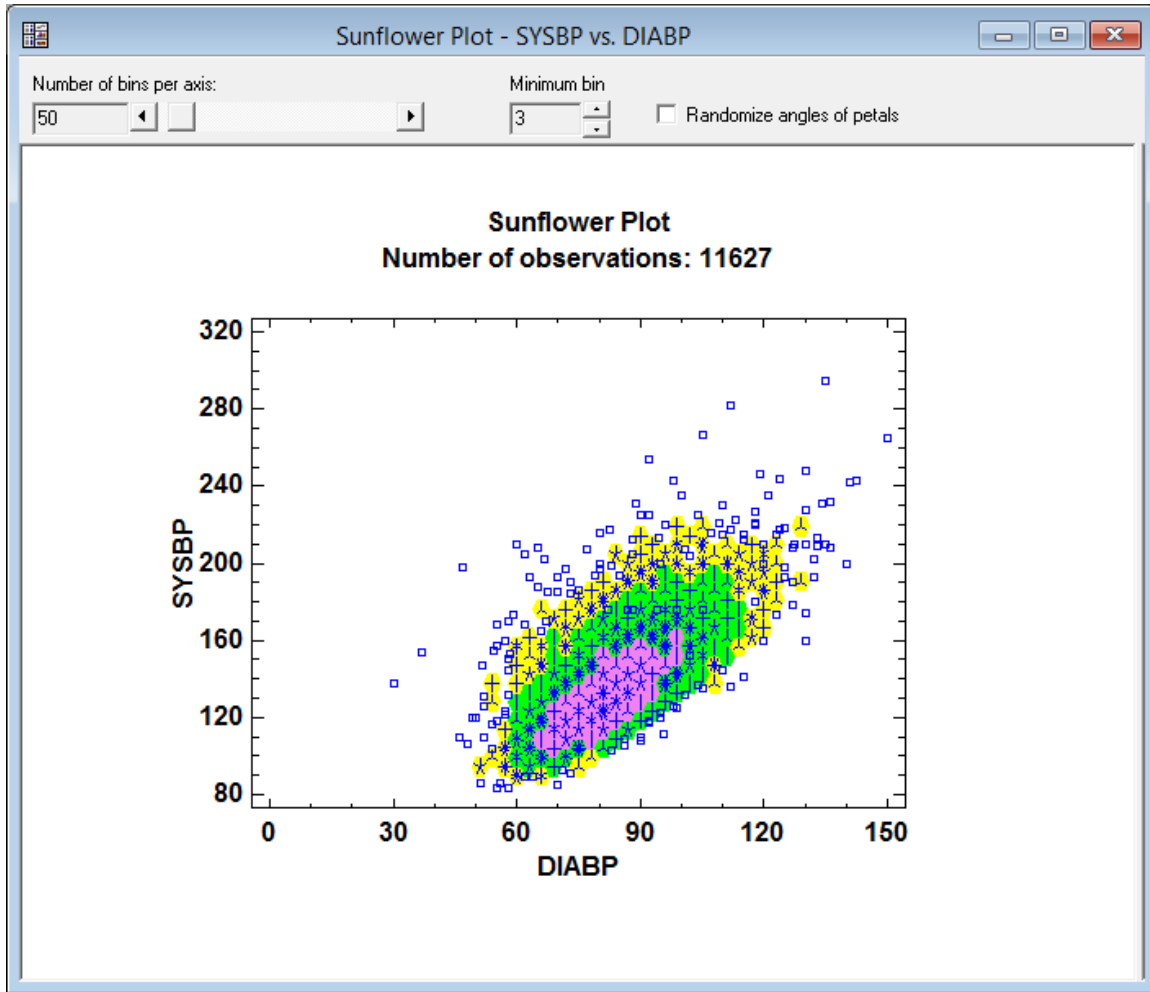
The *Population Pyramid Statlet* is designed to compare the distribution of population counts (or similar values) between 2 groups. It may be used to display that distribution at a single point in time, or it may show changes over time in a dynamic manner. In the latter case, various options are offered for smoothing the data and for dealing with missing values.



Reference: *Population Pyramid Statlet*

Sunflower Plot

The *Sunflower Plot Statlet* is used to display an X-Y scatterplot when the number of observations is large. To avoid the problem of overplotting point symbols with large amounts of data, glyphs in the shape of sunflowers are used to display the number of observations in small regions of the X-Y space.



Reference: *Sunflower Plot*

Text Mining

The *Text Mining* procedure analyzes one or more text columns or documents to determine how frequently various words are used.

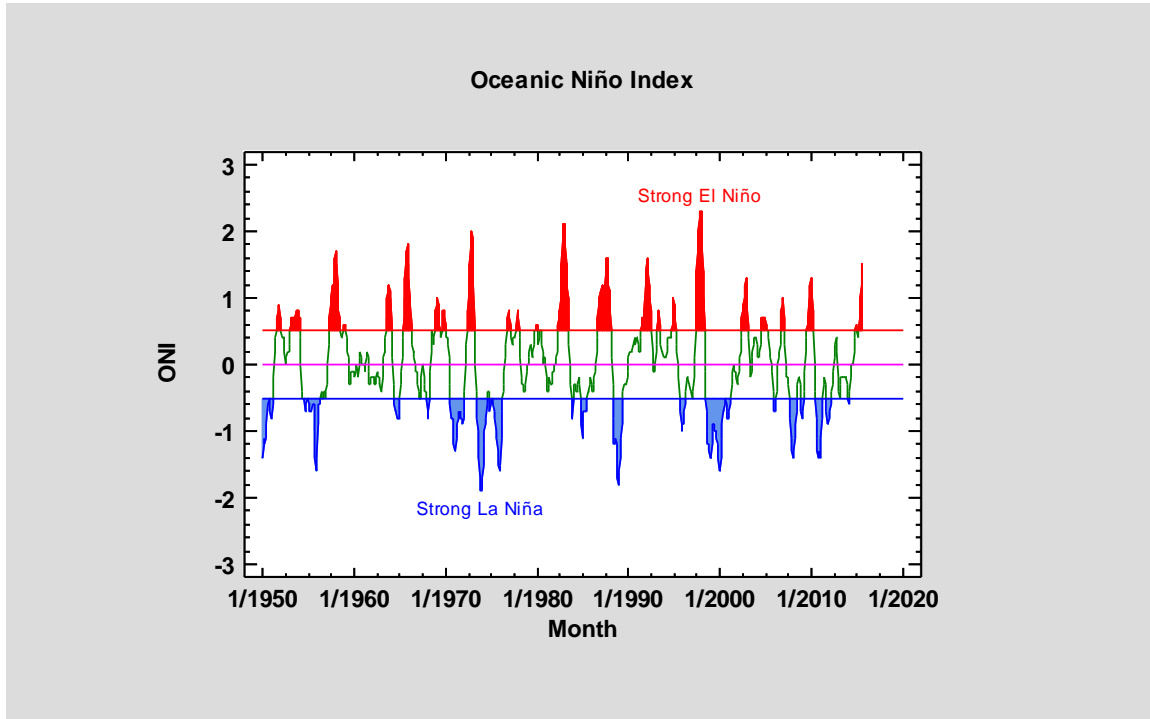
The calculations are performed by the “tm” package in R. To run the procedure, R must be installed on your computer together with the *tm*, *wordcloud*, and *RColorBrewer* packages.

The main output of this procedure is an identification of those words that occur most frequently. Both tabular and graphical summaries are provided.

Reference: *Text Mining*

Time Series Baseline Plot

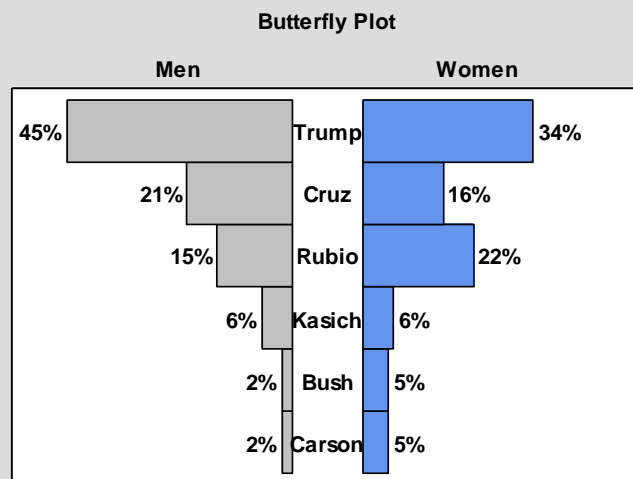
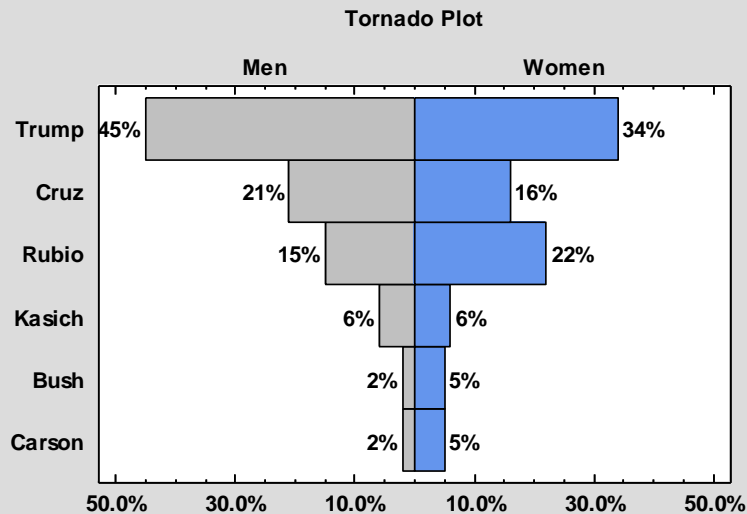
This procedure plots a time series in sequential order, identifying points that are beyond lower and/or upper limits. It is widely used to plot monthly data such as the Oceanic Niño Index.



Reference: *Time Series Baseline Plot*

Tornado and Butterfly Plots

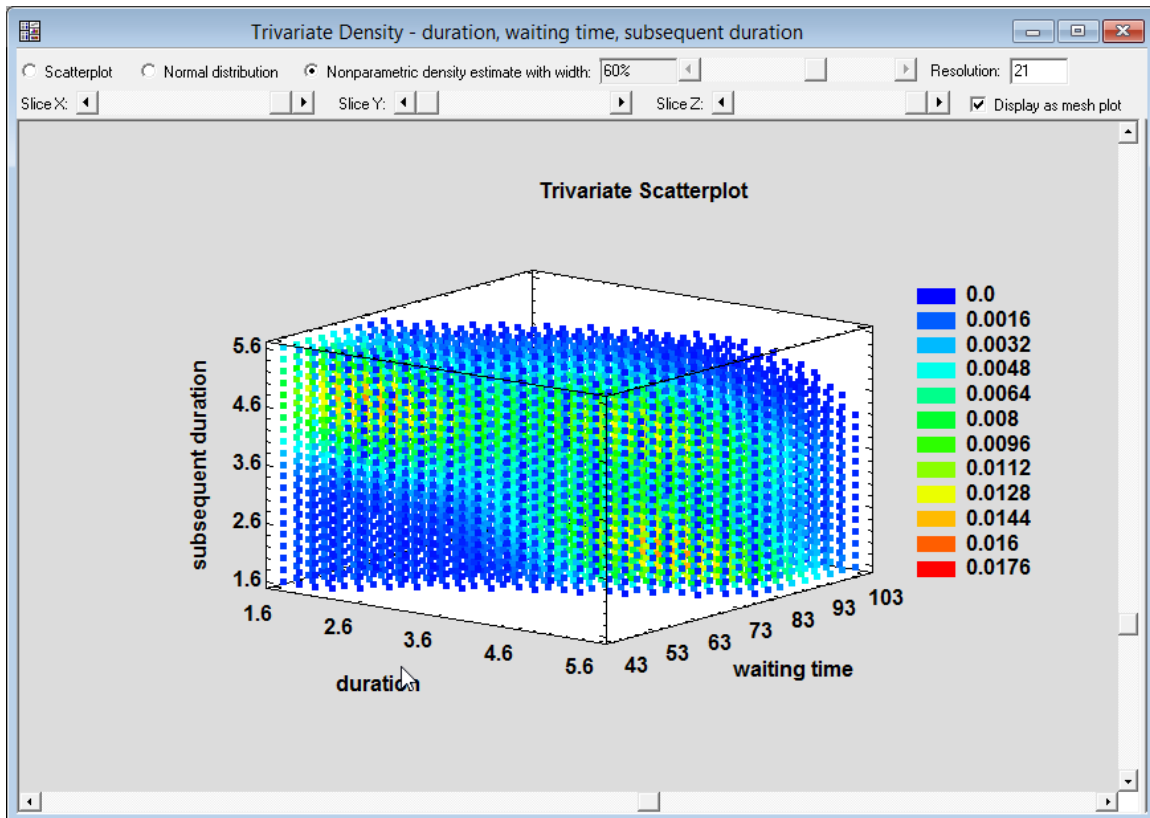
The **Tornado and Butterfly Plots** procedure creates two similar plots that compare 2 samples of attribute data. Each plot consists of 2 sets of bars that show the frequency distribution of each sample over a set of categories. The only difference between the plots is where the labels are placed.



Reference: *Tornado and Butterfly Plots*

Trivariate Density Statlet

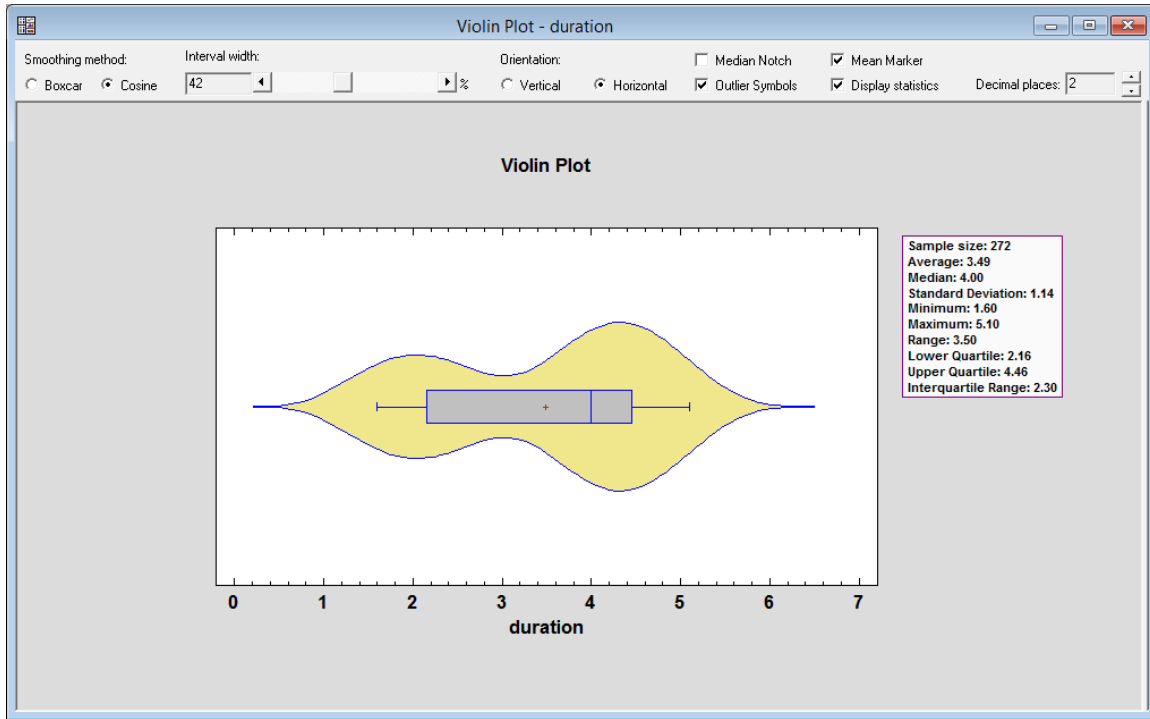
The **Trivariate Density Statlet** displays the estimated density function for 3 columns of numeric data. It does so using either a 3-dimensional contour plot or a 3-dimensional mesh plot. The joint distribution of the 3 variables may either be assumed to be multivariate normal or be estimated using a nonparametric approach.



Reference: *Trivariate Density Statlet*

Violin Plot Statlet

The *Violin Plot Statlet* displays data for a single quantitative sample using a combination of a box-and-whisker plot and a nonparametric density estimator. It is very useful for visualizing the shape of the probability density function for the population from which the data came.



Reference: *Violin Plot Statlet*

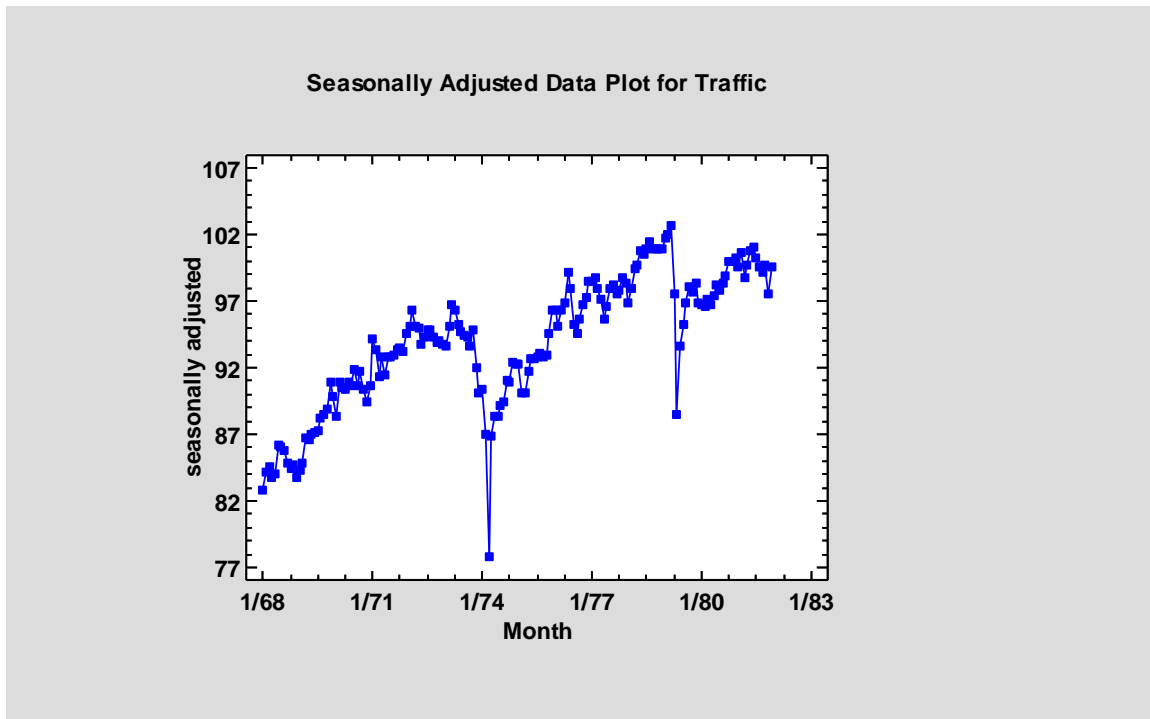
X-13ARIMA-SEATS Seasonal Adjustment

This procedure performs a seasonal adjustment of time series data using the procedure currently employed by the United States Census Bureau. As part of the procedure, the time series is decomposed into 3 components:

1. a trend-cycle component
2. a seasonal component
3. an irregular component

Each component may be plotted separately or saved, together with the seasonally adjusted data.

The seasonal adjustment calculations are performed by the “seasonal” package in R.



Reference: *Seasonal Adjustment using X-13ARIMA-SEATS*