IBM SPSS Statistics Base 29

Note

Before using this information and the product it supports, read the information in “Notices” on page 259.

Product Information

This edition applies to version 29, release 0, modification 2 of IBM® SPSS® Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

© Copyright International Business Machines Corporation .
US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Chapter 1. Core features.....1
Power Analysis.....1
Means.....2
Proportions.....15
Correlations.....24
Regression.....32
Power Analysis: Precision.....34
Power Analysis: Grid Values.....35
Meta-Analysis.....35
Meta-Analysis Continuous.....35
Meta-Analysis Continuous Effect Size.....43
Meta-Analysis Binary.....51
Meta-Analysis Binary Effect Size.....59
Meta-Analysis Regression.....67
Meta-Analysis plotting options.....71
Codebook.....77
Codebook Output Tab.....78
Codebook Statistics Tab.....79
Frequencies.....80
Frequencies Statistics.....81
Frequencies Charts.....82
Frequencies Format.....82
Descriptives.....82
Descriptives Options.....83
DESCRIPTIVES Command Additional Features.....84
Percentiles.....84
Percentiles Missing Values.....84
Explore.....85
Explore Statistics.....85
Explore Plots.....86
Explore Options.....87
EXAMINE Command Additional Features.....87
Crosstabs.....87
Crosstabs layers.....88
Crosstabs clustered bar charts.....88
Crosstabs displaying layer variables in table layers.....88
Crosstabs statistics.....89
Crosstabs cell display.....90
Crosstabs table format.....91
Summarize.....91
Summarize Options.....92
Summarize Statistics.....92
Means.....93
Means Options.....94
OLAP Cubes.....96
OLAP Cubes Statistics.....96
OLAP Cubes Differences.....98
OLAP Cubes Title.....98
Proportions.....98
Proportions introduction.....98
One-Sample Proportions.....99

Paired-Samples Proportions....101
Independent-Samples Proportions....104
T Tests....106
T Tests....106
Independent-Samples T Test....107
Paired-Samples T Test....108
One-Sample T Test....109
T TEST Command Additional Features....110
One-Way ANOVA....111
One-Way ANOVA Contrasts....112
One-Way ANOVA Post Hoc Tests....112
One-Way ANOVA Options....114
ONEWAY Command Additional Features....114
GLM Univariate Analysis....115
GLM Model....116
GLM Contrasts....117
GLM Profile Plots....118
GLM Post Hoc Comparisons....120
GLM Save....123
GLM Estimated Marginal Means....123
GLM Options....124
UNIANOVA Command Additional Features....125
Bivariate Correlations....126
Bivariate Correlations Options....127
Bivariate Correlations Confidence Interval....127
CORRELATIONS and NONPAR CORR Command Additional Features....128
Partial Correlations....128
Partial Correlations Options....129
PARTIAL CORR Command Additional Features....129
Distances....129
Distances Dissimilarity Measures....130
Distances Similarity Measures....130
PROXIMITIES Command Additional Features....131
Linear models....131
To obtain a linear model....131
Objectives....131
Basics....132
Model Selection....132
Ensembles....133
Advanced....133
Model Options....134
Model Summary....134
Automatic Data Preparation....134
Predictor Importance....134
Predicted By Observed....135
Residuals....135
Outliers....135
Effects....135
Coefficients....135
Estimated Means....136
Model Building Summary....136
Linear Regression....137
Linear Regression Variable Selection Methods....137
Linear Regression Set Rule....138
Linear Regression Plots....138
Linear Regression: Saving New Variables....139
Linear Regression Statistics....140
Linear Regression Options....141

REGRESSION Command Additional Features....141
Ordinal Regression....142
Ordinal Regression Options....142
Ordinal Regression Output....143
Ordinal Regression Location Model....144
Ordinal Regression Scale Model....144
PLUM Command Additional Features....145
Linear Elastic Net Regression....145
Linear Elastic Net Regression: Options....146
Linear Lasso Regression....148
Linear Lasso Regression: Options....149
Linear Ridge Regression....151
Linear Ridge Regression: Options....152
Curve Estimation....154
Curve Estimation Models....155
Curve Estimation Save....155
Partial Least Squares Regression....155
Model....157
Options....158
Nearest Neighbor Analysis....158
Neighbors....160
Features....160
Partitions....161
Save....162
Output....162
Options....162
Model View....162
Discriminant Analysis....165
Discriminant Analysis Define Range....166
Discriminant Analysis Select Cases....166
Discriminant Analysis Statistics....167
Discriminant Analysis Stepwise Method....167
Discriminant Analysis Classification....168
Discriminant Analysis Save....168
DISCRIMINANT Command Additional Features....169
Factor Analysis....169
Factor Analysis Select Cases....170
Factor Analysis Descriptives....170
Factor Analysis Extraction....170
Factor Analysis Rotation....171
Factor Analysis Scores....172
Factor Analysis Options....172
FACTOR Command Additional Features....172
Choosing a Procedure for Clustering....172
TwoStep Cluster Analysis....173
TwoStep Cluster Analysis Options....174
TwoStep Cluster Analysis Output....175
The Cluster Viewer....176
Hierarchical Cluster Analysis....180
Hierarchical Cluster Analysis Method....181
Hierarchical Cluster Analysis Statistics....181
Hierarchical Cluster Analysis Plots....181
Hierarchical Cluster Analysis Save New Variables....182
CLUSTER Command Syntax Additional Features....182
K-Means Cluster Analysis....182
K-Means Cluster Analysis Efficiency....183
K-Means Cluster Analysis Iterate....183
K-Means Cluster Analysis Save....183

K-Means Cluster Analysis Options....184
QUICK CLUSTER Command Additional Features....184
Nonparametric Tests....184
One-Sample Nonparametric Tests....184
One-Sample Wilcoxon Test....188
Independent-Samples Nonparametric Tests....189
Related-Samples Nonparametric Tests....191
NPTESTS command additional features....194
Legacy Dialogs....194
Multiple Response Analysis....204
Multiple Response Analysis....204
Multiple Response Define Sets....205
Multiple Response Frequencies....205
Multiple Response Crosstabs....206
Reporting Results....208
Reporting Results....208
Report Summaries in Rows....208
Report Summaries in Columns....210
REPORT Command Additional Features....212
Reliability Analysis....212
Reliability Analysis: Statistics....213
RELIABILITY Command Additional Features....216
Weighted Kappa....216
Weighted Kappa: Criteria....217
Weighted Kappa: Print....217
Multidimensional Scaling....218
Multidimensional Scaling Shape of Data....218
Multidimensional Scaling Create Measure....219
Multidimensional Scaling Model....219
Multidimensional Scaling Options....219
ALSCAL Command Additional Features....220
Ratio Statistics....220
Ratio Statistics: Statistics....221
P-P Plot....222
Q-Q Plot....224
ROC Analysis....227
ROC Analysis: Options....228
ROC Analysis: Display....228
ROC Analysis: Define Groups (string)....229
ROC Analysis: Define Groups (numeric)....229
ROC Curve....230
ROC Curve Options....230
Simulation....231
To design a simulation based on a model file....231
To design a simulation based on custom equations....232
To design a simulation without a predictive model....232
To run a simulation from a simulation plan....233
Simulation Builder....234
Run Simulation dialog....243
Working with chart output from Simulation....246
Geospatial Modeling....247
Selecting Maps....247
Data Sources....249
Geospatial Association Rules....250
Spatial Temporal Prediction....254
Finish....257

Notices **259**
Trademarks 260

Index **261**

Chapter 1. Core features

The following core features are included in IBM SPSS Statistics Base Edition.

Power Analysis

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

IBM SPSS Statistics provides the following Power Analysis procedures:

One Sample T-Test
In one-sample analysis, the observed data are collected as a single random sample. It is assumed that the sample data independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the mean parameter.

Paired Sample T-Test
In paired-sample analysis, the observed data contain two paired and correlated samples, and each case has two measurements. It is assumed that the data in each sample independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the difference of the two means.

Independent Sample T-Test
In independent-sample analysis, the observed data contain two independent samples. It is assumed that the data in each sample independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the difference of the two means.

One-way ANOVA
Analysis of variance (ANOVA) is a statistical method of estimating the means of several populations which are often assumed to be normally distributed. The One-way ANOVA, a common type of ANOVA, is an extension of the two-sample *t*-test.

Example. The power of a hypothesis test to detect a correct alternative hypothesis is the probability that the test will reject the test hypothesis. Since the probability of a type II error is the probability of accepting the null hypothesis when the alternative hypothesis is true, the power can be expressed as (1-probability of a type II error), which is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

Statistics and plots. One-sided test, two-sided test, significance level, Type I error rate, test assumptions, population standard deviation, population mean under testing, hypothesized value, two-dimensional power by sample size, two-dimensional power by effect size, three-dimensional power by sample size, three-dimensional power by effect size, rotation degrees, group pairs, Pearson product-moment correlation coefficient, mean difference, plot range of the effect size, pooled population standard deviation, contrasts and pairwise differences, contrast coefficients, contrast test, BONFERRONI, SIDAK, LSD, power by total sample size, two-dimensional power by pooled standard deviation, three-dimensional power by total sample, three-dimensional power by total sample size, pooled standard deviation, Student's t-distribution, non-central t-distribution,

Power Analysis data considerations

Data
In one-sample analysis, the observed data are collected as a single random sample.

In paired-sample analysis, the observed data contain two paired and correlated samples, and each case has two measurements.

In independent-sample analysis, the observed data contain two independent samples.

Analysis of variance (ANOVA) is a statistical method of estimating the means of several populations which are often assumed to be normally distributed.

Assumptions

In one-sample analysis, it is assumed that the sample data independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the mean parameter.

In paired-sample analysis, it is assumed that the data in each sample independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the difference of the two means.

In independent-sample analysis, it is assumed that the data in each sample independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the difference of the two means.

In one-way ANOVA, the statistical method of estimating the means of several populations are often assumed to be normally distributed.

Obtaining a Power Analysis

1. From the menus choose:

   Analyze > **Power Analysis** > **Compare Means** > **One-Sample T-Test**, or **Paired-Sample T-Test**, or **Independent-Sample T-Test**, or **One-way ANOVA**
2. Define the required test assumptions.
3. Click **OK**.

Means

The following statistics features are included in IBM SPSS Statistics Base Edition.

Power Analysis of One-Sample T Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

In one-sample analysis, the observed data are collected as a single random sample. It is assumed that the sample data independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the mean parameter.

1. From the menus choose:

   Analyze > **Power Analysis** > **Means** > **One-Sample T Test**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. When **Power** is selected as the test assumption **Estimate** method, enter an appropriate **Sample size** for power estimation value. The value must be an integer greater than 1.
5. Optionally, select an option from the **Specify** field.

Hypothesized Values
The default setting provides the **Population mean** and **Null value** settings.

Population mean
Enter a value that specifies the population mean under testing. The value must be a single numeric.

Null value
Optionally, enter a value that specifies the null hypothesis value to be tested. The value must be a single numeric.

Effect Size
Cohen's f is used to estimate the effect size as an input to the estimation of the power or sample size. The defined effect size **Value** is passed to the intermediate step in the procedure and calculates the desired power or sample size.

6. Enter a **Population standard deviation** value. The value must be a single numeric greater than 0.
7. Select whether the test is one or two-sided.

Nondirectional (two-sided) analysis
When selected, a two-sided test is used. This is the default setting.

Directional (one-sided) analysis
When selected, power is computed for a one-sided test.

8. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
9. You can optionally click **Plot** to specify "Power Analysis of One-Sample T Test: Plot" on page 3 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

Note: Plot is available only when **Power** is selected as the test assumption.

10. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method, **Hypothesized Values** is selected from the **Specify** list, and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of One-Sample T Test: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power by sample and effect size charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot

Power estimation versus sample size
When enabled, this optional setting provides options for controlling the two-dimensional power by sample size chart. The setting is disabled by default.

Range of sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound
Controls the lower bound for the two-dimensional power by sample size chart. The value must be greater than 1, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Power estimation versus effect size

By default, this optional setting is disabled. When enabled, the chart displays in the output. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size used.

Range of effect size

When selected, the lower and upper bound options are available.

Lower bound

Controls the lower bound for the two-dimensional power by effect size chart. The value must be greater than, or equal to, -5.0 and cannot be greater than the **Upper bound** value.

Upper bound

Controls the upper bound for the two-dimensional power by effect size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5.0.

Three-Dimensional Plot

Power estimation versus

Provides options for controlling the three-dimensional power by sample size (x-axis) and effect size (y-axis) chart, the vertical/horizontal rotation settings, and the user specified plot range of sample/effect size. This setting is disabled by default.

Effect size on x-axis and sample size on y-axis

The optional setting controls the three-dimensional power by sample size (x-axis) and effect size (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Effect size on y-axis and sample size on x-axis

The optional setting controls the three-dimensional power by sample size (y-axis) and effect size (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Vertical rotation

The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation

The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Range of sample size

When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound

Controls the lower bound for the three-dimensional power by sample size chart. The value must be greater than 1, and cannot be greater than the **Upper bound** value.

Upper bound

Controls the upper bound for the three-dimensional power by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Range of effect size

When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size is used.

Lower bound

Controls the lower bound for the three-dimensional power by effect size chart. The value must be greater than, or equal to, -5.0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by effect size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5.0.

Power Analysis of Paired-Samples T Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

In paired-samples analysis, the observed data contain two paired and correlated samples, and each case has two measurements. It is assumed that the data in each sample independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the difference of the two means.

1. From the menus choose:

   Analyze > **Power Analysis** > **Means** > **Paired-Samples T Test**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. When **Power** is selected as the test assumption **Estimate** method, enter the appropriate **Sample size** value. The value must be a positive integer greater than, or equal to, 2.
5. Optionally, select an option from the **Specify** field.

   Hypothesized Values
   The default setting provides the **Mean** and **Standard Deviation** settings.

   Population mean difference
   When a single population mean is required, enter a **Population mean difference** value. When single value is specified, it denotes the population mean difference μ_d.

   Note: The value cannot be 0 when **Sample size** is selected.

   Population mean for group 1 and group 2
   When multiple population means are required for the specified group pairs, enter values for **Population mean for group 1** and **Population mean for group 2**. When multiple values are specified, they denote the population mean difference μ_1 and μ_2.

   Note: The two values cannot be the same when **Sample size** is selected.

   Population standard deviation for mean difference
   When a single population mean is specified, enter the population standard deviation for mean difference value. When a single value is specified, it denotes the population standard deviation of the group difference σ_d. The value must be a single numeric greater than 0.

   Population standard deviation for group 1 and group 2
   When multiple population means are specified, enter the population standard deviation for group 1 and group 2 values. When multiple values are specified, they denote the population standard deviation of the group difference σ_1 and σ_2. The values must be a single numerics greater than 0.

   Pearson product-moment correlation coefficient
   Optionally, enter a value that specifies the Pearson product-moment correlation coefficient ρ. The value must be a single numeric value between -1 and 1. The value cannot be 0.

Note: When a single **Population standard deviation for mean difference** value is specified, this setting is ignored. Otherwise, the values for **Population standard deviation for group 1 and group 2** are used to compute σ_d.

Effect Size
Estimates the effect size as an input to the estimation of the power or sample size. The defined effect size **Value** is passed to the intermediate step in the procedure and calculates the desired power or sample size.

6. Select whether the test is one or two-sided.

   Nondirectional (two-sided) analysis
   When selected, a two-sided test is used. This is the default setting.

   Directional (one-sided) analysis
   When selected, power is computed for a one-sided test.

7. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
8. Optionally, click **Plot** to specify "Power Analysis of Paired-Samples T Test: Plot" on page 6 settings (chart output, two-dimensional plot settings, three-dimensional plot settings, and tooltips).

   Note: Plot is available only when **Power** is selected as the test assumption.

9. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method and **Hypothesized Values** is selected from the **Specify** list.

Power Analysis of Paired-Samples T Test: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power by sample and effect size charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot

Power estimation versus sample size
When enabled, this optional setting provides options for controlling the two-dimensional power by sample size chart. The setting is disabled by default.

Range of sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound
Controls the lower bound for the two-dimensional power by sample size chart. The value must be greater than 1, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Power estimation versus effect size
By default, this optional setting is disabled. When enabled, the chart displays in the output. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size used.

Range of effect size
When selected, the lower and upper bound options are available.

Lower bound
Controls the lower bound for the two-dimensional power by effect size chart. The value must be greater than, or equal to, -5.0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power by effect size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5.0.

Three-Dimensional Plot

Power estimation versus
Provides options for controlling the three-dimensional power by sample size (x-axis) and effect size (y-axis) chart, the vertical/horizontal rotation settings, and the user specified plot range of sample/effect size. This setting is disabled by default.

Effect size on x-axis and sample size on y-axis
The optional setting controls the three-dimensional power by sample size (x-axis) and effect size (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Effect size on y-axis and sample size on x-axis
The optional setting controls the three-dimensional power by sample size (y-axis) and effect size (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Range of sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound
Controls the lower bound for the three-dimensional power by sample size chart. The value must be greater than 1, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Range of effect size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size is used.

Lower bound
Controls the lower bound for the three-dimensional power by effect size chart. The value must be greater than, or equal to, -5.0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by effect size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5.0.

Power Analysis of Independent-Samples T Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low,

there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

In independent-samples analysis, the observed data contain two independent samples. It is assumed that the data in each sample independently and identically follow a normal distribution with a fixed mean and variance, and draws statistical inference about the difference of the two means.

1. From the menus choose:

   Analyze > **Power Analysis** > **Means** > **Independent-Samples T Test**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.

   Enter a **Group size ratio** value for specifying the ratio of the sample sizes (the value must be a single value between 0.01 and 100.
4. When **Power** is selected as the test assumption **Estimate** method, enter values to specify the sample size for **Sample size for group 1 and group 2** for comparison. The values must be an integers greater than 1.
5. Optionally, select an option from the **Specify** field.

   Hypothesized Values
   The default setting provides the **Population mean** and **Population standard deviations** settings.

   Population mean difference
   When a single population mean is specified, enter the population mean difference value. When a single value is specified, it denotes the population of the group mean difference σ_d. The value must be a single numeric greater than 0.

   Population mean for group 1 and group 2
   When multiple population means are specified, enter the population mean for group 1 and group 2 values. When multiple values are specified, they denote the population mean of the group difference σ_1 and σ_2. The values must be a single numerics greater than 0.

   Population standard deviations are
   Specify whether the population standard deviations are **Equal for two groups** or **Not equal for two groups**.

   - When the population standard deviations are equal for two groups, enter a value for **Pooled standard deviation** that denotes σ, and assumes that the two group variances are equal, or $\sigma_1 = \sigma_2 = \sigma$.
   - When the population standard deviations are not equal for two groups, enter values for **Standard deviation for group 1 and group 2** that denote σ_1 and σ_2.

     Note: When the values for **Standard deviation for group 1 and group 2** are identical, they are treated as a single value.

   Effect Size
   Estimates the effect size as an input to the estimation of the power or sample size. The defined effect size **Value** is passed to the intermediate step in the procedure and calculates the desired power or sample size.
6. Select whether the test is one or two-sided.

   Nondirectional (two-sided) analysis
   When selected, a two-sided test is used. This is the default setting.

   Directional (one-sided) analysis
   When selected, power is computed for a one-sided test.
7. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.

8. Optionally, click **Plot** to specify "Power Analysis of Independent-Samples T Test: Plot" on page 9 settings (chart output, two-dimensional plot settings, three-dimensional plot settings, and tooltips).

   Note: Plot is available only when **Power** is selected as the test assumption.

9. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method, **Hypothesized Values** is selected from the **Specify** list, and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of Independent-Samples T Test: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power by sample ratio, effect size, or mean difference charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot

Power estimation versus sample size ratio/Power estimation versus sample size
When enabled, this optional setting provides options for controlling the two-dimensional power by sample size ratio chart. The additional settings are not available when multiple power values have been specified (**Power estimation versus sample size**). The setting is disabled by default.

Range of sample size ratio
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound
Controls the lower bound for the two-dimensional power by sample size ratio chart. The value must be between 0.01 and 100 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power by sample size ratio chart. The value must be between 0.01 and 100 and must be greater than the **Lower bound** value.

Power estimation versus effect size (or mean difference)
By default, this optional setting is disabled. When enabled, the chart displays in the output. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size (or mean difference) is used.

Range of effect size (or mean difference)
When selected, the lower and upper bound options are available.

Lower bound
Controls the lower bound for the two-dimensional power by effect size chart. The value must be greater than, or equal to, -5.0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power by effect size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5.0.

Three-Dimensional Plot

Power estimation versus
Provides options for controlling the three-dimensional power by sample size ratio (x-axis) and effect size (y-axis) chart, the vertical/horizontal rotation settings, and the user specified plot range of sample/effect size. This setting is disabled by default.

Effect size (or mean difference) on x-axis and sample size on y-axis
The optional setting controls the three-dimensional power by sample size ratio (x-axis) and effect size (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Effect size (or mean difference) on y-axis and sample size on x-axis
The optional setting controls the three-dimensional power by sample size (y-axis) and effect size (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Range of sample size ratio
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound
Controls the lower bound for the three-dimensional power by sample size ratio chart. The value must be between 0.01 and 100 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by sample size ratio chart. The value must be between 0.01 and 100 and must be greater than the **Lower bound** value.

Range of effect size (or mean difference)
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size is used.

Lower bound
Controls the lower bound for the three-dimensional power by effect size chart. The value must be greater than, or equal to, -5.0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by effect size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Power Analysis of One-Way ANOVA

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

Analysis of variance (ANOVA) is a statistical method of estimating the means of several populations which are often assumed to be normally distributed. The One-way ANOVA, a common type of ANOVA, is an extension of the two-sample *t*-test. The procedure provides approaches for estimating the power for two types of hypothesis to compare the multiple group means, the overall test, and the test with specified contrasts. The over test focuses on the null hypothesis that all group means are equal. The test with specified contrasts breaks down the overall ANOVA hypotheses into smaller but more describable and useful pieces of the means.

1. From the menus choose:

   Analyze > **Power Analysis** > **Means** > **One-way ANOVA**

2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. Optionally, select an option from the **Specify** field.

   Hypothesized Values
   The default setting provides the **Pooled population standard deviation** setting. Enter a single numeric value greater than 0. The default value is 1.

   Effect Size: Cohen's f
   Cohen's f is used to estimate the effect size as an input to the estimation of the power or sample size. The defined effect size **Value** is passed to the intermediate step in the procedure and calculates the desired power or sample size.

   Effect Size: Eta-squared
   Eta-squared (η^2) is used to estimate the effect size as an input to the estimation of the power or sample size. The defined effect size **Value** is passed to the intermediate step in the procedure and calculates the desired power or sample size.
5. When **Power** is selected as the test assumption **Estimate** setting, specify the **Group sizes** and **Group means** values. At least two group size values must be specified, and each of the group size values must be greater than or equal to 2 (but cannot exceed 2,147, 483, 647). At least two group mean values must also be specified (the number of specified values must equal the group size values).

   When **Sample size** is selected as the test assumption **Estimate** setting, specify the **Group weights** (optional) and **Group means** values. Group weights assign the group size weights when **Power** is selected.

   Note: The **Group weights** settings are ignored when **Group sizes** values are specified.

   Click **Add** to include more **Group sizes**, **Group weights**, and **Group means** values. Click **Delete** to remove existing values.
6. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
7. Optionally, click **Contrast** to specify "Power Analysis of One-way ANOVA: Contrast" on page 11 settings (contrast test, pairwise differences, effect sizes, and estimate the sample size based on confidence intervals).
8. Optionally, click **Plot** to specify "Power Analysis of One-way ANOVA: Plot" on page 12 settings (chart output, two-dimensional plot settings, three-dimensional plot settings, and tooltips).

   Note: Plot is available only when **Group sizes** values are specified and **Power** is selected as the **Estimate** setting.
9. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method.

Power Analysis of One-way ANOVA: Contrast

The **Contrast** dialog provides options for specifying contrast, coefficient, effect size, pairwise differences, and confidence interval half-width settings for the Power Analysis of One-way ANOVA analysis.

Contrast Test

Test with linear contrasts
When enabled, the contrast and coefficient settings are available.

Test direction

Nondirectional (two-sided) analysis
When selected, a two-sided test is used. This is the default setting.

Directional (one-sided) analysis
When selected, power is computed for a one-sided test.

Coefficients
Use the table to specify the contrast coefficients and request the contrast test. The table values are optional. The number of specified values must be equal to the values specified for **Group sizes** and **Group means**. The sum for all specified values must equal 0, otherwise the last value will automatically be adjusted.

Effect size
Specify the effect size of the contrast test that is measured by δ. You must specify a single numeric value. When the **Test Direction** is set to **Directional (one-sided) analysis**, the specified value must be ≥ 0. When **Power** is specified as the **Estimate** test assumption, the specified value cannot be 0.

Pairwise Differences

Estimate the power of testing for pairwise differences
Controls whether or not to estimate the power of testing for the pairwise differences. Be default, the optional setting is disabled, which suppresses output for the pairwise differences.

Bonferroni correction
Uses the Bonferroni correction in estimating the power of pairwise differences. This is the default setting.

Sidak correction
Uses the Sidak correction in estimating the power of pairwise differences.

Least significant difference (LSD)
Uses the LSD correction in estimating the power of pairwise differences.

Specify the half-width of confidence interval
Estimates the sample size based on the confidence interval half-width value. Enter a value in the range 0 - 1. For the One-Sample Binomial test, the value must be in the range 0 - 0.5.

Note: Duplicate values are ignored.

- Click **Add** to add the specified half-width value to the list.
- Highlight an existing half-width value and click **Change** to update the value.
- Highlight an existing half-width value and click **Remove** to remove the value from the list.

Power Analysis of One-way ANOVA: Plot

The **Plot** dialog controls the plots that are output to illustrate the two and three-dimensional power by sample and effect size charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot

Power estimation versus total sample size
When enabled, this optional setting provides options for controlling the two-dimensional power by total sample size chart. The setting is disabled by default.

Range of total sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the total sample size is used.

Lower bound
Controls the lower bound for the two-dimensional power by total sample size chart. The value must be greater than, or equal to:

- 2 x the number of integers specified for **Group sizes**
- 2 x the sum of the integers specified for **Group sizes** / by the smallest integer value for **Group sizes**

The value cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power by total sample size chart. The value must be less than, or equal to:

- 5000 / by the largest integer value specified for **Group sizes** x the sum of the integers specified for **Group sizes**

The value must be greater than the **Lower bound** value and cannot be greater than 2147483647.

Power estimation versus pooled standard deviation
By default, this optional setting is disabled. The setting controls the two-dimensional power by pooled standard deviation chart. When enabled, the chart displays in the output. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the pooled standard deviation is used.

Note:

The plot is disabled when the specified **Group means** values are all the same.

Range of pooled standard deviation
When selected, the lower and upper bound options are available.

Lower bound
Controls the lower bound for the two-dimensional power by pooled standard deviation chart. The value must be greater than 0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power by pooled standard deviation chart. The value must be greater than the **Lower bound** value.

Three-Dimensional Plot

Power estimation versus
Provides options for controlling the three-dimensional power by total sample size (x-axis) and effect size (y-axis) chart, the vertical and horizontal rotation settings, and the user specified plot range of sample and effect size. This setting is disabled by default.

Note:

The plot is disabled when the specified **Group means** values are all the same.

Pooled standard deviation on x-axis and total sample size on y-axis
The optional setting controls the three-dimensional power by total sample size (x-axis) and pooled standard deviation (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Pooled standard deviation on y-axis and total sample size on x-axis
The optional setting controls the three-dimensional power by total sample size (y-axis) and pooled standard deviation (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Range of total sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound
Controls the lower bound for the three-dimensional power by total sample size chart. The value must be greater than 0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by total sample size chart. The value must be greater than the **Lower bound** value.

Range of pooled standard deviation
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size is used.

Lower bound
Controls the lower bound for the three-dimensional power by pooled standard deviation chart. The value must be greater than 0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by pooled standard deviation chart. The value must be greater than the **Lower bound** value.

Power estimation versus sample size
Provides options for controlling the three-dimensional power by effect size and the user specified plot range of sample and effect size. This setting is disabled by default.

Note:

The plot is disabled when the specified **Group means** values are all the same.

Pooled standard deviation on x-axis and total sample size on y-axis
The optional setting controls the three-dimensional power by total sample size (x-axis) and effect size (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Pooled standard deviation on y-axis and total sample size on x-axis
The optional setting controls the three-dimensional power by total sample size (y-axis) and effect size (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Range of total sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the sample size is used.

Lower bound
Controls the lower bound for the three-dimensional power by total sample size chart. The value must be greater than 0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by total sample size chart. The value must be greater than the **Lower bound** value.

Range of effect size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range of the effect size is used.

Lower bound
Controls the lower bound for the three-dimensional power by total sample size chart. The value must be greater than 0 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power by total sample size chart. The value must be greater than the **Lower bound** value.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect

when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Proportions

The following statistics features are included in IBM SPSS Statistics Base Edition.

Power Analysis of One-Sample Binomial Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

Binomial distribution is based on a sequence of Bernoulli trials. It can be used to model experiments, including a fixed number of total trials, which are assumed to be independent of each other. Each trial leads to a dichotomous result, with the same probability for a successful outcome.

The one-sample binomial test makes statistical inference about the proportion parameter by comparing it with a hypothesized value. The methods for estimating the power for such a test are either the normal approximation or the binomial enumeration.

1. From the menus choose:

   Analyze > **Power Analysis** > **Proportions** > **One-Sample Binomial Test**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. When **Power** is selected as the test assumption **Estimate** setting, enter the appropriate **Total number of trials** value. The value must be an integer greater than, or equal to, 1.
5. Enter a value that specifies the alternative hypothesis value of the proportion parameter in the **Population proportion** field. The value must be a single numeric.

   Note: When a **Power** value is specified, the **Population proportion** value cannot be equal to the **Null value**.
6. Optionally, enter a value that specifies the null hypothesis value of the proportion parameter to be tested in the **Null value** field. The value must be a single numeric in the range 0 - 1. The default value is 0.50.
7. Select a method for estimating the power.

   Normal approximation
   Enables normal approximation. This is the default setting.

   Apply continuity correction
   Control whether or not the continuity correction is used for the normal approximation method.

Binomial enumeration
Enables the binomial enumeration method. Optionally, use the **Time limit** field to specify the maximum number of minutes allowed to estimate the sample size. When the time limit is reached, the analysis is terminated and a warning message is displayed. When specified, the value must be a single positive integer to denote the number of minutes. The default setting is 5 minutes.

Note: The selected power estimation assumption has no effect when the **Total number of trials** value exceeds 500.

8. Select whether the test is one or two-sided.

   Nondirectional (two-sided) analysis
   When selected, a two-sided test is used. This is the default setting.

   Directional (one-sided) analysis
   When selected, power is computed for a one-sided test.

9. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
10. Optionally, click **Plot** to specify "Power Analysis of One-Sample Binomial: Plot" on page 16 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

    Note: Plot is available only when **Power** is selected as the test assumption **Estimate** and **Binomial enumeration** is not selected.

11. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of One-Sample Binomial: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power by charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot
Provides options for controlling the two-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus null hypothesis value
When enabled, this optional setting controls the two-dimensional power by null value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus alternative hypothesis value
When enabled, this optional setting controls the two-dimensional power by alternative value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus the difference between hypothesized values
When enabled, this optional setting controls the two-dimensional power by difference between hypothesized values chart. The setting is disabled by default.

Power estimation versus total number of trials
When enabled, this optional setting controls the two-dimensional power by total number of trials chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of total number of trials
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus total number of trials chart. The value must be greater than 0, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus total number of trials chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Three-Dimensional Plot
Provides options for controlling the three-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus total number of trials
When selected, this setting enables the following options.

on x-axis and the difference between hypothesized values on y-axis
The optional setting controls the three-dimensional power by total number of trials (x -axis) and difference between hypothesized values (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and the difference between hypothesized values on x-axis
The optional setting controls the three-dimensional power by total number of trials (y-axis) and difference between hypothesized values (x -axis) chart. By default, the chart is suppressed. When specified, the chart displays.

Plot range of total number of trials
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the three-dimensional power estimation versus total number of trials chart. The value must be greater than 0, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power estimation versus total number of trials chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Power estimation versus null hypothesis value
When selected, this setting enables the following options.

on x-axis and alternative hypothesis value on y-axis
The optional setting controls the three-dimensional power by null (x -axis) and alternative value (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and alternative hypothesis value on x-axis
The optional setting controls the three-dimensional power by null (y-axis) and alternative value (x -axis) chart.. By default, the chart is suppressed. When specified, the chart displays.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Power Analysis of Related-Sample Binomial Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will

be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

Binomial distribution is based on a sequence of Bernoulli trials. It can be used to model experiments, including a fixed number of total trials, which are assumed to be independent of each other. Each trial leads to a dichotomous result, with the same probability for a successful outcome.

The related-sample binomial estimates the power of McNemar's test to compare two proportion parameters based on the matched pair subjects sampled from two related binomial populations.

1. From the menus choose:

   Analyze > **Power Analysis** > **Proportions** > **Related-Sample Binomial Test**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. When **Power** is selected as the test assumption **Estimate** method, enter the appropriate **Total number of pairs** value.
5. Select to specify testing values for either **proportions** or **counts**.
   - When **proportions** is selected, enter values in the **Proportion 1** and **Proportion 2** fields. The values must be in the range 0 - 1.
   - When **counts** is selected, enter values in the **Count 1** and **Count 2** fields. The values must be between 0 and value specified for **Total number of pairs**.

   Proportions notes:

   - **Proportions** is the only available option when a **Power** value is specified.
   - When **Test values are marginal** is not selected: 0 < **Proportion 1** + **Proportion 2** ≤ 1
   - When **Test values are marginal** is selected:
     - **Proportion 1** * **Proportion 2** > 0
     - **Proportion 1** < 1
     - **Proportion 2** < 1
     - The values for **Proportion 1** and **Proportion 2** cannot be the same.

   Counts notes:

   - The **counts** settings are available only when **Power** is selected as the test assumption **Estimate** setting.
   - When **Test values are marginal** is not selected: 0 < **Count 1** + **Count 2** ≤ **Total number of pairs**
   - When **Test values are marginal** is selected:
     - **Count 1** * **Count 2** > 0
     - **Count 1** < **Total number of pairs**
     - **Count 2** < **Total number of pairs**
6. You can optionally select **Test values are marginal** to control whether or not the specified proportions or counts values are marginal. When **Test values are marginal** is enabled, you must specify a **Correlation between matched pairs** value. The value must be a single value between -1 and 1.
7. Select a method for estimating the power.

   Normal approximation
   Enables normal approximation. This is the default setting.

Binomial enumeration
Enables the binomial enumeration method. Optionally, use the **Time limit** field to specify the maximum number of minutes allowed to estimate the sample size. When the time limit is reached, the analysis is terminated and a warning message is displayed. When specified, the value must be a single positive integer to denote the number of minutes. The default setting is 5 minutes.

8. Select whether the test is one or two-sided.

   Nondirectional (two-sided) analysis
   When selected, a two-sided test is used. This is the default setting.

   Directional (one-sided) analysis
   When selected, power is computed for a one-sided test.

9. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
10. Optionally, click **Plot** to specify "Power Analysis of Related-Sample Binomial Test" on page 17 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

    Note: Plot is available only when **Power** is selected as the test assumption **Estimate** and **Binomial enumeration** is not selected.

11. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of Related-Sample Binomial: Plot

The **Plot** dialog controls the plots that are output to illustrate the two and three-dimensional power estimation charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot
Provides options for controlling the two-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus total number of pairs
When enabled, this optional setting controls the two-dimensional power by total number of pairs chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of total number of pairs
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus total number of pairs chart. The value must be greater than 1, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus total number of pairs chart. The value must be greater than the **Lower bound** value and cannot be greater than 2500.

Power estimation versus risk difference
When enabled, this optional setting controls the two-dimensional power by risk difference chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus risk ratio
When enabled, this optional setting control the two-dimensional power by risk ratio chart. The setting is disabled by default.

Plot range of risk ratio
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus risk ratio chart. The value cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus risk ratio chart. The value must be greater than the **Lower bound** value and cannot be greater than 10.

Power estimation versus odds ratio
When enabled, this optional setting controls the two-dimensional power by odds ratio chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of odds ratio
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus odds ratio chart. The value cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus odds ratio chart. The value must be greater than the **Lower bound** value and cannot be greater than 10.

Power estimation versus correlation between matched pairs
Controls the estimated power Ω versus the correlation between the matched pairs chart. The chart is created only when the marginal proportions or counts are specified (rather than the discordant proportions).

Three-Dimensional Plot
Provides options for controlling the three-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus discordant proportions
When enabled, this optional setting controls the three-dimensional power by discordant proportions chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus marginal proportions
When enabled, this optional setting controls the three-dimensional power by marginal proportions chart. The setting is disabled by default. When selected, this setting displays the chart.

Note: This setting is available only when **Test values are marginal** is selected.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Power Analysis of Independent-Sample Binomial Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

The binomial distribution is based on a sequence of Bernoulli trials. It can be used to model those experiments including a fixed number of total trials that are assumed to be independent of each other. Each trial leads to a dichotomous result, with the same probability for a "successful" outcome. The independent-sample binomial test compares two independent proportion parameters.

1. From the menus choose:

   Analyze > **Power Analysis** > **Proportions** > **Independent-Samples Binomial Test**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.

   Optionally, specify a **Group size ratio** value. The default value is 1.
4. When **Power** is selected as the test assumption **Estimate** setting, enter values to specify the **Total number of trials for group 1 and group 2**. The values must be an integers greater than 1.
5. Specify the proportion parameters for the two groups. Both values must be in the range 0 - 1.

   Note: The two values cannot be the same when a **Power** value is specified.
6. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
7. Select the desired test method:

   Chi-squared test
   Estimates the power based on Pearson's chi-squared test. This is the default setting.

   Standard deviation is pooled
   This optional setting controls whether the estimation of the standard deviation is pooled or unpooled. The setting is enabled by default.

   Apply continuity correction
   This optional setting controls whether or not the continuity correction is used. The setting is disabled by default.

   T-test
   Estimates the power based on Student's t-test.

   Standard deviation is pooled
   This optional setting controls whether the estimation of the standard deviation is pooled or unpooled. The setting is enabled by default.

   Likelihood ratio test
   Estimates the power based on the likelihood ratio test.

   Fisher's exact test
   Estimates the power based on Fisher's exact test.

   Notes:

   - In some cases, Fisher's exact test may take an extended amount of time to complete.
   - All plots are blocked when Fisher's exact test is selected.
8. Select a method for estimating the power.

   Normal approximation
   Enables normal approximation. This is the default setting.

Binomial enumeration
Enables the binomial enumeration method. Optionally, use the **Time limit** field to specify the maximum number of minutes allowed to estimate the sample size. When the time limit is reached, the analysis is terminated and a warning message is displayed. When specified, the value must be a single positive integer to denote the number of minutes. The default setting is 5 minutes.

9. Select whether the test is one or two-sided.

   Nondirectional (two-sided) analysis
   When selected, a two-sided test is used. This is the default setting.

   Directional (one-sided) analysis
   When selected, power is computed for a one-sided test.

10. You can optionally click **Plot** to specify "Power Analysis of Independent-Samples Binomial Test: Plot" on page 22 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

    Note: Plot is available only when **Power** is selected as the test assumption **Estimate** and **Binomial enumeration** is not selected.

11. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of Independent-Samples Binomial Test: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power estimation charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot
Provides options for controlling the two-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus group size ratio/Power estimation versus group size
When enabled, this optional setting controls the two-dimensional power by group size ratio chart. When selected, this setting displays the chart. The additional settings are not available when multiple power values have been specified (**Power estimation versus group size**). The setting is disabled by default.

Plot range of group size ratio
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus total number of pairs chart. The value must be greater than .01, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus total number of pairs chart. The value must be greater than the **Lower bound** value and cannot be greater than 100.

Power estimation versus risk difference
When enabled, this optional setting controls the two-dimensional power by risk difference chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus risk ratio
When enabled, this optional setting control the two-dimensional power by risk ratio chart. The setting is disabled by default.

Plot range of risk ratio
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus risk ratio chart. The value cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus risk ratio chart. The value must be greater than the **Lower bound** value and cannot be greater than 10.

Power estimation versus odds ratio
When enabled, this optional setting controls the two-dimensional power by odds ratio chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of odds ratio
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus odds ratio chart. The value cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus odds ratio chart. The value must be greater than the **Lower bound** value and cannot be greater than 10.

Three-Dimensional Plot
Provides options for controlling the three-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus proportions
When selected, this optional setting provides the following power by proportion options:

proportion of group 1 on x-axis and proportion of group 2 on y-axis
Controls the three-dimensional power by proportion of Group 1 (x -axis) and proportion of Group 2 (y-axis) chart. The setting is disabled by default. When selected, this setting displays the chart.

proportion of group 1 on y-axis and proportion of group 2 on x-axis
Controls the three-dimensional power by proportion of Group 2 (x -axis) and proportion of Group 1 (y-axis) chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus group sizes
When selected, this optional setting provides the following power by group sizes options:

size of group 1 on x-axis and size of group 2 on y-axis
Controls the three-dimensional power by number of trials in Group 1 (x -axis) and number of trials in Group 2 (y-axis) chart. The setting is disabled by default. When selected, this setting displays the chart.

size of group 1 on y-axis and size of group 2 on x-axis
Controls the three-dimensional power by number of trials in Group 2 (x -axis) and number of trials in Group 1 (y-axis) chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of size of group 1
When selected, the lower and upper bound options for the group 1 plot range are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus odds ratio chart. The value must be greater than or equal to 2 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus odds ratio chart. The value must be greater than the **Lower bound** value and cannot be greater than 2500.

Plot range of size of group 2
When selected, the lower and upper bound options for the group 2 plot range are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation versus odds ratio chart. The value must be greater than or equal to 2 and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation versus odds ratio chart. The value must be greater than the **Lower bound** value and cannot be greater than 2500.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Correlations

The following statistics features are included in IBM SPSS Statistics Base Edition.

Power Analysis of One-Sample Pearson Correlation Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

Pearson's product-moment correlation coefficient measures the strength of linear association between two scale random variables that are assumed to follow a bivariate normal distribution. By convention, it is a dimensionless quantity and obtained by standardizing the covariance between two continuous variables, thereby ranging between -1 and 1.

The test uses Fisher's asymptotic method to estimate the power for the one-sample Pearson correlation.

1. From the menus choose:

   Analyze > **Power Analysis** > **Correlations** > **Pearson Product-Moment**

2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).

3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.

4. When **Power** is selected as the test assumption **Estimate** setting, enter the appropriate **Sample size in pairs** value. The value must be a single integer greater than 3.
5. Enter a value that specifies the alternative hypothesis value of the correlation parameter in the **Pearson correlation parameter** field. The value must be a single numeric between -1 and 1.

   Note: When **Power** is specified, the **Pearson correlation parameter** value cannot be -1 or 1 and cannot be equal to the **Null value**.

6. Optionally, enter a value that specifies the null hypothesis value of the correlation parameter to be tested in the **Null value** field. The value must be a single numeric between -1 and 1. The default value is 0.

   Note: When **Power** is specified, **Null value** cannot be -1 or 1.

7. Optionally, select **Use bias-correction formiula in the power estimation** to specify whether the bias adjustment is involved or ignored. The setting is enabled by default, which includes the bias adjustment term in the power estimation. When the setting is not selected, the bias adjustment term is ignored.
8. Select whether the test is one or two-sided.

   Nondirectional (two-sided) analysis
   When selected, a two-sided test is used. This is the default setting.

   Directional (one-sided) analysis
   When selected, power is computed for a one-sided test.

9. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
10. You can optionally click **Plot** to specify "Power Analysis of One-Sample Pearson Correlation: Plot" on page 25 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

    Note: Plot is available only when **Power** is selected as the test assumption.

11. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of One-Sample Pearson Correlation: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power by charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot
Provides options for controlling the two-dimensional power estimation versus charts. The settings are disabled by default.

Power estimation versus null hypothesis value
When enabled, this optional setting controls the two-dimensional power by null value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus alternative hypothesis value
When enabled, this optional setting controls the two-dimensional power by alternative value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus the difference between hypothesized values
When enabled, this optional setting controls the two-dimensional power by difference between hypothesized values chart. The setting is disabled by default.

Power estimation versus sample size (in pairs)
When enabled, this optional setting controls the two-dimensional power by sample size chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Three-Dimensional Plot
Provides options for controlling the three-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus sample size
When selected, this setting enables the following options.

on x-axis and the difference between hypothesized values on y-axis
The optional setting controls the three-dimensional power by sample size (x -axis) and difference between hypothesized values (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and the difference between hypothesized values on x-axis
The optional setting controls the three-dimensional power by sample size (y -axis) and difference between hypothesized values (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays

Plot range of sample size (in pairs)
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the three-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Power estimation versus null hypothesis value
When selected, this setting enables the following options.

on x-axis and alternative hypothesis value on y-axis
The optional setting controls the three-dimensional power by null (x -axis) and alternative value (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and alternative hypothesis value on x-axis
The optional setting controls the three-dimensional power by null (y-axis) and alternative value (x -axis) chart.. By default, the chart is suppressed. When specified, the chart displays.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes

effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Power Analysis of One-Sample Spearman Correlation Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

Spearman rank-order correlation coefficient is a rank-based nonparametric statistic to measure the monotonic relationship between two variables that are usually censored and not normally distributed. The Spearman rank-order correlation is equal to the Pearson correlation between the rank values of the two variables, thereby also ranging between -1 and 1. Detecting the power of the Spearman rank correlation test is an important topic in the analysis of hydrological time series data.

The test uses Fisher's asymptotic method to estimate the power for the one-sample Spearman rank-order correlation.

1. From the menus choose:

   Analyze > **Power Analysis** > **Correlations** > **Spearman Rank-Order**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. When **Power** is selected as the test assumption **Estimate** setting, enter the appropriate **Sample size in pairs** value. The value must be a single integer greater than 3.
5. Enter a value that specifies the alternative hypothesis value of the correlation parameter in the **Spearman correlation parameter** field. The value must be a single numeric between -1 and 1.

   Note: When **Power** is specified, the **Spearman correlation parameter** value cannot be -1 or 1 and cannot be equal to the **Null value**.
6. Optionally, enter a value that specifies the null hypothesis value of the correlation parameter to be tested in the **Null value** field. The value must be a single numeric between -1 and 1. The default value is 0.

   Note: When **Power** is specified, **Null value** cannot be -1 or 1.
7. Optionally, select an option that determines how the asymptotic variance is estimated for the power analysis.

   Bonett and Wright
   Estimates the variance suggested by Bonett and Wright. This is the default setting.

   Fieller, Hartley and Pearson
   Estimates the variance suggested by Fieller, Hartley and Pearson.

   Caruso and Cliff
   Estimates the variance suggested by Caruso and Cliff.
8. Select whether the test is one or two-sided.

Nondirectional (two-sided) analysis
When selected, a two-sided test is used. This is the default setting.

Directional (one-sided) analysis
When selected, power is computed for a one-sided test.

9. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
10. Optionally, click **Plot** to specify "Power Analysis of One-Sample Spearman Correlation: Plot" on page 28 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

    Note: Plot is available only when **Power** is selected as the test assumption.
11. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of One-Sample Spearman Correlation: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power by charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot
Provides options for controlling the two-dimensional power estimation versus charts. The settings are disabled by default.

Power estimation versus null hypothesis value
When enabled, this optional setting controls the two-dimensional power by null value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus alternative hypothesis value
When enabled, this optional setting controls the two-dimensional power by alternative value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus the difference between hypothesized values
When enabled, this optional setting controls the two-dimensional power by difference between hypothesized values chart. The setting is disabled by default.

Power estimation versus sample size (in pairs)
When enabled, this optional setting controls the two-dimensional power by sample size chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Three-Dimensional Plot
Provides options for controlling the three-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus sample size
When selected, this setting enables the following options.

on x-axis and the difference between hypothesized values on y-axis
The optional setting controls the three-dimensional power by sample size (x -axis) and difference between hypothesized values (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and the difference between hypothesized values on x-axis
The optional setting controls the three-dimensional power by sample size (y -axis) and difference between hypothesized values (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays

Plot range of sample size (in pairs)
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the three-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Power estimation versus null hypothesis value
When selected, this setting enables the following options.

on x-axis and alternative hypothesis value on y-axis
The optional setting controls the three-dimensional power by null (x -axis) and alternative value (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and alternative hypothesis value on x-axis
The optional setting controls the three-dimensional power by null (y-axis) and alternative value (x -axis) chart.. By default, the chart is suppressed. When specified, the chart displays.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Power Analysis of Partial Pearson Correlation Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

Partial correlation can be explained as the association between two random variables after eliminating the effect of another or several other variables. It is a useful measurement in the presence of confounding. Similar to the Pearson correlation coefficient, partial correlation coefficient is also a dimensionless quantity ranging between -1 and 1.

The test uses Fisher's asymptotic method to estimate the power for the one-sample Pearson correlation.

1. From the menus choose:

   Analyze > **Power Analysis** > **Correlations** > **Partial**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. When **Power** is selected as the test assumption **Estimate** setting, enter the appropriate **Sample size** value. The value must be a single integer greater than 1.
5. Enter a value that specifies the **Number of the variables assumed to be partialled out**. The value must be a single integer greater than or equal to 0.
6. Enter a value that specifies the alternative hypothesis value of the **Partial correlation parameter**. The value must be a single numeric between -1 and 1.

   Note: When **Power** is specified, the **Partial correlation parameter** value cannot be -1 or 1 and cannot be equal to the **Null value**.
7. Optionally, enter a value that specifies the null hypothesis value of the partial correlation parameter to be tested in the **Null value** field. The value must be a single numeric between -1 and 1. The default value is 0.

   Note: When **Power** is specified, **Null value** cannot be -1 or 1.
8. Select whether the test is one or two-sided.

   Nondirectional (two-sided) analysis
   When selected, a two-sided test is used. This is the default setting.

   Directional (one-sided) analysis
   When selected, power is computed for a one-sided test.
9. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
10. Optionally, click **Plot** to specify "Power Analysis of Partial Pearson Correlation: Plot" on page 30 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

    Note: Plot is available only when **Power** is selected as the test assumption.
11. Optionally, click **Precision** to estimate the sample size based on confidence intervals by specifying the values of the confidence interval half-widths. For more information, see "Power Analysis: Precision" on page 34.

Note: Precision is available only when **Sample size** is selected as the test assumption **Estimate** method and **Non-directional (two-sided)** analysis is selected as the **Test Direction**.

Power Analysis of Partial Pearson Correlation: Plot

The **Plot** dialog provides options for controlling the plots that are output to illustrate the two and three-dimensional power by charts. The dialog also controls the vertical and horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot
Provides options for controlling the two-dimensional power estimation versus charts. The settings are disabled by default.

Power estimation versus null hypothesis value
When enabled, this optional setting controls the two-dimensional power by null value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus alternative hypothesis value
When enabled, this optional setting controls the two-dimensional power by alternative value chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus the number of variables partialled out
When enabled, this optional setting controls the two-dimensional power by number of partialled-out variables chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus the difference between hypothesized values
When enabled, this optional setting controls the two-dimensional power by difference between hypothesized values chart. The setting is disabled by default.

Power estimation versus sample size
When enabled, this optional setting controls the two-dimensional power by sample size chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Three-Dimensional Plot
Provides options for controlling the three-dimensional power estimation versus charts. This setting is disabled by default.

Power estimation versus sample size
When selected, this setting enables the following options.

on x-axis and the difference between hypothesized values on y-axis
The optional setting controls the three-dimensional power by sample size (x -axis) and difference between hypothesized values (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and the difference between hypothesized values on x-axis
The optional setting controls the three-dimensional power by sample size (y -axis) and difference between hypothesized values (x-axis) chart. By default, the chart is suppressed. When specified, the chart displays

Plot range of sample size (in pairs)
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the three-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the three-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Power estimation versus null hypothesis value
When selected, this setting enables the following options.

on x-axis and alternative hypothesis value on y-axis
The optional setting controls the three-dimensional power by null (x -axis) and alternative value (y-axis) chart. By default, the chart is suppressed. When specified, the chart displays.

on y-axis and alternative hypothesis value on x-axis
The optional setting controls the three-dimensional power by null (y-axis) and alternative value (x -axis) chart.. By default, the chart is suppressed. When specified, the chart displays.

Vertical rotation
The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation
The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Regression

The following statistics features are included in IBM SPSS Statistics Base Edition.

Power Analysis of Univariate Linear Regression Test

This feature requires IBM SPSS Statistics Base Edition.

Power analysis plays a pivotal role in a study plan, design, and conduction. The calculation of power is usually before any sample data have been collected, except possibly from a small pilot study. The precise estimation of the power may tell investigators how likely it is that a statistically significant difference will be detected based on a finite sample size under a true alternative hypothesis. If the power is too low, there is little chance of detecting a significant difference, and non-significant results are likely even if real differences truly exist.

Univariate linear regression is a basic and standard statistical approach in which researchers use the values of several variables to explain or predict values of a scale outcome.

The Power Analysis of Univariate Linear Regression test estimates the power of the type III *F*-test in univariate multiple linear regression models. With the effect size represented by multiple (partial) correlations, approaches for both fixed and random predictors are provided. For fixed predictors, the power estimation is based on the non-central *F*-distribution. For random predictors, it is assumed that the target variable and the predictors jointly follow a multivariate normal distribution. In which case, power estimation is based on the distribution of the sample multiple correlation coefficient.

1. From the menus choose:

   Analyze > **Power Analysis** > **Regression** > **Univariate Linear**
2. Select a test assumption **Estimate** setting (**Sample size** or **Power**).
3. When **Sample size** is selected, enter either a **Single power value** for sample size estimation value (the value must be a single value between 0 and 1), or select **Grid power values** and then click **Grid** to view projected sample sizes for a range of specific Power values.

   For more information, see "Power Analysis: Grid Values" on page 35.
4. When **Estimate power** is selected, enter an appropriate **Sample size** for power estimation value. The value must be a single integer greater than or equal to the total number of model predictors +2 (when **Include the intercept term in the model** is enabled). Otherwise the value must be a single integer greater than or equal to the total number of model predictors +1.
5. Specify the value of the multiple partial correlation coefficient in the **Population multiple partial correlation** field. The value must be a single value between -1 and 1.

   Note: When a **Power** value is specified, the **Population multiple partial correlation** value cannot be 0.

   The following settings are enabled when **Population multiple partial correlation** is selected:

   Total number of predictors in the model
   Specify the number of either the total predictors, or the predictors in the full model (not including the intercept, if applicable). The value must be a single integer greater than or equal to 1.

Number of test predictors
Specify the number of either the test predictors, or the predictors in the nested model (not including the intercept, if applicable). The value must be greater than or equal to 1, but no larger than the **Total number of predictors in the model** value.

6. Specify **R-squared values for** multiple correlation coefficients for both **Full model** and **Nested model**. The values must be a single values between 0 and 1.

   Note: When a **Power** value is specified, the **Full model** value must be greater than the **Nested model** value.

   The following settings are enabled when **R-squared values for** is selected:

   Total number of predictors - Full model
   Specify the number of total predictors for the full model (not including the intercept, if applicable). The value must be a single integer greater than or equal to 1.

   Total number of predictors - Nested model
   Specify the number of total predictors for the nested model (not including the intercept, if applicable). The value must be greater than or equal to 1, but less than the **Total number of predictors - Full model** value.

7. Optionally, specify the significance level of the Type I error rate for the test in the **Significance level** field. The value must be a single double value between 0 and 1. The default value is 0.05.
8. You can optionally select the **Include the intercept term in the model** setting. The setting is enabled by default. When not selected, the intercept term is excluded from the power analysis.
9. You can optionally select whether model predictors are **Fixed** or **Random**. **Fixed** is the default setting.
10. You can optionally click **Plot** to specify "Power Analysis of Univariate Linear Regression: Plot" on page 33 settings (chart output, two-dimensional plot settings, and three-dimensional plot settings).

    Note: Plot is available only when **Power** is selected as the test assumption.

Power Analysis of Univariate Linear Regression: Plot

You can control the plots that are output to illustrate the two and three-dimensional power by charts. You can also control the display of tool tips and the vertical/horizontal rotation degrees for three-dimensional charts.

Two-Dimensional Plot
Provides options for controlling the two-dimensional power estimation versus charts. The settings are disabled by default.

Power estimation versus the multiple partial correlation
When enables, this optional setting controls the two-dimensional power by multiple partial correlation coefficient chart. The setting is disabled by default. When selected, this setting displays the chart.

Power estimation versus sample size
When enabled, this optional setting controls the two-dimensional power by sample size chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of sample size
When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound
Controls the lower bound for the two-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound
Controls the upper bound for the two-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Three-Dimensional Plot

Provides options for controlling the three-dimensional power estimation versus charts, the vertical/horizontal rotation settings, and the user specified plot range of sample size. This setting is disabled by default.

Power estimation versus sample size

When enabled, this optional setting controls the three-dimensional power by sample size charts. The setting is disabled by default. When selected, this setting displays the chart.

on x-axis and the multiple partial correlation on y-axis

This optional setting controls the three-dimensional power by sample size (x-axis) and multiple partial correlation coefficient (y-axis) chart. The setting is disabled by default. When selected, this setting displays the chart.

on y-axis and the multiple partial correlation on x-axis

This optional setting controls the three-dimensional power by sample size (y-axis) and multiple partial correlation coefficient (x-axis) chart. The setting is disabled by default. When selected, this setting displays the chart.

Plot range of sample size

When selected, the lower and upper bound options are available. When no integer values are specified for the **Lower bound** or **Upper bound** fields, the default plot range is used.

Lower bound

Controls the lower bound for the two-dimensional power estimation by sample size chart. The value must be greater than or equal to 4, and cannot be greater than the **Upper bound** value.

Upper bound

Controls the upper bound for the two-dimensional power estimation by sample size chart. The value must be greater than the **Lower bound** value and cannot be greater than 5000.

Vertical rotation

The optional setting sets the vertical rotation degrees (clockwise from the left) for the three-dimensional chart. You can use the mouse to rotate the chart vertically. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 10.

Horizontal rotation

The optional setting sets the horizontal rotation degrees (clockwise from the front) for the three-dimensional chart. You can use the mouse to rotate the chart horizontally. The setting takes effect when the three-dimensional plot is requested. The value must be a single integer value less than or equal to 359. The default value is 325.

Power Analysis: Precision

The **Precision** dialog is available for every Power Analysis procedure (except for Univariate Linear Regression), and provides options for estimating the sample size based on specified confidence interval half-widths. The dialog is available when **Single power value** or **Grid power values** are specified for the **Sample size** estimate and when the **Test Direction** setting is set to **Nondirectional (two-sided analysis)**.

Note: The **Test Direction** setting does not apply to One-Way ANOVA.

when the **Estimate sample size** and **Grid power values** options are selected (click the **Grid** control to display the dialog).

Confidence Interval Type(s)

Select the appropriate confidence intervals types and whether continuity correction is applied.

Note: Confidence Interval Type(s) are available only for the Power Analysis proportion procedures (One-Sample Proportion, Related-Samples Binomial Test, and Independent-Samples Binomial Test).

Specify the half-width of confidence interval

Estimates the sample size based on the confidence interval half-width value. Enter a value in the range 0 - 1. For the One-Sample Binomial test, the value must be in the range 0 - 0.5.

Note: Duplicate values are ignored.

- Click **Add** to add the specified half-width value to the list.
- Highlight an existing half-width value and click **Change** to update the value.
- Highlight an existing half-width value and click **Remove** to remove the value from the list.

Power Analysis: Grid Values

The **Grid Values** dialog provides options for specifying a range of **POWER** values for the purpose of viewing projected sample sizes in a grid format for every specified **POWER** range value.

The **Grid Values** dialog is available for every Power Analysis procedure when the **Estimate sample size** and **Grid power values** options are selected (click the **Grid** control to display the dialog).

Specify single power

When selected, at least one value is required to run the analysis. Multiple values are allowed and each value must belong to `[0, 1]`. You can specify multiple values with each value separated by a blank space (or blank spaces). Use the **Add**, **Change**, and **Remove** controls to work with the values in the power value list.

All values must be unique (duplicate values are not allowed).

Specify power range

When selected, a range of power values can be specified from a **Start** value (`value1`) to an **End** value (`value2`) with the increment of **By** (`value3`). If specified, only one valid set of `[value1 TO value2 BY value3]` is allowed. It must satisfy that `0 ≤ value1 ≤ value2 ≤ 1`. In cases where `value1 = value2`, it is equivalent to specifying a single `value1`, regardless of `value3`.

Note: The **Specify single power** and **Specify power range** options are independent; you can select a single option or both options.

Meta-Analysis

Meta analysis is the analysis of the data obtained from a collection of studies that answer similar research questions. These studies are known as primary studies. Meta analysis uses statistical methods to produce an overall estimate of an effect, explore between-study heterogeneity, and investigate the impact of publication bias or, more generally, small-study effects on the final results.

IBM SPSS Statistics supports standard effect sizes and generic (pre-calculated) effect sizes for both binary data (such as the log odds-ratio) and for continuous data (such as Hedges's g). Meta analysis information (such as the study-specific effect sizes and their corresponding standard errors and the meta-analysis model and method) is specified during the meta analysis declaration step. The information is automatically used by all subsequent meta analysis.

Random effects, common effect, and fixed effects meta analysis models are supported. Depending on the chosen meta analysis model, the various estimation methods (for example inverse-variance and Mantel–Haenszel) are available for the common effect and fixed effects models. Several different estimators are also available for the between study variance parameter for the random effects model.

Meta-Analysis Continuous

The Meta-Analysis Continuous procedure performs meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Example

Several research studies were conducted in history to investigate a faddish but debatable medicine to help treat type II diabetes. The oral medicine was claimed to be able to reduce the blood glucose level after meals. Data were collected from different research sites from 1979 to 1986.

A principal investigator would like to draw statistical inference about the effect of the oral medicine. Due to the fact that the data were generated from different studies, she proposed the idea of synthesizing the results across the studies to reach an overall understanding of the effect and to identify those underlying sources of variation in the outcomes.

Statistics

Confidence level, iterative method, step-halving, convergence tolerance, sample means, sample variance, standard deviation, estimated effect size, Cohen's *d*, Hedges' *g*, Glass's Delta, mean difference, cumulative analysis, estimation method, trim-and-fill, regression-based test, random-effects model, fixed-effects model, restricted maximum likelihood estimator, empirical Bayes estimator, Hedges estimator, Hunter-Schmidt estimator, DerSimonian-Laird estimator, Sidik-Jonkman estimator, Knapp-Hartung standard-error adjustment, truncated Knapp-Hartung standard-error adjustment, coefficients, EGGER's regression-based test, intercept, multiplicative model, multiplicative dispersion parameter, quadratic estimator, homogeneity test, heterogeneity measures, prediction interval, estimated standard error, estimated *p*-value, cumulative overall effect size, estimated study weight.

Obtaining a Meta-Analysis Continuous analysis

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. Under the **Treatment Group** section, select a **Study Size** variable to represent the sample size for the treatment group. The selected variable must be numeric (string variables are not supported).
3. Select a **Mean** variable to represent the sample means for the treatment group. The selected variable must be numeric (string variables are not supported).
4. Select either **Standard deviation** to specify the sample standard deviation, or **Variance** to specify the sample variance and then select a variable to represent the standard deviation/variance for the treatment group.
5. Under the **Control Group** section, select a **Study Size** variable to represent the sample size for the control group. The selected variable must be numeric (string variables are not supported).
6. Select a **Mean** variable to represent the sample means for the control group. The selected variable must be numeric (string variables are not supported).
7. Select either **Standard deviation** to specify the sample standard deviation, or **Variance** to specify the sample variance and then select a variable to represent the standard deviation/variance for the control group.
8. Optionally, select **Study ID** and/or **Study Label** variables. The selected **Study ID** variable cannot be the same as the selected **Study Label** variable.
9. Optionally, select an **Effect Size** setting.

   Cohen's d
   The default setting estimates the Cohen's *d*. When **Adjusted standard error** is selected, the setting estimates the Cohen's *d* and its variance using an alternative formula divided by $2(N_{treatment} + N_{control} - 2)$.

   Hedges' g
   Estimates the Hedges' *g*. When **Adjusted standard error** is selected, the setting estimates the Hedges' *g* and its variance using an alternative formula divided by $2(N_{treatment} + N_{control} - 3.94)$.

   Glass's Delta
   Estimates the Glass's Delta based on the control group. When **Standardized based on treatment group** is selected, Glass's Delta is standardized based on the standard deviation of the treatment group.

   Unstandardized Mean Difference
   Estimates the mean difference by assuming the two population standard deviations are equal. When **Unequal group variances** is selected the mean difference is estimated by assuming the two population standard deviations are not equal.
10. Optionally, select a **Model** setting. When **Trim-and-Fill** settings are enabled, the setting also controls the model that is used by pooling in the trim-and-fill analysis. When **Bias** settings are enabled, the setting also controls the model that is used by the regression-based test.

Random-effects
The default setting builds the random-effects model.

Fixed-effects
Builds the fixed-effects model.

11. Optionally, you can:
    - Click **Criteria...** to specify the general criteria.
    - Click **Analysis** to specify the subgroup and cumulative analysis.
    - Click **Inference** to specify the estimation methods.
    - Click **Contrast** to control the contrast test.
    - Click **Bias** to access the publication bias by conducting the EGGER's regression-based test.
    - Click **Trim-and-Fill** to implement the trim-and-fill analysis of publication bias.
    - Click **Print** to control the table outputs.
    - Click **Save** to save the estimated statistics to the active data set.
    - Click **Plot** to specify which plots to include in the output.
12. Click **OK**.

Meta-Analysis Continuous: Criteria

The **Criteria** dialog provides settings for specifying criteria for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Confidence Interval
The optional setting specifies the confidence level. The value must be a numeric value between 0 and 100. The default setting is 95.

Missing Data Scope
The optional settings control how the procedure handles missing data.

Exclude cases analysis by analysis
The default setting includes all cases with sufficient data on the variables used in each particular analysis.

Exclude cases listwise
Includes all cases with sufficient data on all variables used across all analyses specified by the procedure.

User Missing Values
The optional settings control how user missing values are treated.

Exclude
The default setting treats user missing values as valid.

Include
Ignores user missing value designations and treats user missing values as valid.

Iterations

Maximum iterations
The optional setting specifies the maximum number of iterations in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 100. A value of 0 means that no iterations are performed.

Maximum step-halvings
The optional setting specifies the maximum step-halving in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 5. A value of 0 means that step-halving is not applied.

Convergence
The optional setting specifies the convergence tolerance. The value must be a single positive value. The setting is available when iterative methods are used. The default value is 1E-6.

Defining Meta-Analysis Continuous criteria

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Continuous** dialog, click **Criteria**.
3. Select and define the appropriate criteria settings.
4. Click **Continue**.

Meta-Analysis Continuous: Analysis

The **Analysis** dialog provides settings for specifying subgroup and cumulative analysis for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Subgroup Analysis
Select a variable that evokes the subgroup analysis. The variable cannot be the same as the variable defined for **Cumulative Analysis**.

Cumulative Analysis
Select a variable that evokes the cumulative analysis and upon which the cumulative meta-analysis is conducted. The variable cannot be the same as the variable defined for **Subgroup Analysis**. When **Ascending** is selected, cumulative analysis is based upon the specified variable in ascending order. When **Descending** is selected, cumulative analysis is based upon the specified variable in descending order.

Cumulative Statistics
Provides options for saving the estimated cumulative overall effect size. The settings are available only when a **Cumulative Analysis** variable is selected.

Cumulative effect size
Saves the estimated cumulative overall effect size.

Standard error
Saves the estimated standard error of the cumulative overall effect size.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the cumulative overall effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the cumulative overall effect size.

P-value
Saves the estimated *p*-value of the cumulative overall effect size.

Destination
Provides options for specifying a save dataset or data file. When **Dataset** is selected, you can specify a new dataset name (you can retain the default dataset name). When **Data file** is selected, click **Browse...** to select a save file name and location.

Defining Meta-Analysis Continuous analysis settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Continuous** dialog, click **Analysis**.
3. Select and define the appropriate analysis settings.
4. Click **Continue**.

Meta-Analysis Continuous: Inference

The **Inference** dialog provides settings for specifying the estimation methods for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Note: The **Inference** dialog is available only when a **Random-effects** model is selected.

Estimator
Provides settings for specifying the estimator.

Notes:

- When Trim and Fill settings are specified, also controls the estimator that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the estimator that is used by the regression-based test.

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculates the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings that control whether to apply the Knapp-Hartung standard-error adjustment.

Notes:

- When Trim and Fill settings are specified, also controls the standard error adjustment that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the standard error adjustment that is used by the regression-based test.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Continuous inference settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Continuous** dialog, click **Inference**.
3. Select and define the appropriate inference settings.

4. Click **Continue**.

Meta-Analysis Continuous: Contrast

The **Contrast** dialog provides settings for controlling the contrast test for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Variables
The list display all available dataset variables. Select variables from the list and move them to the **Contrast(s)** list.

Contrast(s)
The list identifies coefficients that are stored as variables in the active data set. Multiple variables are allowed. String variables are not supported.

User-Input Coefficient Values
Provides settings for specifying coefficients of a user-supplied contrast. Only numeric values are allowed. To formulate a valid contrast test, the number of specified values must match that of the valid studies.

Defining Meta-Analysis Continuous contrast settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Continuous** dialog, click **Contrast**.
3. Select and define the appropriate contrast settings.
4. Click **Continue**.

Meta-Analysis Continuous: Bias

The **Bias** dialog provides settings for enabling the publication bias by conducting the Eggers's regression-based test for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Egger's regression-based test
Selecting this setting enables the publication bias by conducting the Eggers's regression-based test.

Variables
The list provides all available dataset variables.

Covariate(s)
Variables selected from the **Variables** list are treated as covariates. Multiple covariates are allowed.

Factor(s)
Variables selected from the **Variables** list are treated as factors. Multiple factors are allowed.

Include intercept in regression
Controls the intercept term in the regression-based test.

Include dispersion parameter in fixed-effects model
Controls the multiplicative model setting and introduces the multiplicative dispersion parameter to the analysis. The setting is available only when a fixed-effects model is selected.

Estimate statistics based on t-distribution
Controls the distribution used in the regression-based tests. The setting is enabled by default, which estimates the statistics based on the *t*-distribution. When the setting is not selected, the statistics are estimated based on the normal distribution.

Defining Meta-Analysis Continuous bias settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**

2. In the **Meta-Analysis Continuous** dialog, click **Bias**.
3. Select and define the appropriate bias settings.
4. Click **Continue**.

Meta-Analysis Continuous: Trim-and-Fill

The **Trim-and-Fill** dialog provides settings for implementing the trim-and-fill analysis of publication bias for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Estimate number of missing studies
Controls the trim-and-fill analysis of publication bias. Selecting this setting enables the other dialog settings.

Side to Impute Studies
Provides options for specifying the side of the funnel plot on which the missing studies are imputed.

Determined by the slopes of Egger's test
The default setting determines the side based on the estimated slope of the Egger’s test.

Left
Imputes the left side of the funnel plot.

Right
Imputes the right side of the funnel plot.

Method
Specifies the method to estimate the number of missing studies.

Linear
The default setting calculates the linear estimator.

Run
Calculates the run estimator.

Quadratic
Calculates the quadratic estimator.

Iteration process
Provides settings for specifying the iteration estimator and standard error adjustment.

Fixed-effects model
When selected, a fixed-effects model is employed and the iteration estimation and standard error adjustment options are not available.

Random-effects model
When selected, a random effects model is employed and the following settings are available.

Estimator
Provides settings for specifying the iteration estimator

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculate the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings for controlling whether to apply the Knapp-Hartung standard-error adjustment to the iterations of the trim-and-fill algorithm.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Continuous trim-and-fill settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Continuous** dialog, click **Trim-and-Fill**.
3. Select and define the appropriate trim-and-fill settings.
4. Click **Continue**.

Meta-Analysis Continuous: Print

The **Print** dialog provides settings for controlling table outputs for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Homogeneity/Heterogeneity
Provides settings for controlling the homogeneity and heterogeneity tests.

Test of homogeneity
When selected, the corresponding test of homogeneity is provided in the output.

Heterogeneity measures
When selected, the heterogeneity measures are provided in the output.

Effect Sizes
Provides the following effect size settings.

Individual studies
Controls the display of individual studies. When selected, the corresponding output is provided.

Cumulative effect sizes
Controls the display of cumulative analysis. When selected, the corresponding output is provided. The setting is available only when a **Cumulative Analysis** variable is selected on the **Analysis** dialog.

Prediction interval under random-effects model
Controls the display of the prediction interval. When selected, the corresponding output is provided. The setting is available only when a random-effects model is specified.

Defining Meta-Analysis Continuous print settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Continuous** dialog, click **Print**.
3. Select and define the appropriate print settings.
4. Click **Continue**.

Meta-Analysis Continuous: Save

The **Save** dialog provides settings for saving the estimated statistics to the active data set for meta-analysis with continuous outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Individual Studies
Provides options for saving the estimated effect size.

Individual effect size
Saves the estimated effect size.

Standard error
Saves the estimated standard error of the effect size.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the effect size.

P-value
Saves the estimated *p*-value of the effect size.

Study weight
Saves the estimated study weight.

Percentage of study weight
Saves the normalized study weight as a percentage.

Defining Meta-Analysis Continuous save settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Continuous** dialog, click **Save**.
3. Select and define the appropriate estimated statistics save settings.
4. Click **Continue**.

Meta-Analysis Continuous: Plot

The **Plot** dialog provides settings for the following plot types:

- "Forest Plot" on page 71
- "Cumulative Forest Plot" on page 73
- "Bubble Plot" on page 74
- "Funnel Plot" on page 75
- "Galbraith Plot" on page 76

Meta-Analysis Continuous Effect Size

The Meta-Analysis Continuous Effect Size procedure performs meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Example
Several research studies were conducted in history to investigate a faddish but debatable medicine to help treat type II diabetes. The oral medicine was claimed to be able to reduce the blood glucose level after meals. Data were collected from different research sites from 1979 to 1986.

A principal investigator would like to draw statistical inference about the effect of the oral medicine. Due to the fact that the data were generated from different studies, she proposed the idea of synthesizing the results across the studies to reach an overall understanding of the effect and to identify those underlying sources of variation in the outcomes.

Statistics

Confidence level, iterative method, step-halving, convergence tolerance, sample means, sample variance, standard deviation, estimated effect size, Cohen's *d*, Hedges' *g*, Glass's Delta, mean difference, cumulative analysis, estimation method, trim-and-fill, regression-based test, random-effects model, fixed-effects model, restricted maximum likelihood estimator, empirical Bayes estimator, Hedges estimator, Hunter-Schmidt estimator, DerSimonian-Laird estimator, Sidik-Jonkman estimator, Knapp-Hartung standard-error adjustment, truncated Knapp-Hartung standard-error adjustment, coefficients, EGGER's regression-based test, intercept, multiplicative model, multiplicative dispersion parameter, quadratic estimator, homogeneity test, heterogeneity measures, prediction interval, estimated standard error, estimated *p*-value, cumulative overall effect size, estimated study weight,

Obtaining a Meta-Analysis Continuous Effect Size analysis

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. Select an **Effect Size** variable that denotes the effect size. The selected variable must be numeric (string variables are not supported).
3. Select either **Standard deviation** to specify the effect size standard deviation, or **Variance** to specify the effect size variance and then select a variable to represent the standard deviation/variance for the effect size.
4. Optionally, select **Study ID** and/or **Study Label** variables. The selected **Study ID**variable cannot be the same as the selected **Study Label** variable.
5. Optionally, select a **Model** setting. When **Trim-and-Fill** settings are enabled, the setting also controls the model that is used by pooling in the trim-and-fill analysis. When **Bias** settings are enabled, the setting also controls the model that is used by the regression-based test.

   Random-effects
   The default setting builds the random-effects model.

   Fixed-effects
   Builds the fixed-effects model.
6. Optionally, you can:
   - Click **Criteria...** to specify the general criteria.
   - Click **Analysis** to specify the subgroup and cumulative analysis.
   - Click **Inference** to specify the estimation methods.
   - Click **Contrast** to control the contrast test.
   - Click **Bias** to access the publication bias by conducting the EGGER's regression-based test.
   - Click **Trim-and-Fill** to implement the trim-and-fill analysis of publication bias.
   - Click **Print** to control the table outputs.
   - Click **Save** to save the estimated statistics to the active data set.
   - Click **Plot** to specify which plots to include in the output.
7. Click **OK**.

Meta-Analysis Continuous Effect Size: Criteria

The **Criteria** dialog provides settings for specifying criteria for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Confidence Interval

The optional setting specifies the confidence level. The value must be a numeric value between 0 and 100. The default setting is 95.

Missing Data Scope

The optional settings control how the procedure handles missing data.

Exclude cases analysis by analysis
The default setting includes all cases with sufficient data on the variables used in each particular analysis.

Exclude cases listwise
Includes all cases with sufficient data on all variables used across all analyses specified by the procedure.

User Missing Values
The optional settings control how user missing values are treated.

Exclude
The default setting treats user missing values as valid.

Include
Ignores user missing value designations and treats user missing values as valid.

Iterations

Maximum iterations
The optional setting specifies the maximum number of iterations in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 100. A value of 0 means that no iterations are performed.

Maximum step-halvings
The optional setting specifies the maximum step-halving in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 5. A value of 0 means that step-halving is not applied.

Convergence
The optional setting specifies the convergence tolerance. The value must be a single positive value. The setting is available when iterative methods are used. The default value is 1E-6.

Defining Meta-Analysis Continuous Effect Size criteria

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Criteria**.
3. Select and define the appropriate criteria settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Analysis

The **Analysis** dialog provides settings for specifying subgroup and cumulative analysis for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Subgroup Analysis
Select a variable that evokes the subgroup analysis. The variable cannot be the same as the variable defined for **Cumulative Analysis**.

Cumulative Analysis
Select a variable that evokes the cumulative analysis and upon which the cumulative meta-analysis is conducted. The variable cannot be the same as the variable defined for **Subgroup Analysis**. When **Ascending** is selected, cumulative analysis is based upon the specified variable in ascending order. When **Descending** is selected, cumulative analysis is based upon the specified variable in descending order.

Cumulative Statistics
Provides options for saving the estimated cumulative overall effect size. The settings are available only when a **Cumulative Analysis** variable is selected.

Cumulative effect size
Saves the estimated cumulative overall effect size.

Standard error
Saves the estimated standard error of the cumulative overall effect size.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the cumulative overall effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the cumulative overall effect size.

P-value
Saves the estimated *p*-value of the cumulative overall effect size.

Destination
Provides options for specifying a save dataset or data file. When **Dataset** is selected, you can specify a new dataset name (you can retain the default dataset name). When **Data file** is selected, click **Browse...** to select a save file name and location.

Defining Meta-Analysis Continuous Effect Size analysis settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Analysis**.
3. Select and define the appropriate analysis settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Inference

The **Inference** dialog provides settings for specifying the estimation methods for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Note: The **Inference** dialog is available only when a **Random-effects** model is selected.

Estimator
Provides settings for specifying the estimator.

Notes:

- When Trim and Fill settings are specified, also controls the estimator that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the estimator that is used by the regression-based test.

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculates the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings that control whether to apply the Knapp-Hartung standard-error adjustment.

Notes:

- When Trim and Fill settings are specified, also controls the standard error adjustment that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the standard error adjustment that is used by the regression-based test.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Continuous Effect Size inference settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Inference**.
3. Select and define the appropriate inference settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Contrast

The **Contrast** dialog provides settings for controlling the contrast test for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Variables
The list display all available dataset variables. Select variables from the list and move them to the **Contrast(s)** list.

Contrast(s)
The list identifies coefficients that are stored as variables in the active data set. Multiple variables are allowed. String variables are not supported.

User-Input Coefficient Values
Provides settings for specifying coefficients of a user-supplied contrast. Only numeric values are allowed. To formulate a valid contrast test, the number of specified values must match that of the valid studies.

Defining Meta-Analysis Continuous Effect Size contrast settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Contrast**.
3. Select and define the appropriate contrast settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Bias

The **Bias** dialog provides settings for enabling the publication bias by conducting the Eggers's regression-based test for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Egger's regression-based test
Selecting this setting enables the publication bias by conducting the Eggers's regression-based test.

Variables
The list provides all available dataset variables.

Covariate(s)
Variables selected from the **Variables** list are treated as covariates. Multiple covariates are allowed.

Factor(s)
Variables selected from the **Variables** list are treated as factors. Multiple factors are allowed.

Include intercept in regression
Controls the intercept term in the regression-based test.

Include dispersion parameter in fixed-effects model
Controls the multiplicative model setting and introduces the multiplicative dispersion parameter to the analysis. The setting is available only when a fixed-effects model is selected.

Estimate statistics based on t-distribution
Controls the distribution used in the regression-based tests. The setting is enabled by default, which estimates the statistics based on the *t*-distribution. When the setting is not selected, the statistics are estimated based on the normal distribution.

Defining Meta-Analysis Continuous Effect Size bias settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Bias**.
3. Select and define the appropriate bias settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Trim-and-Fill

The **Trim-and-Fill** dialog provides settings for implementing the trim-and-fill analysis of publication bias for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Estimate number of missing studies
Controls the trim-and-fill analysis of publication bias. Selecting this setting enables the other dialog settings.

Side to Impute Studies
Provides options for specifying the side of the funnel plot on which the missing studies are imputed.

Determined by the slopes of Egger's test
The default setting determines the side based on the estimated slope of the Egger’s test.

Left
Imputes the left side of the funnel plot.

Right
Imputes the right side of the funnel plot.

Method
Specifies the method to estimate the number of missing studies.

Linear
The default setting calculates the linear estimator.

Run
Calculates the run estimator.

Quadratic
Calculates the quadratic estimator.

Iteration process
Provides settings for specifying the iteration estimator and standard error adjustment.

Fixed-effects model
When selected, a fixed-effects model is employed and the iteration estimation and standard error adjustment options are not available.

Random-effects model
When selected, a random effects model is employed and the following settings are available.

Estimator
Provides settings for specifying the iteration estimator

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculate the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings for controlling whether to apply the Knapp-Hartung standard-error adjustment to the iterations of the trim-and-fill algorithm.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Continuous Effect Size trim-and-fill settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Trim-and-Fill**.
3. Select and define the appropriate trim-and-fill settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Print

The **Print** dialog provides settings for controlling table outputs for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Homogeneity/Heterogeneity
Provides settings for controlling the homogeneity and heterogeneity tests.

Test of homogeneity
When selected, the corresponding test of homogeneity is provided in the output.

Heterogeneity measures
When selected, the heterogeneity measures are provided in the output.

Effect Sizes
Provides the following effect size settings.

Individual studies
Controls the display of individual studies. When selected, the corresponding output is provided.

Cumulative effect sizes
Controls the display of cumulative analysis. When selected, the corresponding output is provided. The setting is available only when a **Cumulative Analysis** variable is selected on the **Analysis** dialog.

Prediction interval under random-effects model
Controls the display of the prediction interval. When selected, the corresponding output is provided. The setting is available only when a random-effects model is specified.

Defining Meta-Analysis Continuous Effect Size print settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Print**.
3. Select and define the appropriate print settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Save

The **Save** dialog provides settings for saving the estimated statistics to the active data set for meta-analysis with continuous outcomes when the pre-calculated effect size data are provided in the active data set.

Individual Studies
Provides options for saving the estimated effect size.

Standard error
Saves the estimated standard error of the effect size. The setting is available only when an **Effect Size** variable is not specified.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the effect size.

P-value
Saves the estimated *p*-value of the effect size.

Study weight
Saves the estimated study weight.

Percentage of study weight
Saves the normalized study weight as a percentage.

Defining Meta-Analysis Continuous Effect Size save settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes** > **Pre-Calculated Effect Size**
2. In the **Meta-Analysis Continuous Effect Size** dialog, click **Save**.
3. Select and define the appropriate estimated statistics save settings.
4. Click **Continue**.

Meta-Analysis Continuous Effect Size: Plot

The **Plot** dialog provides settings for the following plot types:

- "Forest Plot" on page 71
- "Cumulative Forest Plot" on page 73
- "Bubble Plot" on page 74
- "Funnel Plot" on page 75
- "Galbraith Plot" on page 76

Meta-Analysis Binary

The Meta-Analysis Binary procedure performs meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Example

Several research studies were conducted in history to investigate a faddish but debatable medicine to help treat type II diabetes. The oral medicine was claimed to be able to reduce the blood glucose level after meals. Data were collected from different research sites from 1979 to 1986.

A principal investigator would like to draw statistical inference about the effect of the oral medicine. Due to the fact that the data were generated from different studies, she proposed the idea of synthesizing the results across the studies to reach an overall understanding of the effect and to identify those underlying sources of variation in the outcomes.

Statistics

Confidence interval, Log Odds Ratio, Peto's Log Odds Ratio, Log Risk Ratio, risk difference, random effects, fixed effects, inverse variance, Mantel-Haenszel, iterations, step-halving, convergence, cumulative statistics, cumulative effect size, restricted maximum likelihood, REML, maximum likelihood, ML, Empirical Bayes, Hedges, Hunter-Schmidt, DerSimonian-Laird, Sidik-Jonkman, Knapp-Hartung, Egger's Test, Harbord's Test, Peters' Test, intercept in regression, dispersion parameter, homogeniety, heterogeniety, exponentiated statistics, standard error, *p*-value, study weight.

Obtaining a Meta-Analysis Binary analysis

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. Under the **Treatment Group** section, select a **Success** variable to represent the "success" counts for the treatment group. The selected variable must be numeric (string variables are not supported).
3. Select a **Failure** variable to represent the "failure" counts for the treatment group. The selected variable must be numeric (string variables are not supported).
4. Under the **Control Group** section, select a **Success** variable to represent the "success" counts for the control group. The selected variable must be numeric (string variables are not supported).
5. Select a **Failure** variable to represent the "failure" counts for the control group. The selected variable must be numeric (string variables are not supported).
6. Optionally, select **Study ID** and/or **Study Label** variables. The selected **Study ID**variable cannot be the same as the selected **Study Label** variable.
7. Optionally, select an **Effect Size** setting. Available options are **Log Odds Ratio**, **Peto's Log Odds Ratio**, **Log Risk Ratio**, and **Risk Difference**.
8. Optionally, select a **Model** setting. When **Trim-and-Fill** settings are enabled, the setting also controls the model that is used by pooling in the trim-and-fill analysis. When **Bias** settings are enabled, the setting also controls the model that is used by the regression-based test.

   Random-effects

   The default setting builds the random-effects model.

Fixed-effects
Builds the fixed-effects model. **Inverse-variance** estimates the inverse-variance weight. **Mantel-Haenszel** estimates the Mantel-Haenszel weight.

9. Optionally, you can:
   - Click **Criteria...** to specify the general criteria.
   - Click **Analysis** to specify the subgroup and cumulative analysis.
   - Click **Inference** to specify the estimation methods.
   - Click **Contrast** to control the contrast test.
   - Click **Bias** to access the publication bias by conducting the EGGER's regression-based test.
   - Click **Trim-and-Fill** to implement the trim-and-fill analysis of publication bias.
   - Click **Print** to control the table outputs.
   - Click **Save** to save the estimated statistics to the active data set.
   - Click **Plot** to specify which plots to include in the output.
10. Click **OK**.

Meta-Analysis Binary: Criteria

The **Criteria** dialog provides settings for specifying criteria for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Confidence Interval
The optional setting specifies the confidence level. The value must be a numeric value between 0 and 100. The default setting is 95.

Missing Data Scope
The optional settings control how the procedure handles missing data.

Exclude cases analysis by analysis
The default setting includes all cases with sufficient data on the variables used in each particular analysis.

Exclude cases listwise
Includes all cases with sufficient data on all variables used across all analyses specified by the procedure.

User Missing Values
The optional settings control how user missing values are treated.

Exclude
The default setting treats user missing values as valid.

Include
Ignores user missing value designations and treats user missing values as valid.

Iterations

Maximum iterations
The optional setting specifies the maximum number of iterations in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 100. A value of 0 means that no iterations are performed.

Maximum step-halvings
The optional setting specifies the maximum step-halving in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 5. A value of 0 means that step-halving is not applied.

Convergence
The optional setting specifies the convergence tolerance. The value must be a single positive value. The setting is available when iterative methods are used. The default value is 1E-6.

Adjust Zero Counts
The optional settings control how to adjust zero-count data, if applicable, in the effect size estimation.

Adjust each value of only those studies containing at least one zero
The default setting adjusts only zero-count data.

Adjust each value of all studies only if there exists at least one study containing a zero value
Adjusts all frequencies when at least one zero count exists in a study.

Adjust each value of all studies
Adjusts all data regardless of the existence of zero-counts.

Make no adjustments and keep zero(s)
Makes no data adjustment.

Added Value
The optional setting specifies the value added to the zero-count data. The default value is 0.5. The specified value must be a single numeric value greater than 0 and less than or equal to 1.

Defining Meta-Analysis Binary criteria

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Binary** dialog, click **Criteria**.
3. Select and define the appropriate criteria settings.
4. Click **Continue**.

Meta-Analysis Binary: Analysis

The **Analysis** dialog provides settings for specifying subgroup and cumulative analysis for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Subgroup Analysis
Select a variable that evokes the subgroup analysis. The variable cannot be the same as the variable defined for **Cumulative Analysis**.

Cumulative Analysis
Select a variable that evokes the cumulative analysis and upon which the cumulative meta-analysis is conducted. The variable cannot be the same as the variable defined for **Subgroup Analysis**. When **Ascending** is selected, cumulative analysis is based upon the specified variable in ascending order. When **Descending** is selected, cumulative analysis is based upon the specified variable in descending order.

Cumulative Statistics
Provides options for saving the estimated cumulative overall effect size. The settings are available only when a **Cumulative Analysis** variable is selected.

Cumulative effect size
Saves the estimated cumulative overall effect size.

Cumulative effect size (exponentiated form)
Saves the estimated cumulative overall effect size in the exponential form.

Standard error
Saves the estimated standard error of the cumulative overall effect size.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the cumulative overall effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the cumulative overall effect size.

Confidence interval lower bound (exponentiated form)
Saves the estimated confidence interval lower limit of the cumulative overall effect size in the exponential form.

Confidence interval upper bound (exponentiated form)
Saves the estimated confidence interval upper limit of the cumulative overall effect size in the exponential form.

P-value
Saves the estimated *p*-value of the cumulative overall effect size.

Destination
Provides options for specifying a save dataset or data file. When **Dataset** is selected, you can specify a new dataset name (you can retain the default dataset name). When **Data file** is selected, click **Browse...** to select a save file name and location.

Defining Meta-Analysis Binary analysis settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Binary** dialog, click **Analysis**.
3. Select and define the appropriate analysis settings.
4. Click **Continue**.

Meta-Analysis Binary: Inference

The **Inference** dialog provides settings for specifying the estimation methods for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Note: The **Inference** dialog is available only when a **Random-effects** model is selected.

Estimator
Provides settings for specifying the estimator.

Notes:

- When Trim and Fill settings are specified, also controls the estimator that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the estimator that is used by the regression-based test.

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculates the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings that control whether to apply the Knapp-Hartung standard-error adjustment.

Notes:

- When Trim and Fill settings are specified, also controls the standard error adjustment that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the standard error adjustment that is used by the regression-based test.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Binary inference settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Binary** dialog, click **Inference**.
3. Select and define the appropriate inference settings.
4. Click **Continue**.

Meta-Analysis Binary: Contrast

The **Contrast** dialog provides settings for controlling the contrast test for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Variables
The list display all available dataset variables. Select variables from the list and move them to the **Contrast(s)** list.

Contrast(s)
The list identifies coefficients that are stored as variables in the active data set. Multiple variables are allowed. String variables are not supported.

User-Input Coefficient Values
Provides settings for specifying coefficients of a user-supplied contrast. Only numeric values are allowed. To formulate a valid contrast test, the number of specified values must match that of the valid studies.

Display exponentiated statistics
Controls the inclusion of the exponential-form statistics. When the setting is selected, the output includes the exponential-form statistics, including the exponential effect size and confidence interval limits. The setting is available when the **Effect Size** is specified as **Logs Odd Ratio**, **Peto's Logs Odd Ratio**, or **Log Risk Ratio**.

Defining Meta-Analysis Binary contrast settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Binary** dialog, click **Contrast**.
3. Select and define the appropriate contrast settings.
4. Click **Continue**.

Meta-Analysis Binary: Bias

The **Bias** dialog provides settings for enabling the publication bias by conducting regression-based tests for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Regression-based Tests
Provides options for specifying the regression based tests. More than one test can be selected.

Egger's test
When selected, conducts Egger's test.

Harbord's test
When selected, conducts Harbord's test. The test is available when the **Effect Size** is specified as **Logs Odd Ratio** or **Log Risk Ratio**.

Peters' test
When selected, conducts Peters' test. The test is available when the **Effect Size** is specified as **Logs Odd Ratio**.

Variables
The list provides all available dataset variables.

Covariate(s)
Variables selected from the **Variables** list are treated as covariates. Multiple covariates are allowed.

Factor(s)
Variables selected from the **Variables** list are treated as factors. Multiple factors are allowed.

Include intercept in regression
Controls the intercept term in the regression-based test.

Include dispersion parameter in fixed-effects model
Controls the multiplicative model setting and introduces the multiplicative dispersion parameter to the analysis. The setting is available only when a fixed-effects model is selected.

Estimate statistics based on t-distribution
Controls the distribution used in the regression-based tests. The setting is enabled by default, which estimates the statistics based on the *t*-distribution. When the setting is not selected, the statistics are estimated based on the normal distribution.

Defining Meta-Analysis Binary bias settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Binary** dialog, click **Bias**.
3. Select and define the appropriate bias settings.
4. Click **Continue**.

Meta-Analysis Binary: Trim-and-Fill

The **Trim-and-Fill** dialog provides settings for implementing the trim-and-fill analysis of publication bias for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Estimate number of missing studies
Controls the trim-and-fill analysis of publication bias. Selecting this setting enables the other dialog settings.

Side to Impute Studies
Provides options for specifying the side of the funnel plot on which the missing studies are imputed.

Determined by the slopes of Egger's test
The default setting determines the side based on the estimated slope of the Egger’s test.

Left
Imputes the left side of the funnel plot.

Right
Imputes the right side of the funnel plot.

Method
Specifies the method to estimate the number of missing studies.

Linear
The default setting calculates the linear estimator.

Run
Calculates the run estimator.

Quadratic
Calculates the quadratic estimator.

Iteration process
Provides settings for specifying the iteration estimator and standard error adjustment.

Fixed-effects model
When selected, a fixed-effects model is employed and the iteration estimation and standard error adjustment options are not available.

Random-effects model
When selected, a random effects model is employed and the following settings are available.

Estimator
Provides settings for specifying the iteration estimator

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculate the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings for controlling whether to apply the Knapp-Hartung standard-error adjustment to the iterations of the trim-and-fill algorithm.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Binary trim-and-fill settings

1. From the menus choose:

Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**

2. In the **Meta-Analysis Binary** dialog, click **Trim-and-Fill**.
3. Select and define the appropriate trim-and-fill settings.
4. Click **Continue**.

Meta-Analysis Binary: Print

The **Print** dialog provides settings for controlling table outputs for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Homogeneity/Heterogeneity
Provides settings for controlling the homogeneity and heterogeneity tests.

Test of homogeneity
When selected, the corresponding test of homogeneity is provided in the output.

Heterogeneity measures
When selected, the heterogeneity measures are provided in the output.

Effect Sizes
Provides the following effect size settings.

Individual studies
Controls the display of individual studies. When selected, the corresponding output is provided.

Cumulative effect sizes
Controls the display of cumulative analysis. When selected, the corresponding output is provided. The setting is available only when a **Cumulative Analysis** variable is selected on the **Analysis** dialog.

Prediction interval under random-effects model
Controls the display of the prediction interval. When selected, the corresponding output is provided. The setting is available only when a random-effects model is specified.

Display exponentiated statistics
Controls the inclusion of the exponential-form statistics. When the setting is selected, the exponential-form statistics, including the exponential effect size and confidence interval limits are output. The process applies to both required and optional effect size estimation and prediction tables. The setting is available when the **Effect Size** is specified as **Logs Odd Ratio**, **Peto's Logs Odd Ratio**, or **Log Risk Ratio**.

Defining Meta-Analysis Binary print settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Binary** dialog, click **Print**.
3. Select and define the appropriate print settings.
4. Click **Continue**.

Meta-Analysis Binary: Save

The **Save** dialog provides settings for saving the estimated statistics to the active data set for meta-analysis with binary outcomes on raw data that are provided in the active dataset for the estimation of the effect size.

Individual Studies
Provides options for saving the estimated effect size.

Individual effect size
Saves the estimated effect size.

Individual effect size (exponentiated form)
Saves the estimated effect size in the exponential form.

Standard error
Saves the estimated standard error of the effect size.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the effect size.

Confidence interval lower bound (exponentiated form)
Saves the estimated confidence interval lower limit of the effect size in the exponential form.

Confidence interval upper bound (exponentiated form)
Saves the estimated confidence interval upper limit of the effect size in the exponential form.

P-value
Saves the estimated *p*-value of the effect size.

Study weight
Saves the estimated study weight.

Percentage of study weight
Saves the normalized study weight as a percentage.

Defining Meta-Analysis Binary save settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis Binary** dialog, click **Save**.
3. Select and define the appropriate estimated statistics save settings.
4. Click **Continue**.

Meta-Analysis Binary: Plot

The **Plot** dialog provides settings for the following plot types:

- "Forest Plot" on page 71
- "Cumulative Forest Plot" on page 73
- "Bubble Plot" on page 74
- "Funnel Plot" on page 75
- "Galbraith Plot" on page 76
- "L'Abb´e Plot" on page 76

Meta-Analysis Binary Effect Size

The Meta-Analysis Binary Effect Size procedure performs meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Example

Several research studies were conducted in history to investigate a faddish but debatable medicine to help treat type II diabetes. The oral medicine was claimed to be able to reduce the blood glucose level after meals. Data were collected from different research sites from 1979 to 1986.

A principal investigator would like to draw statistical inference about the effect of the oral medicine. Due to the fact that the data were generated from different studies, she proposed the idea of synthesizing the results across the studies to reach an overall understanding of the effect and to identify those underlying sources of variation in the outcomes.

Statistics
Confidence interval, Log Odds Ratio, Peto's Log Odds Ratio, Log Risk Ratio, risk difference, random effects, fixed effects, inverse variance, Mantel-Haenszel, iterations, step-halving, convergence, cumulative statistics, cumulative effect size, restricted maximum likelihood, REML, maximum likelihood, ML, Empirical Bayes, Hedges, Hunter-Schmidt, DerSimonian-Laird, Sidik-Jonkman, Knapp-Hartung, Egger's Test, Harbord's Test, Peters' Test, intercept in regression, dispersion parameter, homogeniety, heterogeniety, exponentiated statistics, standard error, *p*-value, study weight.

Obtaining a Meta-Analysis Binary Effect Size analysis

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**
2. Select an **Effect Size** variable that denotes the effect size. The selected variable must be numeric (string variables are not supported).
3. Select either **Standard deviation** to specify the effect size standard deviation, or **Variance** to specify the effect size variance and then select a variable to represent the standard deviation/variance for the effect size.
4. Optionally, select **Study ID** and/or **Study Label** variables. The selected **Study ID**variable cannot be the same as the selected **Study Label** variable.
5. Optionally, select an **Effect Size** setting. Available options are **Log Odds Ratio**, **Peto's Log Odds Ratio**, **Log Risk Ratio**, and **Risk Difference**.
6. Optionally, select a **Model** setting. When **Trim-and-Fill** settings are enabled, the setting also controls the model that is used by pooling in the trim-and-fill analysis. When **Bias** settings are enabled, the setting also controls the model that is used by the regression-based test.

   Random-effects
   The default setting builds the random-effects model.

   Fixed-effects
   Builds the fixed-effects model.
7. Optionally, you can:
   - Click **Criteria...** to specify the general criteria.
   - Click **Analysis** to specify the subgroup and cumulative analysis.
   - Click **Inference** to specify the estimation methods.
   - Click **Contrast** to control the contrast test.
   - Click **Bias** to access the publication bias by conducting the EGGER’s regression-based test.
   - Click **Trim-and-Fill** to implement the trim-and-fill analysis of publication bias.
   - Click **Print** to control the table outputs.
   - Click **Save** to save the estimated statistics to the active data set.
   - Click **Plot** to specify which plots to include in the output.
8. Click **OK**.

Meta-Analysis Binary Effect Size: Criteria

The **Criteria** dialog provides settings for specifying criteria for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Confidence Interval
The optional setting specifies the confidence level. The value must be a numeric value between 0 and 100. The default setting is 95.

Missing Data Scope
The optional settings control how the procedure handles missing data.

Exclude cases analysis by analysis
The default setting includes all cases with sufficient data on the variables used in each particular analysis.

Exclude cases listwise
Includes all cases with sufficient data on all variables used across all analyses specified by the procedure.

User Missing Values
The optional settings control how user missing values are treated.

Exclude
The default setting treats user missing values as valid.

Include
Ignores user missing value designations and treats user missing values as valid.

Iterations

Maximum iterations
The optional setting specifies the maximum number of iterations in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 100. A value of 0 means that no iterations are performed.

Maximum step-halvings
The optional setting specifies the maximum step-halving in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 5. A value of 0 means that step-halving is not applied.

Convergence
The optional setting specifies the convergence tolerance. The value must be a single positive value. The setting is available when iterative methods are used. The default value is 1E-6.

Defining Meta-Analysis Effect Size Binary criteria

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis Binary Effect Size** dialog, click **Criteria**.
3. Select and define the appropriate criteria settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Analysis

The **Analysis** dialog provides settings for specifying subgroup and cumulative analysis for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Subgroup Analysis
Select a variable that evokes the subgroup analysis. The variable cannot be the same as the variable defined for **Cumulative Analysis**.

Cumulative Analysis
Select a variable that evokes the cumulative analysis and upon which the cumulative meta-analysis is conducted. The variable cannot be the same as the variable defined for **Subgroup Analysis**. When **Ascending** is selected, cumulative analysis is based upon the specified variable in ascending order. When **Descending** is selected, cumulative analysis is based upon the specified variable in descending order.

Cumulative Statistics
Provides options for saving the estimated cumulative overall effect size. The settings are available only when a **Cumulative Analysis** variable is selected.

Cumulative effect size
Saves the estimated cumulative overall effect size.

Cumulative effect size (exponentiated form)
Saves the estimated cumulative overall effect size in the exponential form.

Standard error
Saves the estimated standard error of the cumulative overall effect size.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the cumulative overall effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the cumulative overall effect size.

Confidence interval lower bound (exponentiated form)
Saves the estimated confidence interval lower limit of the cumulative overall effect size in the exponential form.

Confidence interval upper bound (exponentiated form)
Saves the estimated confidence interval upper limit of the cumulative overall effect size in the exponential form.

P-value
Saves the estimated *p*-value of the cumulative overall effect size.

Destination
Provides options for specifying a save dataset or data file. When **Dataset** is selected, you can specify a new dataset name (you can retain the default dataset name). When **Data file** is selected, click **Browse...** to select a save file name and location.

Defining Meta-Analysis Binary Effect Size analysis settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis Binary Effect Size** dialog, click **Analysis**.
3. Select and define the appropriate analysis settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Inference

The **Inference** dialog provides settings for specifying the estimation methods for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Estimator
Provides settings for specifying the estimator.

Notes:

- When Trim and Fill settings are specified, also controls the estimator that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the estimator that is used by the regression-based test.

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculates the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings that control whether to apply the Knapp-Hartung standard-error adjustment.

Notes:

- When Trim and Fill settings are specified, also controls the standard error adjustment that is used by pooling in the trim-and-fill analysis.
- When Bias settings are specified, also controls the standard error adjustment that is used by the regression-based test.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Binary Effect Size inference settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis Binary Effect Size** dialog, click **Inference**.
3. Select and define the appropriate inference settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Contrast

The **Contrast** dialog provides settings for controlling the contrast test for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Variables
The list display all available dataset variables. Select variables from the list and move them to the **Contrast(s)** list.

Contrast(s)
The list identifies coefficients that are stored as variables in the active data set. Multiple variables are allowed. String variables are not supported.

User-Input Coefficient Values
Provides settings for specifying coefficients of a user-supplied contrast. Only numeric values are allowed. To formulate a valid contrast test, the number of specified values must match that of the valid studies.

Display exponentiated statistics
Controls the inclusion of the exponential-form statistics. When the setting is selected, the output includes the exponential-form statistics, including the exponential effect size and confidence interval limits. The setting is available when the **Effect Size** is specified as **Logs Odd Ratio**, **Peto's Logs Odd Ratio**, or **Log Risk Ratio**.

Defining Meta-Analysis Binary Effect Size contrast settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis Binary Effect Size** dialog, click **Contrast**.
3. Select and define the appropriate contrast settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Bias

The **Bias** dialog provides settings for enabling the publication bias by conducting regression-based tests for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Egger's regression-based test
Selecting this setting enables the publication bias by conducting the Eggers's regression-based test.

Variables
The list provides all available dataset variables.

Covariate(s)
Variables selected from the **Variables** list are treated as covariates. Multiple covariates are allowed.

Factor(s)
Variables selected from the **Variables** list are treated as factors. Multiple factors are allowed.

Include intercept in regression
Controls the intercept term in the regression-based test.

Include dispersion parameter in fixed-effects model
Controls the multiplicative model setting and introduces the multiplicative dispersion parameter to the analysis. The setting is available only when a fixed-effects model is selected.

Estimate statistics based on t-distribution
Controls the distribution used in the regression-based tests. The setting is enabled by default, which estimates the statistics based on the *t*-distribution. When the setting is not selected, the statistics are estimated based on the normal distribution.

Defining Meta-Analysis Binary Effect Size bias settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis Binary Effect Size** dialog, click **Bias**.
3. Select and define the appropriate bias settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Trim-and-Fill

The **Trim-and-Fill** dialog provides settings for implementing the trim-and-fill analysis of publication bias for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Estimate number of missing studies
Controls the trim-and-fill analysis of publication bias. Selecting this setting enables the other dialog settings.

Side to Impute Studies
Provides options for specifying the side of the funnel plot on which the missing studies are imputed.

Determined by the slopes of Egger's test
The default setting determines the side based on the estimated slope of the Egger's test.

Left
Imputes the left side of the funnel plot.

Right
Imputes the right side of the funnel plot.

Method
Specifies the method to estimate the number of missing studies.

Linear
The default setting calculates the linear estimator.

Run
Calculates the run estimator.

Quadratic
Calculates the quadratic estimator.

Iteration process
Provides settings for specifying the iteration estimator and standard error adjustment.

Fixed-effects model
When selected, a fixed-effects model is employed and the iteration estimation and standard error adjustment options are not available.

Random-effects model
When selected, a random effects model is employed and the following settings are available.

Estimator
Provides settings for specifying the iteration estimator

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculate the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings for controlling whether to apply the Knapp-Hartung standard-error adjustment to the iterations of the trim-and-fill algorithm.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Binary Effect Size trim-and-fill settings

1. From the menus choose:

Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**

2. In the **Meta-Analysis Binary Effect Size** dialog, click **Trim-and-Fill**.
3. Select and define the appropriate trim-and-fill settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Print

The **Print** dialog provides settings for controlling table outputs for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Homogeneity/Heterogeneity
Provides settings for controlling the homogeneity and heterogeneity tests.

Test of homogeneity
When selected, the corresponding test of homogeneity is provided in the output.

Heterogeneity measures
When selected, the heterogeneity measures are provided in the output.

Effect Sizes
Provides the following effect size settings.

Individual studies
Controls the display of individual studies. When selected, the corresponding output is provided.

Cumulative effect sizes
Controls the display of cumulative analysis. When selected, the corresponding output is provided. The setting is available only when a **Cumulative Analysis** variable is selected on the **Analysis** dialog.

Prediction interval under random-effects model
Controls the display of the prediction interval. When selected, the corresponding output is provided. The setting is available only when a random-effects model is specified.

Display exponentiated statistics
Controls the inclusion of the exponential-form statistics. When the setting is selected, the exponential-form statistics, including the exponential effect size and confidence interval limits are output. The process applies to both required and optional effect size estimation and prediction tables. The setting is available when the **Effect Size** is specified as **Logs Odd Ratio**, **Peto's Logs Odd Ratio**, or **Log Risk Ratio**.

Defining Meta-Analysis Binary Effect Size print settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**

2. In the **Meta-Analysis Binary Effect Size** dialog, click **Print**.
3. Select and define the appropriate print settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Save

The **Save** dialog provides settings for saving the estimated statistics to the active data set for meta-analysis with binary outcomes when the pre-calculated effect size data are provided in the active data set.

Individual Studies
Provides options for saving the estimated effect size.

Individual effect size (exponentiated form)
Saves the estimated effect size in the exponential form.

Standard error
Saves the estimated standard error of the effect size.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the effect size.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the effect size.

Confidence interval lower bound (exponentiated form)
Saves the estimated confidence interval lower limit of the effect size in the exponential form.

Confidence interval upper bound (exponentiated form)
Saves the estimated confidence interval upper limit of the effect size in the exponential form.

P-value
Saves the estimated *p*-value of the effect size.

Study weight
Saves the estimated study weight.

Percentage of study weight
Saves the normalized study weight as a percentage.

Defining Meta-Analysis Binary Effect Size save settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Binary Outcomes** > **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis Binary Effect Size** dialog, click **Save**.
3. Select and define the appropriate estimated statistics save settings.
4. Click **Continue**.

Meta-Analysis Binary Effect Size: Plot

The **Plot** dialog provides settings for the following plot types:

- "Forest Plot" on page 71
- "Cumulative Forest Plot" on page 73
- "Bubble Plot" on page 74
- "Funnel Plot" on page 75
- "Galbraith Plot" on page 76
- "L'Abb´e Plot" on page 76

Meta-Analysis Regression

The Meta-Analysis Regression procedure performs meta-regression analysis.

Example
Several research studies were conducted in history to investigate a faddish but debatable medicine to help treat type II diabetes. The oral medicine was claimed to be able to reduce the blood glucose level after meals. Data were collected from different research sites from 1979 to 1986.

A principal investigator would like to draw statistical inference about the effect of the oral medicine. Due to the fact that the data were generated from different studies, she proposed the idea of synthesizing the results across the studies to reach an overall understanding of the effect and to identify those underlying sources of variation in the outcomes.

Statistics
Confidence level, iterative method, step-halving, convergence tolerance, sample means, sample variance, standard deviation, estimated effect size, estimation method, regression-based test, random-effects model, fixed-effects model, dispersion parameter, restricted maximum likelihood estimator, empirical Bayes estimator, Hedges estimator, Hunter-Schmidt estimator, DerSimonian-Laird estimator, Sidik-Jonkman estimator, Knapp-Hartung standard-error adjustment, truncated

Knapp-Hartung standard-error adjustment, coefficients, model coefficient test, exponentiated statistics.

Obtaining a Meta-Analysis Regression analysis

1. From the menus choose:

   Analyze > **Meta Analysis** > **Meta Regression**

2. Select a single dependent **Effect Size** variable that denotes the effect size. The selected variable must be numeric (string variables are not supported).
3. Select one of the following settings and then select a corresponding single numeric variable:

   Standard error
   Select a variable that specifies the standard error that is converted to the weight. This is the default setting.

   Variance
   Select a variable that specifies the variance that is converted to the weight.

   Weight
   Select a variable that specifies the weight.

4. Optionally, add factor variables to the **Factor(s)** list. For each selected factor variable, an optional value can be specified for each variable to designate custom **Last Categories**. You can click **Reset** to restore the **Last Categories** values to their default settings.

   Note: When there are no cases that match the specified **Last Categories** values, the last occurring values are treated as the last categories.

5. Optionally, select numeric covariate variables.
6. Optionally, select a **Model** setting.

   Random-effects
   The default setting builds the random-effects model.

   Fixed-effects
   Builds the fixed-effects model. You can optionally select the **Include dispersion parameter** setting.

7. Optionally, you can:
   - Click **Criteria...** to specify the general criteria.
   - Click **Inference** to specify the estimation methods.
   - Click **Print** to control the table outputs.
   - Click **Save** to predict and save the estimated statistics to the active dataset.
   - Click **Plot** to specify which plots to include in the output.
8. Click **OK**.

Meta-Analysis Regression: Criteria

The **Criteria** dialog provides settings for specifying criteria for meta-analysis regression.

Confidence Interval
The optional setting specifies the confidence level. The value must be a numeric value between 0 and 100. The default setting is 95.

User Missing Values
The optional settings control how user missing values are treated.

Exclude
The default setting treats user missing values as valid.

Include
Ignores user missing value designations and treats user missing values as valid.

Iterations

Maximum iterations
The optional setting specifies the maximum number of iterations in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 100. A value of 0 means that no iterations are performed.

Maximum step-halvings
The optional setting specifies the maximum step-halving in the iterative methods. The setting is available when iterative methods are used. The value must be a single positive integer. The default value is 5. A value of 0 means that step-halving is not applied.

Convergence
The optional setting specifies the convergence tolerance. The value must be a single positive value. The setting is available when iterative methods are used. The default value is 1E-6.

Defining Meta-Analysis Regression criteria

1. From the menus choose:

   Analyze > **Meta Analysis** > **Meta Regression**
2. In the **Meta-Analysis Regression** dialog, click **Criteria**.
3. Select and define the appropriate criteria settings.
4. Click **Continue**.

Meta-Analysis Regression: Inference

The **Inference** dialog provides settings for specifying the estimation methods for meta-analysis regression.

Include intercept in regression
The setting is enabled be default.

Estimate statistics based on t-distribution
Controls the distribution used in the regression-based tests. The setting is enabled by default, which estimates the statistics based on the *t*-distribution. When the setting is not selected, the statistics are estimated based on the normal distribution.

Estimator
Provides settings for specifying the estimator.

Restricted maximum likelihood (REML)
The default setting applies the iterative method and calculates the restricted maximum likelihood estimator.

Maximum likelihood (ML)
Applies the iterative method and calculates the maximum likelihood estimator.

Empirical Bayes
Applies the iterative method and calculates the empirical Bayes estimator.

Hedges
Applies the non-iterative method and calculates the Hedges estimator.

Hunter-Schmidt
Applies the non-iterative method and calculates the Hunter-Schmidt estimator.

DerSimonian-Laird
Applies the non-iterative method and calculates the DerSimonian-Laird estimator.

Sidik-Jonkman
Applies the non-iterative method and calculates the Sidik-Jonkman estimator.

Standard Error Adjustment
Provides settings that control whether to apply the Knapp-Hartung standard-error adjustment.

No adjustment
The default setting does not apply the adjustment.

Apply the Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method.

Apply the truncated Knapp-Hartung adjustment
Applies the Knapp-Hartung adjustment method and truncates the value if less than 1 when estimating the variance-covariance matrix.

Defining Meta-Analysis Regression inference settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Meta Regression**
2. In the **Meta-Analysis Regression** dialog, click **Inference**.
3. Select and define the appropriate inference settings.
4. Click **Continue**.

Meta-Analysis Regression: Print

The **Print** dialog provides settings for controlling table outputs for meta-analysis regression.

Model coefficient test
Controls the model coefficient test. The setting is not specified by default, which suppresses the test. When the setting is specified, the test is provided in the output.

Display exponentiated statistics
Controls the parameter estimates. The setting is not specified by default, which suppresses parameter estimates. When the setting is specified, parameter estimates are provided in the output.

Defining Meta-Analysis Regression print settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Meta Regression**
2. In the **Meta-Analysis Regression** dialog, click **Print**.
3. Select and define the appropriate print settings.
4. Click **Continue**.

Meta-Analysis Regression: Save

The **Save** dialog provides settings for saving the estimated statistics to the active data set for meta-analysis regression.

Individual Studies
Provides options for saving the estimated effect size.

Predicted values
Predicts and saves the fitted values.

Standard error of predicted values
Predicts and saves the standard error of fitted values.

Confidence interval lower bound
Saves the estimated confidence interval lower limit of the fitted values.

Confidence interval upper bound
Saves the estimated confidence interval upper limit of the fitted values.

Residuals
Predicts and saves the residuals.

Standard error of residuals
Predicts and saves the standard error of the residuals.

Leverages
Predicts and saves the leverages.

Cumulative Statistics
Provides options for saving linear predictions.

Note: The cumulative statistics settings are ignored when **Fixed-effects** is selected as the **Model** setting.

Fixed linear predictions
Predicts and saves the linear prediction.

Standard error of fixed linear predictions
Predicts and saves the standard error of the linear prediction.

Best linear unbiased predictions
Predicts and saves the best linear unbiased prediction of the random effects.

Standard error of BLUPs
Predicts and saves the standard error of the best linear unbiased prediction of the random effects.

Defining Meta-Analysis Regression save settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Meta Regression**
2. In the **Meta-Analysis Regression** dialog, click **Save**.
3. Select and define the appropriate save settings.
4. Click **Continue**.

Meta-Analysis Regression: Plot

The **Plot** dialog provides settings for defining bubble plots. For more information, see "Bubble Plot" on page 74.

Meta-Analysis plotting options

The following plot options are available for the Meta-Analysis procedures and are accessible in the **Plot** dialogs.

Forest Plot

The **Forest Plot** tab provides settings for controlling forest plot charts that display in the output for meta-analysis with continuous and binary outcomes when raw data and pre-calculated effect size data are provided in the active data set.

Forest plot
The optional setting enables and disables forest plot output. The setting is disabled by default.

Display Columns
The optional settings specify the estimated statistics that are tabulated.

Effect size
Includes the estimated effect size.

Standard Error
Includes the standard error of the estimated effect size.

Confidence interval limits
Includes the confidence interval lower and upper limits of the estimated effect size.

P-value
Includes the *p*-value of the estimated effect size for each individual study.

Weight
Includes the weight of the primary studies.

Display exponentiated form

Note: This setting applies only to meta-analysis with binary outcomes.

Controls the plot column and the display of effect size and confidence interval. When the setting is selected, plots the odds ratio, Peto's odds ratio, or risk ratio, and tabulates the exponential-form statistics. When the setting is not selected, plots the log odds ratio, Peto's log odds ratio, or log risk ratio, and tabulates the log-transformed statistics.

Variables
Lists the available dataset variables.

Additional Columns
Optionally, select variables to display as additional columns. The variable order determines the column display order.

Sort
Optionally, specify a variable upon which the forest plot is sorted. When a variable is specified, the forest plot is sorted in **Ascending** order (the default setting). When **Descending** is selected, the forest plot is sorted in descending order.

Position of plot column
Controls the placement of the plot column.

Right
The default setting places the plot column to the right of the other table columns.

Left
Places the plot column to the left of the other table columns.

Reference Lines
Controls the reference lines that are added to the forest plot. **Overall effect size** adds the line to denote the estimated overall effect size. **Null effect size** adds the line to denote the null effect size.

Annotations
Controls the display of the annotations.

Homogeneity
Prints the homogeneity test statistics.

Heterogeneity
Prints the heterogeneity test statistics.

Test
Prints the test of the overall effect size and the test of between-subgroup homogeneity in the presence of the subgroup analysis.

Crop Range
Specifies the crop range. When the setting is selected, two numeric values must be specified and satisfy `Upper bound > Lower bound`.

For meta-analysis with binary outcomes, when the **Effect Size** setting is set to **Risk Difference**, it must also satisfy `Upper bound ≤ 1` and `Lower bound ≥ −1`. When the **Display exponentiated form** setting is selected, it must also satisfy `Lower bound ≥ 0`.

Defining forest plot settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes or Binary Outcomes** > **Raw Data...** or **Pre-Calculated Effect Size...**

2. In the **Meta-Analysis** dialog, click **Plot** and then click the **Forest Plot** tab.
3. Select and define the appropriate forest plot settings.
4. Click **Continue**.

Cumulative Forest Plot

The **Cumulative Forest Plot** tab provides settings for controlling cumulative forest plot charts that display in the output for meta-analysis with continuous and binary outcomes when raw data and pre-calculated effect size data are provided in the active data set.

Cumulative forest plot
The optional setting enables and disables forest plot output. The setting is disabled by default.

Display Columns
The optional settings specify the estimated statistics that are tabulated.

Cumulative effect size
Includes the cumulative overall effect size.

Standard Error
Includes the standard error of the cumulative overall effect size.

Confidence interval limits
Includes the confidence interval lower and upper limits of the cumulative overall effect size.

P-value
Includes the *p*-value of the cumulative overall effect size for each individual study.

Display exponentiated form

Note: This setting applies only to meta-analysis with binary outcomes.

Controls the plot column and the display of effect size and confidence interval. When the setting is selected, it plots the odds ratio, Peto's odds ratio, or risk ratio, and tabulates the exponential-form statistics. When the setting is not enabled, it plots the log odds ratio, Peto's log odds ratio, or log risk ratio, and tabulates the log-transformed statistics.

Variables
Lists the available dataset variables.

Additional Columns
Optionally, select variables to display as additional columns. The variable order determines the column display order.

Position of plot column
Controls the placement of the plot column.

Right
The default setting places the plot column to the right of the other table columns.

Left
Places the plot column to the left of the other table columns.

Crop Range
Specifies the crop range. When the setting is selected, two numeric values must be specified and satisfy `Upper bound > Lower bound`.

For meta-analysis with binary outcomes, when the **Effect Size** setting is set to **Risk Difference**, it must also satisfy `Upper bound ≤ 1` and `Lower bound ≥ −1`. When the **Display exponentiated form** setting is selected, it must also satisfy `Lower bound ≥ 0`.

Defining cumulative forest plot settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes or Binary Outcomes** > **Raw Data...** or **Pre-Calculated Effect Size...**

2. In the **Meta-Analysis** dialog, click **Plot** and then click the **Cumulative Forest Plot** tab.
3. Select and define the appropriate cumulative forest plot settings.
4. Click **Continue**.

Bubble Plot

The **Bubble Plot** tab provides settings for controlling bubble plot charts that display in the output for meta-analysis with continuous and binary outcomes when raw data and pre-calculated effect size data are provided in the active data set.

Note: Bubble Plot is the only available plot type for meta-analysis regression analysis.

Bubble plot
The optional setting enables and disables bubble plot output. The setting is disabled by default.

Note: This setting is not available for meta-analysis regression analysis.

Variables
Lists the available dataset variables.

Predictor(s)
Specify variables that are treated as continuous predictors on the x-axis. Multiple variables are allowed with each creating a separate bubble plot.

Note: For meta-analysis regression analysis, this setting is labeled **Bubble Plot Predictor(s)** and allows you to select covariates for which to create bubble plots.

Mean-centered predictor(s)
The optional setting controls the continuous predictor on the x-axis. The setting is disabled by default, which does not center the predictor. When the setting is enabled, the predictor is mean-centered.

Label
Optionally specify a variable upon which the bubble plot is labeled. The **Placement** list provides options for placing labels automatically (**Auto**) or on the **Right**, **Left**, **Upper**, and **Bottom**.

Draw bubbles proportional to weights
Controls how bubbles are drawn in relation to the weights. The setting is enabled by default, which draws the bubbles proportional to the weights. When the setting is disabled, all bubbles are drawn to the same size.

Display fitted line
Controls the display of the fitted regression line. The setting is enabled by default, which adds the fitted regression line to the bubble plot. When the setting is disabled, the fitted regression line is not added to the bubble plot.

Display confidence interval bounds
Controls the display of confidence interval bounds. The setting is enabled by default, which adds the confidence interval bounds to the bubble plot. When the setting is disabled, the confidence interval bounds are not added to the bubble plot.

Range of x-axis
Specifies the x-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound`. Note that the specified values apply to all bubble plots.

Range of y-axis
Specifies the y-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound`. Note that the specified values apply to all bubble plots.

For meta-analysis with binary outcomes, when the **Effect Size** setting is set to **Risk Difference**, it must also satisfy `Upper bound ≤ 1` and `Lower bound ≥ −1`.

Defining bubble plot settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes or Binary Outcomes** > **Raw Data...** or **Pre-Calculated Effect Size...**

   or

   Analyze > **Meta Analysis** > **Meta Regression**

2. In the **Meta-Analysis** dialog, click **Plot** and then click the **Bubble Plot** tab.
3. Select and define the appropriate bubble plot settings.
4. Click **Continue**.

Funnel Plot

The **Funnel Plot** tab provides settings for controlling funnel plot charts that display in the output for meta-analysis with continuous and binary outcomes when raw data and pre-calculated effect size data are provided in the active data set.

Funnel plot
The optional setting enables and disables funnel plot output. The setting is disabled by default.

Y-Axis Values
Controls the y-axis values.

Standard error
Plots the standard error.

Inverse standard error
Plots the inverse standard error.

Variance
Plots the variance.

Inverse variance
Plots the inverse variance.

Range of x-axis
Specifies the x-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound`. Note that the specified values apply to all funnel plots.

For meta-analysis with binary outcomes, when the **Effect Size** setting is set to **Risk Difference**, it must also satisfy `Upper bound ≤ 1` and `Lower bound ≥ −1`.

Range of y-axis
Specifies the y-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound ≥ 0`. Note that the specified values apply to all funnel plots.

Include imputed studies with trim-and-fill
Controls whether the imputed studies are included in the funnel plot when the **Estimate number of missing studies** setting is enabled in the **Trim-and-Fill** dialog. The settings is disabled by default, which excludes the imputed studies. When **Display overall effect size of observed studies** is selected, a vertical reference line that denotes the estimated overall effect size is added. The setting takes into account both observed and imputed studies.

Variables
Lists the available dataset variables.

Label
Optionally specify a variable upon which the funnel plot is labeled. The **Placement** list provides options for placing labels automatically (**Auto**) or on the **Right**, **Left**, **Upper**, and **Bottom**.

Defining funnel plot settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes or Binary Outcomes** > **Raw Data...** or **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis** dialog, click **Plot** and then click the **Funnel Plot** tab.
3. Select and define the appropriate funnel plot settings.
4. Click **Continue**.

Galbraith Plot

The **Galbraith Plot** tab provides settings for controlling Galbraith plot charts that display in the output for meta-analysis with continuous and binary outcomes when raw data and pre-calculated effect size data are provided in the active data set.

Galbraith plot
The optional setting enables and disables bubble plot output. The setting is disabled by default.

Display confidence interval bounds
Controls the display of confidence interval bounds. The setting is enabled by default, which adds the confidence interval bounds to the bubble plot. When the setting is disabled, the confidence interval bounds are not added to the bubble plot.

Range of x-axis
Specifies the x-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound ≥ 0`. Note that the specified values apply to all Galbraith plots.

Range of y-axis
Specifies the y-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound`. Note that the specified values apply to all Galbraith plots.

Variables
Lists the available dataset variables.

Label
Optionally specify a variable upon which the bubble plot is labeled. The **Placement** list provides options for placing labels automatically (**Auto**) or on the **Right**, **Left**, **Upper**, and **Bottom**.

Defining Galbraith plot settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes or Binary Outcomes** > **Raw Data...** or **Pre-Calculated Effect Size...**
2. In the **Meta-Analysis** dialog, click **Plot** and then click the **Galbraith Plot** tab.
3. Select and define the appropriate Galbraith plot settings.
4. Click **Continue**.

L'Abb´e Plot

The **L'Abb´e Plot** tab provides settings for controlling L'Abb´e plot charts that display in the output for meta-analysis with continuous and binary outcomes when raw data and pre-calculated effect size data are provided in the active data set.

L'Abb´e plot
The optional setting enables and disables L'Abb´e plot output. The setting is disabled by default.

Draw dots proportional to weights
Controls how dots are drawn in relation to the weights. The setting is enabled by default, which draws the dots proportional to the weights. When the setting is disabled, all dots are drawn to the same size.

Reference lines

Overall effect size
Controls the display of the reference line that represents the estimated overall effect size. The setting is disabled by default, which suppresses the reference line. The reference line displays when the setting is enabled.

Null effect size
Controls the display of the reference line that represents no effect. The setting is enabled by default, which displays the reference line. The reference line is suppressed when the setting is not enabled.

Range of x-axis

Specifies the x-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound`. Note that the specified values apply to all Galbraith plots.

For meta-analysis with binary outcomes, when the **Effect Size** setting is set to **Risk Difference**, it must also satisfy `Lower bound ≥ 0` and `Upper bound ≤ 1`. When the **Effect Size** setting is set to **Log Risk Ratio**, it must satisfy `Upper bound ≤ 0`.

Range of y-axis

Specifies the y-axis plot range. When the setting is enabled, two numeric values must be specified and satisfy `Upper bound > Lower bound`. Note that the specified values apply to all Galbraith plots.

For meta-analysis with binary outcomes, when the **Effect Size** setting is set to **Risk Difference**, it must also satisfy `Lower bound ≥ 0` and `Upper bound ≤ 1`. When the **Effect Size** setting is set to **Log Risk Ratio**, it must satisfy `Upper bound ≤ 0`.

Variables

Lists the available dataset variables.

Label

Optionally specify a variable upon which the bubble plot is labeled. The **Placement** list provides options for placing labels automatically (**Auto**) or on the **Right**, **Left**, **Upper**, and **Bottom**.

Defining L'Abb´e plot settings

1. From the menus choose:

   Analyze > **Meta Analysis** > **Continuous Outcomes or Binary Outcomes** > **Raw Data...**
2. In the **Meta-Analysis** dialog, click **Plot** and then click the **L'Abb´e Plot** tab.
3. Select and define the appropriate L'Abb´e plot settings.
4. Click **Continue**.

Codebook

Codebook reports the dictionary information -- such as variable names, variable labels, value labels, missing values -- and summary statistics for all or specified variables and multiple response sets in the active dataset. For nominal and ordinal variables and multiple response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation, and quartiles.

Note: Codebook ignores split file status. This includes split-file groups created for multiple imputation of missing values (available in the Missing Values add-on option).

To Obtain a Codebook

1. From the menus choose:

   Analyze > **Reports** > **Codebook**
2. Click the Variables tab.
3. Select one or more variables and/or multiple response sets.

Optionally, you can:

- Control the variable information that is displayed.
- Control the statistics that are displayed (or exclude all summary statistics).
- Control the order in which variables and multiple response sets are displayed.
- Change the measurement level for any variable in the source list in order to change the summary statistics displayed. See the topic "Codebook Statistics Tab" on page 79 for more information.

Changing Measurement Level

You can temporarily change the measurement level for variables. (You cannot change the measurement level for multiple response sets. They are always treated as nominal.)

1. Right-click a variable in the source list.
2. Select a measurement level from the pop-up menu.

This changes the measurement level temporarily. In practical terms, this is only useful for numeric variables. The measurement level for string variables is restricted to nominal or ordinal, which are both treated the same by the Codebook procedure.

Codebook Output Tab

The Output tab controls the variable information included for each variable and multiple response set, the order in which the variables and multiple response sets are displayed, and the contents of the optional file information table.

Variable Information

This controls the dictionary information displayed for each variable.

Position. An integer that represents the position of the variable in file order. This is not available for multiple response sets.

Label. The descriptive label associated with the variable or multiple response set.

Type. Fundamental data type. This is either *Numeric*, *String*, or *Multiple Response Set*.

Format. The display format for the variable, such as *A4*, *F8.2*, or *DATE11*. This is not available for multiple response sets.

Measurement level. The possible values are *Nominal*, *Ordinal*, *Scale*, and *Unknown*. The value displayed is the measurement level stored in the dictionary and is not affected by any temporary measurement level override specified by changing the measurement level in the source variable list on the Variables tab. This is not available for multiple response sets.

Note: The measurement level for numeric variables may be "unknown" prior to the first data pass when the measurement level has not been explicitly set, such as data read from an external source or newly created variables. See the topic for more information.

Role. Some dialogs support the ability to pre-select variables for analysis based on defined roles.

Value labels. Descriptive labels associated with specific data values.

- If Count or Percent is selected on the Statistics tab, defined value labels are included in the output even if you don't select Value labels here.
- For multiple dichotomy sets, "value labels" are either the variable labels for the elementary variables in the set or the labels of counted values, depending on how the set is defined. See the topic for more information.

Missing values. User-defined missing values. If Count or Percent is selected on the Statistics tab, defined value labels are included in the output even if you don't select Missing values here. This is not available for multiple response sets.

Custom attributes. User-defined custom variable attributes. Output includes both the names and values for any custom variable attributes associated with each variable. See the topic for more information. This is not available for multiple response sets.

Reserved attributes. Reserved system variable attributes. You can display system attributes, but you should not alter them. System attribute names start with a dollar sign ($) . Non-display attributes, with names that begin with either "@" or "$@", are not included. Output includes both the names and values for any system attributes associated with each variable. This is not available for multiple response sets.

File Information

The optional file information table can include any of the following file attributes:

File name. Name of the IBM SPSS Statistics data file. If the dataset has never been saved in IBM SPSS Statistics format, then there is no data file name. (If there is no file name displayed in the title bar of the Data Editor window, then the active dataset does not have a file name.)

Location. Directory (folder) location of the IBM SPSS Statistics data file. If the dataset has never been saved in IBM SPSS Statistics format, then there is no location.

Number of cases. Number of cases in the active dataset. This is the total number of cases, including any cases that may be excluded from summary statistics due to filter conditions.

Label. This is the file label (if any) defined by the `FILE LABEL` command.

Documents. Data file document text.

Weight status. If weighting is on, the name of the weight variable is displayed. See the topic for more information.

Custom attributes. User-defined custom data file attributes. Data file attributes defined with the `DATAFILE ATTRIBUTE` command.

Reserved attributes. Reserved system data file attributes. You can display system attributes, but you should not alter them. System attribute names start with a dollar sign ($) . Non-display attributes, with names that begin with either "@" or "$@", are not included. Output includes both the names and values for any system data file attributes.

Variable Display Order

The following alternatives are available for controlling the order in which variables and multiple response sets are displayed.

Alphabetical. Alphabetic order by variable name.

File. The order in which variables appear in the dataset (the order in which they are displayed in the Data Editor). In ascending order, multiple response sets are displayed last, after all selected variables.

Measurement level. Sort by measurement level. This creates four sorting groups: nominal, ordinal, scale, and unknown. Multiple response sets are treated as nominal.

Note: The measurement level for numeric variables may be "unknown" prior to the first data pass when the measurement level has not been explicitly set, such as data read from an external source or newly created variables.

Variable list. The order in which variables and multiple response sets appear in the selected variables list on the Variables tab.

Custom attribute name. The list of sort order options also includes the names of any user-defined custom variable attributes. In ascending order, variables that don't have the attribute sort to the top, followed by variables that have the attribute but no defined value for the attribute, followed by variables with defined values for the attribute in alphabetic order of the values.

Maximum Number of Categories

If the output includes value labels, counts, or percents for each unique value, you can suppress this information from the table if the number of values exceeds the specified value. By default, this information is suppressed if the number of unique values for the variable exceeds 200.

Codebook Statistics Tab

The Statistics tab allows you to control the summary statistics that are included in the output, or suppress the display of summary statistics entirely.

Counts and Percents

For nominal and ordinal variables, multiple response sets, and labeled values of scale variables, the available statistics are:

Count. The count or number of cases having each value (or range of values) of a variable.

Percent. The percentage of cases having a particular value.

Central Tendency and Dispersion

For scale variables, the available statistics are:

Mean. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.

Standard Deviation. A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.

Quartiles. Displays values corresponding to the 25th, 50th, and 75th percentiles.

Note: You can temporarily change the measurement level associated with a variable (and thereby change the summary statistics displayed for that variable) in the source variable list on the Variables tab.

Frequencies

The Frequencies procedure provides statistics and graphical displays that are useful for describing many types of variables. The Frequencies procedure is a good place to start looking at your data.

For a frequency report and bar chart, you can arrange the distinct values in ascending or descending order, or you can order the categories by their frequencies. The frequencies report can be suppressed when a variable has many distinct values. You can label charts with frequencies (the default) or percentages.

Example
What is the distribution of a company's customers by industry type? From the output, you might learn that 37.5% of your customers are in government agencies, 24.9% are in corporations, 28.1% are in academic institutions, and 9.4% are in the healthcare industry. For continuous, quantitative data, such as sales revenue, you might learn that the average product sale is $3,576, with a standard deviation of $1,078.

Statistics and plots
Frequency counts, percentages, cumulative percentages, mean, median, mode, sum, standard deviation, variance, range, minimum and maximum values, standard error of the mean, skewness and kurtosis (both with standard errors), quartiles, user-specified percentiles, bar charts, pie charts, and histograms.

Data considerations

Data
Use numeric codes or strings to code categorical variables (nominal or ordinal level measurements).

Assumptions
The tabulations and percentages provide a useful description for data from any distribution, especially for variables with ordered or unordered categories. Most of the optional summary statistics, such as the mean and standard deviation, are based on normal theory and are appropriate for quantitative variables with symmetric distributions. Robust statistics, such as the median, quartiles, and percentiles, are appropriate for quantitative variables that may or may not meet the assumption of normality.

Obtaining frequency tables

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Frequencies...**
2. Select one or more categorical or quantitative variables.
3. Optionally, select **Create APA style tables** to create output tables that adhere to APA style guidelines.
4. Optionally, you can:
   - Click **Statistics** for descriptive statistics for quantitative variables.
   - Click **Charts** for bar charts, pie charts, and histograms.
   - Click **Format** for the order in which results are displayed.

- Click **Style** to specify conditions for automatically changing properties of pivot tables based on specific conditions.
- Click **Bootstrap** to derive robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient. It may also be used for constructing hypothesis tests.

Frequencies Statistics

Percentile Values. Values of a quantitative variable that divide the ordered data into groups so that a certain percentage is above and another percentage is below. Quartiles (the 25th, 50th, and 75th percentiles) divide the observations into four groups of equal size. If you want an equal number of groups other than four, select **Cut points for n equal groups.** You can also specify individual percentiles (for example, the 95th percentile, the value below which 95% of the observations fall).

Central Tendency. Statistics that describe the location of the distribution include the mean, median, mode, and sum of all the values.

- *Mean*. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.
- *Median*. The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).
- *Mode*. The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode. The Frequencies procedure reports only the smallest of such multiple modes.
- *Sum*. The sum or total of the values, across all cases with nonmissing values.

Dispersion. Statistics that measure the amount of variation or spread in the data include the standard deviation, variance, range, minimum, maximum, and standard error of the mean.

- *Std. deviation*. A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
- *Variance*. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.
- *Range*. The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.
- *Minimum*. The smallest value of a numeric variable.
- *Maximum*. The largest value of a numeric variable.
- *S. E. mean*. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

Distribution. Skewness and kurtosis are statistics that describe the shape and symmetry of the distribution. These statistics are displayed with their standard errors.

- *Skewness*. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.
- *Kurtosis*. A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal distribution. The definition of kurtosis that is used, where the value is 0 for a normal distribution, is

sometimes referred to as excess kurtosis. Some software may report kurtosis such that the value is 3 for a normal distribution.

Values are group midpoints. If the values in your data are midpoints of groups (for example, ages of all people in their thirties are coded as 35), select this option to estimate the median and percentiles for the original, ungrouped data.

Frequencies Charts

Note: Charts are not produced in the output when **Perform bootstrapping** is enabled on the Bootstrap dialog.

Chart Type

A pie chart displays the contribution of parts to a whole. Each slice of a pie chart corresponds to a group that is defined by a single grouping variable. A bar chart displays the count for each distinct value or category as a separate bar, allowing you to compare categories visually. A histogram also has bars, but they are plotted along an equal interval scale. The height of each bar is the count of values of a quantitative variable falling within the interval. A histogram shows the shape, center, and spread of the distribution. A normal curve superimposed on a histogram helps you judge whether the data are normally distributed.

Chart Values

For bar charts, the scale axis can be labeled by frequency counts or percentages.

Frequencies Format

Order by. The frequency table can be arranged according to the actual values in the data or according to the count (frequency of occurrence) of those values, and the table can be arranged in either ascending or descending order. However, if you request a histogram or percentiles, Frequencies assumes that the variable is quantitative and displays its values in ascending order.

Multiple Variables. If you produce statistics tables for multiple variables, you can either display all variables in a single table (**Compare variables**) or display a separate statistics table for each variable (**Organize output by variables)**.

Suppress tables with many categories. This option prevents the display of tables with more than the specified number of values.

Descriptives

The Descriptives procedure displays univariate summary statistics for several variables in a single table and calculates standardized values (*z* scores). Variables can be ordered by the size of their means (in ascending or descending order), alphabetically, or by the order in which you select the variables (the default).

When *z* scores are saved, they are added to the data in the Data Editor and are available for charts, data listings, and analyses. When variables are recorded in different units (for example, gross domestic product per capita and percentage literate), a *z*-score transformation places variables on a common scale for easier visual comparison.

Example. If each case in your data contains the daily sales totals for each member of the sales staff (for example, one entry for Bob, one entry for Kim, and one entry for Brian) collected each day for several months, the Descriptives procedure can compute the average daily sales for each staff member and can order the results from highest average sales to lowest average sales.

Statistics. Sample size, mean, minimum, maximum, standard deviation, variance, range, sum, standard error of the mean, and kurtosis and skewness with their standard errors.

Descriptives Data Considerations

Data. Use numeric variables after you have screened them graphically for recording errors, outliers, and distributional anomalies. The Descriptives procedure is very efficient for large files (thousands of cases).

Assumptions. Most of the available statistics (including *z* scores) are based on normal theory and are appropriate for quantitative variables (interval- or ratio-level measurements) with symmetric distributions. Avoid variables with unordered categories or skewed distributions. The distribution of *z* scores has the same shape as that of the original data; therefore, calculating *z* scores is not a remedy for problem data.

To Obtain Descriptive Statistics

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Descriptives...**
2. Select one or more variables.

Optionally, you can:

- Select **Save standardized values as variables** to save *z* scores as new variables.
- Click **Options** for optional statistics and display order.

Descriptives Options

Mean and Sum. The mean, or arithmetic average, is displayed by default.

Dispersion. Statistics that measure the spread or variation in the data include the standard deviation, variance, range, minimum, maximum, and standard error of the mean.

- *Std. deviation*. A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
- *Variance*. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.
- *Range*. The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.
- *Minimum*. The smallest value of a numeric variable.
- *Maximum*. The largest value of a numeric variable.
- *S.E. mean*. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

Distribution. Kurtosis and skewness are statistics that characterize the shape and symmetry of the distribution. These statistics are displayed with their standard errors.

- *Kurtosis*. A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal distribution. The definition of kurtosis that is used, where the value is 0 for a normal distribution, is sometimes referred to as excess kurtosis. Some software may report kurtosis such that the value is 3 for a normal distribution.
- *Skewness*. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

Display Order. By default, the variables are displayed in the order in which you selected them. Optionally, you can display variables alphabetically, by ascending means, or by descending means.

DESCRIPTIVES Command Additional Features

The command syntax language also allows you to:

- Save standardized scores (*z* scores) for some but not all variables (with the `VARIABLES` subcommand).
- Specify names for new variables that contain standardized scores (with the `VARIABLES` subcommand).
- Exclude from the analysis cases with missing values for any variable (with the `MISSING` subcommand).
- Sort the variables in the display by the value of any statistic, not just the mean (with the `SORT` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Percentiles

The Percentiles procedure displays tables of percentiles.

Percentiles Values
When you select **Percentile Values**, by default displays the values for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. Values of a quantitative variable that divide the ordered data into groups so that a certain percentage is above and another percentage is below.

Quartiles
Quartiles (the 25th, 50th, and 75th percentiles) divide the observations into four groups of equal size.

Custom
When you select **Custom**, enter at least one value to run the analysis. The input values must be a number in the range 0 - 100. Use the **Add**, **Change**, and **Remove** buttons to work with the values in the percentile value list.

Percentile Method
By default, the HAVERAGE method is selected for computing percentiles.

Obtaining percentile tables

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Percentiles..**
2. Select one or more numeric variables and move them onto the Variables field.
3. Select **Percentile Values** to specify values to compute the percentile.
4. Select **Percentile Method** to select the percentile computing method.
5. Optionally, you can:
   - Click **Missing** to control the treatment of missing values.
   - Click **Bootstrap** to derive robust estimates of standard errors and confidence intervals for estimates of percentiles.

Percentiles Missing Values

Missing Values
Controls the treatment of missing values.

Exclude Cases Listwise

Cases with missing values for any specified variable are excluded from all analyses. This is the default.

By Variable
Exclude cases with missing values on a variable-by-variable basis.

Explore

The Explore procedure produces summary statistics and graphical displays, either for all of your cases or separately for groups of cases. There are many reasons for using the Explore procedure--data screening, outlier identification, description, assumption checking, and characterizing differences among subpopulations (groups of cases). Data screening may show that you have unusual values, extreme values, gaps in the data, or other peculiarities. Exploring the data can help to determine whether the statistical techniques that you are considering for data analysis are appropriate. The exploration may indicate that you need to transform the data if the technique requires a normal distribution. Or you may decide that you need nonparametric tests.

Example. Look at the distribution of maze-learning times for rats under four different reinforcement schedules. For each of the four groups, you can see if the distribution of times is approximately normal and whether the four variances are equal. You can also identify the cases with the five largest and five smallest times. The boxplots and stem-and-leaf plots graphically summarize the distribution of learning times for each of the groups.

Statistics and plots. Mean, median, 5% trimmed mean, standard error, variance, standard deviation, minimum, maximum, range, interquartile range, skewness and kurtosis and their standard errors, confidence interval for the mean (and specified confidence level), percentiles, Huber's M-estimator, Andrews' wave estimator, Hampel's redescending M-estimator, Tukey's biweight estimator, the five largest and five smallest values, the Kolmogorov-Smirnov statistic with a Lilliefors significance level for testing normality, and the Shapiro-Wilk statistic. Boxplots, stem-and-leaf plots, histograms, normality plots, and spread-versus-level plots with Levene tests and transformations.

Explore Data Considerations

Data. The Explore procedure can be used for quantitative variables (interval- or ratio-level measurements). A factor variable (used to break the data into groups of cases) should have a reasonable number of distinct values (categories). These values may be short string or numeric. The case label variable, used to label outliers in boxplots, can be short string, long string (first 15 bytes), or numeric.

Assumptions. The distribution of your data does not have to be symmetric or normal.

To Explore Your Data

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Explore...**
2. Select one or more dependent variables.

Optionally, you can:

- Select one or more factor variables, whose values will define groups of cases.
- Select an identification variable to label cases.
- Click **Statistics** for robust estimators, outliers, percentiles, and frequency tables.
- Click **Plots** for histograms, normal probability plots and tests, and spread-versus-level plots with Levene's statistics.
- Click **Options** for the treatment of missing values.

Explore Statistics

Descriptives. These measures of central tendency and dispersion are displayed by default. Measures of central tendency indicate the location of the distribution; they include the mean, median, and 5% trimmed mean. Measures of dispersion show the dissimilarity of the values; these include standard error, variance, standard deviation, minimum, maximum, range, and interquartile range. The descriptive statistics also include measures of the shape of the distribution; skewness and kurtosis are displayed with their standard errors. The 95% level confidence interval for the mean is also displayed; you can specify a different confidence level.

M-estimators. Robust alternatives to the sample mean and median for estimating the location. The estimators who are calculated differ in the weights that they apply to cases. Huber's M-estimator, Andrews' wave estimator, Hampel's redescending M-estimator, and Tukey's biweight estimator are displayed.

Outliers. Displays the five largest and five smallest values with case labels.

Percentiles

The Percentiles procedure displays tables of percentiles. When you select **Percentile**, by default displays the values for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. Values of a quantitative variable that divide the ordered data into groups so that a certain percentage is above and another percentage is below.

Quartiles

Quartiles (the 25th, 50th, and 75th percentiles) divide the observations into four groups of equal size.

Custom

When you select **Custom**, enter at least one value to run the analysis. The input values must be a number in the range 0 - 100. Use the **Add**, **Change**, and **Remove** buttons to work with the values in the percentile value list.

Method

By default, HAVERAGE method is selected for computing percentiles.

Explore Plots

Boxplots. These alternatives control the display of boxplots when you have more than one dependent variable. **Factor levels together** generates a separate display for each dependent variable. Within a display, boxplots are shown for each of the groups defined by a factor variable. **Dependents together** generates a separate display for each group defined by a factor variable. Within a display, boxplots are shown side by side for each dependent variable. This display is particularly useful when the different variables represent a single characteristic measured at different times.

Descriptive. The Descriptive group allows you to choose stem-and-leaf plots and histograms.

Normality plots with tests. Displays normal probability and detrended normal probability plots. The Kolmogorov-Smirnov statistic, with a Lilliefors significance level for testing normality, is displayed. If non-integer weights are specified, the Shapiro-Wilk statistic is calculated when the weighted sample size lies between 3 and 50. For no weights or integer weights, the statistic is calculated when the weighted sample size lies between 3 and 5,000.

Spread vs. Level with Levene Test. Controls data transformation for spread-versus-level plots. For all spread-versus-level plots, the slope of the regression line and Levene's robust tests for homogeneity of variance are displayed. If you select a transformation, Levene's tests are based on the transformed data. If no factor variable is selected, spread-versus-level plots are not produced. **Power estimation** produces a plot of the natural logs of the interquartile ranges against the natural logs of the medians for all cells, as well as an estimate of the power transformation for achieving equal variances in the cells. A spread-versus-level plot helps to determine the power for a transformation to stabilize (make more equal) variances across groups. **Transformed** allows you to select one of the power alternatives, perhaps following the recommendation from power estimation, and produces plots of transformed data. The interquartile range and median of the transformed data are plotted. **Untransformed** produces plots of the raw data. This is equivalent to a transformation with a power of 1.

Explore Power Transformations

These are the power transformations for spread-versus-level plots. To transform data, you must select a power for the transformation. You can choose one of the following alternatives:

- **Natural log.** Natural log transformation. This is the default.
- **1/square root.** For each data value, the reciprocal of the square root is calculated.
- **Reciprocal.** The reciprocal of each data value is calculated.

- **Square root.** The square root of each data value is calculated.
- **Square.** Each data value is squared.
- **Cube.** Each data value is cubed.

Explore Options

Missing Values. Controls the treatment of missing values.

- **Exclude cases listwise.** Cases with missing values for any dependent or factor variable are excluded from all analyses. This is the default.
- **Exclude cases pairwise.** Cases with no missing values for variables in a group (cell) are included in the analysis of that group. The case may have missing values for variables used in other groups.
- **Report values.** Missing values for factor variables are treated as a separate category. All output is produced for this additional category. Frequency tables include categories for missing values. Missing values for a factor variable are included but labeled as missing.

EXAMINE Command Additional Features

The Explore procedure uses `EXAMINE` command syntax. The command syntax language also allows you to:

- Request total output and plots in addition to output and plots for groups defined by the factor variables (with the `TOTAL` subcommand).
- Specify a common scale for a group of boxplots (with the `SCALE` subcommand).
- Specify interactions of the factor variables (with the `VARIABLES` subcommand).
- Specify percentiles other than the defaults (with the `PERCENTILES` subcommand).
- Calculate percentiles according to any of five methods (with the `PERCENTILES` subcommand).
- Specify any power transformation for spread-versus-level plots (with the `PLOT` subcommand).
- Specify the number of extreme values to be displayed (with the `STATISTICS` subcommand).
- Specify parameters for the M-estimators, robust estimators of location (with the `MESTIMATORS` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Crosstabs

The Crosstabs procedure forms two-way and multiway tables and provides a variety of tests and measures of association for two-way tables. The structure of the table and whether categories are ordered determine what test or measure to use.

With the exception of partial gamma coefficients, Crosstabs' statistics and measures of association are computed separately for each two-way table. If you specify a row, a column, and a layer factor (control variable), the Crosstabs procedure forms one panel of associated statistics and measures for each value of the layer factor (or a combination of values for two or more control variables). For example, if *gender* is a layer factor for a table of *married* (yes, no) against *life* (is life exciting, routine, or dull), the results for a two-way table for the females are computed separately from those for the males and printed as panels following one another.

Example. Are customers from small companies more likely to be profitable in sales of services (for example, training and consulting) than those from larger companies? From a crosstabulation, you might learn that the majority of small companies (fewer than 500 employees) yield high service profits, while the majority of large companies (more than 2,500 employees) yield low service profits.

Statistics and measures of association. Pearson chi-square, likelihood-ratio chi-square, linear-by-linear association test, Fisher's exact test, Yates' corrected chi-square, Pearson's *r*, Spearman's rho, contingency coefficient, phi, Cramér's *V*, symmetric and asymmetric lambdas, Goodman and Kruskal's tau, uncertainty coefficient, gamma, Somers' *d*, Kendall's tau-*b*, Kendall's tau-*c*, eta coefficient, Cohen's

kappa, relative risk estimate, odds ratio, McNemar test, Cochran's and Mantel-Haenszel statistics, and column proportions statistics.

Crosstabs Data Considerations

Data. To define the categories of each table variable, use values of a numeric or string variable. For example, for *gender*, you could code the data as 1 and 2 or as *male* and *female*.

Assumptions. Some statistics and measures assume ordered categories (ordinal data) or quantitative values (interval or ratio data), as discussed in the section on statistics. Others are valid when the table variables have unordered categories (nominal data). For the chi-square-based statistics (phi, Cramér's *V*, and contingency coefficient), the data should be a random sample from a multinomial distribution.

Note: Ordinal variables can be either numeric codes that represent categories (for example, 1 = *low*, 2 = *medium*, 3 = *high*) or string values. However, the alphabetic order of string values is assumed to reflect the true order of the categories. For example, for a string variable with the values of *low*, *medium*, *high*, the order of the categories is interpreted as *high*, *low*, *medium*--which is not the correct order. In general, it is more reliable to use numeric codes to represent ordinal data.

To Obtain Crosstabulations

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Crosstabs...**
2. Select one or more row variables and one or more column variables.

Optionally, you can:

- Select one or more control variables.
- Click **Statistics** for tests and measures of association for two-way tables or subtables.
- Click **Cells** for observed and expected values, percentages, and residuals.
- Click **Format** for controlling the order of categories.

Crosstabs layers

If you select one or more layer variables, a separate crosstabulation is produced for each category of each layer variable (control variable). For example, if you have one row variable, one column variable, and one layer variable with two categories, you get a two-way table for each category of the layer variable. To make another layer of control variables, click **Next**. Subtables are produced for each combination of categories for each first-layer variable, each second-layer variable, and so on. If statistics and measures of association are requested, they apply to two-way subtables only.

Crosstabs clustered bar charts

Display clustered bar charts. A clustered bar chart helps summarize your data for groups of cases. There is one cluster of bars for each value of the variable you specified under Rows. The variable that defines the bars within each cluster is the variable you specified under Columns. There is one set of differently colored or patterned bars for each value of this variable. If you specify more than one variable under Columns or Rows, a clustered bar chart is produced for each combination of two variables.

Crosstabs displaying layer variables in table layers

Display layer variables in table layers. You can choose to display the layer variables (control variables) as table layers in the crosstabulation table. This allows you to create views that show the overall statistics for row and column variables as well as permitting drill down on categories of layer variables.

An example that uses the data file *demo.sav* (available in the Samples directory of the installation directory) is shown below and was obtained as follows:

1. Select *Income category in thousands (inccat)* as the row variable, *Owns PDA (ownpda)* as the column variable and *Level of Education (ed)* as the layer variable.
2. Select **Display layer variables in table layers**.

3. Select **Column** in the Cell Display subdialog.
4. Run the Crosstabs procedure, double-click the crosstabulation table and select **College degree** from the Level of education drop down list.

The selected view of the crosstabulation table shows the statistics for respondents who have a college degree.

Crosstabs statistics

Chi-square. For tables with two rows and two columns, select **Chi-square** to calculate the Pearson chi-square, the likelihood-ratio chi-square, Fisher's exact test, and Yates' corrected chi-square (continuity correction). For 2 × 2 tables, Fisher's exact test is computed when a table that does not result from missing rows or columns in a larger table has a cell with an expected frequency of less than 5. Yates' corrected chi-square is computed for all other 2 × 2 tables. For tables with any number of rows and columns, select **Chi-square** to calculate the Pearson chi-square and the likelihood-ratio chi-square. When both table variables are quantitative, **Chi-square** yields the linear-by-linear association test.

Correlations. For tables in which both rows and columns contain ordered values, **Correlations** yields Spearman's correlation coefficient, rho (numeric data only). Spearman's rho is a measure of association between rank orders. When both table variables (factors) are quantitative, **Correlations** yields the Pearson correlation coefficient, *r*, a measure of linear association between the variables.

Nominal. For nominal data (no intrinsic order, such as Catholic, Protestant, and Jewish), you can select **Contingency coefficient**, **Phi** (coefficient) **and Cramér's V**, **Lambda** (symmetric and asymmetric lambdas and Goodman and Kruskal's tau), and **Uncertainty coefficient**.

- *Contingency coefficient*. A measure of association based on chi-square. The value ranges between 0 and 1, with 0 indicating no association between the row and column variables and values close to 1 indicating a high degree of association between the variables. The maximum value possible depends on the number of rows and columns in a table.
- *Phi and Cramer's V*. Phi is a chi-square-based measure of association that involves dividing the chi-square statistic by the sample size and taking the square root of the result. Cramer's V is a measure of association based on chi-square.
- *Lambda*. A measure of association that reflects the proportional reduction in error when values of the independent variable are used to predict values of the dependent variable. A value of 1 means that the independent variable perfectly predicts the dependent variable. A value of 0 means that the independent variable is no help in predicting the dependent variable.
- *Uncertainty coefficient*. A measure of association that indicates the proportional reduction in error when values of one variable are used to predict values of the other variable. For example, a value of 0.83 indicates that knowledge of one variable reduces error in predicting values of the other variable by 83%. The program calculates both symmetric and asymmetric versions of the uncertainty coefficient.

Ordinal. For tables in which both rows and columns contain ordered values, select **Gamma** (zero-order for 2-way tables and conditional for 3-way to 10-way tables), **Kendall's tau-b**, and **Kendall's tau-c**. For predicting column categories from row categories, select **Somers' d**.

- *Gamma*. A symmetric measure of association between two ordinal variables that ranges between -1 and 1. Values close to an absolute value of 1 indicate a strong relationship between the two variables. Values close to 0 indicate little or no relationship. For 2-way tables, zero-order gammas are displayed. For 3-way to n-way tables, conditional gammas are displayed.
- *Somers' d*. A measure of association between two ordinal variables that ranges from -1 to 1. Values close to an absolute value of 1 indicate a strong relationship between the two variables, and values close to 0 indicate little or no relationship between the variables. Somers' d is an asymmetric extension of gamma that differs only in the inclusion of the number of pairs not tied on the independent variable. A symmetric version of this statistic is also calculated.
- *Kendall's tau-b*. A nonparametric measure of correlation for ordinal or ranked variables that take ties into account. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. Possible values range from -1 to 1, but a value of -1 or +1 can be obtained only from square tables.

- *Kendall's tau-c*. A nonparametric measure of association for ordinal variables that ignores ties. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. Possible values range from -1 to 1, but a value of -1 or +1 can be obtained only from square tables.

Nominal by Interval. When one variable is categorical and the other is quantitative, select **Eta**. The categorical variable must be coded numerically.

- *Eta*. A measure of association that ranges from 0 to 1, with 0 indicating no association between the row and column variables and values close to 1 indicating a high degree of association. Eta is appropriate for a dependent variable measured on an interval scale (for example, income) and an independent variable with a limited number of categories (for example, gender). Two eta values are computed: one treats the row variable as the interval variable, and the other treats the column variable as the interval variable.

Kappa. Cohen's kappa measures the agreement between the evaluations of two raters when both are rating the same object. A value of 1 indicates perfect agreement. A value of 0 indicates that agreement is no better than chance. Kappa is based on a square table in which row and column values represent the same scale. Any cell that has observed values for one variable but not the other is assigned a count of 0. Kappa is not computed if the data storage type (string or numeric) is not the same for the two variables. For string variable, both variables must have the same defined length.

Risk. For 2 x 2 tables, a measure of the strength of the association between the presence of a factor and the occurrence of an event. If the confidence interval for the statistic includes a value of 1, you cannot assume that the factor is associated with the event. The odds ratio can be used as an estimate or relative risk when the occurrence of the factor is rare.

McNemar. A nonparametric test for two related dichotomous variables. Tests for changes in responses using the chi-square distribution. Useful for detecting changes in responses due to experimental intervention in "before-and-after" designs. For larger square tables, the McNemar-Bowker test of symmetry is reported.

Cochran's and Mantel-Haenszel statistics. Cochran's and Mantel-Haenszel statistics can be used to test for independence between a dichotomous factor variable and a dichotomous response variable, conditional upon covariate patterns defined by one or more layer (control) variables. Note that while other statistics are computed layer by layer, the Cochran's and Mantel-Haenszel statistics are computed once for all layers.

Crosstabs cell display

To help you uncover patterns in the data that contribute to a significant chi-square test, the Crosstabs procedure displays expected frequencies and three types of residuals (deviates) that measure the difference between observed and expected frequencies. Each cell of the table can contain any combination of counts, percentages, and residuals selected.

Counts. The number of cases actually observed and the number of cases expected if the row and column variables are independent of each other. You can choose to hide counts that are less than a specified integer. Hidden values will be displayed as **<N**, where **N** is the specified integer. The specified integer must be greater than or equal to 2, although the value 0 is permitted and specifies that no counts are hidden.

Compare column proportions. This option computes pairwise comparisons of column proportions and indicates which pairs of columns (for a given row) are significantly different. Significant differences are indicated in the crosstabulation table with APA-style formatting using subscript letters and are calculated at the 0.05 significance level. *Note*: If this option is specified without selecting observed counts or column percentages, then observed counts are included in the crosstabulation table, with the APA-style subscript letters indicating the results of the column proportions tests.

- **Adjust p-values (Bonferroni method).** Pairwise comparisons of column proportions make use of the Bonferroni correction, which adjusts the observed significance level for the fact that multiple comparisons are made.

Percentages. The percentages can add up across the rows or down the columns. The percentages of the total number of cases represented in the table (one layer) are also available.

Note: If **Hide small counts** is selected in the Counts group, then percentages associated with hidden counts are also hidden.

Residuals. Raw unstandardized residuals give the difference between the observed and expected values. Standardized and adjusted standardized residuals are also available.

- *Unstandardized*. The difference between an observed value and the expected value. The expected value is the number of cases you would expect in the cell if there were no relationship between the two variables. A positive residual indicates that there are more cases in the cell than there would be if the row and column variables were independent.
- *Standardized*. The residual divided by an estimate of its standard deviation. Standardized residuals, which are also known as Pearson residuals, have a mean of 0 and a standard deviation of 1.
- *Adjusted standardized*. The residual for a cell (observed minus expected value) divided by an estimate of its standard error. The resulting standardized residual is expressed in standard deviation units above or below the mean.

Create APA style table. Creates output tables that adhere to APA style guidelines.

Note: The **Observed**, **Expected**, **Row**, **Column**, and **Total** options are not available when **Create APA style table** is selected.

Noninteger Weights. Cell counts are normally integer values, since they represent the number of cases in each cell. But if the data file is currently weighted by a weight variable with fractional values (for example, 1.25), cell counts can also be fractional values. You can truncate or round either before or after calculating the cell counts or use fractional cell counts for both table display and statistical calculations.

- *Round cell counts*. Case weights are used as is but the accumulated weights in the cells are rounded before computing any statistics.
- *Truncate cell counts*. Case weights are used as is but the accumulated weights in the cells are truncated before computing any statistics.
- *Round case weights*. Case weights are rounded before use.
- *Truncate case weights*. Case weights are truncated before use.
- *No adjustments*. Case weights are used as is and fractional cell counts are used. However, when Exact Statistics (available only with Sampling and Testing) are requested, the accumulated weights in the cells are either truncated or rounded before computing the Exact test statistics.

Crosstabs table format

You can arrange rows in ascending or descending order of the values of the row variable.

Summarize

The Summarize procedure calculates subgroup statistics for variables within categories of one or more grouping variables. All levels of the grouping variable are crosstabulated. You can choose the order in which the statistics are displayed. Summary statistics for each variable across all categories are also displayed. Data values in each category can be listed or suppressed. With large datasets, you can choose to list only the first *n* cases.

Example. What is the average product sales amount by region and customer industry? You might discover that the average sales amount is slightly higher in the western region than in other regions, with corporate customers in the western region yielding the highest average sales amount.

Statistics. Sum, number of cases, mean, median, grouped median, standard error of the mean, minimum, maximum, range, variable value of the first category of the grouping variable, variable value of the last category of the grouping variable, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of total sum, percentage of total *N*, percentage of sum in, percentage of *N* in, geometric mean, and harmonic mean.

Summarize Data Considerations

Data. Grouping variables are categorical variables whose values can be numeric or string. The number of categories should be reasonably small. The other variables should be able to be ranked.

Assumptions. Some of the optional subgroup statistics, such as the mean and standard deviation, are based on normal theory and are appropriate for quantitative variables with symmetric distributions. Robust statistics, such as the median and the range, are appropriate for quantitative variables that may or may not meet the assumption of normality.

To Obtain Case Summaries

1. From the menus choose:

   Analyze > **Reports** > **Case Summaries...**
2. Select one or more variables.

Optionally, you can:

- Select one or more grouping variables to divide your data into subgroups.
- Click **Options** to change the output title, add a caption below the output, or exclude cases with missing values.
- Click **Statistics** for optional statistics.
- Select **Display cases** to list the cases in each subgroup. By default, the system lists only the first 100 cases in your file. You can raise or lower the value for **Limit cases to first** *n* or deselect that item to list all cases.

Summarize Options

Summarize allows you to change the title of your output or add a caption that will appear below the output table. You can control line wrapping in titles and captions by typing \n wherever you want to insert a line break in the text.

You can also choose to display or suppress subheadings for totals and to include or exclude cases with missing values for any of the variables used in any of the analyses. Often it is desirable to denote missing cases in output with a period or an asterisk. Enter a character, phrase, or code that you would like to have appear when a value is missing; otherwise, no special treatment is applied to missing cases in the output.

Summarize Statistics

You can choose one or more of the following subgroup statistics for the variables within each category of each grouping variable: sum, number of cases, mean, median, grouped median, standard error of the mean, minimum, maximum, range, variable value of the first category of the grouping variable, variable value of the last category of the grouping variable, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of total sum, percentage of total *N*, percentage of sum in, percentage of *N* in, geometric mean, harmonic mean. The order in which the statistics appear in the Cell Statistics list is the order in which they will be displayed in the output. Summary statistics are also displayed for each variable across all categories.

First. Displays the first data value encountered in the data file.

Geometric Mean. The nth root of the product of the data values, where n represents the number of cases.

Grouped Median. Median that is calculated for data that is coded into groups. For example, with age data, if each value in the 30s is coded 35, each value in the 40s is coded 45, and so on, the grouped median is the median calculated from the coded data.

Harmonic Mean. Used to estimate an average group size when the sample sizes in the groups are not equal. The harmonic mean is the total number of samples divided by the sum of the reciprocals of the sample sizes.

Kurtosis. A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal

distribution. The definition of kurtosis that is used, where the value is 0 for a normal distribution, is sometimes referred to as excess kurtosis. Some software may report kurtosis such that the value is 3 for a normal distribution.

Last. Displays the last data value encountered in the data file.

Maximum. The largest value of a numeric variable.

Mean. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.

Median. The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).

Minimum. The smallest value of a numeric variable.

N. The number of cases (observations or records).

Percent of Total N. Percentage of the total number of cases in each category.

Percent of Total Sum. Percentage of the total sum in each category.

Range. The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.

Skewness. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

Standard Deviation. A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.

Standard Error of Kurtosis. The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for kurtosis indicates that the tails of the distribution are longer than those of a normal distribution; a negative value for kurtosis indicates shorter tails (becoming like those of a box-shaped uniform distribution).

Standard Error of Mean. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

Standard Error of Skewness. The ratio of skewness to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for skewness indicates a long right tail; an extreme negative value indicates a long left tail.

Sum. The sum or total of the values, across all cases with nonmissing values.

Variance. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.

Means

The Means procedure calculates subgroup means and related univariate statistics for dependent variables within categories of one or more independent variables. Optionally, you can obtain a one-way analysis of variance, eta, and tests for linearity.

Example. Measure the average amount of fat absorbed by three different types of cooking oil, and perform a one-way analysis of variance to see whether the means differ.

Statistics. Sum, number of cases, mean, median, grouped median, standard error of the mean, minimum, maximum, range, variable value of the first category of the grouping variable, variable value of the last category of the grouping variable, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of total sum, percentage of total *N*, percentage of sum in, percentage of *N* in, geometric mean, and harmonic mean. Options include analysis of variance, eta, eta squared, and tests for linearity *R* and R^2.

Means Data Considerations

Data. The dependent variables are quantitative, and the independent variables are categorical. The values of categorical variables can be numeric or string.

Assumptions. Some of the optional subgroup statistics, such as the mean and standard deviation, are based on normal theory and are appropriate for quantitative variables with symmetric distributions. Robust statistics, such as the median, are appropriate for quantitative variables that may or may not meet the assumption of normality. Analysis of variance is robust to departures from normality, but the data in each cell should be symmetric. Analysis of variance also assumes that the groups come from populations with equal variances. To test this assumption, use Levene's homogeneity-of-variance test, available in the One-Way ANOVA procedure.

To Obtain Subgroup Means

1. From the menus choose:

   Analyze > **Compare Means** > **Means...**
2. Select one or more dependent variables.
3. Use one of the following methods to select categorical independent variables:
   - Select one or more independent variables. Separate results are displayed for each independent variable.
   - Select one or more layers of independent variables. Each layer further subdivides the sample. If you have one independent variable in Layer 1 and one independent variable in Layer 2, the results are displayed in one crossed table, as opposed to separate tables for each independent variable.
4. Optionally, click **Options** for optional statistics, an analysis of variance table, eta, eta squared, *R*, and R^2.

Means Options

You can choose one or more of the following subgroup statistics for the variables within each category of each grouping variable: sum, number of cases, mean, median, grouped median, standard error of the mean, minimum, maximum, range, variable value of the first category of the grouping variable, variable value of the last category of the grouping variable, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of total sum, percentage of total *N*, percentage of sum in, percentage of *N* in, geometric mean, and harmonic mean. You can change the order in which the subgroup statistics appear. The order in which the statistics appear in the Cell Statistics list is the order in which they are displayed in the output. Summary statistics are also displayed for each variable across all categories.

First. Displays the first data value encountered in the data file.

Geometric Mean. The nth root of the product of the data values, where n represents the number of cases.

Grouped Median. Median that is calculated for data that is coded into groups. For example, with age data, if each value in the 30s is coded 35, each value in the 40s is coded 45, and so on, the grouped median is the median calculated from the coded data.

Harmonic Mean. Used to estimate an average group size when the sample sizes in the groups are not equal. The harmonic mean is the total number of samples divided by the sum of the reciprocals of the sample sizes.

Kurtosis. A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a

normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal distribution. The definition of kurtosis that is used, where the value is 0 for a normal distribution, is sometimes referred to as excess kurtosis. Some software may report kurtosis such that the value is 3 for a normal distribution.

Last. Displays the last data value encountered in the data file.

Maximum. The largest value of a numeric variable.

Mean. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.

Median. The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).

Minimum. The smallest value of a numeric variable.

N. The number of cases (observations or records).

Percent of total N. Percentage of the total number of cases in each category.

Percent of total sum. Percentage of the total sum in each category.

Range. The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.

Skewness. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

Standard Deviation. A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.

Standard Error of Kurtosis. The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for kurtosis indicates that the tails of the distribution are longer than those of a normal distribution; a negative value for kurtosis indicates shorter tails (becoming like those of a box-shaped uniform distribution).

Standard Error of Mean. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

Standard Error of Skewness. The ratio of skewness to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for skewness indicates a long right tail; an extreme negative value indicates a long left tail.

Sum. The sum or total of the values, across all cases with nonmissing values.

Variance. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.

Statistics for First Layer

Anova table and eta. Displays a one-way analysis-of-variance table and calculates eta and eta-squared (measures of association) for each independent variable in the first layer.

Test for linearity. Calculates the sum of squares, degrees of freedom, and mean square associated with linear and nonlinear components, as well as the F ratio, R and R-squared. Linearity is not calculated if the independent variable is a short string.

OLAP Cubes

The OLAP (Online Analytical Processing) Cubes procedure calculates totals, means, and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.

Example. Total and average sales for different regions and product lines within regions.

Statistics. Sum, number of cases, mean, median, grouped median, standard error of the mean, minimum, maximum, range, variable value of the first category of the grouping variable, variable value of the last category of the grouping variable, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of total cases, percentage of total sum, percentage of total cases within grouping variables, percentage of total sum within grouping variables, geometric mean, and harmonic mean.

OLAP Cubes Data Considerations

Data. The summary variables are quantitative (continuous variables measured on an interval or ratio scale), and the grouping variables are categorical. The values of categorical variables can be numeric or string.

Assumptions. Some of the optional subgroup statistics, such as the mean and standard deviation, are based on normal theory and are appropriate for quantitative variables with symmetric distributions. Robust statistics, such as the median and range, are appropriate for quantitative variables that may or may not meet the assumption of normality.

To Obtain OLAP Cubes

1. From the menus choose:

   Analyze > **Reports** > **OLAP Cubes...**
2. Select one or more continuous summary variables.
3. Select one or more categorical grouping variables.

Optionally:

- Select different summary statistics (click **Statistics**). You must select one or more grouping variables before you can select summary statistics.
- Calculate differences between pairs of variables and pairs of groups that are defined by a grouping variable (click **Differences**).
- Create custom table titles (click **Title**).
- Hide counts that are less than a specified integer. Hidden values will be displayed as **<N**, where **N** is the specified integer. The specified integer must be greater than or equal to 2.

OLAP Cubes Statistics

You can choose one or more of the following subgroup statistics for the summary variables within each category of each grouping variable: sum, number of cases, mean, median, grouped median, standard error of the mean, minimum, maximum, range, variable value of the first category of the grouping variable, variable value of the last category of the grouping variable, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of total cases, percentage of total sum, percentage of total cases within grouping variables, percentage of total sum within grouping variables, geometric mean, and harmonic mean.

You can change the order in which the subgroup statistics appear. The order in which the statistics appear in the Cell Statistics list is the order in which they are displayed in the output. Summary statistics are also displayed for each variable across all categories.

First. Displays the first data value encountered in the data file.

Geometric Mean. The nth root of the product of the data values, where n represents the number of cases.

Grouped Median. Median that is calculated for data that is coded into groups. For example, with age data, if each value in the 30s is coded 35, each value in the 40s is coded 45, and so on, the grouped median is the median calculated from the coded data.

Harmonic Mean. Used to estimate an average group size when the sample sizes in the groups are not equal. The harmonic mean is the total number of samples divided by the sum of the reciprocals of the sample sizes.

Kurtosis. A measure of the extent to which there are outliers. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the data exhibit more extreme outliers than a normal distribution. Negative kurtosis indicates that the data exhibit less extreme outliers than a normal distribution. The definition of kurtosis that is used, where the value is 0 for a normal distribution, is sometimes referred to as excess kurtosis. Some software may report kurtosis such that the value is 3 for a normal distribution.

Last. Displays the last data value encountered in the data file.

Maximum. The largest value of a numeric variable.

Mean. A measure of central tendency. The arithmetic average, the sum divided by the number of cases.

Median. The value above and below which half of the cases fall, the 50th percentile. If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).

Minimum. The smallest value of a numeric variable.

N. The number of cases (observations or records).

Percent of N in. Percentage of the number of cases for the specified grouping variable within categories of other grouping variables. If you only have one grouping variable, this value is identical to percentage of total number of cases.

Percent of Sum in. Percentage of the sum for the specified grouping variable within categories of other grouping variables. If you only have one grouping variable, this value is identical to percentage of total sum.

Percent of Total N. Percentage of the total number of cases in each category.

Percent of Total Sum. Percentage of the total sum in each category.

Range. The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.

Skewness. A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

Standard Deviation. A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.

Standard Error of Kurtosis. The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for kurtosis indicates that the tails of the distribution are longer than those of a normal distribution; a negative value for kurtosis indicates shorter tails (becoming like those of a box-shaped uniform distribution).

Standard Error of Mean. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude the two values are different if the ratio of the difference to the standard error is less than -2 or greater than +2).

Standard Error of Skewness. The ratio of skewness to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2). A large positive value for skewness indicates a long right tail; an extreme negative value indicates a long left tail.

Sum. The sum or total of the values, across all cases with nonmissing values.

Variance. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.

OLAP Cubes Differences

This dialog box allows you to calculate percentage and arithmetic differences between summary variables or between groups that are defined by a grouping variable. Differences are calculated for all measures that are selected in the OLAP Cubes Statistics dialog box.

Differences between Variables. Calculates differences between pairs of variables. Summary statistics values for the second variable (the Minus variable) in each pair are subtracted from summary statistics values for the first variable in the pair. For percentage differences, the value of the summary variable for the Minus variable is used as the denominator. You must select at least two summary variables in the main dialog box before you can specify differences between variables.

Differences between Groups of Cases. Calculates differences between pairs of groups defined by a grouping variable. Summary statistics values for the second category in each pair (the Minus category) are subtracted from summary statistics values for the first category in the pair. Percentage differences use the value of the summary statistic for the Minus category as the denominator. You must select one or more grouping variables in the main dialog box before you can specify differences between groups.

OLAP Cubes Title

You can change the title of your output or add a caption that will appear below the output table. You can also control line wrapping of titles and captions by typing \n wherever you want to insert a line break in the text.

Proportions

Proportions introduction

Proportions computes tests and confidence intervals for binomial proportions or differences of proportions. Statistics are available for one-sample proportions (tested against a specified value), paired-samples (different variables), or independent samples (different groups of cases). Various options for types of test statistics and confidence intervals are offered. Other procedures that provide some of the same functionality include **CROSSTABS**, **NPAR TESTS**, and **NPTESTS**.

One-Sample Proportions
One-sample tests and confidence intervals. Output includes the observed proportion, the estimate of the difference between the population proportion and the hypothesized population proportion, asymptotic standard errors under null and alternative hypotheses, specified test statistics with two-sided probabilities, and specified confidence intervals for proportions.

Paired-Samples Proportions
Paired-samples tests and confidence intervals for differences among proportions. Output includes observed proportions, estimates of differences in population proportions, asymptotic standard errors of population differences under null and alternative hypotheses, specified test statistics with two-sided probabilities, and specified confidence intervals for differences in proportions.

Independent-Samples Proportions
Independent-samples tests and confidence intervals. Output includes observed proportions, estimates of differences in population proportions, asymptotic standard errors of population differences under null and alternative hypotheses, specified test statistics with two-sided probabilities, and specified confidence intervals for differences in proportions.

One-Sample Proportions

The One-Sample Proportions procedure provides tests and confidence intervals for individual binomial proportions. The data are assumed to be from a simple random sample, and each hypothesis test or confidence interval is a separate test or individual interval, based on a binomial proportion. Output includes the observed proportion, the estimate of the difference between the population proportion and the hypothesized population proportion, asymptotic standard errors under null and alternative hypotheses, specified test statistics with two-sided probabilities, and specified confidence intervals for proportions.

Example

Statistics

Agresti-Coull, Anscombe, Clopper-Pearson (Exact), Jeffreys, Logit, Wald, Wald (continuity corrected), Wilson Score, Wilson Score (continuity corrected), Exact Binomial, Mid-p Adjusted Binomial, Score, Score (continuity corrected).

Data considerations

Data

The procedure displays requested test statistics and two-sided probabilities, confidence intervals for differences of proportions, as well as proportions, standard errors, and counts for each group or variable. The procedure is limited to maximum of one test value.

Assumptions

Obtaining One-Sample Proportions tests

1. From the menus choose:

   Analyze > **Compare Means** > **One-Sample Proportions...**
2. Select one or more quantitative test variables.
3. Optionally, you can:
   - Select success criteria settings under the **Define Success** section:

     Last Value
     The last or highest value among the sorted distinct values in the data is used. This applies to numeric or string variables. This is the default setting.

     First Value
     The first or lowest value among the sorted distinct values in the data is used. This applies to numeric or string variables.

     Value(s)
     One or more parenthesized specific values. Multiple values must be separated by spaces. This applies to numeric or string variables. String variable values should be enclosed in single quotes.

     Midpoint
     Values at or above the middle of the range of observed values in the data. This applies only to numeric data.

     Cut Point
     Values at or above a specified value. This applies only to numeric data.
   - Click **Confidence Intervals...** to specify which types of confidence intervals are displayed, or to suppress all confidence intervals.
   - Click **Tests...** to specify which types of test statistics are displayed, or to suppress all tests.
   - Click **Missing Values...** to control the treatment of missing data.
   - Click **Bootstrap...** for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.
4. Click **OK**.

One-Sample Proportions: Confidence Intervals

The Confidence Intervals dialog provides options for specifying the coverage level and for selecting which types of confidence intervals are displayed.

Coverage Level
Specifies the confidence interval percentage. A numeric value in the range (0,100) must be specified. 95 is the default setting.

Interval Type(s)
Provides options for specifying which types of confidence intervals are displayed. Available options include:

- Agresti-Coull
- Anscombe
- Clopper-Pearson (Exact)
- Jeffreys
- Logit
- Wald
- Wald (Continuity Corrected)
- Wilson Score
- Wilson Score (Continuity Corrected)

Specifying Confidence Intervals for One-Sample Proportions

1. From the menus choose:

   Analyze > **Compare Means** > **One-Sample Proportions...**
2. Click **Confidence Intervals** to specify which types of confidence intervals are displayed, or to suppress all confidence intervals.

One-Sample Proportions: Tests

The Tests dialog provides options for specifying which types of test statistics are displayed.

All
All test statistics display in the output.

None
No test statistics display in the output.

Exact Binomial
Displays exact binomial probabilities.

Mid-p Adjusted Binomial
Displays mid-p adjusted binomial probabilities. This is a default setting.

Score
Displays the Score Z test statistic. This is a default setting.

Score (Continuity Corrected)
Displays the continuity-corrected score Z test statistic.

Wald
Displays the Wald Z test statistic.

Wald (Continuity Corrected)
Displays the continuity-corrected Wald Z test statistic.

Test Value
Specifies a test value between 0 and 1. The default value is 0.5.

Obtaining One-Sample Proportions tests

1. From the menus choose:

   Analyze > **Compare Means** > **One-Sample Proportions...**
2. In the One-Sample Proportions dialog, click **Tests**.
3. Select one or more of the available tests.

One-Sample Proportions: Missing Values

The Missing Values dialog provides options for dealing with missing values.

Missing Data Scope

Exclude cases analysis by analysis
Indicates inclusion of all cases with sufficient data on the variables used in each particular analysis. This is the default setting.

Exclude cases listwise
Indicates inclusion of all cases with sufficient data on all variables used across all analyses.

User Missing Values
Exclude treats user missing values as missing. **Include** ignores user missing value designations and treats user missing values as valid.

Defining missing value settings for One-Sample Proportions

1. From the menus choose:

   Analyze > **Compare Means** > **One-Sample Proportions...**
2. In the One-Sample Proportions dialog, click **Missing Values**.
3. Select the desired missing values settings.

Paired-Samples Proportions

The Paired-Samples Proportions procedure provides tests and confidence intervals for the difference in two related or paired binomial proportions. The data are assumed to be from a simple random sample, and each hypothesis test or confidence interval is a separate test or individual interval. Output includes observed proportions, estimates of differences in population proportions, asymptotic standard errors of population differences under null and alternative hypotheses, specified test statistics with two-sided probabilities, and specified confidence intervals for differences in proportions.

Example

Statistics
Agresti-Min, Bonett-Price, Newcombe, Wald, Wald (continuity corrected), Exact Binomial, Mid-p Adjusted Binomial, McNemar, McNemar (continuity corrected).

Data Considerations

Data

- A variable list containing at least two variables is required.
- If a single list of variables is specified, each member of the list is paired with every other member of the list.

Assumptions

- If two lists of variables are separated by **WITH** without the (**PAIRED**) keyword, each member of the first list is paired with each member of the second list.
- If two lists of variables are separated by **WITH** and the second list is followed by (**PAIRED**), members of the two lists in order are paired: the first member of the first list is paired with the first

member of the second list, the second members of each list are paired, etc. Unmatched variables are ignored and a warning message is issued.

Obtaining Paired-Samples Proportions tests

1. From the menus choose:

   Analyze > **Compare Means** > **Paired-Samples Proportions...**
2. Select one or more quantitative test variables.
3. Optionally, you can:
   - Select success criteria settings under the **Define Success** section:

     Last Value
     The last or highest value among the sorted distinct values in the data is used. This applies to numeric or string variables. This is the default setting.

     First Value
     The first or lowest value among the sorted distinct values in the data is used. This applies to numeric or string variables.

     Value(s)
     One or more parenthesized specific values. Multiple values must be separated by spaces. This applies to numeric or string variables. String variable values should be enclosed in single quotes.

     Midpoint
     Values at or above the middle of the range of observed values in the data. This applies only to numeric data.

     Cut Point
     Values at or above a specified value. This applies only to numeric data.
   - Click **Confidence Intervals...** to specify which types of confidence intervals are displayed, or to suppress all confidence intervals.
   - Click **Tests...** to specify which types of test statistics are displayed, or to suppress all tests.
   - Click **Missing Values...** to control the treatment of missing data.
   - Click **Bootstrap...** for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.
4. Click **OK**.

Paired-Samples Proportions: Confidence Intervals

The Confidence Intervals dialog provides options for specifying the coverage level and for selecting which types of confidence intervals are displayed.

Coverage Level
Specifies the confidence interval percentage. A numeric value in the range (0,100) must be specified. 95 is the default setting.

Interval Type(s)
Provides options for specifying which types of confidence intervals are displayed. Available options include:

- Agresti-Min
- Bonett-Price
- Newcombe
- Wald
- Wald (Continuity Corrected)

Specifying Confidence Intervals for Paired-Samples Proportions

1. From the menus choose:

   Analyze > **Compare Means** > **Paired-Samples Proportions...**

2. Click **Confidence Intervals** to specify which types of confidence intervals are displayed, or to suppress all confidence intervals.

Paired-Samples Proportions: Tests

The Tests dialog provides options for specifying which types of test statistics are displayed.

All
All test statistics display in the output.

None
No test statistics display in the output.

Exact Binomial
Displays exact binomial probabilities.

Mid-p Adjusted Binomial
Displays mid-p adjusted binomial probabilities. This is a default setting.

McNemar
Displays the McNemar Z test statistic. This is a default setting.

McNemar (Continuity Corrected)
Displays the continuity-corrected McNemar Z test statistic.

Wald
Displays the Wald Z test statistic.

Wald (Continuity Corrected)
Displays the continuity-corrected Wald Z test statistic.

Obtaining Paired-Samples Proportions tests

1. From the menus choose:

   Analyze > **Compare Means** > **Paired-Samples Proportions...**

2. In the Paired-Samples Proportions dialog, click **Tests**.
3. Select one or more of the available tests.

Paired-Samples Proportions: Missing Values

The Missing Values dialog provides options for dealing with missing values.

Missing Data Scope

Exclude cases analysis by analysis
Indicates inclusion of all cases with sufficient data on the variables used in each particular analysis. This is the default setting.

Exclude cases listwise
Indicates inclusion of all cases with sufficient data on all variables used across all analyses.

User Missing Values
Exclude treats user missing values as missing. **Include** ignores user missing value designations and treats user missing values as valid.

Defining missing value settings for Paired-Samples Proportions

1. From the menus choose:

   Analyze > **Compare Means** > **Paired-Samples Proportions...**

2. In the Paired-Samples Proportions dialog, click **Missing Values**.
3. Select the desired missing values settings.

Independent-Samples Proportions

The Independent-Samples Proportions procedure provides tests and confidence intervals for the difference in two independent binomial proportions. The data are assumed to be from a simple random sample, and each hypothesis test or confidence interval is a separate test or individual interval. Output includes observed proportions, estimates of differences in population proportions, asymptotic standard errors of population differences under null and alternative hypotheses, specified test statistics with two-sided probabilities, and specified confidence intervals for differences in proportions.

Example

Statistics

Agresti-Min, Bonett-Price, Newcombe, Wald, Wald (continuity corrected), Exact Binomial, Mid-p Adjusted Binomial, McNemar, McNemar (continuity corrected).

Data Considerations

Data

- At least one dependent variable and a single variable to identify the two groups to be compared are required.
- The grouping variable can be either numeric or string.

Assumptions

Obtaining Independent-Samples Proportions tests

1. From the menus choose:

   Analyze > **Compare Means** > **Independent-Samples Proportions...**
2. Select one or more quantitative test variables.
3. Select a single **Grouping Variable** that identifies the two groups to be compared.
4. Optionally, specify settings for the selected **Grouping Variable**.
   - When **Value(s)** is selected, you can specify two numeric or string values within parentheses for the values to be compared. String values should be enclosed in single quotes. Cases with other values are ignored.
   - **Midpoint** applies only to numeric variables. Cases at or above the midpoint of the distribution of the grouping variable are assigned to the second group, cases below the midpoint are assigned to the first group.
   - **Cut Point** applies only the numeric variables and allows specification with parentheses of a single numeric value. Cases at or above the cut point on the grouping variable are assigned to the second group, cases below the cut point are assigned to the first group.
5. Optionally, you can:
   - Select success criteria settings under the **Define Success** section:

     Last Value
     The last or highest value among the sorted distinct values in the data is used. This applies to numeric or string variables. This is the default setting.

     First Value
     The first or lowest value among the sorted distinct values in the data is used. This applies to numeric or string variables.

     Value(s)
     One or more parenthesized specific values. Multiple values must be separated by spaces. This applies to numeric or string variables. String variable values should be enclosed in single quotes.

Midpoint
Values at or above the middle of the range of observed values in the data. This applies only to numeric data.

Cut Point
Values at or above a specified value. This applies only to numeric data.

- Click **Confidence Intervals...** to specify which types of confidence intervals are displayed, or to suppress all confidence intervals.
- Click **Tests...** to specify which types of test statistics are displayed, or to suppress all tests.
- Click **Missing Values...** to control the treatment of missing data.
- Click **Bootstrap...** for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.

6. Click **OK**.

Independent-Samples Proportions: Confidence Intervals

The Confidence Intervals dialog provides options for specifying the coverage level and for selecting which types of confidence intervals are displayed.

Coverage Level
Specifies the confidence interval percentage. A numeric value in the range (0,100) must be specified. 95 is the default setting.

Interval Type(s)
Provides options for specifying which types of confidence intervals are displayed. Available options include:

- Agresti-Caffo
- Brown-Li-Jeffreys
- Hauck-Anderson
- Newcombe
- Newcombe (Continuity Corrected)
- Wald
- Wald (Continuity Corrected)

Specifying Confidence Intervals for Independent-Samples Proportions

1. From the menus choose:

   Analyze > **Compare Means** > **Independent-Samples Proportions...**

2. Click **Confidence Intervals** to specify which types of confidence intervals are displayed, or to suppress all confidence intervals.

Independent-Samples Proportions: Tests

The Tests dialog provides options for specifying which types of test statistics are displayed.

All
All test statistics display in the output.

None
No test statistics display in the output.

Hauck-Anderson
Displays the Hauck-Anderson Z test statistic.

Wald
Displays the Wald Z test statistic.

Wald (Continuity Corrected)
Displays the continuity-corrected Wald Z test statistic.

Wald H0
Displays the Wald Z test statistic using variance estimates under H_0.

Wald H0 (Continuity Corrected)
Displays the continuity-corrected Wald Z test statistic using variance estimates under H_0.

Obtaining Independent-Samples Proportions tests

1. From the menus choose:

   Analyze > **Compare Means** > **Independent-Samples Proportions...**
2. In the Independent-Samples Proportions dialog, click **Tests**.
3. Select one or more of the available tests.

Independent-Samples Proportions: Missing Values

The Missing Values dialog provides options for dealing with missing values.

Missing Data Scope

Exclude cases analysis by analysis
Indicates inclusion of all cases with sufficient data on the variables used in each particular analysis. This is the default setting.

Exclude cases listwise
Indicates inclusion of all cases with sufficient data on all variables used across all analyses.

User Missing Values
Exclude treats user missing values as missing. **Include** ignores user missing value designations and treats user missing values as valid.

Defining missing value settings for Independent-Samples Proportions

1. From the menus choose:

   Analyze > **Compare Means** > **Independent-Samples Proportions...**
2. In the Independent-Samples Proportions dialog, click **Missing Values**.
3. Select the desired missing values settings.

T Tests

T Tests

Three types of *t* tests are available:

Independent-samples t test (two-sample t test). Compares the means of one variable for two groups of cases. Descriptive statistics for each group and Levene's test for equality of variances are provided, as well as both equal- and unequal-variance *t* values and a 95% confidence interval for the difference in means.

Paired-samples t test (dependent t test). Compares the means of two variables for a single group. This test is also for matched pairs or case-control study designs. The output includes descriptive statistics for the test variables, the correlation between the variables, descriptive statistics for the paired differences, the *t* test, and a 95% confidence interval.

One-sample t test. Compares the mean of one variable with a known or hypothesized value. Descriptive statistics for the test variables are displayed along with the *t* test. A 95% confidence interval for the difference between the mean of the test variable and the hypothesized test value is part of the default output.

Independent-Samples T Test

The Independent-Samples T Test procedure compares means for two groups of cases and automates the *t*-test effect size computation. Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors. This is not the case if you compare average income for males and females. A person is not randomly assigned to be a male or female. In such situations, you should ensure that differences in other factors are not masking or enhancing a significant difference in means. Differences in average income may be influenced by factors such as education (and not by sex alone).

Example

Patients with high blood pressure are randomly assigned to a placebo group and a treatment group. The placebo subjects receive an inactive pill, and the treatment subjects receive a new drug that is expected to lower blood pressure. After the subjects are treated for two months, the two-sample *t* test is used to compare the average blood pressures for the placebo group and the treatment group. Each patient is measured once and belongs to one group.

Statistics

For each variable: sample size, mean, standard deviation, standard error of the mean, and the estimation of the effect size for the *t*-test. For the difference in means: mean, standard error, and confidence interval (you can specify the confidence level). Tests: Levene's test for equality of variances and both pooled-variances and separate-variances *t* tests for equality of means.

Data Considerations

Data

The values of the quantitative variable of interest are in a single column in the data file. The procedure uses a grouping variable with two values to separate the cases into two groups. The grouping variable can be numeric (values such as 1 and 2 or 6.25 and 12.5) or short string (such as *yes* and *no*). As an alternative, you can use a quantitative variable, such as *age,* to split the cases into two groups by specifying a cutpoint (cutpoint 21 splits *age* into an under-21 group and a 21-and-over group).

Assumptions

For the equal-variance *t* test, the observations should be independent, random samples from normal distributions with the same population variance. For the unequal-variance *t* test, the observations should be independent, random samples from normal distributions. The two-sample *t* test is fairly robust to departures from normality. When checking distributions graphically, look to see that they are symmetric and have no outliers.

Obtaining an Independent-Samples T Test

1. From the menus choose:

   Analyze > **Compare Means** > **Independent-Samples T Test...**
2. Select one or more quantitative test variables. A separate *t* test is computed for each variable.
3. Select a single grouping variable, and then click **Define Groups** to specify two codes for the groups that you want to compare.
4. Optionally, you can:
   - Select **Estimate effect sizes** to control the estimation of the *t*-test effect size.
   - Click **Options** to control the treatment of missing data and the level of the confidence interval.
   - Click **Bootstrap** for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.

Independent-Samples T-Test Define Groups

For numeric grouping variables, define the two groups for the *t*-test by specifying two values or a cutpoint:

- **Use specified values.** Enter a value for Group 1 and another value for Group 2. Cases with any other values are excluded from the analysis. Numbers need not be integers (for example, 6.25 and 12.5 are valid).
- **Cutpoint.** Enter a number that splits the values of the grouping variable into two sets. All cases with values that are less than the cutpoint form one group, and cases with values that are greater than or equal to the cutpoint form the other group.

For string grouping variables, enter a string for Group 1 and another value for Group 2, such as *yes* and *no*. Cases with other strings are excluded from the analysis.

Independent-Samples T Test Options

Confidence Interval. By default, a 95% confidence interval for the difference in means is displayed. Enter a value between 1 and 99 to request a different confidence level.

Missing Values. When you test several variables, and data are missing for one or more variables, you can tell the procedure which cases to include (or exclude).

- **Exclude cases analysis by analysis.** Each *t* test uses all cases that have valid data for the tested variables. Sample sizes may vary from test to test.
- **Exclude cases listwise.** Each *t* test uses only cases that have valid data for all variables that are used in the requested *t* tests. The sample size is constant across tests.

Paired-Samples T Test

The Paired-Samples T Test procedure compares the means of two variables for a single group. The procedure computes the differences between values of the two variables for each case and tests whether the average differs from 0. The procedure also automates the *t*-test effect size computation.

Example

In a study on high blood pressure, all patients are measured at the beginning of the study, given a treatment, and measured again. Thus, each subject has two measures, often called *before* and *after* measures. An alternative design for which this test is used is a matched-pairs or case-control study, in which each record in the data file contains the response for the patient and also for his or her matched control subject. In a blood pressure study, patients and controls might be matched by age (a 75-year-old patient with a 75-year-old control group member).

Statistics

For each variable: mean, sample size, standard deviation, and standard error of the mean. For each pair of variables: correlation, average difference in means, *t* test, confidence interval for mean difference (you can specify the confidence level), and the estimation of the effect size for the *t*-test. Standard deviation and standard error of the mean difference.

Data considerations

Data

For each paired test, specify two quantitative variables (interval level of measurement or ratio level of measurement). For a matched-pairs or case-control study, the response for each test subject and its matched control subject must be in the same case in the data file.

Assumptions

Observations for each pair should be made under the same conditions. The mean differences should be normally distributed. Variances of each variable can be equal or unequal.

Obtaining a Paired-Samples T Test

1. From the menus choose:

   Analyze > **Compare Means** > **Paired-Samples T Test...**
2. Select one or more pairs of variables.

3. Optionally, change/select a **Estimate effect sizes** option. The settings control how the standardizer is computed in estimating the Cohen's *d* and Hedges' correction for each variable pair.

   Standard deviation of the difference
   The denominator used in estimating the effect size. Cohen's *d* uses the sample standard deviation of the mean difference. Hedges' correction uses the sample standard deviation of the mean difference adjusted by a correction factor.

   Corrected standard deviation of the difference
   The denominator used in estimating the effect size. Cohen's *d* uses the sample standard deviation of the mean difference adjusted by the correlation between measures. Hedges' correction uses the sample standard deviation of the mean difference adjusted by the correlation between measures, plus a correction factor.

   Average of variances
   The denominator used in estimating the effect size. Cohen's *d* uses the square root of the average variance of measures. Hedges' correction uses the square root of the average variance of measures, plus a correction factor.

4. Optionally, you can:
   - Select **Estimate effect sizes** to control the estimation of the *t*-test effect size. When the setting is selected, you can further control how the standardizer is computed in estimating the Cohen's *d* and Hedges' correction for each variable pair.
   - Click **Options** to control the treatment of missing data and the level of the confidence interval.
   - Click **Bootstrap** for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.

Paired-Samples T Test Options

Confidence Interval. By default, a 95% confidence interval for the difference in means is displayed. Enter a value between 1 and 99 to request a different confidence level.

Missing Values. When you test several variables, and data are missing for one or more variables, you can tell the procedure which cases to include (or exclude):

- **Exclude cases analysis by analysis.** Each *t* test uses all cases that have valid data for the tested pair of variables. Sample sizes may vary from test to test.
- **Exclude cases listwise.** Each *t* test uses only cases that have valid data for all pairs of tested variables. The sample size is constant across tests.

T TEST Command Additional Features

The command syntax language also allows you to:

- Produce both one-sample and independent-samples t tests by running a single command.
- Test a variable against each variable on a list in a paired t test (with the `PAIRS` subcommand).
- Control the estimation of the *t*-test effect size (with the ES subcommand).

See the *Command Syntax Reference* for complete syntax information.

One-Sample T Test

The One-Sample T Test procedure tests whether the mean of a single variable differs from a specified constant and automates the *t*-test effect size computation.

Examples
A researcher might want to test whether the average IQ score for a group of students differs from 100. Or a cereal manufacturer can take a sample of boxes from the production line and check whether the mean weight of the samples differs from 1.3 pounds at the 95% confidence level.

Statistics

For each test variable: mean, standard deviation, standard error of the mean, and the estimation of the effect size for the *t*-test. The average difference between each data value and the hypothesized test value, a *t* test that tests that this difference is 0, and a confidence interval for this difference (you can specify the confidence level).

Data Considerations

Data

To test the values of a quantitative variable against a hypothesized test value, choose a quantitative variable and enter a hypothesized test value.

Assumptions

This test assumes that the data are normally distributed; however, this test is fairly robust to departures from normality.

Obtaining a One-Sample T Test

1. From the menus choose:

   Analyze > **Compare Means** > **One-Sample T Test...**
2. Select one or more variables to be tested against the same hypothesized value.
3. Enter a numeric test value against which each sample mean is compared.
4. Optionally, you can:
   - Select **Estimate effect sizes** to control the estimation of the *t*-test effect size.
   - Click **Options** to control the treatment of missing data and the level of the confidence interval.

One-Sample T Test Options

Confidence Interval. By default, a 95% confidence interval for the difference between the mean and the hypothesized test value is displayed. Enter a value between 1 and 99 to request a different confidence level.

Missing Values. When you test several variables, and data are missing for one or more variables, you can tell the procedure which cases to include (or exclude).

- **Exclude cases analysis by analysis.** Each *t* test uses all cases that have valid data for the tested variable. Sample sizes may vary from test to test.
- **Exclude cases listwise.** Each *t* test uses only cases that have valid data for all variables that are used in any of the requested *t* tests. The sample size is constant across tests.

T TEST Command Additional Features

The command syntax language also allows you to:

- Produce both one-sample and independent-samples t tests by running a single command.
- Test a variable against each variable on a list in a paired t test (with the PAIRS subcommand).
- Control the estimation of the *t*-test effect size (with the ES subcommand).

See the *Command Syntax Reference* for complete syntax information.

T TEST Command Additional Features

The command syntax language also allows you to:

- Produce both one-sample and independent-samples t tests by running a single command.
- Test a variable against each variable on a list in a paired t test (with the PAIRS subcommand).
- Control the estimation of the *t*-test effect size (with the ES subcommand).

See the *Command Syntax Reference* for complete syntax information.

One-Way ANOVA

The One-Way ANOVA procedure produces a one-way analysis of variance for a quantitative dependent variable by a single factor (independent) variable and estimates the effect size in one-way ANOVA. Analysis of variance is used to test the hypothesis that several means are equal. This technique is an extension of the two-sample *t* test.

In addition to determining that differences exist among the means, you may want to know which means differ. There are two types of tests for comparing means: a priori contrasts and post hoc tests. Contrasts are tests set up *before* running the experiment, and post hoc tests are run *after* the experiment has been conducted. You can also test for trends across categories.

Example

Doughnuts absorb fat in various amounts when they are cooked. An experiment is set up involving three types of fat: peanut oil, corn oil, and lard. Peanut oil and corn oil are unsaturated fats, and lard is a saturated fat. Along with determining whether the amount of fat absorbed depends on the type of fat used, you could set up an a priori contrast to determine whether the amount of fat absorption differs for saturated and unsaturated fats.

Statistics

For each group: number of cases, mean, standard deviation, standard error of the mean, minimum, maximum, 95% confidence interval for the mean, and the estimation of the effect size for in a one-way ANOVA. Levene tests for homogeneity of variance, analysis-of-variance table and robust tests of the equality of means for each dependent variable, user-specified a priori contrasts, and post hoc range tests and multiple comparisons: Bonferroni, Sidak, Tukey's honestly significant difference, Hochberg's GT2, Gabriel, Dunnett, Ryan-Einot-Gabriel-Welsch *F* test (R-E-G-W *F*), Ryan-Einot-Gabriel-Welsch range test (R-E-G-W *Q*), Tamhane's T2, Dunnett's T3, Games-Howell, Dunnett's *C*, Duncan's multiple range test, Student-Newman-Keuls (S-N-K), Tukey's *b*, Waller-Duncan, Scheffé, and least-significant difference.

Data considerations

Data

The dependent variable should be quantitative (interval level of measurement).

Assumptions

Each group is an independent random sample from a normal population. Analysis of variance is robust to departures from normality, although the data should be symmetric. The groups should come from populations with equal variances. To test this assumption, use Levene's homogeneity-of-variance test.

Obtaining a One-Way analysis of variance

1. From the menus choose:

   Analyze > **Compare Means** > **One-Way ANOVA...**
2. Select one or more dependent variables.
3. Select a single independent factor variable.

Optionally, you can:

- Select **Estimate effect size for overall tests** to control the calculation of the effect size for the overall test. When selected, the “ANOVA Effect Sizes” table displays in the output.
- Click **Contrasts** to partition the between-groups sums of squares into trend components or specify a priori contrasts.
- Click **Post Hoc** to use post hoc range tests and pairwise multiple comparisons to determine which means differ.
- Click **Options** to control the treatment of missing data and the level of the confidence interval.

- Click **Bootstrap** for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.

One-Way ANOVA Contrasts

You can partition the between-groups sums of squares into trend components or specify a priori contrasts.

Polynomial
Partitions the between-groups sums of squares into trend components. You can test for a trend of the dependent variable across the ordered levels of the factor variable. For example, you could test for a linear trend (increasing or decreasing) in salary across the ordered levels of highest degree earned.

- **Degree.** You can choose a 1st, 2nd, 3rd, 4th, or 5th degree polynomial.

Coefficients
User-specified a priori contrasts to be tested by the *t* statistic. Enter a coefficient for each group (category) of the factor variable and click **Add** after each entry. Each new value is added to the bottom of the coefficient list. To specify additional sets of contrasts, click **Next**. Use **Next** and **Previous** to move between sets of contrasts.

Estimate effect size for contrasts
Controls the calculation of the effect size for the overall test. When this setting is enabled, at least one of the following options must be selected to calculate the effect sizes. This setting is enabled when at least one contrast is specified and results in an ANOVA Effect Sizes table in the output.

Use pooled standard deviation for all the groups as the standardizer
Uses the pooled standard deviation for all the groups as the standardizer in estimating the effect size. This is the default setting and is available when **Estimate effect size for contrasts** is selected.

Use pooled standard deviation for those groups involved in the contrast as the standardizer
Uses the pooled standard deviation for the groups involved in the contrast as the standardizer. The setting is available when **Estimate effect size for contrasts** is selected.

The order of the coefficients is important because it corresponds to the ascending order of the category values of the factor variable. The first coefficient on the list corresponds to the lowest group value of the factor variable, and the last coefficient corresponds to the highest value. For example, if there are six categories of the factor variable, the coefficients –1, 0, 0, 0, 0.5, and 0.5 contrast the first group with the fifth and sixth groups. For most applications, the coefficients should sum to 0. Sets that do not sum to 0 can also be used, but a warning message is displayed.

One-Way ANOVA Post Hoc Tests

Once you have determined that differences exist among the means, post hoc range tests and pairwise multiple comparisons can determine which means differ. Range tests identify homogeneous subsets of means that are not different from each other. Pairwise multiple comparisons test the difference between each pair of means and yield a matrix where asterisks indicate significantly different group means at an alpha level of 0.05.

Equal Variances Assumed

Tukey's honestly significant difference test, Hochberg's GT2, Gabriel, and Scheffé are multiple comparison tests and range tests. Other available range tests are Tukey's *b*, S-N-K (Student-Newman-Keuls), Duncan, R-E-G-W *F* (Ryan-Einot-Gabriel-Welsch *F* test), R-E-G-W *Q* (Ryan-Einot-Gabriel-Welsch range test), and Waller-Duncan. Available multiple comparison tests are Bonferroni, Tukey's honestly significant difference test, Sidak, Gabriel, Hochberg, Dunnett, Scheffé, and LSD (least significant difference).

- *LSD*. Uses t tests to perform all pairwise comparisons between group means. No adjustment is made to the error rate for multiple comparisons.

- *Bonferroni*. Uses *t* tests to perform pairwise comparisons between group means, but controls overall error rate by setting the error rate for each test to the experimentwise error rate divided by the total number of tests. Hence, the observed significance level is adjusted for the fact that multiple comparisons are being made.
- *Sidak*. Pairwise multiple comparison test based on a t statistic. Sidak adjusts the significance level for multiple comparisons and provides tighter bounds than Bonferroni.
- *Scheffe*. Performs simultaneous joint pairwise comparisons for all possible pairwise combinations of means. Uses the F sampling distribution. Can be used to examine all possible linear combinations of group means, not just pairwise comparisons.
- *R-E-G-W F*. Ryan-Einot-Gabriel-Welsch multiple stepdown procedure based on an F test.
- *R-E-G-W Q*. Ryan-Einot-Gabriel-Welsch multiple stepdown procedure based on the Studentized range.
- *S-N-K*. Makes all pairwise comparisons between means using the Studentized range distribution. With equal sample sizes, it also compares pairs of means within homogeneous subsets, using a stepwise procedure. Means are ordered from highest to lowest, and extreme differences are tested first.
- *Tukey*. Uses the Studentized range statistic to make all of the pairwise comparisons between groups. Sets the experimentwise error rate at the error rate for the collection for all pairwise comparisons.
- *Tukey's b*. Uses the Studentized range distribution to make pairwise comparisons between groups. The critical value is the average of the corresponding value for the Tukey's honestly significant difference test and the Student-Newman-Keuls.
- *Duncan*. Makes pairwise comparisons using a stepwise order of comparisons identical to the order used by the Student-Newman-Keuls test, but sets a protection level for the error rate for the collection of tests, rather than an error rate for individual tests. Uses the Studentized range statistic.
- *Hochberg's GT2*. Multiple comparison and range test that uses the Studentized maximum modulus. Similar to Tukey's honestly significant difference test.
- *Gabriel*. Pairwise comparison test that used the Studentized maximum modulus and is generally more powerful than Hochberg's GT2 when the cell sizes are unequal. Gabriel's test may become liberal when the cell sizes vary greatly.
- *Waller-Duncan*. Multiple comparison test based on a t statistic; uses a Bayesian approach.
- *Dunnett*. Pairwise multiple comparison *t* test that compares a set of treatments against a single control mean. The last category is the default control category. Alternatively, you can choose the first category. **2-sided** tests that the mean at any level (except the control category) of the factor is not equal to that of the control category. **< Control** tests if the mean at any level of the factor is smaller than that of the control category. **> Control** tests if the mean at any level of the factor is greater than that of the control category.

Equal Variances Not Assumed

Multiple comparison tests that do not assume equal variances are Tamhane's T2, Dunnett's T3, Games-Howell, and Dunnett's *C*.

- *Tamhane's T2*. Conservative pairwise comparisons test based on a t test. This test is appropriate when the variances are unequal.
- *Dunnett's T3*. Pairwise comparison test based on the Studentized maximum modulus. This test is appropriate when the variances are unequal.
- *Games-Howell*. Pairwise comparison test that is sometimes liberal. This test is appropriate when the variances are unequal.
- *Dunnett's C*. Pairwise comparison test based on the Studentized range. This test is appropriate when the variances are unequal.

Note: You may find it easier to interpret the output from post hoc tests if you deselect **Hide empty rows and columns** in the Table Properties dialog box (in an activated pivot table, choose **Table Properties** from the Format menu).

Null Hypothesis test

Specifies how the significance level (alpha) is handled for the post hoc test.

Use the same significance level (alpha) as the settings in Options
When selected, uses same setting that is specified in the Options dialog.

Specify the significance level (alpha) for the post hoc test
When selected, you can specify the significance level (alpha) in the **Level** field.

Obtaining Post Hoc Tests for One-Way ANOVA

One-Way ANOVA Options

Statistics
Choose one or more of the following:

Descriptive
Calculates the number of cases, mean, standard deviation, standard error of the mean, minimum, maximum, and 95% confidence intervals for each dependent variable for each group.

Fixed and random effects
Displays the standard deviation, standard error, and 95% confidence interval for the fixed-effects model, and the standard error, 95% confidence interval, and estimate of between-components variance for the random-effects model.

Homogeneity of variance test
Calculates the Levene statistic to test for the equality of group variances. This test is not dependent on the assumption of normality.

Brown-Forsythe
Calculates the Brown-Forsythe statistic to test for the equality of group means. This statistic is preferable to the *F* statistic when the assumption of equal variances does not hold.

Welch
Calculates the Welch statistic to test for the equality of group means. This statistic is preferable to the *F* statistic when the assumption of equal variances does not hold.

Missing Values
Controls the treatment of missing values.

Exclude cases analysis by analysis
A case with a missing value for either the dependent or the factor variable for a given analysis is not used in that analysis. Also, a case outside the range specified for the factor variable is not used.

Exclude cases listwise
Cases with missing values for the factor variable or for any dependent variable included on the dependent list in the main dialog box are excluded from all analyses. If you have not specified multiple dependent variables, this has no effect.

Confidence Interval
By default, a 95% confidence interval for the difference between the mean and the hypothesized test value is displayed. Enter a value between 1 and 99 to request a different confidence level.

Means plot
Displays a chart that plots the subgroup means (the means for each group defined by values of the factor variable).

Specifying Options for One-Way ANOVA

ONEWAY Command Additional Features

The command syntax language also allows you to:

- Obtain fixed- and random-effects statistics. Standard deviation, standard error of the mean, and 95% confidence intervals for the fixed-effects model. Standard error, 95% confidence intervals, and estimate of between-components variance for random-effects model (using `STATISTICS=EFFECTS`).
- Specify alpha levels for the least significance difference, Bonferroni, Duncan, and Scheffé multiple comparison tests (with the `RANGES` subcommand).
- Write a matrix of means, standard deviations, and frequencies, or read a matrix of means, frequencies, pooled variances, and degrees of freedom for the pooled variances. These matrices can be used in place of raw data to obtain a one-way analysis of variance (with the `MATRIX` subcommand).

See the *Command Syntax Reference* for complete syntax information.

GLM Univariate Analysis

The GLM Univariate procedure provides regression analysis and analysis of variance for one dependent variable by one or more factors and/or variables. The factor variables divide the population into groups. Using this General Linear Model procedure, you can test null hypotheses about the effects of other variables on the means of various groupings of a single dependent variable. You can investigate interactions between factors as well as the effects of individual factors, some of which may be random. In addition, the effects of covariates and covariate interactions with factors can be included. For regression analysis, the independent (predictor) variables are specified as covariates.

Both balanced and unbalanced models can be tested. A design is balanced if each cell in the model contains the same number of cases. In addition to testing hypotheses, GLM Univariate produces estimates of parameters.

Commonly used a priori contrasts are available to perform hypothesis testing. Additionally, after an overall *F* test has shown significance, you can use post hoc tests to evaluate differences among specific means. Estimated marginal means give estimates of predicted mean values for the cells in the model, and profile plots (interaction plots) of these means allow you to easily visualize some of the relationships.

Residuals, predicted values, Cook's distance, and leverage values can be saved as new variables in your data file for checking assumptions.

WLS Weight allows you to specify a variable used to give observations different weights for a weighted least-squares (WLS) analysis, perhaps to compensate for a different precision of measurement.

Example. Data are gathered for individual runners in the Chicago marathon for several years. The time in which each runner finishes is the dependent variable. Other factors include weather (cold, pleasant, or hot), number of months of training, number of previous marathons, and gender. Age is considered a covariate. You might find that gender is a significant effect and that the interaction of gender with weather is significant.

Methods. Type I, Type II, Type III, and Type IV sums of squares can be used to evaluate different hypotheses. Type III is the default.

Statistics. Post hoc range tests and multiple comparisons: least significant difference, Bonferroni, Sidak, Scheffé, Ryan-Einot-Gabriel-Welsch multiple *F*, Ryan-Einot-Gabriel-Welsch multiple range, Student-Newman-Keuls, Tukey's honestly significant difference, Tukey's *b*, Duncan, Hochberg's GT2, Gabriel, Waller-Duncan *t* test, Dunnett (one-sided and two-sided), Tamhane's T2, Dunnett's T3, Games-Howell, and Dunnett's *C*. Descriptive statistics: observed means, standard deviations, and counts for all of the dependent variables in all cells. Levene tests for homogeneity of variance.

Plots. Spread-versus-level, residual, and profile (interaction).

GLM Univariate Data Considerations

Data. The dependent variable is quantitative. Factors are categorical. They can have numeric values or string values of up to eight characters. Covariates are quantitative variables that are related to the dependent variable.

Assumptions. The data are a random sample from a normal population; in the population, all cell variances are the same. Analysis of variance is robust to departures from normality, although the data

should be symmetric. To check assumptions, you can use homogeneity of variances tests and spread-versus-level plots. You can also examine residuals and residual plots.

To Obtain GLM Univariate Tables

1. From the menus choose:

   Analyze > **General Linear Model** > **Univariate...**
2. Select a dependent variable.
3. Select variables for Fixed Factor(s), Random Factor(s), and Covariate(s), as appropriate for your data.
4. Optionally, you can use WLS Weight to specify a weight variable for weighted least-squares analysis. If the value of the weighting variable is zero, negative, or missing, the case is excluded from the analysis. A variable already used in the model cannot be used as a weighting variable.

GLM Model

Figure 1. Univariate Model dialog box

Specify Model. A full factorial model contains all factor main effects, all covariate main effects, and all factor-by-factor interactions. It does not contain covariate interactions. Select **Custom** to specify only a subset of interactions or to specify factor-by-covariate interactions. You must indicate all of the terms to be included in the model.

Factors and Covariates. The factors and covariates are listed.

Model. The model depends on the nature of your data. After selecting **Custom**, you can select the main effects and interactions that are of interest in your analysis.

Sum of squares. The method of calculating the sums of squares. For balanced or unbalanced models with no missing cells, the Type III sum-of-squares method is most commonly used.

Include intercept in model. The intercept is usually included in the model. If you can assume that the data pass through the origin, you can exclude the intercept.

Build Terms and Custom Terms

Build terms

Use this choice when you want to include non-nested terms of a certain type (such as main effects) for all combinations of a selected set of factors and covariates.

Build custom terms

Use this choice when you want to include nested terms or when you want to explicitly build any term variable by variable. Building a nested term involves the following steps:

Sum of Squares

For the model, you can choose a type of sums of squares. Type III is the most commonly used and is the default.

Type I. This method is also known as the hierarchical decomposition of the sum-of-squares method. Each term is adjusted for only the term that precedes it in the model. Type I sums of squares are commonly used for:

- A balanced ANOVA model in which any main effects are specified before any first-order interaction effects, any first-order interaction effects are specified before any second-order interaction effects, and so on.
- A polynomial regression model in which any lower-order terms are specified before any higher-order terms.
- A purely nested model in which the first-specified effect is nested within the second-specified effect, the second-specified effect is nested within the third, and so on. (This form of nesting can be specified only by using syntax.)

Type II. This method calculates the sums of squares of an effect in the model adjusted for all other "appropriate" effects. An appropriate effect is one that corresponds to all effects that do not contain the effect being examined. The Type II sum-of-squares method is commonly used for:

- A balanced ANOVA model.
- Any model that has main factor effects only.
- Any regression model.
- A purely nested design. (This form of nesting can be specified by using syntax.)

Type III. The default. This method calculates the sums of squares of an effect in the design as the sums of squares, adjusted for any other effects that do not contain the effect, and orthogonal to any effects (if any) that contain the effect. The Type III sums of squares have one major advantage in that they are invariant with respect to the cell frequencies as long as the general form of estimability remains constant. Hence, this type of sums of squares is often considered useful for an unbalanced model with no missing cells. In a factorial design with no missing cells, this method is equivalent to the Yates' weighted-squares-of-means technique. The Type III sum-of-squares method is commonly used for:

- Any models listed in Type I and Type II.
- Any balanced or unbalanced model with no empty cells.

Type IV. This method is designed for a situation in which there are missing cells. For any effect *F* in the design, if *F* is not contained in any other effect, then Type IV = Type III = Type II. When *F* is contained in other effects, Type IV distributes the contrasts being made among the parameters in *F* to all higher-level effects equitably. The Type IV sum-of-squares method is commonly used for:

- Any models listed in Type I and Type II.
- Any balanced model or unbalanced model with empty cells.

GLM Contrasts

Contrasts are used to test for differences among the levels of a factor. You can specify a contrast for each factor in the model (in a repeated measures model, for each between-subjects factor). Contrasts represent linear combinations of the parameters.

GLM Univariate. Hypothesis testing is based on the null hypothesis **LB** = 0, where **L** is the contrast coefficients matrix and **B** is the parameter vector. When a contrast is specified, an **L** matrix is created. The columns of the **L** matrix corresponding to the factor match the contrast. The remaining columns are adjusted so that the **L** matrix is estimable.

The output includes an *F* statistic for each set of contrasts. Also displayed for the contrast differences are Bonferroni-type simultaneous confidence intervals based on Student's *t* distribution.

Available Contrasts

Available contrasts are deviation, simple, difference, Helmert, repeated, and polynomial. For deviation contrasts and simple contrasts, you can choose whether the reference category is the last or first category.

Contrast Types

Deviation. Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean). The levels of the factor can be in any order.

Simple. Compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group. You can choose the first or last category as the reference.

Difference. Compares the mean of each level (except the first) to the mean of previous levels. (Sometimes called reverse Helmert contrasts.)

Helmert. Compares the mean of each level of the factor (except the last) to the mean of subsequent levels.

Repeated. Compares the mean of each level (except the last) to the mean of the subsequent level.

Polynomial. Compares the linear effect, quadratic effect, cubic effect, and so on. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are often used to estimate polynomial trends.

GLM Profile Plots

Profile plots (interaction plots) are useful for comparing marginal means in your model. A profile plot is a line plot in which each point indicates the estimated marginal mean of a dependent variable (adjusted for any covariates) at one level of a factor. The levels of a second factor can be used to make separate lines. Each level in a third factor can be used to create a separate plot. All fixed and random factors, if any, are available for plots. For multivariate analyses, profile plots are created for each dependent variable. In a repeated measures analysis, both between-subjects factors and within-subjects factors can be used in profile plots. GLM Multivariate and GLM Repeated Measures are available only if you have the Advanced Statistics option installed.

A profile plot of one factor shows whether the estimated marginal means are increasing or decreasing across levels. For two or more factors, parallel lines indicate that there is no interaction between factors, which means that you can investigate the levels of only one factor. Nonparallel lines indicate an interaction.

Figure 2. Nonparallel plot (left) and parallel plot (right)

After a plot is specified by selecting factors for the horizontal axis and, optionally, factors for separate lines and separate plots, the plot must be added to the Plots list.

Chart Type

The chart can be a line chart or a bar chart

Error Bars
You can include error bars that represent the confidence interval or a number of standard errors. The confidence interval is based on the significance level specified on the Options dialog.

Include reference line for grand mean
Includes a reference line that represents the overall grand mean.

Y axis starts at 0
For line charts with all positive or all negative values, forces the Y axis to start at 0. Bar charts always start (or include) 0.

GLM Options

Optional statistics are available from this dialog box. Statistics are calculated using a fixed-effects model.

Display. Select **Descriptive statistics** to produce observed means, standard deviations, and counts for all of the dependent variables in all cells. **Estimates of effect size** gives a partial eta-squared value for each effect and each parameter estimate. The eta-squared statistic describes the proportion of total variability attributable to a factor. Select **Observed power** to obtain the power of the test when the alternative hypothesis is set based on the observed value. Select **Parameter estimates** to produce the parameter estimates, standard errors, *t* tests, confidence intervals, and the observed power for each test. Select **Contrast coefficient matrix** to obtain the **L** matrix.

Homogeneity tests produces Levene tests of the homogeneity of variance for each dependent variable across all level combinations of the between-subjects factors, for between-subjects factors only. The spread-versus-level and residual plots options are useful for checking assumptions about the data. This item is disabled if there are no factors. Select **Residual plot** to produce an observed-by-predicted-by-standardized residual plot for each dependent variable. These plots are useful for investigating the assumption of equal variance. Select **Lack of fit** to check if the relationship between the dependent variable and the independent variables can be adequately described by the model. **General estimable function(s)** allows you to construct custom hypothesis tests based on the general estimable function(s). Rows in any contrast coefficient matrix are linear combinations of the general estimable function(s).

Heteroskedasticity Tests are available for testing whether the variance of the errors (for each dependent variable) depends on the values of the independent variables. For the **Breusch-Pagan test**, **Modified Breusch-Pagan test**, and **F test** you can specify the model on which the test is based. By default, the model consists of a constant term, a term that is linear in the predicted values, a term that is quadratic in the predicted values, and an error term.

Parameter estimates with robust standard errors displays a table of parameter estimates, along with robust or heteroskedasticity-consistent (HC) standard errors; and *t* statistics, significance values, and confidence intervals that use the robust standard errors. Five different methods are available for the robust covariance matrix estimation.

HC0
Based on the original asymptotic or large sample robust, empirical, or "sandwich" estimator of the covariance matrix of the parameter estimates. The middle part of the sandwich contains squared OLS (ordinary least squares) or squared weighted WLS (weighted least squares) residuals.

HC1
A finite-sample modification of HC0, multiplying it by N/(N-p), where N is the sample size and p is the number of non-redundant parameters in the model.

HC2
A modification of HC0 that involves dividing the squared residual by 1-h, where h is the leverage for the case.

HC3
A modification of HC0 that approximates a jackknife estimator. Squared residuals are divided by the square of 1-h.

HC4
A modification of HC0 that divides the squared residuals by 1-h to a power that varies according to h, N, and p, with an upper limit of 4.

Significance level. You might want to adjust the significance level used in post hoc tests and the confidence level used for constructing confidence intervals. The specified value is also used to calculate the observed power for the test. When you specify a significance level, the associated level of the confidence intervals is displayed in the dialog box.

UNIANOVA Command Additional Features

The command syntax language also allows you to:

- Specify nested effects in the design (using the `DESIGN` subcommand).
- Specify tests of effects versus a linear combination of effects or a value (using the `TEST` subcommand).
- Specify multiple contrasts (using the `CONTRAST` subcommand).
- Include user-missing values (using the `MISSING` subcommand).
- Specify EPS criteria (using the `CRITERIA` subcommand).
- Construct a custom **L** matrix, **M** matrix, or **K** matrix (using the `LMATRIX`, `MMATRIX`, and `KMATRIX` subcommands).
- For deviation or simple contrasts, specify an intermediate reference category (using the `CONTRAST` subcommand).
- Specify metrics for polynomial contrasts (using the `CONTRAST` subcommand).
- Specify error terms for post hoc comparisons (using the `POSTHOC` subcommand).
- Compute estimated marginal means for any factor or factor interaction among the factors in the factor list (using the `EMMEANS` subcommand).
- Specify names for temporary variables (using the `SAVE` subcommand).
- Construct a correlation matrix data file (using the `OUTFILE` subcommand).
- Construct a matrix data file that contains statistics from the between-subjects ANOVA table (using the `OUTFILE` subcommand).
- Save the design matrix to a new data file (using the `OUTFILE` subcommand).

See the *Command Syntax Reference* for complete syntax information.

GLM Post Hoc Comparisons

Post hoc multiple comparison tests. Once you have determined that differences exist among the means, post hoc range tests and pairwise multiple comparisons can determine which means differ. Comparisons are made on unadjusted values. These tests are used for fixed between-subjects factors only. In GLM Repeated Measures, these tests are not available if there are no between-subjects factors, and the post hoc multiple comparison tests are performed for the average across the levels of the within-subjects factors. For GLM Multivariate, the post hoc tests are performed for each dependent variable separately. GLM Multivariate and GLM Repeated Measures are available only if you have the Advanced Statistics option installed.

The Bonferroni and Tukey's honestly significant difference tests are commonly used multiple comparison tests. The **Bonferroni test**, based on Student's *t* statistic, adjusts the observed significance level for the fact that multiple comparisons are made. **Sidak's t test** also adjusts the significance level and provides tighter bounds than the Bonferroni test. **Tukey's honestly significant difference test** uses the Studentized range statistic to make all pairwise comparisons between groups and sets the experimentwise error rate to the error rate for the collection for all pairwise comparisons. When testing a large number of pairs of means, Tukey's honestly significant difference test is more powerful than the Bonferroni test. For a small number of pairs, Bonferroni is more powerful.

Hochberg's GT2 is similar to Tukey's honestly significant difference test, but the Studentized maximum modulus is used. Usually, Tukey's test is more powerful. **Gabriel's pairwise comparisons test** also uses the Studentized maximum modulus and is generally more powerful than Hochberg's GT2 when the cell sizes are unequal. Gabriel's test may become liberal when the cell sizes vary greatly.

Dunnett's pairwise multiple comparison t test compares a set of treatments against a single control mean. The last category is the default control category. Alternatively, you can choose the first category. You can also choose a two-sided or one-sided test. To test that the mean at any level (except the control category) of the factor is not equal to that of the control category, use a two-sided test. To test whether the mean at any level of the factor is smaller than that of the control category, select **< Control**. Likewise, to test whether the mean at any level of the factor is larger than that of the control category, select **> Control**.

Ryan, Einot, Gabriel, and Welsch (R-E-G-W) developed two multiple step-down range tests. Multiple step-down procedures first test whether all means are equal. If all means are not equal, subsets of means are tested for equality. **R-E-G-W F** is based on an *F* test and **R-E-G-W Q** is based on the Studentized range. These tests are more powerful than Duncan's multiple range test and Student-Newman-Keuls (which are also multiple step-down procedures), but they are not recommended for unequal cell sizes.

When the variances are unequal, use **Tamhane's T2** (conservative pairwise comparisons test based on a *t* test), **Dunnett's T3** (pairwise comparison test based on the Studentized maximum modulus), **Games-Howell pairwise comparison test** (sometimes liberal), or **Dunnett's C** (pairwise comparison test based on the Studentized range). Note that these tests are not valid and will not be produced if there are multiple factors in the model.

Duncan's multiple range test, Student-Newman-Keuls (**S-N-K**), and **Tukey's b** are range tests that rank group means and compute a range value. These tests are not used as frequently as the tests previously discussed.

The **Waller-Duncan t test** uses a Bayesian approach. This range test uses the harmonic mean of the sample size when the sample sizes are unequal.

The significance level of the **Scheffé** test is designed to allow all possible linear combinations of group means to be tested, not just pairwise comparisons available in this feature. The result is that the Scheffé test is often more conservative than other tests, which means that a larger difference between means is required for significance.

The least significant difference (**LSD**) pairwise multiple comparison test is equivalent to multiple individual *t* tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.

Tests displayed. Pairwise comparisons are provided for LSD, Sidak, Bonferroni, Games-Howell, Tamhane's T2 and T3, Dunnett's *C*, and Dunnett's T3. Homogeneous subsets for range tests are provided for S-N-K, Tukey's *b*, Duncan, R-E-G-W *F*, R-E-G-W *Q*, and Waller. Tukey's honestly significant difference test, Hochberg's GT2, Gabriel's test, and Scheffé's test are both multiple comparison tests and range tests.

GLM Options

Optional statistics are available from this dialog box. Statistics are calculated using a fixed-effects model.

Display. Select **Descriptive statistics** to produce observed means, standard deviations, and counts for all of the dependent variables in all cells. **Estimates of effect size** gives a partial eta-squared value for each effect and each parameter estimate. The eta-squared statistic describes the proportion of total variability attributable to a factor. Select **Observed power** to obtain the power of the test when the alternative hypothesis is set based on the observed value. Select **Parameter estimates** to produce the parameter estimates, standard errors, *t* tests, confidence intervals, and the observed power for each test. Select **Contrast coefficient matrix** to obtain the **L** matrix.

Homogeneity tests produces Levene tests of the homogeneity of variance for each dependent variable across all level combinations of the between-subjects factors, for between-subjects factors only. The spread-versus-level and residual plots options are useful for checking assumptions about the data. This item is disabled if there are no factors. Select **Residual plot** to produce an observed-by-predicted-by-standardized residual plot for each dependent variable. These plots are useful for investigating the assumption of equal variance. Select **Lack of fit** to check if the relationship between the dependent variable and the independent variables can be adequately described by the model. **General estimable**

function(s) allows you to construct custom hypothesis tests based on the general estimable function(s). Rows in any contrast coefficient matrix are linear combinations of the general estimable function(s).

Heteroskedasticity Tests are available for testing whether the variance of the errors (for each dependent variable) depends on the values of the independent variables. For the **Breusch-Pagan test**, **Modified Breusch-Pagan test**, and **F test** you can specify the model on which the test is based. By default, the model consists of a constant term, a term that is linear in the predicted values, a term that is quadratic in the predicted values, and an error term.

Parameter estimates with robust standard errors displays a table of parameter estimates, along with robust or heteroskedasticity-consistent (HC) standard errors; and *t* statistics, significance values, and confidence intervals that use the robust standard errors. Five different methods are available for the robust covariance matrix estimation.

HC0
: Based on the original asymptotic or large sample robust, empirical, or "sandwich" estimator of the covariance matrix of the parameter estimates. The middle part of the sandwich contains squared OLS (ordinary least squares) or squared weighted WLS (weighted least squares) residuals.

HC1
: A finite-sample modification of HC0, multiplying it by N/(N-p), where N is the sample size and p is the number of non-redundant parameters in the model.

HC2
: A modification of HC0 that involves dividing the squared residual by 1-h, where h is the leverage for the case.

HC3
: A modification of HC0 that approximates a jackknife estimator. Squared residuals are divided by the square of 1-h.

HC4
: A modification of HC0 that divides the squared residuals by 1-h to a power that varies according to h, N, and p, with an upper limit of 4.

Significance level. You might want to adjust the significance level used in post hoc tests and the confidence level used for constructing confidence intervals. The specified value is also used to calculate the observed power for the test. When you specify a significance level, the associated level of the confidence intervals is displayed in the dialog box.

UNIANOVA Command Additional Features

The command syntax language also allows you to:

- Specify nested effects in the design (using the `DESIGN` subcommand).
- Specify tests of effects versus a linear combination of effects or a value (using the `TEST` subcommand).
- Specify multiple contrasts (using the `CONTRAST` subcommand).
- Include user-missing values (using the `MISSING` subcommand).
- Specify EPS criteria (using the `CRITERIA` subcommand).
- Construct a custom **L** matrix, **M** matrix, or **K** matrix (using the `LMATRIX`, `MMATRIX`, and `KMATRIX` subcommands).
- For deviation or simple contrasts, specify an intermediate reference category (using the `CONTRAST` subcommand).
- Specify metrics for polynomial contrasts (using the `CONTRAST` subcommand).
- Specify error terms for post hoc comparisons (using the `POSTHOC` subcommand).
- Compute estimated marginal means for any factor or factor interaction among the factors in the factor list (using the `EMMEANS` subcommand).
- Specify names for temporary variables (using the `SAVE` subcommand).
- Construct a correlation matrix data file (using the `OUTFILE` subcommand).

- Construct a matrix data file that contains statistics from the between-subjects ANOVA table (using the `OUTFILE` subcommand).
- Save the design matrix to a new data file (using the `OUTFILE` subcommand).

See the *Command Syntax Reference* for complete syntax information.

GLM Save

You can save values predicted by the model, residuals, and related measures as new variables in the Data Editor. Many of these variables can be used for examining assumptions about the data. To save the values for use in another IBM SPSS Statistics session, you must save the current data file.

Predicted Values. The values that the model predicts for each case.

- *Unstandardized*. The value the model predicts for the dependent variable.
- *Weighted*. Weighted unstandardized predicted values. Available only if a WLS variable was previously selected.
- *Standard error*. An estimate of the standard deviation of the average value of the dependent variable for cases that have the same values of the independent variables.

Diagnostics. Measures to identify cases with unusual combinations of values for the independent variables and cases that may have a large impact on the model.

- *Cook's distance*. A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. A large Cook's D indicates that excluding a case from computation of the regression statistics changes the coefficients substantially.
- *Leverage values*. Uncentered leverage values. The relative influence of each observation on the model's fit.

Residuals. An unstandardized residual is the actual value of the dependent variable minus the value predicted by the model. Standardized, Studentized, and deleted residuals are also available. If a WLS variable was chosen, weighted unstandardized residuals are available.

- *Unstandardized*. The difference between an observed value and the value predicted by the model.
- *Weighted*. Weighted unstandardized residuals. Available only if a WLS variable was previously selected.
- *Standardized*. The residual divided by an estimate of its standard deviation. Standardized residuals, which are also known as Pearson residuals, have a mean of 0 and a standard deviation of 1.
- *Studentized*. The residual divided by an estimate of its standard deviation that varies from case to case, depending on the distance of each case's values on the independent variables from the means of the independent variables. Sometimes referred to as internally studentized residuals.
- *Deleted*. The residual for a case when that case is excluded from the calculation of the regression coefficients. It is the difference between the value of the dependent variable and the adjusted predicted value.

Coefficient Statistics. Writes a variance-covariance matrix of the parameter estimates in the model to a new dataset in the current session or an external IBM SPSS Statistics data file. Also, for each dependent variable, there will be a row of parameter estimates, a row of standard errors of the parameter estimates, a row of significance values for the *t* statistics corresponding to the parameter estimates, and a row of residual degrees of freedom. For a multivariate model, there are similar rows for each dependent variable. When Heteroskedasticity-consistent statistics is selected (only available for univariate models), the variance-covariance matrix is calculated using a robust estimator, the row of standard errors displays the robust standard errors, and the significance values reflect the robust errors. You can use this matrix file in other procedures that read matrix files.

GLM Estimated Marginal Means

Select the factors and interactions for which you want estimates of the population marginal means in the cells. These means are adjusted for the covariates, if any.

Compare main effects
Provides uncorrected pairwise comparisons among estimated marginal means for any main effect in the model, for both between- and within-subjects factors. This item is available only if main effects are selected under the Display Means For list.

Compare simple main effects
The setting is enabled whenever the target list contains one or more product or interaction effects (for example, A*B, A*B*C). The setting supports the specification of comparisons among simple main effects, which are main effects nested within the levels of other factors.

Confidence interval adjustment
Select least significant difference (LSD), Bonferroni, or Sidak adjustment to the confidence intervals and significance. This item is available only when **Compare main effects** and/or **Compare simple main effects** is selected.

Specifying Estimated Marginal Means

1. From the menus choose one of the procedures available under > **Analyze** > **General Linear Model**.
2. In the main dialog, click **EM Means**.

GLM Options

Optional statistics are available from this dialog box. Statistics are calculated using a fixed-effects model.

Display. Select **Descriptive statistics** to produce observed means, standard deviations, and counts for all of the dependent variables in all cells. **Estimates of effect size** gives a partial eta-squared value for each effect and each parameter estimate. The eta-squared statistic describes the proportion of total variability attributable to a factor. Select **Observed power** to obtain the power of the test when the alternative hypothesis is set based on the observed value. Select **Parameter estimates** to produce the parameter estimates, standard errors, *t* tests, confidence intervals, and the observed power for each test. Select **Contrast coefficient matrix** to obtain the **L** matrix.

Homogeneity tests produces Levene tests of the homogeneity of variance for each dependent variable across all level combinations of the between-subjects factors, for between-subjects factors only. The spread-versus-level and residual plots options are useful for checking assumptions about the data. This item is disabled if there are no factors. Select **Residual plot** to produce an observed-by-predicted-by-standardized residual plot for each dependent variable. These plots are useful for investigating the assumption of equal variance. Select **Lack of fit** to check if the relationship between the dependent variable and the independent variables can be adequately described by the model. **General estimable function(s)** allows you to construct custom hypothesis tests based on the general estimable function(s). Rows in any contrast coefficient matrix are linear combinations of the general estimable function(s).

Heteroskedasticity Tests are available for testing whether the variance of the errors (for each dependent variable) depends on the values of the independent variables. For the **Breusch-Pagan test**, **Modified Breusch-Pagan test**, and **F test** you can specify the model on which the test is based. By default, the model consists of a constant term, a term that is linear in the predicted values, a term that is quadratic in the predicted values, and an error term.

Parameter estimates with robust standard errors displays a table of parameter estimates, along with robust or heteroskedasticity-consistent (HC) standard errors; and *t* statistics, significance values, and confidence intervals that use the robust standard errors. Five different methods are available for the robust covariance matrix estimation.

HC0
Based on the original asymptotic or large sample robust, empirical, or "sandwich" estimator of the covariance matrix of the parameter estimates. The middle part of the sandwich contains squared OLS (ordinary least squares) or squared weighted WLS (weighted least squares) residuals.

HC1
A finite-sample modification of HC0, multiplying it by N/(N-p), where N is the sample size and p is the number of non-redundant parameters in the model.

HC2
A modification of HC0 that involves dividing the squared residual by 1-h, where h is the leverage for the case.

HC3
A modification of HC0 that approximates a jackknife estimator. Squared residuals are divided by the square of 1-h.

HC4
A modification of HC0 that divides the squared residuals by 1-h to a power that varies according to h, N, and p, with an upper limit of 4.

Significance level. You might want to adjust the significance level used in post hoc tests and the confidence level used for constructing confidence intervals. The specified value is also used to calculate the observed power for the test. When you specify a significance level, the associated level of the confidence intervals is displayed in the dialog box.

GLM Auxiliary Regression Model

The Auxiliary Regression Model dialog box specifies the model that is used to test for heteroskedasticity.

Use predicted values
Uses a model that consists of a constant term, a term that is linear in the predicted values, a term that is quadratic in the predicted values, and an error term.

Use univariate model
Uses the model that is specified on the Model subdialog. An intercept term is included if the specified model does not contain one.

Custom model
Uses the model that you explicitly specify.

Build terms
Use this choice when you want to include non-nested terms of a certain type (such as main effects) for all combinations of a selected set of factors and covariates.

Build custom terms
Use this choice when you want to include nested terms or when you want to explicitly build any term variable by variable. Building a nested term involves the following steps:

UNIANOVA Command Additional Features

The command syntax language also allows you to:

- Specify nested effects in the design (using the `DESIGN` subcommand).
- Specify tests of effects versus a linear combination of effects or a value (using the `TEST` subcommand).
- Specify multiple contrasts (using the `CONTRAST` subcommand).
- Include user-missing values (using the `MISSING` subcommand).
- Specify EPS criteria (using the `CRITERIA` subcommand).
- Construct a custom **L** matrix, **M** matrix, or **K** matrix (using the `LMATRIX`, `MMATRIX`, and `KMATRIX` subcommands).
- For deviation or simple contrasts, specify an intermediate reference category (using the `CONTRAST` subcommand).
- Specify metrics for polynomial contrasts (using the `CONTRAST` subcommand).
- Specify error terms for post hoc comparisons (using the `POSTHOC` subcommand).
- Compute estimated marginal means for any factor or factor interaction among the factors in the factor list (using the `EMMEANS` subcommand).
- Specify names for temporary variables (using the `SAVE` subcommand).
- Construct a correlation matrix data file (using the `OUTFILE` subcommand).

- Construct a matrix data file that contains statistics from the between-subjects ANOVA table (using the `OUTFILE` subcommand).
- Save the design matrix to a new data file (using the `OUTFILE` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Bivariate Correlations

The Bivariate Correlations procedure computes Pearson's correlation coefficient, Spearman's rho, and Kendall's tau-*b* with their significance levels. Correlations measure how variables or rank orders are related. Before calculating a correlation coefficient, screen your data for outliers (which can cause misleading results) and evidence of a linear relationship. Pearson's correlation coefficient is a measure of linear association. Two variables can be perfectly related, but if the relationship is not linear, Pearson's correlation coefficient is not an appropriate statistic for measuring their association.

Confidence interval settings are available for Pearson and Spearman.

Example
Is the number of games won by a basketball team correlated with the average number of points scored per game? A scatterplot indicates that there is a linear relationship. Analyzing data from the 1994–1995 NBA season yields that Pearson's correlation coefficient (0.581) is significant at the 0.01 level. You might suspect that the more games won per season, the fewer points the opponents scored. These variables are negatively correlated (–0.401), and the correlation is significant at the 0.05 level.

Statistics
For each variable: number of cases with nonmissing values, mean, and standard deviation. For each pair of variables: Pearson's correlation coefficient, Spearman's rho, Kendall's tau-*b*, cross-product of deviations, and covariance.

Data considerations

Data
Use symmetric quantitative variables for Pearson's correlation coefficient and quantitative variables or variables with ordered categories for Spearman's rho and Kendall's tau-*b*.

Assumptions
Pearson's correlation coefficient assumes that each pair of variables is bivariate normal.

Obtaining Bivariate Correlations

From the menus choose:

Analyze > **Correlate** > **Bivariate...**

1. Select two or more numeric variables.

   The following options are also available:

   Correlation Coefficients
   For quantitative, normally distributed variables, choose the **Pearson** correlation coefficient. If your data are not normally distributed or have ordered categories, choose **Kendall's tau-b** or **Spearman**, which measure the association between rank orders. Correlation coefficients range in value from –1 (a perfect negative relationship) and +1 (a perfect positive relationship). A value of 0 indicates no linear relationship. When interpreting your results, be careful not to draw any cause-and-effect conclusions due to a significant correlation.

   Test of Significance
   You can select two-tailed or one-tailed probabilities. If the direction of association is known in advance, select **One-tailed**. Otherwise, select **Two-tailed**.

Flag significant correlations
Correlation coefficients significant at the 0.05 level are identified with a single asterisk, and those significant at the 0.01 level are identified with two asterisks.

Show only the lower triangle
When selected, only the correlation matrix table's lower triangle is presented in the output. When not selected, the full correlation matrix table is presented in the output. The setting allows table output to adhere to APA style guidelines.

Show diagonal
When selected, the correlation matrix table's lower triangle along with diagonal values are presented in the output. The setting allows table output to adhere to APA style guidelines.

2. You can optionally select the following:
   - Click **Options...** to specify Pearson correlation statistics and missing values settings.
   - Click **Style...** to specify conditions for automatically changing properties of pivot tables based on specific conditions.
   - Click **Bootstrap...** for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient.
   - Click **Confidence Interval...** to set the options for the estimation of the confidence intervals.

Bivariate Correlations Options

Statistics
For Pearson correlations, you can choose one or both of the following:

Means and standard deviations
Displayed for each variable. The number of cases with nonmissing values is also shown. Missing values are handled on a variable-by-variable basis regardless of your missing values setting.

Cross-product deviations and covariances
Displayed for each pair of variables. The cross-product of deviations is equal to the sum of the products of mean-corrected variables. This is the numerator of the Pearson correlation coefficient. The covariance is an unstandardized measure of the relationship between two variables, equal to the cross-product deviation divided by N–1.

Missing Values
You can choose one of the following:

Exclude cases pairwise
Cases with missing values for one or both of a pair of variables for a correlation coefficient are excluded from the analysis. Since each coefficient is based on all cases that have valid codes on that particular pair of variables, the maximum information available is used in every calculation. This can result in a set of coefficients based on a varying number of cases.

Exclude cases listwise
Cases with missing values for any variable are excluded from all correlations.

Bivariate Correlations Confidence Interval

The Confidence Interval dialog provides options for the estimation of the confidence intervals. The dialog is available when **Pearson**, **Kendall's tau-b**, or **Spearman** is selected on the Bivariate Correlations dialog.

Estimate confidence interval of bivariate correlation parameter
Controls the confidence interval estimation of bivariate correlation parameter. When selected, confidence interval estimation occurs.

Confidence interval (%)
Specifies the confidence level for all confidence intervals produced. Specify a numeric value between 0 and 100. 95 is the default value.

Pearson Correlation

The **Apply the bias adjustment** setting controls whether the bias adjustment is applied. By default, the setting is not selected, which does not take the bias term into consideration. When selected, the bias adjustment to the estimation of the confidence limits is applied. The setting is available when **Pearson** is selected on the Bivariate Correlations dialog.

Spearman Correlation

The setting is available when **Spearman** is selected on the Bivariate Correlations dialog and provides options for estimating the Spearman Correlation variance via the following methods:

- **Fieller, Hartley and Pearson**
- **Bonett and Wright**
- **Coruso and Cliff**

CORRELATIONS and NONPAR CORR Command Additional Features

The command syntax language also allows you to:

- Write a correlation matrix for Pearson correlations that can be used in place of raw data to obtain other analyses such as factor analysis (with the `MATRIX` subcommand).
- Obtain correlations of each variable on a list with each variable on a second list (using the keyword `WITH` on the `VARIABLES` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Partial Correlations

The Partial Correlations procedure computes partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables. Correlations are measures of linear association. Two variables can be perfectly related, but if the relationship is not linear, a correlation coefficient is not an appropriate statistic for measuring their association.

Example

Is there a relationship between healthcare funding and disease rates? Although you might expect any such relationship to be a negative one, a study reports a significant *positive* correlation: as healthcare funding increases, disease rates appear to increase. Controlling for the rate of visits to healthcare providers, however, virtually eliminates the observed positive correlation. Healthcare funding and disease rates only appear to be positively related because more people have access to healthcare when funding increases, which leads to more reported diseases by doctors and hospitals.

Statistics

For each variable: number of cases with nonmissing values, mean, and standard deviation. Partial and zero-order correlation matrices, with degrees of freedom and significance levels.

Data considerations

Data

Use symmetric, quantitative variables.

Assumptions

The Partial Correlations procedure assumes that each pair of variables is bivariate normal.

Obtaining Partial Correlations

1. From the menus choose:

   Analyze > **Correlate** > **Partial...**
2. Select two or more numeric variables for which partial correlations are to be computed.
3. Select one or more numeric control variables.

The following options are also available:

Test of Significance
You can select two-tailed or one-tailed probabilities. If the direction of association is known in advance, select **One-tailed**. Otherwise, select **Two-tailed**.

Display actual significance level
By default, the probability and degrees of freedom are shown for each correlation coefficient. If you deselect this item, coefficients significant at the 0.05 level are identified with a single asterisk, coefficients significant at the 0.01 level are identified with a double asterisk, and degrees of freedom are suppressed. This setting affects both partial and zero-order correlation matrices.

Partial Correlations Options

Statistics. You can choose one or both of the following:

- **Means and standard deviations.** Displayed for each variable. The number of cases with nonmissing values is also shown.
- **Zero-order correlations.** A matrix of simple correlations between all variables, including control variables, is displayed.

Missing Values. You can choose one of the following alternatives:

- **Exclude cases listwise.** Cases having missing values for any variable, including a control variable, are excluded from all computations.
- **Exclude cases pairwise.** For computation of the zero-order correlations on which the partial correlations are based, a case having missing values for both or one of a pair of variables is not used. Pairwise deletion uses as much of the data as possible. However, the number of cases may differ across coefficients. When pairwise deletion is in effect, the degrees of freedom for a particular partial coefficient are based on the smallest number of cases used in the calculation of any of the zero-order correlations.

PARTIAL CORR Command Additional Features

The command syntax language also allows you to:

- Read a zero-order correlation matrix or write a partial correlation matrix (with the `MATRIX` subcommand).
- Obtain partial correlations between two lists of variables (using the keyword `WITH` on the `VARIABLES` subcommand).
- Obtain multiple analyses (with multiple `VARIABLES` subcommands).
- Specify order values to request (for example, both first- and second-order partial correlations) when you have two control variables (with the `VARIABLES` subcommand).
- Suppress redundant coefficients (with the `FORMAT` subcommand).
- Display a matrix of simple correlations when some coefficients cannot be computed (with the `STATISTICS` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Distances

This procedure calculates any of a wide variety of statistics measuring either similarities or dissimilarities (distances), either between pairs of variables or between pairs of cases. These similarity or distance measures can then be used with other procedures, such as factor analysis, cluster analysis, or multidimensional scaling, to help analyze complex datasets.

Example. Is it possible to measure similarities between pairs of automobiles based on certain characteristics, such as engine size, MPG, and horsepower? By computing similarities between autos, you can gain a sense of which autos are similar to each other and which are different from each other. For

a more formal analysis, you might consider applying a hierarchical cluster analysis or multidimensional scaling to the similarities to explore the underlying structure.

Statistics. Dissimilarity (distance) measures for interval data are Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, or customized; for count data, chi-square or phi-square; for binary data, Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, shape, or Lance and Williams. Similarity measures for interval data are Pearson correlation or cosine; for binary data, Russel and Rao, simple matching, Jaccard, dice, Rogers and Tanimoto, Sokal and Sneath 1, Sokal and Sneath 2, Sokal and Sneath 3, Kulczynski 1, Kulczynski 2, Sokal and Sneath 4, Hamann, Lambda, Anderberg's *D*, Yule's *Y*, Yule's *Q*, Ochiai, Sokal and Sneath 5, phi 4-point correlation, or dispersion.

To Obtain Distance Matrices

1. From the menus choose:

   Analyze > **Correlate** > **Distances...**
2. Select at least one numeric variable to compute distances between cases, or select at least two numeric variables to compute distances between variables.
3. Select an alternative in the Compute Distances group to calculate proximities either between cases or between variables.

Distances Dissimilarity Measures

From the Measure group, select the alternative that corresponds to your type of data (interval, count, or binary); then, from the drop-down list, select one of the measures that corresponds to that type of data. Available measures, by data type, are:

- **Interval data.** Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, or customized.
- **Count data.** Chi-square measure or phi-square measure.
- **Binary data.** Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, shape, or Lance and Williams. (Enter values for Present and Absent to specify which two values are meaningful; Distances will ignore all other values.)

The Transform Values group allows you to standardize data values for either cases or variables *before* computing proximities. These transformations are not applicable to binary data. Available standardization methods are *z* scores, range –1 to 1, range 0 to 1, maximum magnitude of 1, mean of 1, or standard deviation of 1.

The Transform Measures group allows you to transform the values generated by the distance measure. They are applied after the distance measure has been computed. Available options are absolute values, change sign, and rescale to 0–1 range.

Distances Similarity Measures

From the Measure group, select the alternative that corresponds to your type of data (interval or binary); then, from the drop-down list, select one of the measures that corresponds to that type of data. Available measures, by data type, are:

- **Interval data.** Pearson correlation or cosine.
- **Binary data.** Russell and Rao, simple matching, Jaccard, Dice, Rogers and Tanimoto, Sokal and Sneath 1, Sokal and Sneath 2, Sokal and Sneath 3, Kulczynski 1, Kulczynski 2, Sokal and Sneath 4, Hamann, Lambda, Anderberg's *D*, Yule's *Y*, Yule's *Q*, Ochiai, Sokal and Sneath 5, phi 4-point correlation, or dispersion. (Enter values for Present and Absent to specify which two values are meaningful; Distances will ignore all other values.)

The Transform Values group allows you to standardize data values for either cases or variables before computing proximities. These transformations are not applicable to binary data. Available standardization methods are *z* scores, range –1 to 1, range 0 to 1, maximum magnitude of 1, mean of 1, and standard deviation of 1.

The Transform Measures group allows you to transform the values generated by the distance measure. They are applied after the distance measure has been computed. Available options are absolute values, change sign, and rescale to 0–1 range.

PROXIMITIES Command Additional Features

The Distances procedure uses `PROXIMITIES` command syntax. The command syntax language also allows you to:

- Specify any integer as the power for the Minkowski distance measure.
- Specify any integers as the power and root for a customized distance measure.

See the *Command Syntax Reference* for complete syntax information.

Linear models

Linear models predict a continuous target based on linear relationships between the target and one or more predictors.

Linear models are relatively simple and give an easily interpreted mathematical formula for scoring. The properties of these models are well understood and can typically be built very quickly compared to other model types (such as neural networks or decision trees) on the same dataset.

Example. An insurance company with limited resources to investigate homeowners' insurance claims wants to build a model for estimating claims costs. By deploying this model to service centers, representatives can enter claim information while on the phone with a customer and immediately obtain the "expected" cost of the claim based on past data.

Field requirements. There must be a Target and at least one Input. By default, fields with predefined roles of Both or None are not used. The target must be continuous (scale). There are no measurement level restrictions on predictors (inputs); categorical (nominal, and ordinal) fields are used as factors in the model and continuous fields are used as covariates.

Note: If a categorical field has more than 1000 categories, the procedure does not run and no model is built.

To obtain a linear model

This feature requires the Statistics Base option.

From the menus choose:

Analyze > **Regression** > **Automatic Linear Models...**

1. Make sure there is at least one target and one input.
2. Click **Build Options** to specify optional build and model settings.
3. Click **Model Options** to save scores to the active dataset and export the model to an external file.
4. Click **Run** to run the procedure and create the Model objects.

Objectives

What is your main objective? Select the appropriate objective.

- **Create a standard model.** The method builds a single model to predict the target using the predictors. Generally speaking, standard models are easier to interpret and can be faster to score than boosted, bagged, or large dataset ensembles.
- **Enhance model accuracy (boosting).** The method builds an ensemble model using boosting, which generates a sequence of models to obtain more accurate predictions. Ensembles can take longer to build and to score than a standard model.

Boosting produces a succession of "component models", each of which is built on the entire dataset. Prior to building each successive component model, the records are weighted based on the previous component model's residuals. Cases with large residuals are given relatively higher analysis weights so that the next component model will focus on predicting these records well. Together these component models form an ensemble model. The ensemble model scores new records using a combining rule; the available rules depend upon the measurement level of the target.

- **Enhance model stability (bagging).** The method builds an ensemble model using bagging (bootstrap aggregating), which generates multiple models to obtain more reliable predictions. Ensembles can take longer to build and to score than a standard model.

  Bootstrap aggregation (bagging) produces replicates of the training dataset by sampling with replacement from the original dataset. This creates bootstrap samples of equal size to the original dataset. Then a "component model" is built on each replicate. Together these component models form an ensemble model. The ensemble model scores new records using a combining rule; the available rules depend upon the measurement level of the target.
- **Create a model for very large datasets (requires IBM SPSS Statistics Server).** The method builds an ensemble model by splitting the dataset into separate data blocks. Choose this option if your dataset is too large to build any of the models above, or for incremental model building. This option can take less time to build, but can take longer to score than a standard model. This option requires IBM SPSS Statistics Server connectivity.

See “Ensembles ” on page 133 for settings related to boosting, bagging, and very large datasetsboosting and bagging.

Basics

Automatically prepare data. This option allows the procedure to internally transform the target and predictors in order to maximize the predictive power of the model; any transformations are saved with the model and applied to new data for scoring. The original versions of transformed fields are excluded from the model. By default, the following automatic data preparation are performed.

- **Date and Time handling.** Each date predictor is transformed into new a continuous predictor containing the elapsed time since a reference date (1970-01-01). Each time predictor is transformed into a new continuous predictor containing the time elapsed since a reference time (00:00:00).
- **Adjust measurement level.** Continuous predictors with less than 5 distinct values are recast as ordinal predictors. Ordinal predictors with greater than 10 distinct values are recast as continuous predictors.
- **Outlier handling.** Values of continuous predictors that lie beyond a cutoff value (3 standard deviations from the mean) are set to the cutoff value.
- **Missing value handling.** Missing values of nominal predictors are replaced with the mode of the training partition. Missing values of ordinal predictors are replaced with the median of the training partition. Missing values of continuous predictors are replaced with the mean of the training partition.
- **Supervised merging.** This makes a more parsimonious model by reducing the number of fields to be processed in association with the target. Similar categories are identified based upon the relationship between the input and the target. Categories that are not significantly different (that is, having a p-value greater than 0.1) are merged. If all categories are merged into one, the original and derived versions of the field are excluded from the model because they have no value as a predictor.

Confidence level. This is the level of confidence used to compute interval estimates of the model coefficients in the Coefficients view. Specify a value greater than 0 and less than 100. The default is 95.

Model Selection

Model selection method. Choose one of the model selection methods (details below) or **Include all predictors**, which simply enters all available predictors as main effects model terms. By default, **Forward stepwise** is used.

Forward Stepwise Selection. This starts with no effects in the model and adds and removes effects one step at a time until no more can be added or removed according to the stepwise criteria.

- **Criteria for entry/removal.** This is the statistic used to determine whether an effect should be added to or removed from the model. **Information Criterion (AICC)** is based on the likelihood of the training set given the model, and is adjusted to penalize overly complex models. **F Statistics** is based on a statistical test of the improvement in model error. **Adjusted R-squared** is based on the fit of the training set, and is adjusted to penalize overly complex models. **Overfit Prevention Criterion (ASE)** is based on the fit (average squared error, or ASE) of the overfit prevention set. The overfit prevention set is a random subsample of approximately 30% of the original dataset that is not used to train the model.

  If any criterion other than **F Statistics** is chosen, then at each step the effect that corresponds to the greatest positive increase in the criterion is added to the model. Any effects in the model that correspond to a decrease in the criterion are removed.

  If **F Statistics** is chosen as the criterion, then at each step the effect that has the smallest *p*-value less than the specified threshold, **Include effects with p-values less than**, is added to the model. The default is 0.05. Any effects in the model with a *p*-value greater than the specified threshold, **Remove effects with p-values greater than**, are removed. The default is 0.10.
- **Customize maximum number of effects in the final model.** By default, all available effects can be entered into the model. Alternatively, if the stepwise algorithm ends a step with the specified maximum number of effects, the algorithm stops with the current set of effects.
- **Customize maximum number of steps.** The stepwise algorithm stops after a certain number of steps. By default, this is 3 times the number of available effects. Alternatively, specify a positive integer maximum number of steps.

Best Subsets Selection. This checks "all possible" models, or at least a larger subset of the possible models than forward stepwise, to choose the best according to the best subsets criterion. **Information Criterion (AICC)** is based on the likelihood of the training set given the model, and is adjusted to penalize overly complex models. **Adjusted R-squared** is based on the fit of the training set, and is adjusted to penalize overly complex models. **Overfit Prevention Criterion (ASE)** is based on the fit (average squared error, or ASE) of the overfit prevention set. The overfit prevention set is a random subsample of approximately 30% of the original dataset that is not used to train the model.

The model with the greatest value of the criterion is chosen as the best model.

Note: Best subsets selection is more computationally intensive than forward stepwise selection. When best subsets is performed in conjunction with boosting, bagging, or very large datasets, it can take considerably longer to build than a standard model built using forward stepwise selection.

Ensembles

These settings determine the behavior of ensembling that occurs when boosting, bagging, or very large datasets are requested in Objectives. Options that do not apply to the selected objective are ignored.

Bagging and Very Large Datasets. When scoring an ensemble, this is the rule used to combine the predicted values from the base models to compute the ensemble score value.

- **Default combining rule for continuous targets.** Ensemble predicted values for continuous targets can be combined using the mean or median of the predicted values from the base models.

Note that when the objective is to enhance model accuracy, the combining rule selections are ignored. Boosting always uses a weighted majority vote to score categorical targets and a weighted median to score continuous targets.

Boosting and Bagging. Specify the number of base models to build when the objective is to enhance model accuracy or stability; for bagging, this is the number of bootstrap samples. It should be a positive integer.

Advanced

Replicate results. Setting a random seed allows you to replicate analyses. The random number generator is used to choose which records are in the overfit prevention set. Specify an integer or click **Generate**, which will create a pseudo-random integer between 1 and 2147483647, inclusive. The default is 54752075.

Model Options

Save predicted values to the dataset. The default variable name is *PredictedValue*.

Export model. This writes the model to an external `.zip` file. You can use this model file to apply the model information to other data files for scoring purposes. Specify a unique, valid file name. If the file specification refers to an existing file, then the file is overwritten.

Model Summary

The Model Summary view is a snapshot, at-a-glance summary of the model and its fit.

Table. The table identifies some high-level model settings, including:

- The name of the target specified on the Fields tab,
- Whether automatic data preparation was performed as specified on the Basicssettings,
- The model selection method and selection criterion specified on the Model Selectionsettings. The value of the selection criterion for the final model is also displayed, and is presented in smaller is better format.

Chart. The chart displays the accuracy of the final model, which is presented in larger is better format. The value is 100 × the adjusted R^2 for the final model.

Automatic Data Preparation

This view shows information about which fields were excluded and how transformed fields were derived in the automatic data preparation (ADP) step. For each field that was transformed or excluded, the table lists the field name, its role in the analysis, and the action taken by the ADP step. Fields are sorted by ascending alphabetical order of field names. The possible actions taken for each field include:

- **Derive duration: months** computes the elapsed time in months from the values in a field containing dates to the current system date.
- **Derive duration: hours** computes the elapsed time in hours from the values in a field containing times to the current system time.
- **Change measurement level from continuous to ordinal** recasts continuous fields with less than 5 unique values as ordinal fields.
- **Change measurement level from ordinal to continuous** recasts ordinal fields with more than 10 unique values as continuous fields.
- **Trim outliers** sets values of continuous predictors that lie beyond a cutoff value (3 standard deviations from the mean) to the cutoff value.
- **Replace missing values** replaces missing values of nominal fields with the mode, ordinal fields with the median, and continuous fields with the mean.
- **Merge categories to maximize association with target** identifies "similar" predictor categories based upon the relationship between the input and the target. Categories that are not significantly different (that is, having a *p*-value greater than 0.05) are merged.
- **Exclude constant predictor / after outlier handling / after merging of categories** removes predictors that have a single value, possibly after other ADP actions have been taken.

Predictor Importance

Typically, you will want to focus your modeling efforts on the predictor fields that matter most and consider dropping or ignoring those that matter least. The predictor importance chart helps you do this by indicating the relative importance of each predictor in estimating the model. Since the values are relative, the sum of the values for all predictors on the display is 1.0. Predictor importance does not relate to model accuracy. It just relates to the importance of each predictor in making a prediction, not whether or not the prediction is accurate.

Predicted By Observed

This displays a binned scatterplot of the predicted values on the vertical axis by the observed values on the horizontal axis. Ideally, the points should lie on a 45-degree line; this view can tell you whether any records are predicted particularly badly by the model.

Residuals

This displays a diagnostic chart of model residuals.

Chart styles. There are different display styles, which are accessible from the **Style** dropdown list.

- **Histogram.** This is a binned histogram of the studentized residuals with an overlay of the normal distribution. Linear models assume that the residuals have a normal distribution, so the histogram should ideally closely approximate the smooth line.
- **P-P Plot.** This is a binned probability-probability plot comparing the studentized residuals to a normal distribution. If the slope of the plotted points is less steep than the normal line, the residuals show greater variability than a normal distribution; if the slope is steeper, the residuals show less variability than a normal distribution. If the plotted points have an S-shaped curve, then the distribution of residuals is skewed.

Outliers

This table lists records that exert undue influence upon the model, and displays the record ID (if specified on the Fields tab), target value, and Cook's distance. Cook's distance is a measure of how much the residuals of all records would change if a particular record were excluded from the calculation of the model coefficients. A large Cook's distance indicates that excluding a record from changes the coefficients substantially, and should therefore be considered influential.

Influential records should be examined carefully to determine whether you can give them less weight in estimating the model, or truncate the outlying values to some acceptable threshold, or remove the influential records completely.

Effects

This view displays the size of each effect in the model.

Styles. There are different display styles, which are accessible from the **Style** dropdown list.

- **Diagram.** This is a chart in which effects are sorted from top to bottom by decreasing predictor importance. Connecting lines in the diagram are weighted based on effect significance, with greater line width corresponding to more significant effects (smaller *p*-values). Hovering over a connecting line reveals a tooltip that shows the *p*-value and importance of the effect. This is the default.
- **Table.** This is an ANOVA table for the overall model and the individual model effects. The individual effects are sorted from top to bottom by decreasing predictor importance. Note that by default, the table is collapsed to only show the results for the overall model. To see the results for the individual model effects, click the **Corrected Model** cell in the table.

Predictor importance. There is a Predictor Importance slider that controls which predictors are shown in the view. This does not change the model, but simply allows you to focus on the most important predictors. By default, the top 10 effects are displayed.

Significance. There is a Significance slider that further controls which effects are shown in the view, beyond those shown based on predictor importance. Effects with significance values greater than the slider value are hidden. This does not change the model, but simply allows you to focus on the most important effects. By default the value is 1.00, so that no effects are filtered based on significance.

Coefficients

This view displays the value of each coefficient in the model. Note that factors (categorical predictors) are indicator-coded within the model, so that **effects** containing factors will generally have multiple

associated **coefficients**; one for each category except the category corresponding to the redundant (reference) parameter.

Styles. There are different display styles, which are accessible from the **Style** dropdown list.

- **Diagram.** This is a chart which displays the intercept first, and then sorts effects from top to bottom by decreasing predictor importance. Within effects containing factors, coefficients are sorted by ascending order of data values. Connecting lines in the diagram are colored based on the sign of the coefficient (see the diagram key) and weighted based on coefficient significance, with greater line width corresponding to more significant coefficients (smaller *p*-values). Hovering over a connecting line reveals a tooltip that shows the value of the coefficient, its *p*-value, and the importance of the effect the parameter is associated with. This is the default style.
- **Table.** This shows the values, significance tests, and confidence intervals for the individual model coefficients. After the intercept, the effects are sorted from top to bottom by decreasing predictor importance. Within effects containing factors, coefficients are sorted by ascending order of data values. Note that by default the table is collapsed to only show the coefficient, significance, and importance of each model parameter. To see the standard error, *t* statistic, and confidence interval, click the **Coefficient** cell in the table. Hovering over the name of a model parameter in the table reveals a tooltip that shows the name of the parameter, the effect the parameter is associated with, and (for categorical predictors), the value labels associated with the model parameter. This can be particularly useful to see the new categories created when automatic data preparation merges similar categories of a categorical predictor.

Predictor importance. There is a Predictor Importance slider that controls which predictors are shown in the view. This does not change the model, but simply allows you to focus on the most important predictors. By default, the top 10 effects are displayed.

Significance. There is a Significance slider that further controls which coefficients are shown in the view, beyond those shown based on predictor importance. Coefficients with significance values greater than the slider value are hidden. This does not change the model, but simply allows you to focus on the most important coefficients. By default the value is 1.00, so that no coefficients are filtered based on significance.

Estimated Means

These are charts displayed for significant predictors. The chart displays the model-estimated value of the target on the vertical axis for each value of the predictor on the horizontal axis, holding all other predictors constant. It provides a useful visualization of the effects of each predictor's coefficients on the target.

Note: if no predictors are significant, no estimated means are produced.

Model Building Summary

When a model selection algorithm other than **None** is chosen on the Model Selection settings, this provides some details of the model building process.

Forward stepwise. When forward stepwise is the selection algorithm, the table displays the last 10 steps in the stepwise algorithm. For each step, the value of the selection criterion and the effects in the model at that step are shown. This gives you a sense of how much each step contributes to the model. Each column allows you to sort the rows so that you can more easily see which effects are in the model at a given step.

Best subsets. When best subsets is the selection algorithm, the table displays the top 10 models. For each model, the value of the selection criterion and the effects in the model are shown. This gives you a sense of the stability of the top models; if they tend to have many similar effects with a few differences, then you can be fairly confident in the "top" model; if they tend to have very different effects, then some of the effects may be too similar and should be combined (or one removed). Each column allows you to sort the rows so that you can more easily see which effects are in the model at a given step.

Linear Regression

Linear Regression estimates the coefficients of the linear equation, involving one or more independent variables, that best predict the value of the dependent variable. For example, you can try to predict a salesperson's total yearly sales (the dependent variable) from independent variables such as age, education, and years of experience.

Example. Is the number of games won by a basketball team in a season related to the average number of points the team scores per game? A scatterplot indicates that these variables are linearly related. The number of games won and the average number of points scored by the opponent are also linearly related. These variables have a negative relationship. As the number of games won increases, the average number of points scored by the opponent decreases. With linear regression, you can model the relationship of these variables. A good model can be used to predict how many games teams will win.

Statistics. For each variable: number of valid cases, mean, and standard deviation. For each model: regression coefficients, correlation matrix, part and partial correlations, multiple R, R^2, adjusted R^2, change in R^2, standard error of the estimate, analysis-of-variance table, predicted values, and residuals. Also, 95%-confidence intervals for each regression coefficient, variance-covariance matrix, variance inflation factor, tolerance, Durbin-Watson test, distance measures (Mahalanobis, Cook, and leverage values), DfBeta, DfFit, prediction intervals, and casewise diagnostic information. Plots: scatterplots, partial plots, histograms, and normal probability plots.

Linear Regression Data Considerations

Data. The dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.

Assumptions. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear, and all observations should be independent.

To Obtain a Linear Regression Analysis

1. From the menus choose:

   Analyze > **Regression** > **Linear...**
2. In the Linear Regression dialog box, select a numeric dependent variable.
3. Select one or more numeric independent variables.

Optionally, you can:

- Group independent variables into blocks and specify different entry methods for different subsets of variables.
- Choose a selection variable to limit the analysis to a subset of cases having a particular value(s) for this variable.
- Select a case identification variable for identifying points on plots.
- Select a numeric WLS Weight variable for a weighted least squares analysis.

WLS. Allows you to obtain a weighted least-squares model. Data points are weighted by the reciprocal of their variances. This means that observations with large variances have less impact on the analysis than observations associated with small variances. If the value of the weighting variable is zero, negative, or missing, the case is excluded from the analysis.

Linear Regression Variable Selection Methods

Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables.

- *Enter (Regression)*. A procedure for variable selection in which all variables in a block are entered in a single step.

- *Stepwise*. At each step, the independent variable not in the equation that has the smallest probability of F is entered, if that probability is sufficiently small. Variables already in the regression equation are removed if their probability of F becomes sufficiently large. The method terminates when no more variables are eligible for inclusion or removal.
- *Remove*. A procedure for variable selection in which all variables in a block are removed in a single step.
- *Backward Elimination*. A variable selection procedure in which all variables are entered into the equation and then sequentially removed. The variable with the smallest partial correlation with the dependent variable is considered first for removal. If it meets the criterion for elimination, it is removed. After the first variable is removed, the variable remaining in the equation with the smallest partial correlation is considered next. The procedure stops when there are no variables in the equation that satisfy the removal criteria.
- *Forward Selection*. A stepwise variable selection procedure in which variables are sequentially entered into the model. The first variable considered for entry into the equation is the one with the largest positive or negative correlation with the dependent variable. This variable is entered into the equation only if it satisfies the criterion for entry. If the first variable is entered, the independent variable not in the equation that has the largest partial correlation is considered next. The procedure stops when there are no variables that meet the entry criterion.

The significance values in your output are based on fitting a single model. Therefore, the significance values are generally invalid when a stepwise method (stepwise, forward, or backward) is used.

All variables must pass the tolerance criterion to be entered in the equation, regardless of the entry method specified. The default tolerance level is 0.0001. Also, a variable is not entered if it would cause the tolerance of another variable already in the model to drop below the tolerance criterion.

All independent variables selected are added to a single regression model. However, you can specify different entry methods for different subsets of variables. For example, you can enter one block of variables into the regression model using stepwise selection and a second block using forward selection. To add a second block of variables to the regression model, click **Next**.

Linear Regression Set Rule

Cases defined by the selection rule are included in the analysis. For example, if you select a variable, choose **equals**, and type 5 for the value, then only cases for which the selected variable has a value equal to 5 are included in the analysis. A string value is also permitted.

Linear Regression Plots

Plots can aid in the validation of the assumptions of normality, linearity, and equality of variances. Plots are also useful for detecting outliers, unusual observations, and influential cases. After saving them as new variables, predicted values, residuals, and other diagnostic information are available in the Data Editor for constructing plots with the independent variables. The following plots are available:

Scatterplots. You can plot any two of the following: the dependent variable, standardized predicted values, standardized residuals, deleted residuals, adjusted predicted values, Studentized residuals, or Studentized deleted residuals. Plot the standardized residuals against the standardized predicted values to check for linearity and equality of variances.

Source variable list. Lists the dependent variable (DEPENDNT) and the following predicted and residual variables: Standardized predicted values (*ZPRED), Standardized residuals (*ZRESID), Deleted residuals (*DRESID), Adjusted predicted values (*ADJPRED), Studentized residuals (*SRESID), Studentized deleted residuals (*SDRESID).

Produce all partial plots. Displays scatterplots of residuals of each independent variable and the residuals of the dependent variable when both variables are regressed separately on the rest of the independent variables. At least two independent variables must be in the equation for a partial plot to be produced.

Standardized Residual Plots. You can obtain histograms of standardized residuals and normal probability plots comparing the distribution of standardized residuals to a normal distribution.

If any plots are requested, summary statistics are displayed for standardized predicted values and standardized residuals (***ZPRED** and ***ZRESID**).

Linear Regression: Saving New Variables

You can save predicted values, residuals, and other statistics useful for diagnostic information. Each selection adds one or more new variables to your active data file.

Predicted Values. Values that the regression model predicts for each case.

- *Unstandardized*. The value the model predicts for the dependent variable.
- *Standardized*. A transformation of each predicted value into its standardized form. That is, the mean predicted value is subtracted from the predicted value, and the difference is divided by the standard deviation of the predicted values. Standardized predicted values have a mean of 0 and a standard deviation of 1.
- *Adjusted*. The predicted value for a case when that case is excluded from the calculation of the regression coefficients.
- *S.E. of mean predictions*. Standard errors of the predicted values. An estimate of the standard deviation of the average value of the dependent variable for cases that have the same values of the independent variables.

Distances. Measures to identify cases with unusual combinations of values for the independent variables and cases that may have a large impact on the regression model.

- *Mahalanobis*. A measure of how much a case's values on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables.
- *Cook's*. A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. A large Cook's D indicates that excluding a case from computation of the regression statistics changes the coefficients substantially.
- *Leverage values*. Measures the influence of a point on the fit of the regression. The centered leverage ranges from 0 (no influence on the fit) to (N-1)/N.

Prediction Intervals. The upper and lower bounds for both mean and individual prediction intervals.

- *Mean*. Lower and upper bounds (two variables) for the prediction interval of the mean predicted response.
- *Individual*. Lower and upper bounds (two variables) for the prediction interval of the dependent variable for a single case.
- *Confidence Interval*. Enter a value between 1 and 99.99 to specify the confidence level for the two Prediction Intervals. Mean or Individual must be selected before entering this value. Typical confidence interval values are 90, 95, and 99.

Residuals. The actual value of the dependent variable minus the value predicted by the regression equation.

- *Unstandardized*. The difference between an observed value and the value predicted by the model.
- *Standardized*. The residual divided by an estimate of its standard deviation. Standardized residuals, which are also known as Pearson residuals, have a mean of 0 and a standard deviation of 1.
- *Studentized*. The residual divided by an estimate of its standard deviation that varies from case to case, depending on the distance of each case's values on the independent variables from the means of the independent variables. Sometimes referred to as internally studentized residuals.
- *Deleted*. The residual for a case when that case is excluded from the calculation of the regression coefficients. It is the difference between the value of the dependent variable and the adjusted predicted value.
- *Studentized deleted*. The deleted residual for a case divided by its standard error. The difference between a Studentized deleted residual and its associated Studentized residual indicates how much

difference eliminating a case makes on its own prediction. Sometimes referred to as externally studentized residuals.

Influence Statistics. The change in the regression coefficients (DfBeta[s]) and predicted values (DfFit) that results from the exclusion of a particular case. Standardized DfBetas and DfFit values are also available along with the covariance ratio.

- *DfBetas*. The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant.
- *Standardized DfBeta*. Standardized difference in beta value. The change in the regression coefficient that results from the exclusion of a particular case. You may want to examine cases with absolute values greater than 2 divided by the square root of N, where N is the number of cases. A value is computed for each term in the model, including the constant.
- *DfFit*. The difference in fit value is the change in the predicted value that results from the exclusion of a particular case.
- *Standardized DfFit*. Standardized difference in fit value. The change in the predicted value that results from the exclusion of a particular case. You may want to examine standardized values which in absolute value exceed 2 times the square root of p/N, where p is the number of parameters in the model and N is the number of cases.
- *Covariance ratio*. The ratio of the determinant of the covariance matrix with a particular case excluded from the calculation of the regression coefficients to the determinant of the covariance matrix with all cases included. If the ratio is close to 1, the case does not significantly alter the covariance matrix.

Coefficient Statistics. Saves regression coefficients to a dataset or a data file. Datasets are available for subsequent use in the same session but are not saved as files unless explicitly saved prior to the end of the session. Dataset names must conform to variable naming rules.

Export model information to XML file. Parameter estimates and (optionally) their covariances are exported to the specified file in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.

Linear Regression Statistics

The following statistics are available:

Regression Coefficients - Estimates displays Regression coefficient *B*, standard error of *B*, standardized coefficient beta, *t* value for *B*, and two-tailed significance level of *t*. **Confidence intervals** displays confidence intervals with the specified level of confidence for each regression coefficient or a covariance matrix. **Covariance matrix** displays a variance-covariance matrix of regression coefficients with covariances off the diagonal and variances on the diagonal. A correlation matrix is also displayed.

Model fit - The variables that are entered and removed from the model are listed, and the following 'goodness-of-fit statistics' are displayed: multiple *R*, R^2 and adjusted R^2, standard error of the estimate, and an analysis-of-variance table.

R squared change - The change in the R^2 statistic that is produced by adding or deleting an independent variable. If the R^2 change associated with a variable is large, that means that the variable is a good predictor of the dependent variable.

Descriptive - Provides the number of valid cases, the mean, and the standard deviation for each variable in the analysis. A correlation matrix with a one-tailed significance level and the number of cases for each correlation are also displayed.

Part Correlation. The correlation between the dependent variable and an independent variable when the linear effects of the other independent variables in the model have been removed from the independent variable. It is related to the change in R-squared when a variable is added to an equation. Sometimes called the semipartial correlation.

Partial Correlation. The correlation that remains between two variables after removing the correlation that is due to their mutual association with the other variables. The correlation between the dependent

variable and an independent variable when the linear effects of the other independent variables in the model have been removed from both.

Collinearity diagnostics - Collinearity (or multicollinearity) is the undesirable situation when one independent variable is a linear function of other independent variables. Eigenvalues of the scaled and uncentered cross-products matrix, condition indices, and variance-decomposition proportions are displayed along with variance inflation factors (VIF) and tolerances for individual variables.

Selection criteria - Includes Akaike information criterion (AIC), Ameniya's prediction criterion (PC), Mallows conditional mean squared error of prediction criterion (Cp), and Schwarz Bayesian criterion (SBC). The statistics are displayed in the Model Summary table.

Residuals - You can select **'PRESS Statistic'** to use as a cross-validation statistic to compare different models. This also displays the **'Durban-Watson'** test for serial correlation of the residuals. Choose the **'Case wise diagnostic'** information for the cases that meet the selection criterion (outliers above *n* standard deviations).

Linear Regression Options

The following options are available:

Stepping Method Criteria. These options apply when either the forward, backward, or stepwise variable selection method is specified. Variables can be entered or removed from the model depends on either the significance (probability) of the *F* value or the *F* value itself.

- *Use Probability of F*. A variable is entered into the model if the significance level of its F value is less than the Entry value and is removed if the significance level is greater than the Removal value. Entry must be less than Removal, and both values must be positive. To enter more variables into the model, increase the Entry value. To remove more variables from the model, lower the Removal value.
- *Use F Value*. A variable is entered into the model if its F value is greater than the Entry value and is removed if the F value is less than the Removal value. Entry must be greater than Removal, and both values must be positive. To enter more variables into the model, lower the Entry value. To remove more variables from the model, increase the Removal value.

Tolerance. By default, the value is .0001. Tolerance is the proportion of the variance of a variable in the equation that is not accounted for by other independent variables in the equation. The minimum tolerance for any variable in the equation if the variable under consideration was included in the analysis is the minimum tolerance of a variable that is not included in the equation. Variables must pass tolerance and minimum tolerance tests to enter and remain in a regression equation. If a variable passes the tolerance criteria, it is eligible for inclusion based on the method in effect.

Include constant in equation. By default, the regression model includes a constant term. Deselecting this option forces regression through the origin, which is rarely done. Some results of regression through the origin are not comparable to results of regression that do include a constant. For example, R^2 cannot be interpreted in the usual way.

Missing Values. You can choose one of the following:

- **Exclude cases listwise.** Only cases with valid values for all variables are included in the analyses.
- **Exclude cases pairwise.** Cases with complete data for the pair of variables being correlated are used to compute the correlation coefficient on which the regression analysis is based. Degrees of freedom are based on the minimum pairwise *N*.
- **Replace with mean.** All cases are used for computations, with the mean of the variable substituted for missing observations.

REGRESSION Command Additional Features

The command syntax language also allows you to:

- Write a correlation matrix or read a matrix in place of raw data to obtain your regression analysis (with the `MATRIX` subcommand).

- Specify tolerance levels (with the `CRITERIA` subcommand).
- Obtain multiple models for the same or different dependent variables (with the `METHOD` and `DEPENDENT` subcommands).
- Obtain additional statistics (with the `DESCRIPTIVES` and `STATISTICS` subcommands).

See the *Command Syntax Reference* for complete syntax information.

Ordinal Regression

Ordinal Regression allows you to model the dependence of a polytomous ordinal response on a set of predictors, which can be factors or covariates. The design of Ordinal Regression is based on the methodology of McCullagh (1980, 1998), and the procedure is referred to as `PLUM` in the syntax.

Standard linear regression analysis involves minimizing the sum-of-squared differences between a response (dependent) variable and a weighted combination of predictor (independent) variables. The estimated coefficients reflect how changes in the predictors affect the response. The response is assumed to be numerical, in the sense that changes in the level of the response are equivalent throughout the range of the response. For example, the difference in height between a person who is 150 cm tall and a person who is 140 cm tall is 10 cm, which has the same meaning as the difference in height between a person who is 210 cm tall and a person who is 200 cm tall. These relationships do not necessarily hold for ordinal variables, in which the choice and number of response categories can be quite arbitrary.

Example. Ordinal Regression could be used to study patient reaction to drug dosage. The possible reactions may be classified as *none*, *mild*, *moderate*, or *severe*. The difference between a mild and moderate reaction is difficult or impossible to quantify and is based on perception. Moreover, the difference between a mild and moderate response may be greater or less than the difference between a moderate and severe response.

Statistics and plots. Observed and expected frequencies and cumulative frequencies, Pearson residuals for frequencies and cumulative frequencies, observed and expected probabilities, observed and expected cumulative probabilities of each response category by covariate pattern, asymptotic correlation and covariance matrices of parameter estimates, Pearson's chi-square and likelihood-ratio chi-square, goodness-of-fit statistics, iteration history, test of parallel lines assumption, parameter estimates, standard errors, confidence intervals, and Cox and Snell's, Nagelkerke's, and McFadden's R^2 statistics.

Ordinal Regression Data Considerations

Data. The dependent variable is assumed to be ordinal and can be numeric or string. The ordering is determined by sorting the values of the dependent variable in ascending order. The lowest value defines the first category. Factor variables are assumed to be categorical. Covariate variables must be numeric. Note that using more than one continuous covariate can easily result in the creation of a very large cell probabilities table.

Assumptions. Only one response variable is allowed, and it must be specified. Also, for each distinct pattern of values across the independent variables, the responses are assumed to be independent multinomial variables.

Related procedures. Nominal logistic regression uses similar models for nominal dependent variables.

Obtaining an Ordinal Regression

1. From the menus choose:

   Analyze > **Regression** > **Ordinal...**
2. Select one dependent variable.
3. Click **OK**.

Ordinal Regression Options

The Options dialog box allows you to adjust parameters used in the iterative estimation algorithm, choose a level of confidence for your parameter estimates, and select a link function.

Iterations. You can customize the iterative algorithm.

- **Maximum iterations.** Specify a non-negative integer. If 0 is specified, the procedure returns the initial estimates.
- **Maximum step-halving.** Specify a positive integer.
- **Log-likelihood convergence.** The algorithm stops if the absolute or relative change in the log-likelihood is less than this value. The criterion is not used if 0 is specified.
- **Parameter convergence.** The algorithm stops if the absolute or relative change in each of the parameter estimates is less than this value. The criterion is not used if 0 is specified.

Confidence interval. Specify a value greater than or equal to 0 and less than 100.

Delta. The value added to zero cell frequencies. Specify a non-negative value less than 1.

Singularity tolerance. Used for checking for highly dependent predictors. Select a value from the list of options.

Link function. The link function is a transformation of the cumulative probabilities that allows estimation of the model. The following five link functions are available.

- **Logit.** $f(x)=\log(x/(1-x))$. Typically used for evenly distributed categories.
- **Complementary log-log.** $f(x)=\log(-\log(1-x))$. Typically used when higher categories are more probable.
- **Negative log-log.** $f(x)=-\log(-\log(x))$. Typically used when lower categories are more probable.
- **Probit.** $f(x)=\Phi^{-1}(x)$. Typically used when the latent variable is normally distributed.
- **Cauchit (inverse Cauchy).** $f(x)=\tan(\pi(x-0.5))$. Typically used when the latent variable has many extreme values.

Ordinal Regression Output

The Output dialog box allows you to produce tables for display in the Viewer and save variables to the working file.

Display. Produces tables for:

- **Print iteration history for every n step(s).** The log-likelihood and parameter estimates are printed for the print iteration frequency specified. The first and last iterations are always printed.
- **Goodness of fit statistics.** The Pearson and likelihood-ratio chi-square statistics. They are computed based on the classification specified in the variable list.
- **Summary statistics.** Cox and Snell's, Nagelkerke's, and McFadden's R^2 statistics.
- **Parameter estimates.** Parameter estimates, standard errors, and confidence intervals.
- **Asymptotic correlation of parameter estimates.** Matrix of parameter estimate correlations.
- **Asymptotic covariance of parameter estimates.** Matrix of parameter estimate covariances.
- **Cell information.** Observed and expected frequencies and cumulative frequencies, Pearson residuals for frequencies and cumulative frequencies, observed and expected probabilities, and observed and expected cumulative probabilities of each response category by covariate pattern. Note that for models with many covariate patterns (for example, models with continuous covariates), this option can generate a very large, unwieldy table.
- **Test of parallel lines.** Test of the hypothesis that the location parameters are equivalent across the levels of the dependent variable. This is available only for the location-only model.

Saved Variables. Saves the following variables to the working file:

- **Estimated response probabilities.** Model-estimated probabilities of classifying a factor/covariate pattern into the response categories. There are as many probabilities as the number of response categories.
- **Predicted category.** The response category that has the maximum estimated probability for a factor/covariate pattern.

- **Predicted category probability.** Estimated probability of classifying a factor/covariate pattern into the predicted category. This probability is also the maximum of the estimated probabilities of the factor/covariate pattern.
- **Actual category probability.** Estimated probability of classifying a factor/covariate pattern into the actual category.

Print Log-Likelihood. Controls the display of the log-likelihood. **Including multinomial constant** gives you the full value of the likelihood. To compare your results across products that do not include the constant, you can choose to exclude it.

Ordinal Regression Location Model

The Location dialog box allows you to specify the location model for your analysis.

Specify model. A main-effects model contains the covariate and factor main effects but no interaction effects. You can create a custom model to specify subsets of factor interactions or covariate interactions.

Factors/covariates. The factors and covariates are listed.

Location model. The model depends on the main effects and interaction effects that you select.

For the selected factors and covariates:

Interaction
Creates the highest-level interaction term of all selected variables. This is the default.

Main effects
Creates a main-effects term for each variable selected.

All 2-way
Creates all possible two-way interactions of the selected variables.

All 3-way
Creates all possible three-way interactions of the selected variables.

All 4-way
Creates all possible four-way interactions of the selected variables.

All 5-way
Creates all possible five-way interactions of the selected variables.

Build Terms and Custom Terms

Build terms
Use this choice when you want to include non-nested terms of a certain type (such as main effects) for all combinations of a selected set of factors and covariates.

Build custom terms
Use this choice when you want to include nested terms or when you want to explicitly build any term variable by variable. Building a nested term involves the following steps:

Ordinal Regression Scale Model

The Scale dialog box allows you to specify the scale model for your analysis.

Factors/covariates. The factors and covariates are listed.

Scale model. The model depends on the main and interaction effects that you select.

For the selected factors and covariates:

Interaction
Creates the highest-level interaction term of all selected variables. This is the default.

Main effects
Creates a main-effects term for each variable selected.

All 2-way
Creates all possible two-way interactions of the selected variables.

All 3-way
Creates all possible three-way interactions of the selected variables.

All 4-way
Creates all possible four-way interactions of the selected variables.

All 5-way
Creates all possible five-way interactions of the selected variables.

Build Terms and Custom Terms

Build terms
Use this choice when you want to include non-nested terms of a certain type (such as main effects) for all combinations of a selected set of factors and covariates.

Build custom terms
Use this choice when you want to include nested terms or when you want to explicitly build any term variable by variable. Building a nested term involves the following steps:

PLUM Command Additional Features

You can customize your Ordinal Regression if you paste your selections into a syntax window and edit the resulting PLUM command syntax. The command syntax language also allows you to:

- Create customized hypothesis tests by specifying null hypotheses as linear combinations of parameters.

See the *Command Syntax Reference* for complete syntax information.

Linear Elastic Net Regression

Linear Elastic Net uses the Python sklearn.linear_model.ElasticNet class to estimate regularized linear regression models for a dependent variable on one or more independent variables. Regularization combines L1 (Lasso) and L2 (Ridge) penalties. The extension includes optional modes to display trace plots for different values of alpha for a given L1 ratio, and to select the L1 ratio and alpha hyperparameter values based on crossvalidation. When a single model is fitted or crossvalidation is used to select the penalty ratio and/or alpha, a partition of holdout data can be used to estimate out-of-sample performance.

In addition to fitting a model with specified values of the ratio of L1 penalty and alpha regularization parameter, linear elastic net can display a trace plot of coefficient values for a range of alpha values for a given ratio, or facilitate choice of the hyperparameters value via k-fold crossvalidation on specified grids of values. If a single model is fitted or ratio and/or alpha selection via crossvalidation is performed, the final model can be applied to held-out data that is created by a partition of the input data to obtain a valid estimate of out-of-sample performance of the model.

Obtaining a Linear Elastic Net Regression analysis

1. From the menus choose:

   Analyze > **Regression** > **Linear OLS Alternatives** > **Elastic Net**

   The dialog allows you to specify a variable that assigns each case in the active dataset to the training or holdout sample.
2. Select a numeric target variable. Only one target variable is required to run an analysis.
3. Specify a numeric dependent.
4. Specify at least one categorical factor variable or numeric covariate variable.

Optionally, **Partition** provides a way to create a holdout or test subset of the input data for estimation of out-of-sample performance of the specified or chosen model. All partitioning is performed after

listwise deletion of any cases with invalid data for any variable used by the procedure. Note that for crossvalidation, folds or partitions of the training data are created in Python. The holdout data that is created by the partition is not used in estimation, regardless of the mode in effect.

The partition can be defined either by specifying the ratio of cases that are randomly assigned to each sample (under **Training and Holdout partitions**), or by a variable that assigns each case to the training or holdout sample. You cannot specify both training and variables. If the partition is not specified, a holdout sample is created of approximately 30% of the input data is created.

The **Training %** specifies the relative number of cases in the active dataset to randomly assign to the training sample. The default training is 70%.

Linear Elastic Net Regression: Options

The Options tab provides options for:

Mode

This selection provides options for specifying one of the following modes:

Fit with specified L1 ratio and alpha

When you select this mode, a single model is fitted to the training data by using specified L1 ratio and alpha regularization values. This is the default. If a partition is specified, the single or final model that is fitted is applied to the held-out test data to estimate out-of-sample performance.

Under **Plot**, plots of observed and/or residuals versus predicted values can be selected.

Under **Save**, you can specify predicted values and residuals to save.

Trace plot

When you select this mode, three plots for the training data are displayed as a function of alpha for the specified set of alpha values:

- A trace plot of regression coefficients.
- A plot of R^2.
- A plot of mean squared error (MSE).

Although the partition is honored, no results for held-out test data are provided because no final model results from this mode.

L1 ratio and/or alpha selection via cross validation

When you select this mode, a grid search with cross validation to evaluate models is performed, and the best ratio and alpha values are chosen based on the best average R^2 over the validation folds. The **Number of cross validation folds** field can be used to change the default value of five splits or folds for cross validation. If a partition is specified, the single or final model that is fitted is applied to the held-out test data to estimate out-of-sample performance.

Under **Display**, you can select to show basic information about only the model with the chosen value of ratio and alpha (**Best**), basic information about all models compared (**Compare models**), or complete information on all splits or folds for all models (**Compare models and folds**). **Best** is the default.

Under **Plot**, plots of mean R^2 and/or MSE over validation folds can be selected. Plots of observed and/or residuals versus predicted values can also be selected.

Under **Save**, you can specify predicted values and residuals to save.

Specify individual L1 ratios

When you select this mode for the **Fit with specified L1 ratio and alpha** or **Trace plot** mode, you can specify a single L1 penalty ratio value. When selected for the **L1 ratio and/or alpha selection via cross validation** mode, you can specify multiple values.

Specify grid of alpha values

When you select this mode for the **L1 ratio and/or alpha selection via cross validation** mode, a grid of unique alpha values can be specified from a **Start** value (`value1`) to an **End** value (`value2`) with the increment of **By** (`value3`). If specified, only one valid set of [`value1 TO value2 BY value3`] is allowed. It must satisfy that `0 ≤ value1 ≤ value2 ≤ 1`. In cases where `value1 = value2`, it is equivalent to specifying a single `value1`, regardless of `value3`.

Plots are displayed by using the specified metric for the horizontal X axes of the varying alpha values.

Specify individual alphas

When you select this mode for the **Fit with specified L1 ratio and alpha** mode, you can specify a single alpha regularization value. When selected for the **Trace plot** or **L1 ratio and/or alpha selection via cross validation** mode, you can specify multiple values.

Specify grid of alpha values

When you select this mode for the **Trace plot** or **L1 ratio and/or alpha selection via cross validation** mode, a grid of unique alpha values can be specified from a **Start** value (`value1`) to an **End** value (`value2`) with the increment of **By** (`value3`). If specified, only one valid set of [`value1 TO value2 BY value3`] is allowed. It must satisfy that `0 ≤ value1 ≤ value2 ≤ 1`. In cases where `value1 = value2`, it is equivalent to specifying a single `value1`, regardless of `value3`.

The **Alpha metric** for ranges of values can be either **Linear** or **Base 10 logarithmic** (10 is raised to the power of specified values).

Plots are displayed by using the specified metric for the horizontal X axes of the varying alpha values.

Criteria

Controls analyses.

Include intercept

This criterion includes an intercept in the fitted model(s). Note that the extension procedure does not center or standardize the dependent variable, and the intercept is not penalized during estimation.

Standardize predictors

Standardizes all independent variables.

Number of cross validation folds

The number of splits or folds for cross validation evaluation of models. Must be a positive integer value greater than 1. The default is 5.

Python random state

The value of the random_state setting in Python used while you perform cross validation evaluation of models. It allows replication of results that involve pseudo-random numbers. The value must be an integer in the range of 0 to 2^{32}-1. The default is 0.

Time limit (minutes)

The number of minutes allowed for the model computations to run. If you specify 0, the timer is turned off. The default value is 5.

Display

This option specifies the amount of output to display for the **L1 ratio and/or alpha selection via cross validation** mode.

Best

Displays only basic results for the chosen best model. This is set by default.

Compare models

Displays basic results for all evaluated models.

Compare models and folds

Displays full verbose results for each split or fold for each evaluated model.

Plot

This option specifies plots of observed or residual values versus predicted values, and with cross validation, specification of plots of average mean squared error (MSE) and/or average R^2 over cross validation folds versus alpha values.

Average crossvalidation mean squared error (MSE) versus alpha
: For the **L1 ratio and/or alpha selection via cross validation** mode, displays a line plot of average MSE over cross validation folds versus alpha for the specified or selected best L1 ratio value. For the **Trace plot** mode, a similar plot is automatically produced based on the complete training data.

Average cross-validation R Square versus alpha
: For the **L1 ratio and/or alpha selection via cross validation** mode, displays a line plot of average R^2 over cross validation folds versus alpha for the specified or selected best L1 ratio value. For the **Trace plot** mode, a similar plot is automatically produced based on the complete training data.

Observed versus Predicted
: Displays a scatterplot of observed versus predicted values for the specified or best model.

Residuals versus Predicted
: Displays a scatterplot of residuals versus predicted values for the specified or best model.

Save

Specifies variables to save to the active data set.

Predicted values
: Save predicted values from the specified or best model to the active data set. You can also specify a **Custom variable name**.

Residuals
: Save residuals from the specified or best model predictions to the active data set. You can also specify a **Custom variable name**.

Linear Lasso Regression

Linear Lasso uses the Python sklearn.linear_model.Lasso class to estimate L1 loss regularized linear regression models for a dependent variable on one or more independent variables, and includes optional modes to display trace plots and to select the alpha hyperparameter value based on crossvalidation. When a single model is fitted or crossvalidation is used to select alpha, a partition of holdout data can be used to estimate out-of-sample performance.

In addition to fitting a model with a specified value of the alpha regularization parameter, linear lasso can display a trace plot of coefficient values for a range of alpha values, or facilitate choice of the hyperparameter value via k-fold crossvalidation on specified grids of values. If a single model is fitted or alpha selection via crossvalidation is performed, the final model can be applied to held-out data that is created by a partition of the input data to obtain a valid estimate of out-of-sample performance of the model.

Obtaining a Linear Lasso Regression analysis

1. From the menus choose:

   Analyze > **Regression** > **Linear OLS Alternatives** > **Lasso**

   The dialog allows you to specify a variable that assigns each case in the active dataset to the training or holdout sample.
2. Select a numeric target variable. Only one target variable is required to run an analysis.
3. Specify a numeric dependent.
4. Specify at least one categorical factor variable or numeric covariate variable.

Optionally, **Partition** provides a way to create a holdout or test subset of the input data for estimation of out-of-sample performance of the specified or chosen model. All partitioning is performed after listwise deletion of any cases with invalid data for any variable used by the procedure. Note that for crossvalidation, folds or partitions of the training data are created in Python. The holdout data created by the partition is not used in estimation, regardless of the mode in effect.

The partition can be defined either by specifying the ratio of cases randomly assigned to each sample (under **Training and Holdout partitions**), or by a variable that assigns each case to the training or holdout sample. You cannot specify both training and variables. If the partition is not specified, a holdout sample is created of approximately 30% of the input data is created.

The **Training %** specifies the relative number of cases in the active dataset to randomly assign to the training sample. The default training is 70%.

Linear Lasso Regression: Options

The Options tab provides options for:

Mode

This selection provides options for specifying one of the following modes:

Fit with specified alpha

When you select this mode, a single model is fitted to the training data by using only one alpha regularization value. This is set by default. If a partition is specified, the single or final model that is fitted is applied to the held-out test data to estimate out-of-sample performance.

Under **Plot**, plots of observed and/or residuals versus predicted values can be selected.

Under **Save**, you can specify predicted values and residuals to save.

Trace plot

When you select this mode, three plots for the training data are displayed as a function of alpha for the specified set of alpha values:

- A trace plot of regression coefficients.
- A plot of R^2.
- A plot of mean squared error (MSE).

Although the partition is honored, no results for held-out test data are provided because no final model results from this mode.

Alpha selection via cross validation

When you select a grid search with cross validation to evaluate models, and choose the best alpha based on the best average, R^2 over the validation folds. The **Number of cross validation folds** field can be used to change the default value of five splits or folds for cross validation. If a partition is specified, the single or final model that is fitted is applied to the held-out test data to estimate out-of-sample performance.

Under **Display**, you can select to show basic information about only the model with the chosen value of alpha (**Best**), basic information about all models compared (**Compare models**), or complete information on all splits or folds for all models (**Compare models and folds**). **Best** is the default.

Under **Plot**, plots of mean R^2 and/or MSE over validation folds can be selected. Plots of observed and/or residuals versus predicted values can also be selected.

Under **Save**, you can specify predicted values and residuals to save.

Specify individual alphas

When you select the **Fit with specified alpha** mode, you can specify a single alpha regularization value. When you select the **Trace plot** or **Alpha selection via cross validation** mode, you can specify multiple values.

Value(s)
Specify one or more positive alpha regularization values. Multiple values can be specified individually or as ranges. The default is 1.

Specify grid of alpha values
When you select the **Trace plot** or **Alpha selection via cross validation** mode, a grid of unique alpha values can be specified from a **Start** value (`value1`) to an **End** value (`value2`) with the increment of **By** (`value3`). If specified, only one valid set of [`value1 TO value2 BY value3`] is allowed. It must satisfy that `0 ≤ value1 ≤ value2 ≤ 1`. In cases where `value1 = value2`, it is equivalent to specifying a single `value1`, regardless of `value3`.

The **Alpha metric** for ranges of values can be either **Linear** or **Base 10 logarithmic** (10 is raised to the power of specified values).

Plots are displayed by using the specified metric for the horizontal X axes of the varying alpha values.

Criteria

Controls analyses.

Include intercept
Includes an intercept in the fitted model(s). Note that the extension procedure does not center or standardize the dependent variable, and the intercept is not penalized during estimation.

Standardize predictors
Standardizes all independent variables.

Number of cross validation folds
The number of splits or folds for cross validation evaluation of models. Must be a positive integer value greater than 1. The default is 5.

Python random state
The value of the random_state setting in Python used when performing cross validation evaluation of models. Allows replication of results that involve pseudo-random numbers. Must be an integer in the range of 0 to 2^{32}-1. The default is 0.

Time limit (minutes)
The number of minutes allowed for the model computations to run. If you specify 0, the timer is turned off. The default value is 5.

Display

Specifies the amount of output to display for the **Alpha selection via cross validation** mode.

Best
Displays only basic results for the chosen best model. This is the default.

Compare models
Displays basic results for all evaluated models.

Compare models and folds
Displays full verbose results for each split or fold for each evaluated model.

Plot

Specifies plots of observed or residual values versus predicted values, and with cross validation, specification of plots of average mean squared error (MSE) and/or average R^2 over cross validation folds versus alpha values.

Average cross validation mean squared error (MSE) versus alpha
For the **Alpha selection via cross validation** mode, displays a line plot of average MSE over cross validation folds versus alpha. For the **Trace plot** mode, a similar plot is automatically produced based on the complete training data.

Average cross-validation R Square versus alpha
For the **Alpha selection via cross validation** mode, displays a line plot of average R^2 over cross validation folds versus alpha. For the **Trace plot** mode, a similar plot is automatically produced based on the complete training data.

Observed versus Predicted
Displays a scatterplot of observed versus predicted values for the specified or best model.

Residuals versus Predicted
Displays a scatterplot of residuals versus predicted values for the specified or best model.

Save

Specifies variables to save to the active data set.

Predicted values
Save predicted values from the specified or best model to the active data set. You can also specify a **Custom variable name**.

Residuals
Save residuals from the specified or best model predictions to the active data set. You can also specify a **Custom variable name**.

Linear Ridge Regression

Linear Ridge uses the Python sklearn.linear_model.Ridge class to estimate L2 or squared loss regularized linear regression models for a dependent variable on one or more independent variables, and includes optional modes to display trace plots and to select the alpha hyperparameter value based on crossvalidation. When a single model is fitted or crossvalidation is used to select alpha, a partition of holdout data can be used to estimate out-of-sample performance.

In addition to fitting a model with a specified value of the alpha regularization parameter, linear ridge can display a ridge trace plot of coefficient values for a range of alpha values, or facilitate choice of the hyperparameter value via k-fold crossvalidation on specified grids of values. If a single model is fitted or alpha selection via crossvalidation is performed, the final model can be applied to held-out data that is created by a partition of the input data to obtain a valid estimate of out-of-sample performance of the model.

Obtaining a Linear Ridge Regression analysis

1. From the menus choose:

   Analyze > **Regression** > **Linear OLS Alternatives** > **Ridge**

   The dialog allows you to specify a variable that assigns each case in the active dataset to the training or holdout sample.
2. Select a numeric target variable. Only one target variable is required to run an analysis.
3. Specify a numeric dependent.
4. Specify at least one categorical factor variable or numeric covariate variable.

Optionally, **Partition** provides a way to create a holdout or test subset of the input data for estimation of out-of-sample performance of the specified or chosen model. All partitioning is performed after listwise deletion of any cases with invalid data for any variable used by the procedure. Note that for crossvalidation, folds or partitions of the training data are created in Python. The holdout data that is created by the partition is not used in estimation, regardless of the mode in effect.

The partition can be defined either by specifying the ratio of cases that are randomly assigned to each sample (under **Training and Holdout partitions**), or by a variable that assigns each case to the training or holdout sample. You cannot specify both training and variables. If the partition is not specified, a holdout sample is created of approximately 30% of the input data is created.

The **Training %** specifies the relative number of cases in the active dataset to randomly assign to the training sample. The default training is 70%.

Linear Ridge Regression: Options

The Options tab provides options for:

Mode

This selection provides options for specifying one of the following modes:

Fit with specified alpha

When you select this option, a single model is fitted to the training data that uses only one alpha regularization value. This is set by default. If a partition is specified, the single or final model that is fitted is applied to the held-out test data to estimate out-of-sample performance.

Under **Plot**, plots of observed and/or residuals versus predicted values can be selected.

Under **Save**, you can specify predicted values and residuals to save.

Trace plot

When you select this option, three plots for the training data are displayed as a function of alpha for the specified set of alpha values:

- A ridge trace plot of regression coefficients.
- A plot of R^2.
- A plot of mean squared error (MSE).

Although the partition is honored, no results for held-out test data are provided because no final model results from this mode.

Alpha selection via cross validation

When you select this option, a grid search with cross validation to evaluate models is done, and the best alpha is chosen based on the best average R^2 over the validation folds. The **Number of cross validation folds** field can be used to change the default value of five splits or folds for cross validation. If a partition is specified, the single or final model that is fitted is applied to the held-out test data to estimate out-of-sample performance.

Under **Display**, you can select to show basic information about only the model with the chosen value of alpha (**Best**), basic information about all models compared (**Compare models**), or complete information on all splits or folds for all models (**Compare models and folds**). **Best** is the default.

Under **Plot**, plots of mean R^2 and/or MSE over validation folds can be selected. Plots of observed and/or residuals versus predicted values can also be selected.

Under **Save**, you can specify predicted values and residuals to save.

Specify individual alphas

When you select the **Fit with specified alpha** mode, you can specify a single alpha regularization value. When you select the **Trace plot** or **Alpha selection via cross validation** mode, you can specify multiple values.

Value(s)

Specify one or more positive alpha regularization values. You can specify multiple values individually, or as ranges. The default is 1.

Specify grid of alpha values

When you select the **Trace plot** or **Alpha selection via cross validation** mode, a grid of unique alpha values can be specified from a **Start** value (`value1`) to an **End** value (`value2`) with the increment of **By** (`value3`). If specified, only one valid set of [`value1 TO value2 BY value3`] is allowed. It must satisfy that `0 ≤ value1 ≤ value2 ≤ 1`. In cases where `value1 = value2`, it is equivalent to specifying a single `value1`, regardless of `value3`.

The **Alpha metric** for ranges of values can be either **Linear** or **Base 10 logarithmic** (10 is raised to the power of specified values).

Plots are displayed by using the specified metric for the horizontal X axes of the varying alpha values.

Criteria

Controls analyses.

Include intercept
This criterion includes an intercept in one or more fitted models. Note that the extension procedure does not center or standardize the dependent variable, and the intercept is not penalized during estimation.

Standardize predictors
This criterion standardizes all independent variables.

Number of cross validation folds
Use this criteria to set the number of splits or folds for cross validation evaluation of models. The number must be a positive integer value greater than 1. The default setting is 5.

Python random state
The value of the random_state setting in Python is used while you perform cross validation evaluation of models. Allows replication of results that involve pseudo-random numbers. Must be an integer in the range of 0 to 2^{32}-1. The default is 0.

Time limit (minutes)
The number of minutes allowed for the model computations to run. If you specify 0, the timer is turned off. The default value is 5.

Display

This section specifies the amount of output to display for the **Alpha selection via cross validation** mode.

Best
It displays only basic results for the chosen best model. This is set by default.

Compare models
It also displays basic results for all evaluated models.

Compare models and folds
Lastly it displays full verbose results for each split or fold for each evaluated model.

Plot

Specifies plots of observed or residual values versus predicted values, with cross validation, specification of plots of average mean squared error (MSE) and/or average R^2 over cross validation folds versus alpha values.

Average cross validation mean squared error (MSE) versus alpha
The **Alpha selection via cross validation** mode, displays a line plot of average MSE over cross validation folds versus alpha. For the **Trace plot** mode, a similar plot is automatically produced based on the complete training data.

Average cross-validation R Square versus alpha
The **Alpha selection via cross validation** mode, displays a line plot of average R^2 over cross validation folds versus alpha. For the **Trace plot** mode, a similar plot is automatically produced based on the complete training data.

Observed versus Predicted
It displays a scatterplot of observed versus predicted values for the specified or best model.

Residuals versus Predicted
It also displays a scatterplot of residuals versus predicted values for the specified or best model.

Save

Specifies variables to save to the active data set.

Predicted values
: Save predicted values from the specified or best model to the active data set. Otherwise, you can specify a **Custom variable name**.

Residuals
: Save residuals from the specified or best model predictions to the active data set. Otherwise, you can specify a **Custom variable name**.

Curve Estimation

The Curve Estimation procedure produces curve estimation regression statistics and related plots for 11 different curve estimation regression models. A separate model is produced for each dependent variable. You can also save predicted values, residuals, and prediction intervals as new variables.

Example. An Internet service provider tracks the percentage of virus-infected e-mail traffic on its networks over time. A scatterplot reveals that the relationship is nonlinear. You might fit a quadratic or cubic model to the data and check the validity of assumptions and the goodness of fit of the model.

Statistics. For each model: regression coefficients, multiple R, R^2, adjusted R^2, standard error of the estimate, analysis-of-variance table, predicted values, residuals, and prediction intervals. Models: linear, logarithmic, inverse, quadratic, cubic, power, compound, S-curve, logistic, growth, and exponential.

Curve Estimation Data Considerations

Data. The dependent and independent variables should be quantitative. If you select **Time** from the active dataset as the independent variable (instead of selecting a variable), the Curve Estimation procedure generates a time variable where the length of time between cases is uniform. If **Time** is selected, the dependent variable should be a time-series measure. Time-series analysis requires a data file structure in which each case (row) represents a set of observations at a different time and the length of time between cases is uniform.

Assumptions. Screen your data graphically to determine how the independent and dependent variables are related (linearly, exponentially, etc.). The residuals of a good model should be randomly distributed and normal. If a linear model is used, the following assumptions should be met: For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and the independent variable should be linear, and all observations should be independent.

To Obtain a Curve Estimation

1. From the menus choose:

   Analyze > **Regression** > **Curve Estimation...**
2. Select one or more dependent variables. A separate model is produced for each dependent variable.
3. Select an independent variable (either select a variable in the active dataset or select **Time**).
4. Optionally:

- Select a variable for labeling cases in scatterplots. For each point in the scatterplot, you can use the Point Selection tool to display the value of the Case Label variable.
- Click **Save** to save predicted values, residuals, and prediction intervals as new variables.

The following options are also available:

- **Include constant in equation.** Estimates a constant term in the regression equation. The constant is included by default.
- **Plot models.** Plots the values of the dependent variable and each selected model against the independent variable. A separate chart is produced for each dependent variable.

- **Display ANOVA table.** Displays a summary analysis-of-variance table for each selected model.

Curve Estimation Models

You can choose one or more curve estimation regression models. To determine which model to use, plot your data. If your variables appear to be related linearly, use a simple linear regression model. When your variables are not linearly related, try transforming your data. When a transformation does not help, you may need a more complicated model. View a scatterplot of your data; if the plot resembles a mathematical function you recognize, fit your data to that type of model. For example, if your data resemble an exponential function, use an exponential model.

Linear. Model whose equation is Y = b0 + (b1 * t). The series values are modeled as a linear function of time.

Logarithmic. Model whose equation is Y = b0 + (b1 * ln(t)).

Inverse. Model whose equation is Y = b0 + (b1 / t).

Quadratic. Model whose equation is Y = b0 + (b1 * t) + (b2 * t**2). The quadratic model can be used to model a series that "takes off" or a series that dampens.

Cubic. Model that is defined by the equation Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Power. Model whose equation is Y = b0 * (t**b1) or ln(Y) = ln(b0) + (b1 * ln(t)).

Compound. Model whose equation is Y = b0 * (b1**t) or ln(Y) = ln(b0) + (ln(b1) * t).

S-curve. Model whose equation is Y = e**(b0 + (b1/t)) or ln(Y) = b0 + (b1/t).

Logistic. Model whose equation is Y = 1 / (1/u + (b0 * (b1**t))) or ln(1/y-1/u) = ln (b0) + (ln(b1) * t) where u is the upper boundary value. After selecting Logistic, specify the upper boundary value to use in the regression equation. The value must be a positive number that is greater than the largest dependent variable value.

Growth. Model whose equation is Y = e**(b0 + (b1 * t)) or ln(Y) = b0 + (b1 * t).

Exponential. Model whose equation is Y = b0 * (e**(b1 * t)) or ln(Y) = ln(b0) + (b1 * t).

Curve Estimation Save

Save Variables. For each selected model, you can save predicted values, residuals (observed value of the dependent variable minus the model predicted value), and prediction intervals (upper and lower bounds). The new variable names and descriptive labels are displayed in a table in the output window.

Predict Cases. In the active dataset, if you select **Time** instead of a variable as the independent variable, you can specify a forecast period beyond the end of the time series. You can choose one of the following alternatives:

- **Predict from estimation period through last case.** Predicts values for all cases in the file, based on the cases in the estimation period. The estimation period, displayed at the bottom of the dialog box, is defined with the Range subdialog box of the Select Cases option on the Data menu. If no estimation period has been defined, all cases are used to predict values.
- **Predict through.** Predicts values through the specified date, time, or observation number, based on the cases in the estimation period. This feature can be used to forecast values beyond the last case in the time series. The currently defined date variables determine what text boxes are available for specifying the end of the prediction period. If there are no defined date variables, you can specify the ending observation (case) number.

Use the Define Dates option on the Data menu to create date variables.

Partial Least Squares Regression

The Partial Least Squares Regression procedure estimates partial least squares (PLS, also known as "projection to latent structure") regression models. PLS is a predictive technique that is an alternative

to ordinary least squares (OLS) regression, canonical correlation, or structural equation modeling, and it is particularly useful when predictor variables are highly correlated or when the number of predictors exceeds the number of cases.

PLS combines features of principal components analysis and multiple regression. It first extracts a set of latent factors that explain as much of the covariance as possible between the independent and dependent variables. Then a regression step predicts values of the dependent variables using the decomposition of the independent variables.

Tables
: Proportion of variance explained (by latent factor), latent factor weights, latent factor loadings, independent variable importance in projection (VIP), and regression parameter estimates (by dependent variable) are all produced by default.

Charts
: Variable importance in projection (VIP), factor scores, factor weights for the first three latent factors, and distance to the model are all produced from the Options tab.

Data considerations

Measurement level
: The dependent and independent (predictor) variables can be scale, nominal, or ordinal. The procedure assumes that the appropriate measurement level has been assigned to all variables, although you can temporarily change the measurement level for a variable by right-clicking the variable in the source variable list and selecting a measurement level from the pop-up menu. Categorical (nominal or ordinal) variables are treated equivalently by the procedure.

Categorical variable coding
: The procedure temporarily recodes categorical dependent variables using one-of-*c* coding for the duration of the procedure. If there are *c* categories of a variable, then the variable is stored as *c* vectors, with the first category denoted (1,0,...,0), the next category (0,1,0,...,0), ..., and the final category (0,0,...,0,1). Categorical dependent variables are represented using dummy coding; that is, simply omit the indicator corresponding to the reference category.

Frequency weights
: Weight values are rounded to the nearest whole number before use. Cases with missing weights or weights less than 0.5 are not used in the analyses.

Missing values
: User- and system-missing values are treated as invalid.

Rescaling
: All model variables are centered and standardized, including indicator variables representing categorical variables.

Obtaining Partial Least Squares Regression

From the menus choose:

Analyze > **Regression** > **Partial Least Squares...**

1. Select at least one dependent variable.
2. Select at least one independent variable.

Optionally, you can:

- Specify a reference category for categorical (nominal or ordinal) dependent variables.
- Specify a variable to be used as a unique identifier for casewise output and saved datasets.
- Specify an upper limit on the number of latent factors to be extracted.

Prerequisites

The Partial Least Squares Regression procedure is a Python extension command and requires Python functionality, which is a part of your IBM SPSS Statistics product. It also requires the NumPy and SciPy Python libraries, which are freely available.

Note: For users working in distributed analysis mode (requires IBM SPSS Statistics Server), NumPy and SciPy must be installed on the server. Contact your system administrator for assistance.

Windows and Mac Users

For Windows and Mac, NumPy and SciPy must be installed to a separate version of Python 3.10 from the version that is installed with IBM SPSS Statistics. If you do not have a separate version of Python 3.10, you can download it from http://www.python.org. Then, install NumPy and SciPy for Python version 3.10. The installers are available from http://www.scipy.org/Download.

To enable use of NumPy and SciPy, you must set your Python location to the version of Python 3.10 where you installed NumPy and SciPy. The Python location is set from the File Locations tab in the Options dialog (Edit > Options).

Linux Users

We suggest that you download the source and build NumPy and SciPy yourself. The source is available from http://www.scipy.org/Download. You can install NumPy and SciPy to the version of Python 3.10 that is installed with IBM SPSS Statistics. It is in the `Python` directory under the location where IBM SPSS Statistics is installed.

If you choose to install NumPy and SciPy to a version of Python 3.10 other than the version that is installed with IBM SPSS Statistics, then you must set your Python location to point to that version. The Python location is set from the File Locations tab in the Options dialog (**Edit** > **Options**).

Windows and Unix Server

NumPy and SciPy must be installed, on the server, to a separate version of Python 3.10 from the version that is installed with IBM SPSS Statistics. If there is not a separate version of Python 3.10 on the server, then it can be downloaded from http://www.python.org. NumPy and SciPy for Python 3.10 are available from http://www.scipy.org/Download. To enable use of NumPy and SciPy, the Python location for the server must be set to the version of Python 3.10 where NumPy and SciPy are installed. The Python location is set from the IBM SPSS Statistics Administration Console.

Model

Specify Model Effects. A main-effects model contains all factor and covariate main effects. Select **Custom** to specify interactions. You must indicate all of the terms to be included in the model.

Factors and Covariates. The factors and covariates are listed.

Model. The model depends on the nature of your data. After selecting **Custom**, you can select the main effects and interactions that are of interest in your analysis.

Build Terms

For the selected factors and covariates:

Interaction. Creates the highest-level interaction term of all selected variables. This is the default.

Main effects. Creates a main-effects term for each variable selected.

All 2-way. Creates all possible two-way interactions of the selected variables.

All 3-way. Creates all possible three-way interactions of the selected variables.

All 4-way. Creates all possible four-way interactions of the selected variables.

All 5-way. Creates all possible five-way interactions of the selected variables.

Options

The Options tab allows the user to save and plot model estimates for individual cases, latent factors, and predictors.

For each type of data, specify the name of a dataset. The dataset names must be unique. If you specify the name of an existing dataset, its contents are replaced; otherwise, a new dataset is created.

- **Save estimates for individual cases.** Saves the following casewise model estimates: predicted values, residuals, distance to latent factor model, and latent factor scores. It also plots latent factor scores.
- **Save estimates for latent factors.** Saves latent factor loadings and latent factor weights. It also plots latent factor weights.
- **Save estimates for independent variables.** Saves regression parameter estimates and variable importance to projection (VIP). It also plots VIP by latent factor.

Nearest Neighbor Analysis

Nearest Neighbor Analysis is a method for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

Cases that are near each other are said to be “neighbors.” When a new case (holdout) is presented, its distance from each of the cases in the model is computed. The classifications of the most similar cases – the nearest neighbors – are tallied and the new case is placed into the category that contains the greatest number of nearest neighbors.

You can specify the number of nearest neighbors to examine; this value is called *k*.

Nearest neighbor analysis can also be used to compute values for a continuous target. In this situation, the average or median target value of the nearest neighbors is used to obtain the predicted value for the new case.

Nearest Neighbor Analysis Data Considerations

Target and features. The target and features can be:

- *Nominal*. A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, postal code, and religious affiliation.
- *Ordinal*. A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.
- *Scale*. A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

  Nominal and Ordinal variables are treated equivalently by Nearest Neighbor Analysis. The procedure assumes that the appropriate measurement level has been assigned to each variable; however, you can temporarily change the measurement level for a variable by right-clicking the variable in the source variable list and selecting a measurement level from the pop-up menu.

An icon next to each variable in the variable list identifies the measurement level and data type:

Table 1. Measurement level icons

	Numeric	**String**	**Date**	**Time**
Scale (Continuous)		n/a		

Table 1. Measurement level icons (continued)				
	Numeric	**String**	**Date**	**Time**
Ordinal				
Nominal				

Categorical variable coding. The procedure temporarily recodes categorical predictors and dependent variables using one-of-*c* coding for the duration of the procedure. If there are *c* categories of a variable, then the variable is stored as c vectors, with the first category denoted (1,0,...,0), the next category (0,1,0,...,0), ..., and the final category (0,0,...,0,1).

This coding scheme increases the dimensionality of the feature space. In particular, the total number of dimensions is the number of scale predictors plus the number of categories across all categorical predictors. As a result, this coding scheme can lead to slower training. If your nearest neighbors training is proceeding very slowly, you might try reducing the number of categories in your categorical predictors by combining similar categories or dropping cases that have extremely rare categories before running the procedure.

All one-of-*c* coding is based on the training data, even if a holdout sample is defined (see "Partitions " on page 161). Thus, if the holdout sample contains cases with predictor categories that are not present in the training data, then those cases are not scored. If the holdout sample contains cases with dependent variable categories that are not present in the training data, then those cases are scored.

Rescaling. Scale features are normalized by default. All rescaling is performed based on the training data, even if a holdout sample is defined (see "Partitions " on page 161). If you specify a variable to define partitions, it is important that the features have similar distributions across the training and holdout samples. Use, for example, the Explore procedure to examine the distributions across partitions.

Frequency weights. Frequency weights are ignored by this procedure.

Replicating results. The procedure uses random number generation during random assignment of partitions and cross-validation folds. If you want to replicate your results exactly, in addition to using the same procedure settings, set a seed for the Mersenne Twister (see "Partitions " on page 161), or use variables to define partitions and cross-validation folds.

To obtain a nearest neighbor analysis

From the menus choose:

Analyze > **Classify** > **Nearest Neighbor...**

1. Specify one or more features, which can be thought of independent variables or predictors if there is a target.

   Target (optional). If no target (dependent variable or response) is specified, then the procedure finds the *k* nearest neighbors only – no classification or prediction is done.

   Normalize scale features. Normalized features have the same range of values, which can improve the performance of the estimation algorithm. Adjusted normalization, [2*(x–min)/(max–min)]–1, is used. Adjusted normalized values fall between –1 and 1.

   Focal case identifier (optional). This allows you to mark cases of particular interest. For example, a researcher wants to determine whether the test scores from one school district – the focal case – are comparable to those from similar school districts. He uses nearest neighbor analysis to find the school districts that are most similar with respect to a given set of features. Then he compares the test scores from the focal school district to those from the nearest neighbors.

   Focal cases could also be used in clinical studies to select control cases that are similar to clinical cases. Focal cases are displayed in the *k* nearest neighbors and distances table, feature space chart,

peers chart, and quadrant map. Information on focal cases is saved to the files specified on the Output tab.

Cases with a positive value on the specified variable are treated as focal cases. It is invalid to specify a variable with no positive values.

Case label (optional). Cases are labeled using these values in the feature space chart, peers chart, and quadrant map.

Fields with unknown measurement level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Scan Data. Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

Assign Manually. Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Neighbors

Number of Nearest Neighbors (k). Specify the number of nearest neighbors. Note that using a greater number of neighbors will not necessarily result in a more accurate model.

If a target is specified on the Variables tab, you can alternatively specify a range of values and allow the procedure to choose the "best" number of neighbors within that range. The method for determining the number of nearest neighbors depends upon whether feature selection is requested on the Features tab.

- If feature selection is in effect, then feature selection is performed for each value of *k* in the requested range, and the *k*, and accompanying feature set, with the lowest error rate (or the lowest sum-of-squares error if the target is scale) is selected.
- If feature selection is not in effect, then *V*-fold cross-validation is used to select the “best” number of neighbors. See the Partition tab for control over assignment of folds.

Distance Computation. This is the metric used to specify the distance metric used to measure the similarity of cases.

- **Euclidean metric.** The distance between two cases, x and y, is the square root of the sum, over all dimensions, of the squared differences between the values for the cases.
- **City block metric.** The distance between two cases is the sum, over all dimensions, of the absolute differences between the values for the cases. Also called Manhattan distance.

Optionally, if a target is specified on the Variables tab, you can choose to weight features by their normalized importance when computing distances. Feature importance for a predictor is calculated by the ratio of the error rate or sum-of-squares error of the model with the predictor removed from the model to the error rate or sum-of-squares error for the full model. Normalized importance is calculated by reweighting the feature importance values so that they sum to 1.

Predictions for Scale Target. If a scale target is specified on the Variables tab, this specifies whether the predicted value is computed based upon the mean or the median value of the nearest neighbors.

Features

The Features tab allows you to request and specify options for feature selection when a target is specified on the Variables tab. By default, all features are considered for feature selection, but you can optionally select a subset of features to force into the model.

Stopping Criterion. At each step, the feature whose addition to the model results in the smallest error (computed as the error rate for a categorical target and sum of squares error for a scale target) is considered for inclusion in the model set. Forward selection continues until the specified condition is met.

- **Specified number of features.** The algorithm adds a fixed number of features in addition to those forced into the model. Specify a positive integer. Decreasing values of the number to select creates a more parsimonious model, at the risk of missing important features. Increasing values of the number to select will capture all the important features, at the risk of eventually adding features that actually increase the model error.
- **Minimum change in absolute error ratio.** The algorithm stops when the change in the absolute error ratio indicates that the model cannot be further improved by adding more features. Specify a positive number. Decreasing values of the minimum change will tend to include more features, at the risk of including features that don't add much value to the model. Increasing the value of the minimum change will tend to exclude more features, at the risk of losing features that are important to the model. The "optimal" value of the minimum change will depend upon your data and application. See the Feature Selection Error Log in the output to help you assess which features are most important. See the topic "Feature selection error log " on page 165 for more information.

Partitions

The Partitions tab allows you to divide the dataset into training and holdout sets and, when applicable, assign cases into cross-validation folds

Training and Holdout Partitions. This group specifies the method of partitioning the active dataset into training and holdout samples. The **training sample** comprises the data records used to train the nearest neighbor model; some percentage of cases in the dataset must be assigned to the training sample in order to obtain a model. The **holdout sample** is an independent set of data records used to assess the final model; the error for the holdout sample gives an "honest" estimate of the predictive ability of the model because the holdout cases were not used to build the model.

- **Randomly assign cases to partitions.** Specify the percentage of cases to assign to the training sample. The rest are assigned to the holdout sample.
- **Use variable to assign cases.** Specify a numeric variable that assigns each case in the active dataset to the training or holdout sample. Cases with a positive value on the variable are assigned to the training sample, cases with a value of 0 or a negative value, to the holdout sample. Cases with a system-missing value are excluded from the analysis. Any user-missing values for the partition variable are always treated as valid.

Cross-Validation Folds. *V*-fold cross-validation is used to determine the "best" number of neighbors. It is not available in conjunction with feature selection for performance reasons.

Cross-validation divides the sample into a number of subsamples, or folds. Nearest neighbor models are then generated, excluding the data from each subsample in turn. The first model is based on all of the cases except those in the first sample fold, the second model is based on all of the cases except those in the second sample fold, and so on. For each model, the error is estimated by applying the model to the subsample excluded in generating it. The "best" number of nearest neighbors is the one which produces the lowest error across folds.

- **Randomly assign cases to folds.** Specify the number of folds that should be used for cross-validation. The procedure randomly assigns cases to folds, numbered from 1 to *V*, the number of folds.
- **Use variable to assign cases.** Specify a numeric variable that assigns each case in the active dataset to a fold. The variable must be numeric and take values from 1 to *V*. If any values in this range are missing, and on any splits if split files are in effect, this will cause an error.

Set seed for Mersenne Twister. Setting a seed allows you to replicate analyses. Using this control is similar to setting the Mersenne Twister as the active generator and specifying a fixed starting point on the Random Number Generators dialog, with the important difference that setting the seed in this dialog will preserve the current state of the random number generator and restore that state after the analysis is complete.

Save

Names of Saved Variables. Automatic name generation ensures that you keep all of your work. Custom names allow you to discard/replace results from previous runs without first deleting the saved variables in the Data Editor.

Variables to Save

- **Predicted value or category.** This saves the predicted value for a scale target or the predicted category for a categorical target.
- **Predicted probability.** This saves the predicted probabilities for a categorical target. A separate variable is saved for each of the first *n* categories, where *n* is specified in the **Maximum categories to save for categorical target** control.
- **Training/Holdout partition variables.** If cases are randomly assigned to the training and holdout samples on the Partitions tab, this saves the value of the partition (training or holdout) to which the case was assigned.
- **Cross-validation fold variable.** If cases are randomly assigned to cross-validation folds on the Partitions tab, this saves the value of the fold to which the case was assigned.

Output

Viewer Output

- **Case processing summary.** Displays the case processing summary table, which summarizes the number of cases included and excluded in the analysis, in total and by training and holdout samples.
- **Charts and tables.** Displays model-related output, including tables and charts. Tables in the model view include *k* nearest neighbors and distances for focal cases, classification of categorical response variables, and an error summary. Graphical output in the model view includes a selection error log, feature importance chart, feature space chart, peers chart, and quadrant map. See the topic "Model View " on page 162 for more information.

Files

- **Export model to XML.** You can use this model file to apply the model information to other data files for scoring purposes. This option is not available if split files have been defined.
- **Export distances between focal cases and k nearest neighbors.** For each focal case, a separate variable is created for each of the focal case's *k* nearest neighbors (from the training sample) and the corresponding *k* nearest distances.

Options

User-Missing Values. Categorical variables must have valid values for a case to be included in the analysis. These controls allow you to decide whether user-missing values are treated as valid among categorical variables.

System-missing values and missing values for scale variables are always treated as invalid.

Model View

When you select **Charts and tables** in the Output tab, the procedure creates a Nearest Neighbor Model object in the Viewer. By activating (double-clicking) this object, you gain an interactive view of the model. The model view has a 2-panel window:

- The first panel displays an overview of the model called the main view.
- The second panel displays one of two types of views:

  An auxiliary model view shows more information about the model, but is not focused on the model itself.

A linked view is a view that shows details about one feature of the model when the user drills down on part of the main view.

By default, the first panel shows the feature space and the second panel shows the variable importance chart. If the variable importance chart is not available; that is, when **Weight features by importance** was not selected on the Features tab, the first available view in the View dropdown is shown.

When a view has no available information, its item text in the View dropdown is disabled.

Feature Space

The feature space chart is an interactive graph of the feature space (or a subspace, if there are more than 3 features). Each axis represents a feature in the model, and the location of points in the chart show the values of these features for cases in the training and holdout partitions.

Keys. In addition to the feature values, points in the plot convey other information.

- Shape indicates the partition to which a point belongs, either Training or Holdout.
- The color/shading of a point indicates the value of the target for that case; with distinct color values equal to the categories of a categorical target, and shades indicating the range of values of a continuous target. The indicated value for the training partition is the observed value; for the holdout partition, it is the predicted value. If no target is specified, this key is not shown.
- Heavier outlines indicate a case is focal. Focal cases are shown linked to their *k* nearest neighbors.

Controls and Interactivity. A number of controls in the chart allow you explore the Feature Space.

- You can choose which subset of features to show in the chart and change which features are represented on the dimensions.
- “Focal cases” are simply points selected in the Feature Space chart. If you specified a focal case variable, the points representing the focal cases will initially be selected. However, any point can temporarily become a focal case if you select it. The “usual” controls for point selection apply; clicking on a point selects that point and deselects all others; Control-clicking on a point adds it to the set of selected points. Linked views, such as the Peers Chart, will automatically update based upon the cases selected in the Feature Space.
- You can change the number of nearest neighbors (*k*) to display for focal cases.
- Hovering over a point in the chart displays a tooltip with the value of the case label, or case number if case labels are not defined, and the observed and predicted target values.
- A “Reset” button allows you to return the Feature Space to its original state.

Adding and removing fields/variables

You can add new fields/variables to the feature space or remove the ones that are currently displayed.

Variables Palette

The Variables palette must be displayed before you can add and remove variables. To display the Variables palette, the Model Viewer must be in Edit mode and a case must be selected in the feature space.

1. To put the Model Viewer in Edit mode, from the menus choose:

   View > **Edit Mode**
2. Once in Edit Mode, click any case in the feature space.
3. To display the Variables palette, from the menus choose:

   View > **Palettes** > **Variables**

   The Variables palette lists all of the variables in the feature space. The icon next to the variable name indicates the variable's measurement level.
4. To temporarily change a variable's measurement level, right click the variable in the variables palette and choose an option.

Variable Zones

Variables are added to "zones" in the feature space. To display the zones, start dragging a variable from the Variables palette or select **Show zones**.

The feature space has zones for the *x*, *y*, and *z* axes.

Moving Variables into Zones

Here are some general rules for and tips for moving variables into zones:

- To move a variable into a zone, click and drag the variable from the Variables palette and drop it into the zone. If you choose **Show zones**, you can also right-click a zone and select a variable that you want to add to the zone.
- If you drag a variable from the Variables palette to a zone already occupied by another variable, the old variable is replaced with the new.
- If you drag a variable from one zone to a zone already occupied by another variable, the variables swap positions.
- Clicking the X in a zone removes the variable from that zone.
- If there are multiple graphic elements in the visualization, each graphic element can have its own associated variable zones. First select the graphic element.

Variable Importance

Typically, you will want to focus your modeling efforts on the variables that matter most and consider dropping or ignoring those that matter least. The variable importance chart helps you do this by indicating the relative importance of each variable in estimating the model. Since the values are relative, the sum of the values for all variables on the display is 1.0. Variable importance does not relate to model accuracy. It just relates to the importance of each variable in making a prediction, not whether or not the prediction is accurate.

Peers

This chart displays the focal cases and their *k* nearest neighbors on each feature and on the target. It is available if a focal case is selected in the Feature Space.

Linking behavior. The Peers chart is linked to the Feature Space in two ways.

- Cases selected (focal) in the Feature Space are displayed in the Peers chart, along with their *k* nearest neighbors.
- The value of *k* selected in the Feature Space is used in the Peers chart.

Nearest Neighbor Distances

This table displays the *k* nearest neighbors and distances for focal cases only. It is available if a focal case identifier is specified on the Variables tab, and only displays focal cases identified by this variable.

Each row of:

- The **Focal Case** column contains the value of the case labeling variable for the focal case; if case labels are not defined, this column contains the case number of the focal case.
- The *i*th column under the Nearest Neighbors group contains the value of the case labeling variable for the *i*th nearest neighbor of the focal case; if case labels are not defined, this column contains the case number of the *i*th nearest neighbor of the focal case.
- The *i*th column under the Nearest Distances group contains the distance of the *i*th nearest neighbor to the focal case

Quadrant map

This chart displays the focal cases and their *k* nearest neighbors on a scatterplot (or dotplot, depending upon the measurement level of the target) with the target on the *y*-axis and a scale feature on the *x*-axis, paneled by features. It is available if there is a target and if a focal case is selected in the Feature Space.

- Reference lines are drawn for continuous variables, at the variable means in the training partition.

Feature selection error log

Points on the chart display the error (either the error rate or sum-of-squares error, depending upon the measurement level of the target) on the *y*-axis for the model with the feature listed on the *x*-axis (plus all features to the left on the *x*-axis). This chart is available if there is a target and feature selection is in effect.

k selection error log

Points on the chart display the error (either the error rate or sum-of-squares error, depending upon the measurement level of the target) on the *y*-axis for the model with the number of nearest neighbors (*k*) on the *x*-axis. This chart is available if there is a target and *k* selection is in effect.

k and Feature Selection Error Log

These are feature selection charts (see "Feature selection error log " on page 165), paneled by *k*. This chart is available if there is a target and *k* and feature selection are both in effect.

Classification Table

This table displays the cross-classification of observed versus predicted values of the target, by partition. It is available if there is a target and it is categorical.

- The **(Missing)** row in the Holdout partition contains holdout cases with missing values on the target. These cases contribute to the Holdout Sample: Overall Percent values but not to the Percent Correct values.

Error Summary

This table is available if there is a target variable. It displays the error associated with the model; sum-of-squares for a continuous target and the error rate (100% – overall percent correct) for a categorical target.

Discriminant Analysis

Discriminant analysis builds a predictive model for group membership. The model is composed of a discriminant function (or, for more than two groups, a set of discriminant functions) based on linear combinations of the predictor variables that provide the best discrimination between the groups. The functions are generated from a sample of cases for which group membership is known; the functions can then be applied to new cases that have measurements for the predictor variables but have unknown group membership.

Note: The grouping variable can have more than two values. The codes for the grouping variable must be integers, however, and you need to specify their minimum and maximum values. Cases with values outside of these bounds are excluded from the analysis.

Example. On average, people in temperate zone countries consume more calories per day than people in the tropics, and a greater proportion of the people in the temperate zones are city dwellers. A researcher wants to combine this information into a function to determine how well an individual can discriminate between the two groups of countries. The researcher thinks that population size and economic information may also be important. Discriminant analysis allows you to estimate coefficients of

the linear discriminant function, which looks like the right side of a multiple linear regression equation. That is, using coefficients *a*, *b*, *c*, and *d*, the function is:

```
D = a * climate + b * urban + c * population + d * gross domestic product per capita
```

If these variables are useful for discriminating between the two climate zones, the values of *D* will differ for the temperate and tropic countries. If you use a stepwise variable selection method, you may find that you do not need to include all four variables in the function.

Statistics. For each variable: means, standard deviations, univariate ANOVA. For each analysis: Box's *M*, within-groups correlation matrix, within-groups covariance matrix, separate-groups covariance matrix, total covariance matrix. For each canonical discriminant function: eigenvalue, percentage of variance, canonical correlation, Wilks' lambda, chi-square. For each step: prior probabilities, Fisher's function coefficients, unstandardized function coefficients, Wilks' lambda for each canonical function.

Discriminant Analysis Data Considerations

Data. The grouping variable must have a limited number of distinct categories, coded as integers. Independent variables that are nominal must be recoded to dummy or contrast variables.

Assumptions. Cases should be independent. Predictor variables should have a multivariate normal distribution, and within-group variance-covariance matrices should be equal across groups. Group membership is assumed to be mutually exclusive (that is, no case belongs to more than one group) and collectively exhaustive (that is, all cases are members of a group). The procedure is most effective when group membership is a truly categorical variable; if group membership is based on values of a continuous variable (for example, high IQ versus low IQ), consider using linear regression to take advantage of the richer information that is offered by the continuous variable itself.

To Obtain a Discriminant Analysis

1. From the menus choose:

   Analyze > **Classify** > **Discriminant...**
2. Select an integer-valued grouping variable and click **Define Range** to specify the categories of interest.
3. Select the independent, or predictor, variables. (If your grouping variable does not have integer values, Automatic Recode on the Transform menu will create a variable that does.)
4. Select the method for entering the independent variables.
   - **Enter independents together.** Simultaneously enters all independent variables that satisfy tolerance criteria.
   - **Use stepwise method.** Uses stepwise analysis to control variable entry and removal.
5. Optionally, select cases with a selection variable.

Discriminant Analysis Define Range

Specify the minimum and maximum value of the grouping variable for the analysis. Cases with values outside of this range are not used in the discriminant analysis but are classified into one of the existing groups based on the results of the analysis. The minimum and maximum values must be integers.

Discriminant Analysis Select Cases

To select cases for your analysis:

1. In the Discriminant Analysis dialog box, choose a selection variable.
2. Click **Value** to enter an integer as the selection value.

Only cases with the specified value for the selection variable are used to derive the discriminant functions. Statistics and classification results are generated for both selected and unselected cases. This process provides a mechanism for classifying new cases based on previously existing data or for partitioning your data into training and testing subsets to perform validation on the model generated.

Discriminant Analysis Statistics

Descriptives. Available options are means (including standard deviations), univariate ANOVAs, and Box's *M* test.

- *Means*. Displays total and group means, as well as standard deviations for the independent variables.
- *Univariate ANOVAs*. Performs a one-way analysis-of-variance test for equality of group means for each independent variable.
- *Box's M*. A test for the equality of the group covariance matrices. For sufficiently large samples, a nonsignificant p value means there is insufficient evidence that the matrices differ. The test is sensitive to departures from multivariate normality.

Function Coefficients. Available options are Fisher's classification coefficients and unstandardized coefficients.

- *Fisher's*. Displays Fisher's classification function coefficients that can be used directly for classification. A separate set of classification function coefficients is obtained for each group, and a case is assigned to the group for which it has the largest discriminant score (classification function value).
- *Unstandardized*. Displays the unstandardized discriminant function coefficients.

Matrices. Available matrices of coefficients for independent variables are within-groups correlation matrix, within-groups covariance matrix, separate-groups covariance matrix, and total covariance matrix.

- *Within-groups correlation*. Displays a pooled within-groups correlation matrix that is obtained by averaging the separate covariance matrices for all groups before computing the correlations.
- *Within-groups covariance*. Displays a pooled within-groups covariance matrix, which may differ from the total covariance matrix. The matrix is obtained by averaging the separate covariance matrices for all groups.
- *Separate-groups covariance*. Displays separate covariance matrices for each group.
- *Total covariance*. Displays a covariance matrix from all cases as if they were from a single sample.

Discriminant Analysis Stepwise Method

Method. Select the statistic to be used for entering or removing new variables. Available alternatives are Wilks' lambda, unexplained variance, Mahalanobis distance, smallest *F* ratio, and Rao's *V*. With Rao's *V*, you can specify the minimum increase in *V* for a variable to enter.

- *Wilks' lambda*. A variable selection method for stepwise discriminant analysis that chooses variables for entry into the equation on the basis of how much they lower Wilks' lambda. At each step, the variable that minimizes the overall Wilks' lambda is entered.
- *Unexplained variance*. At each step, the variable that minimizes the sum of the unexplained variation between groups is entered.
- *Mahalanobis distance*. A measure of how much a case's values on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables.
- *Smallest F ratio*. A method of variable selection in stepwise analysis based on maximizing an F ratio computed from the Mahalanobis distance between groups.
- *Rao's V*. A measure of the differences between group means. Also called the Lawley-Hotelling trace. At each step, the variable that maximizes the increase in Rao's V is entered. After selecting this option, enter the minimum value a variable must have to enter the analysis.

Criteria. Available alternatives are **Use F value** and **Use probability of F**. Enter values for entering and removing variables.

- *Use F value*. A variable is entered into the model if its F value is greater than the Entry value and is removed if the F value is less than the Removal value. Entry must be greater than Removal, and both values must be positive. To enter more variables into the model, lower the Entry value. To remove more variables from the model, increase the Removal value.

- *Use probability of F*. A variable is entered into the model if the significance level of its F value is less than the Entry value and is removed if the significance level is greater than the Removal value. Entry must be less than Removal, and both values must be positive. To enter more variables into the model, increase the Entry value. To remove more variables from the model, lower the Removal value.

Display. Summary of steps displays statistics for all variables after each step; **F for pairwise distances** displays a matrix of pairwise *F* ratios for each pair of groups.

Discriminant Analysis Classification

Prior Probabilities. This option determines whether the classification coefficients are adjusted for a priori knowledge of group membership.

- **All groups equal.** Equal prior probabilities are assumed for all groups; this has no effect on the coefficients.
- **Compute from group sizes.** The observed group sizes in your sample determine the prior probabilities of group membership. For example, if 50% of the observations included in the analysis fall into the first group, 25% in the second, and 25% in the third, the classification coefficients are adjusted to increase the likelihood of membership in the first group relative to the other two.

Display. Available display options are casewise results, summary table, and leave-one-out classification.

- *Casewise results*. Codes for actual group, predicted group, posterior probabilities, and discriminant scores are displayed for each case.
- *Summary table*. The number of cases correctly and incorrectly assigned to each of the groups based on the discriminant analysis. Sometimes called the "Confusion Matrix."
- *Leave-one-out classification*. Each case in the analysis is classified by the functions derived from all cases other than that case. It is also known as the "U-method."

Replace missing values with mean. Select this option to substitute the mean of an independent variable for a missing value during the classification phase only.

Use Covariance Matrix. You can choose to classify cases using a within-groups covariance matrix or a separate-groups covariance matrix.

- *Within-groups*. The pooled within-groups covariance matrix is used to classify cases.
- *Separate-groups*. Separate-groups covariance matrices are used for classification. Because classification is based on the discriminant functions (not based on the original variables), this option is not always equivalent to quadratic discrimination.

Plots. Available plot options are combined-groups, separate-groups, and territorial map.

- *Combined-groups*. Creates an all-groups scatterplot of the first two discriminant function values. If there is only one function, a histogram is displayed instead.
- *Separate-groups*. Creates separate-group scatterplots of the first two discriminant function values. If there is only one function, histograms are displayed instead.
- *Territorial map*. A plot of the boundaries used to classify cases into groups based on function values. The numbers correspond to groups into which cases are classified. The mean for each group is indicated by an asterisk within its boundaries. The map is not displayed if there is only one discriminant function.

Discriminant Analysis Save

You can add new variables to your active data file. Available options are predicted group membership (a single variable), discriminant scores (one variable for each discriminant function in the solution), and probabilities of group membership given the discriminant scores (one variable for each group).

You can also export model information to the specified file in XML format. You can use this model file to apply the model information to other data files for scoring purposes.

DISCRIMINANT Command Additional Features

The command syntax language also allows you to:

- Perform multiple discriminant analyses (with one command) and control the order in which variables are entered (with the `ANALYSIS` subcommand).
- Specify prior probabilities for classification (with the `PRIORS` subcommand).
- Display rotated pattern and structure matrices (with the `ROTATE` subcommand).
- Limit the number of extracted discriminant functions (with the `FUNCTIONS` subcommand).
- Restrict classification to the cases that are selected (or unselected) for the analysis (with the `SELECT` subcommand).
- Read and analyze a correlation matrix (with the `MATRIX` subcommand).
- Write a correlation matrix for later analysis (with the `MATRIX` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Factor Analysis

Factor analysis attempts to identify underlying variables, or **factors**, that explain the pattern of correlations within a set of observed variables. Factor analysis is often used in data reduction to identify a small number of factors that explain most of the variance that is observed in a much larger number of manifest variables. Factor analysis can also be used to generate hypotheses regarding causal mechanisms or to screen variables for subsequent analysis (for example, to identify collinearity prior to performing a linear regression analysis).

The factor analysis procedure offers a high degree of flexibility:

- Seven methods of factor extraction are available.
- Five methods of rotation are available, including direct oblimin and promax for nonorthogonal rotations.
- Three methods of computing factor scores are available, and scores can be saved as variables for further analysis.

Example. What underlying attitudes lead people to respond to the questions on a political survey as they do? Examining the correlations among the survey items reveals that there is significant overlap among various subgroups of items--questions about taxes tend to correlate with each other, questions about military issues correlate with each other, and so on. With factor analysis, you can investigate the number of underlying factors and, in many cases, identify what the factors represent conceptually. Additionally, you can compute factor scores for each respondent, which can then be used in subsequent analyses. For example, you might build a logistic regression model to predict voting behavior based on factor scores.

Statistics. For each variable: number of valid cases, mean, and standard deviation. For each factor analysis: correlation matrix of variables, including significance levels, determinant, and inverse; reproduced correlation matrix, including anti-image; initial solution (communalities, eigenvalues, and percentage of variance explained); Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity; unrotated solution, including factor loadings, communalities, and eigenvalues; and rotated solution, including rotated pattern matrix and transformation matrix. For oblique rotations: rotated pattern and structure matrices; factor score coefficient matrix and factor covariance matrix. Plots: scree plot of eigenvalues and loading plot of first two or three factors.

Factor Analysis Data Considerations

Data. The variables should be quantitative at the *interval* or *ratio* level. Categorical data (such as religion or country of origin) are not suitable for factor analysis. Data for which Pearson correlation coefficients can sensibly be calculated should be suitable for factor analysis.

Assumptions. The data should have a bivariate normal distribution for each pair of variables, and observations should be independent. The factor analysis model specifies that variables are determined by common factors (the factors estimated by the model) and unique factors (which do not overlap between

observed variables); the computed estimates are based on the assumption that all unique factors are uncorrelated with each other and with the common factors.

To Obtain a Factor Analysis

1. From the menus choose:

   Analyze > **Dimension Reduction** > **Factor...**
2. Select the variables for the factor analysis.

Factor Analysis Select Cases

To select cases for your analysis:

1. Choose a selection variable.
2. Click **Value** to enter an integer as the selection value.

Only cases with that value for the selection variable are used in the factor analysis.

Factor Analysis Descriptives

Statistics. Univariate descriptives includes the mean, standard deviation, and number of valid cases for each variable. **Initial solution** displays initial communalities, eigenvalues, and the percentage of variance explained.

Correlation Matrix. The available options are coefficients, significance levels, determinant, KMO and Bartlett's test of sphericity, inverse, reproduced, and anti-image.

- *KMO and Bartlett's Test of Sphericity*. The Kaiser-Meyer-Olkin measure of sampling adequacy tests whether the partial correlations among variables are small. Bartlett's test of sphericity tests whether the correlation matrix is an identity matrix, which would indicate that the factor model is inappropriate.
- *Reproduced*. The estimated correlation matrix from the factor solution. Residuals (difference between estimated and observed correlations) are also displayed.
- *Anti-image*. The anti-image correlation matrix contains the negatives of the partial correlation coefficients, and the anti-image covariance matrix contains the negatives of the partial covariances. In a good factor model, most of the off-diagonal elements will be small. The measure of sampling adequacy for a variable is displayed on the diagonal of the anti-image correlation matrix.

Factor Analysis Extraction

Method. Allows you to specify the method of factor extraction. Available methods are principal components, unweighted least squares, generalized least squares, maximum likelihood, principal axis factoring, alpha factoring, and image factoring.

- *Principal Components Analysis*. A factor extraction method used to form uncorrelated linear combinations of the observed variables. The first component has maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other. Principal components analysis is used to obtain the initial factor solution. It can be used when a correlation matrix is singular.
- *Unweighted Least-Squares Method*. A factor extraction method that minimizes the sum of the squared differences between the observed and reproduced correlation matrices (ignoring the diagonals).
- *Generalized Least-Squares Method*. A factor extraction method that minimizes the sum of the squared differences between the observed and reproduced correlation matrices. Correlations are weighted by the inverse of their uniqueness, so that variables with high uniqueness are given less weight than those with low uniqueness.
- *Maximum-Likelihood Method*. A factor extraction method that produces parameter estimates that are most likely to have produced the observed correlation matrix if the sample is from a multivariate normal distribution. The correlations are weighted by the inverse of the uniqueness of the variables, and an iterative algorithm is employed.

- *Principal Axis Factoring*. A method of extracting factors from the original correlation matrix, with squared multiple correlation coefficients placed in the diagonal as initial estimates of the communalities. These factor loadings are used to estimate new communalities that replace the old communality estimates in the diagonal. Iterations continue until the changes in the communalities from one iteration to the next satisfy the convergence criterion for extraction.
- *Alpha Factoring*. A factor extraction method that considers the variables in the analysis to be a sample from the universe of potential variables. This method maximizes the alpha reliability of the factors.
- *Image Factoring*. A factor extraction method developed by Guttman and based on image theory. The common part of the variable, called the partial image, is defined as its linear regression on remaining variables, rather than a function of hypothetical factors.

Analyze. Allows you to specify either a correlation matrix or a covariance matrix.

- **Correlation matrix.** Useful if variables in your analysis are measured on different scales.
- **Covariance matrix.** Useful when you want to apply your factor analysis to multiple groups with different variances for each variable.

Extract. You can either retain all factors whose eigenvalues exceed a specified value, or you can retain a specific number of factors.

Display. Allows you to request the unrotated factor solution and a scree plot of the eigenvalues.

- *Unrotated Factor Solution*. Displays unrotated factor loadings (factor pattern matrix), communalities, and eigenvalues for the factor solution.
- *Scree plot*. A plot of the variance that is associated with each factor. This plot is used to determine how many factors should be kept. Typically the plot shows a distinct break between the steep slope of the large factors and the gradual trailing of the rest (the scree).

Maximum Iterations for Convergence. Allows you to specify the maximum number of steps that the algorithm can take to estimate the solution.

Factor Analysis Rotation

Method. Allows you to select the method of factor rotation and whether to apply Kaiser normalization. Available methods are varimax, direct oblimin, quartimax, equamax, or promax.

- *Varimax Method*. An orthogonal rotation method that minimizes the number of variables that have high loadings on each factor. This method simplifies the interpretation of the factors.
- *Direct Oblimin Method*. A method for oblique (nonorthogonal) rotation. When delta equals 0 (the default), solutions are most oblique. As delta becomes more negative, the factors become less oblique. To override the default delta of 0, enter a number less than or equal to 0.8.
- *Quartimax Method*. A rotation method that minimizes the number of factors needed to explain each variable. This method simplifies the interpretation of the observed variables.
- *Equamax Method*. A rotation method that is a combination of the varimax method, which simplifies the factors, and the quartimax method, which simplifies the variables. The number of variables that load highly on a factor and the number of factors needed to explain a variable are minimized.
- *Promax Rotation*. An oblique rotation, which allows factors to be correlated. This rotation can be calculated more quickly than a direct oblimin rotation, so it is useful for large datasets.
- *Apply Kaiser normalization*. By default, allows you to apply Kaiser normalization when a rotation is specified.

Display. Allows you to include output on the rotated solution, as well as loading plots for the first two or three factors.

- *Rotated Solution*. A rotation method must be selected to obtain a rotated solution. For orthogonal rotations, the rotated pattern matrix and factor transformation matrix are displayed. For oblique rotations, the pattern, structure, and factor correlation matrices are displayed.

- *Factor Loading Plot*. Three-dimensional factor loading plot of the first three factors. For a two-factor solution, a two-dimensional plot is shown. The plot is not displayed if only one factor is extracted. Plots display rotated solutions if rotation is requested.

Maximum Iterations for Convergence. Allows you to specify the maximum number of steps that the algorithm can take to perform the rotation.

Factor Analysis Scores

Save as variables. Creates one new variable for each factor in the final solution.

Method. The alternative methods for calculating factor scores are regression, Bartlett, and Anderson-Rubin.

- *Regression Method*. A method for estimating factor score coefficients. The scores that are produced have a mean of 0 and a variance equal to the squared multiple correlation between the estimated factor scores and the true factor values. The scores may be correlated even when factors are orthogonal.
- *Bartlett Scores*. A method of estimating factor score coefficients. The scores that are produced have a mean of 0. The sum of squares of the unique factors over the range of variables is minimized.
- *Anderson-Rubin Method*. A method of estimating factor score coefficients; a modification of the Bartlett method which ensures orthogonality of the estimated factors. The scores that are produced have a mean of 0, have a standard deviation of 1, and are uncorrelated.

Display factor score coefficient matrix. Shows the coefficients by which variables are multiplied to obtain factor scores. Also shows the correlations between factor scores.

Factor Analysis Options

Missing Values. Allows you to specify how missing values are handled. The available choices are to exclude cases *listwise*, exclude cases *pairwise*, or replace with mean.

Coefficient Display Format. Allows you to control aspects of the output matrices. You sort coefficients by size and suppress coefficients with absolute values that are less than the specified value.

FACTOR Command Additional Features

The command syntax language also allows you to:

- Specify convergence criteria for iteration during extraction and rotation.
- Specify individual rotated-factor plots.
- Specify how many factor scores to save.
- Specify diagonal values for the principal axis factoring method.
- Write correlation matrices or factor-loading matrices to disk for later analysis.
- Read and analyze correlation matrices or factor-loading matrices.

See the *Command Syntax Reference* for complete syntax information.

Choosing a Procedure for Clustering

Cluster analyses can be performed using the TwoStep, Hierarchical, or K-Means Cluster Analysis procedure. Each procedure employs a different algorithm for creating clusters, and each has options not available in the others.

TwoStep Cluster Analysis. For many applications, the TwoStep Cluster Analysis procedure will be the method of choice. It provides the following unique features:

- Automatic selection of the best number of clusters, in addition to measures for choosing between cluster models.
- Ability to create cluster models simultaneously based on categorical and continuous variables.

- Ability to save the cluster model to an external XML file and then read that file and update the cluster model using newer data.

Additionally, the TwoStep Cluster Analysis procedure can analyze large data files.

Hierarchical Cluster Analysis. The Hierarchical Cluster Analysis procedure is limited to smaller data files (hundreds of objects to be clustered) but has the following unique features:

- Ability to cluster cases or variables.
- Ability to compute a range of possible solutions and save cluster memberships for each of those solutions.
- Several methods for cluster formation, variable transformation, and measuring the dissimilarity between clusters.

As long as all the variables are of the same type, the Hierarchical Cluster Analysis procedure can analyze interval (continuous), count, or binary variables.

K-Means Cluster Analysis. The K-Means Cluster Analysis procedure is limited to continuous data and requires you to specify the number of clusters in advance, but it has the following unique features:

- Ability to save distances from cluster centers for each object.
- Ability to read initial cluster centers from and save final cluster centers to an external IBM SPSS Statistics file.

Additionally, the K-Means Cluster Analysis procedure can analyze large data files.

TwoStep Cluster Analysis

The TwoStep Cluster Analysis procedure is an exploratory tool designed to reveal natural groupings (or clusters) within a dataset that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:

- **Handling of categorical and continuous variables.** By assuming variables to be independent, a joint multinomial-normal distribution can be placed on categorical and continuous variables.
- **Automatic selection of number of clusters.** By comparing the values of a model-choice criterion across different clustering solutions, the procedure can automatically determine the optimal number of clusters.
- **Scalability.** By constructing a cluster features (CF) tree that summarizes the records, the TwoStep algorithm allows you to analyze large data files.

Example. Retail and consumer product companies regularly apply clustering techniques to data that describe their customers' buying habits, gender, age, income level, etc. These companies tailor their marketing and product development strategies to each consumer group to increase sales and build brand loyalty.

Distance Measure. This selection determines how the similarity between two clusters is computed.

- **Log-likelihood.** The likelihood measure places a probability distribution on the variables. Continuous variables are assumed to be normally distributed, while categorical variables are assumed to be multinomial. All variables are assumed to be independent.
- **Euclidean.** The Euclidean measure is the "straight line" distance between two clusters. It can be used only when all of the variables are continuous.

Number of Clusters. This selection allows you to specify how the number of clusters is to be determined.

- **Determine automatically.** The procedure will automatically determine the "best" number of clusters, using the criterion specified in the Clustering Criterion group. Optionally, enter a positive integer specifying the maximum number of clusters that the procedure should consider.
- **Specify fixed.** Allows you to fix the number of clusters in the solution. Enter a positive integer.

Count of Continuous Variables. This group provides a summary of the continuous variable standardization specifications made in the Options dialog box. See the topic "TwoStep Cluster Analysis Options" on page 174 for more information.

Clustering Criterion. This selection determines how the automatic clustering algorithm determines the number of clusters. Either the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) can be specified.

TwoStep Cluster Analysis Data Considerations

Data. This procedure works with both continuous and categorical variables. Cases represent objects to be clustered, and the variables represent attributes upon which the clustering is based.

Case Order. Note that the cluster features tree and the final solution may depend on the order of cases. To minimize order effects, randomly order the cases. You may want to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution. In situations where this is difficult due to extremely large file sizes, multiple runs with a sample of cases sorted in different random orders might be substituted.

Assumptions. The likelihood distance measure assumes that variables in the cluster model are independent. Further, each continuous variable is assumed to have a normal (Gaussian) distribution, and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but you should try to be aware of how well these assumptions are met.

Use the Bivariate Correlations procedure to test the independence of two continuous variables. Use the Crosstabs procedure to test the independence of two categorical variables. Use the Means procedure to test the independence between a continuous variable and categorical variable. Use the Explore procedure to test the normality of a continuous variable. Use the Chi-Square Test procedure to test whether a categorical variable has a specified multinomial distribution.

To Obtain a TwoStep Cluster Analysis

1. From the menus choose:

   Analyze > **Classify** > **TwoStep Cluster...**

2. Select one or more categorical or continuous variables.

Optionally, you can:

- Adjust the criteria by which clusters are constructed.
- Select settings for noise handling, memory allocation, variable standardization, and cluster model input.
- Request model viewer output.
- Save model results to the working file or to an external XML file.

TwoStep Cluster Analysis Options

Outlier Treatment. This group allows you to treat outliers specially during clustering if the cluster features (CF) tree fills. The CF tree is full if it cannot accept any more cases in a leaf node and no leaf node can be split.

- If you select noise handling and the CF tree fills, it will be regrown after placing cases in sparse leaves into a "noise" leaf. A leaf is considered sparse if it contains fewer than the specified percentage of cases of the maximum leaf size. After the tree is regrown, the outliers will be placed in the CF tree if possible. If not, the outliers are discarded.
- If you do not select noise handling and the CF tree fills, it will be regrown using a larger distance change threshold. After final clustering, values that cannot be assigned to a cluster are labeled outliers. The outlier cluster is given an identification number of –1 and is not included in the count of the number of clusters.

Memory Allocation. This group allows you to specify the maximum amount of memory in megabytes (MB) that the cluster algorithm should use. If the procedure exceeds this maximum, it will use the disk to store information that will not fit in memory. Specify a number greater than or equal to 4.

- Consult your system administrator for the largest value that you can specify on your system.
- The algorithm may fail to find the correct or specified number of clusters if this value is too low.

Variable standardization. The clustering algorithm works with standardized continuous variables. Any continuous variables that are not standardized should be left as variables in the To be Standardized list. To save some time and computational effort, you can select any continuous variables that you have already standardized as variables in the Assumed Standardized list.

Advanced Options

CF Tree Tuning Criteria. The following clustering algorithm settings apply specifically to the cluster features (CF) tree and should be changed with care:

- **Initial Distance Change Threshold.** This is the initial threshold used to grow the CF tree. If inserting a given case into a leaf of the CF tree would yield tightness less than the threshold, the leaf is not split. If the tightness exceeds the threshold, the leaf is split.
- **Maximum Branches (per leaf node).** The maximum number of child nodes that a leaf node can have.
- **Maximum Tree Depth.** The maximum number of levels that the CF tree can have.
- **Maximum Number of Nodes Possible.** This indicates the maximum number of CF tree nodes that could potentially be generated by the procedure, based on the function $(b^{d+1} - 1) / (b - 1)$, where b is the maximum branches and d is the maximum tree depth. Be aware that an overly large CF tree can be a drain on system resources and can adversely affect the performance of the procedure. At a minimum, each node requires 16 bytes.

Cluster Model Update. This group allows you to import and update a cluster model generated in a prior analysis. The input file contains the CF tree in XML format. The model will then be updated with the data in the active file. You must select the variable names in the main dialog box in the same order in which they were specified in the prior analysis. The XML file remains unaltered, unless you specifically write the new model information to the same filename. See the topic "TwoStep Cluster Analysis Output" on page 175 for more information.

If a cluster model update is specified, the options pertaining to generation of the CF tree that were specified for the original model are used. More specifically, the distance measure, noise handling, memory allocation, or CF tree tuning criteria settings for the saved model are used, and any settings for these options in the dialog boxes are ignored.

Note: When performing a cluster model update, the procedure assumes that none of the selected cases in the active dataset were used to create the original cluster model. The procedure also assumes that the cases used in the model update come from the same population as the cases used to create the original model; that is, the means and variances of continuous variables and levels of categorical variables are assumed to be the same across both sets of cases. If your "new" and "old" sets of cases come from heterogeneous populations, you should run the TwoStep Cluster Analysis procedure on the combined sets of cases for the best results.

TwoStep Cluster Analysis Output

Output. This group provides options for displaying the clustering results.

- **Pivot tables.** Results are displayed in pivot tables.
- **Charts and tables in Model Viewer.** Results are displayed in the Model Viewer.
- **Evaluation fields.** This calculates cluster data for variables that were not used in cluster creation. Evaluation fields can be displayed along with the input features in the model viewer by selecting them in the Display subdialog. Fields with missing values are ignored.

Working Data File. This group allows you to save variables to the active dataset.

- **Create cluster membership variable.** This variable contains a cluster identification number for each case. The name of this variable is *tsc_n*, where *n* is a positive integer indicating the ordinal of the active dataset save operation completed by this procedure in a given session.

XML Files. The final cluster model and CF tree are two types of output files that can be exported in XML format.

- **Export final model.** The final cluster model is exported to the specified file in XML (PMML) format. You can use this model file to apply the model information to other data files for scoring purposes.
- **Export CF tree.** This option allows you to save the current state of the cluster tree and update it later using newer data.

The Cluster Viewer

Cluster models are typically used to find groups (or clusters) of similar records based on the variables examined, where the similarity between members of the same group is high and the similarity between members of different groups is low. The results can be used to identify associations that would otherwise not be apparent. For example, through cluster analysis of customer preferences, income level, and buying habits, it may be possible to identify the types of customers who are more likely to respond to a particular marketing campaign.

There are two approaches to interpreting the results in a cluster display:

- Examine clusters to determine characteristics unique to that cluster. *Does one cluster contain all the high-income borrowers? Does this cluster contain more records than the others?*
- Examine fields across clusters to determine how values are distributed among clusters. *Does one's level of education determine membership in a cluster? Does a high credit score distinguish between membership in one cluster or another?*

Using the main views and the various linked views in the Cluster Viewer, you can gain insight to help you answer these questions.

To see information about the cluster model, activate (double-click) the Model Viewer object in the Viewer.

Cluster Viewer

The Cluster Viewer is made up of two panels, the main view on the left and the linked, or auxiliary, view on the right. There are two main views:

- Model Summary (the default). See the topic "Model Summary View" on page 176 for more information.
- Clusters. See the topic "Clusters View" on page 177 for more information.

There are four linked/auxiliary views:

- Predictor Importance. See the topic "Cluster Predictor Importance View" on page 178 for more information.
- Cluster Sizes (the default). See the topic "Cluster Sizes View" on page 178 for more information.
- Cell Distribution. See the topic "Cell Distribution View" on page 179 for more information.
- Cluster Comparison. See the topic "Cluster Comparison View" on page 179 for more information.

Model Summary View

The Model Summary view shows a snapshot, or summary, of the cluster model, including a Silhouette measure of cluster cohesion and separation that is shaded to indicate poor, fair, or good results. This snapshot enables you to quickly check if the quality is poor, in which case you may decide to return to the modeling node to amend the cluster model settings to produce a better result.

The results of poor, fair, and good are based on the work of Kaufman and Rousseeuw (1990) regarding interpretation of cluster structures. In the Model Summary view, a good result equates to data that reflects Kaufman and Rousseeuw's rating as either reasonable or strong evidence of cluster structure, fair reflects their rating of weak evidence, and poor reflects their rating of no significant evidence.

The silhouette measure averages, over all records, (B−A) / max(A,B), where A is the record's distance to its cluster center and B is the record's distance to the nearest cluster center that it doesn't belong to. A silhouette coefficient of 1 would mean that all cases are located directly on their cluster centers. A value of −1 would mean all cases are located on the cluster centers of some other cluster. A value of 0 means, on average, cases are equidistant between their own cluster center and the nearest other cluster.

The summary includes a table that contains the following information:

- **Algorithm.** The clustering algorithm used, for example, "TwoStep".
- **Input Features.** The number of fields, also known as **inputs** or **predictors**.
- **Clusters.** The number of clusters in the solution.

Clusters View

The Clusters view contains a cluster-by-features grid that includes cluster names, sizes, and profiles for each cluster.

The columns in the grid contain the following information:

- **Cluster.** The cluster numbers created by the algorithm.
- **Label.** Any labels applied to each cluster (this is blank by default). Double-click in the cell to enter a label that describes the cluster contents; for example, "Luxury car buyers".
- **Description.** Any description of the cluster contents (this is blank by default). Double-click in the cell to enter a description of the cluster; for example, "55+ years of age, professionals, earning over $100,000".
- **Size.** The size of each cluster as a percentage of the overall cluster sample. Each size cell within the grid displays a vertical bar that shows the size percentage within the cluster, a size percentage in numeric format, and the cluster case counts.
- **Features.** The individual inputs or predictors, sorted by overall importance by default. If any columns have equal sizes they are shown in ascending sort order of the cluster numbers.

  Overall feature importance is indicated by the color of the cell background shading; the most important feature is darkest; the least important feature is unshaded. A guide above the table indicates the importance attached to each feature cell color.

When you hover your mouse over a cell, the full name/label of the feature and the importance value for the cell is displayed. Further information may be displayed, depending on the view and feature type. In the Cluster Centers view, this includes the cell statistic and the cell value; for example: “Mean: 4.32”. For categorical features the cell shows the name of the most frequent (modal) category and its percentage.

Within the Clusters view, you can select various ways to display the cluster information:

- Transpose clusters and features. See the topic “Transpose Clusters and Features” on page 177 for more information.
- Sort features. See the topic “Sort Features” on page 177 for more information.
- Sort clusters. See the topic “Sort Clusters” on page 178 for more information.
- Select cell contents. See the topic “Cell Contents” on page 178 for more information.

Transpose Clusters and Features

By default, clusters are displayed as columns and features are displayed as rows. To reverse this display, click the **Transpose Clusters and Features** button to the left of the **Sort Features By** buttons. For example you may want to do this when you have many clusters displayed, to reduce the amount of horizontal scrolling required to see the data.

Sort Features

The **Sort Features By** buttons enable you to select how feature cells are displayed:

- **Overall Importance.** This is the default sort order. Features are sorted in descending order of overall importance, and sort order is the same across clusters. If any features have tied importance values, the tied features are listed in ascending sort order of the feature names.
- **Within-Cluster Importance.** Features are sorted with respect to their importance for each cluster. If any features have tied importance values, the tied features are listed in ascending sort order of the feature names. When this option is chosen the sort order usually varies across clusters.
- **Name.** Features are sorted by name in alphabetical order.
- **Data order.** Features are sorted by their order in the dataset.

Sort Clusters

By default clusters are sorted in descending order of size. The **Sort Clusters By** buttons enable you to sort them by name in alphabetical order, or, if you have created unique labels, in alphanumeric label order instead.

Features that have the same label are sorted by cluster name. If clusters are sorted by label and you edit the label of a cluster, the sort order is automatically updated.

Cell Contents

The **Cells** buttons enable you to change the display of the cell contents for features and evaluation fields.

- **Cluster Centers.** By default, cells display feature names/labels and the central tendency for each cluster/feature combination. The mean is shown for continuous fields and the mode (most frequently occurring category) with category percentage for categorical fields.
- **Absolute Distributions.** Shows feature names/labels and absolute distributions of the features within each cluster. For categorical features, the display shows bar charts overlaid with categories ordered in ascending order of the data values. For continuous features, the display shows a smooth density plot which use the same endpoints and intervals for each cluster.

  The solid red colored display shows the cluster distribution, whilst the paler display represents the overall data.

- **Relative Distributions.** Shows feature names/labels and relative distributions in the cells. In general the displays are similar to those shown for absolute distributions, except that relative distributions are displayed instead.

  The solid red colored display shows the cluster distribution, while the paler display represents the overall data.

- **Basic View.** Where there are a lot of clusters, it can be difficult to see all the detail without scrolling. To reduce the amount of scrolling, select this view to change the display to a more compact version of the table.

Cluster Predictor Importance View

The Predictor Importance view shows the relative importance of each field in estimating the model.

Cluster Sizes View

The Cluster Sizes view shows a pie chart that contains each cluster. The percentage size of each cluster is shown on each slice; hover the mouse over each slice to display the count in that slice.

Below the chart, a table lists the following size information:

- The size of the smallest cluster (both a count and percentage of the whole).
- The size of the largest cluster (both a count and percentage of the whole).
- The ratio of size of the largest cluster to the smallest cluster.

Cell Distribution View

The Cell Distribution view shows an expanded, more detailed, plot of the distribution of the data for any feature cell you select in the table in the Clusters main panel.

Cluster Comparison View

The Cluster Comparison view consists of a grid-style layout, with features in the rows and selected clusters in the columns. This view helps you to better understand the factors that make up the clusters; it also enables you to see differences between clusters not only as compared with the overall data, but with each other.

To select clusters for display, click on the top of the cluster column in the Clusters main panel. Use either Ctrl-click or Shift-click to select or deselect more than one cluster for comparison.

Note: You can select up to five clusters for display.

Clusters are shown in the order in which they were selected, while the order of fields is determined by the **Sort Features By** option. When you select **Within-Cluster Importance**, fields are always sorted by overall importance .

The background plots show the overall distributions of each features:

- Categorical features are shown as dot plots, where the size of the dot indicates the most frequent/modal category for each cluster (by feature).
- Continuous features are displayed as boxplots, which show overall medians and the interquartile ranges.

Overlaid on these background views are boxplots for selected clusters:

- For continuous features, square point markers and horizontal lines indicate the median and interquartile range for each cluster.
- Each cluster is represented by a different color, shown at the top of the view.

Navigating the Cluster Viewer

The Cluster Viewer is an interactive display. You can:

- Select a field or cluster to view more details.
- Compare clusters to select items of interest.
- Alter the display.
- Transpose axes.

Using the Toolbars

You control the information shown in both the left and right panels by using the toolbar options. You can change the orientation of the display (top-down, left-to-right, or right-to-left) using the toolbar controls. In addition, you can also reset the viewer to the default settings, and open a dialog box to specify the contents of the Clusters view in the main panel.

The **Sort Features By**, **Sort Clusters By**, **Cells**, and **Display** options are only available when you select the **Clusters** view in the main panel. See the topic "Clusters View" on page 177 for more information.

Table 2. Toolbar icons

Icon	Topic
	See Transpose Clusters and Features
	See Sort Features By
	See Sort Clusters By

Table 2. Toolbar icons (continued)	
Icon	**Topic**
	See Cells

Control Cluster View Display

To control what is shown in the Clusters view on the main panel, click the **Display** button; the Display dialog opens.

Features. Selected by default. To hide all input features, deselect the check box.

Evaluation Fields. Choose the evaluation fields (fields not used to create the cluster model, but sent to the model viewer to evaluate the clusters) to display; none are shown by default. *Note* The evaluation field must be a string with more than one value. This check box is unavailable if no evaluation fields are available.

Cluster Descriptions. Selected by default. To hide all cluster description cells, deselect the check box.

Cluster Sizes. Selected by default. To hide all cluster size cells, deselect the check box.

Maximum Number of Categories. Specify the maximum number of categories to display in charts of categorical features; the default is 20.

Filtering Records

If you want to know more about the cases in a particular cluster or group of clusters, you can select a subset of records for further analysis based on the selected clusters.

1. Select the clusters in the Cluster view of the Cluster Viewer. To select multiple clusters, use Ctrl-click.
2. From the menus choose:

   Generate > **Filter Records...**
3. Enter a filter variable name. Records from the selected clusters will receive a value of 1 for this field. All other records will receive a value of 0 and will be excluded from subsequent analyses until you change the filter status.
4. Click **OK**.

Hierarchical Cluster Analysis

This procedure attempts to identify relatively homogeneous groups of cases (or variables) based on selected characteristics, using an algorithm that starts with each case (or variable) in a separate cluster and combines clusters until only one is left. You can analyze raw variables, or you can choose from a variety of standardizing transformations. Distance or similarity measures are generated by the Proximities procedure. Statistics are displayed at each stage to help you select the best solution.

Example. Are there identifiable groups of television shows that attract similar audiences within each group? With hierarchical cluster analysis, you could cluster television shows (cases) into homogeneous groups based on viewer characteristics. This can be used to identify segments for marketing. Or you can cluster cities (cases) into homogeneous groups so that comparable cities can be selected to test various marketing strategies.

Statistics. Agglomeration schedule, distance (or similarity) matrix, and cluster membership for a single solution or a range of solutions. Plots: dendrograms and icicle plots.

Hierarchical Cluster Analysis Data Considerations

Data. The variables can be quantitative, binary, or count data. Scaling of variables is an important issue--differences in scaling may affect your cluster solution(s). If your variables have large differences in scaling (for example, one variable is measured in dollars and the other is measured in years), you should consider standardizing them (this can be done automatically by the Hierarchical Cluster Analysis procedure).

Case order. If tied distances or similarities exist in the input data or occur among updated clusters during joining, the resulting cluster solution may depend on the order of cases in the file. You may want to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution.

Assumptions. The distance or similarity measures used should be appropriate for the data analyzed (see the Proximities procedure for more information on choices of distance and similarity measures). Also, you should include all relevant variables in your analysis. Omission of influential variables can result in a misleading solution. Because hierarchical cluster analysis is an exploratory method, results should be treated as tentative until they are confirmed with an independent sample.

To Obtain a Hierarchical Cluster Analysis

1. From the menus choose:

   Analyze > **Classify** > **Hierarchical Cluster...**

2. If you are clustering cases, select at least one numeric variable. If you are clustering variables, select at least three numeric variables.

Optionally, you can select an identification variable to label cases.

Hierarchical Cluster Analysis Method

Cluster Method. Available alternatives are between-groups linkage, within-groups linkage, nearest neighbor, furthest neighbor, centroid clustering, median clustering, and Ward's method.

Measure. Allows you to specify the distance or similarity measure to be used in clustering. Select the type of data and the appropriate distance or similarity measure:

- **Interval.** Available alternatives are Euclidean distance, squared Euclidean distance, cosine, Pearson correlation, Chebychev, block, Minkowski, and customized.
- **Counts.** Available alternatives are chi-square measure and phi-square measure.
- **Binary.** Available alternatives are Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, dispersion, shape, simple matching, phi 4-point correlation, lambda, Anderberg's *D*, dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance and Williams, Ochiai, Rogers and Tanimoto, Russel and Rao, Sokal and Sneath 1, Sokal and Sneath 2, Sokal and Sneath 3, Sokal and Sneath 4, Sokal and Sneath 5, Yule's *Y*, and Yule's *Q*.

Transform Values. Allows you to standardize data values for either cases or values before computing proximities (not available for binary data). Available standardization methods are *z* scores, range −1 to 1, range 0 to 1, maximum magnitude of 1, mean of 1, and standard deviation of 1.

Transform Measures. Allows you to transform the values generated by the distance measure. They are applied after the distance measure has been computed. Available alternatives are absolute values, change sign, and rescale to 0–1 range.

Hierarchical Cluster Analysis Statistics

Agglomeration schedule. Displays the cases or clusters combined at each stage, the distances between the cases or clusters being combined, and the last cluster level at which a case (or variable) joined the cluster.

Proximity matrix. Gives the distances or similarities between items.

Cluster Membership. Displays the cluster to which each case is assigned at one or more stages in the combination of clusters. Available options are single solution and range of solutions.

Hierarchical Cluster Analysis Plots

Dendrogram. Displays a *dendrogram*. Dendrograms can be used to assess the cohesiveness of the clusters formed and can provide information about the appropriate number of clusters to keep.

Icicle. Displays an *icicle plot*, including all clusters or a specified range of clusters. Icicle plots display information about how cases are combined into clusters at each iteration of the analysis. Orientation allows you to select a vertical or horizontal plot.

Hierarchical Cluster Analysis Save New Variables

Cluster Membership. Allows you to save cluster memberships for a single solution or a range of solutions. Saved variables can then be used in subsequent analyses to explore other differences between groups.

CLUSTER Command Syntax Additional Features

The Hierarchical Cluster procedure uses CLUSTER command syntax. The command syntax language also allows you to:

- Use several clustering methods in a single analysis.
- Read and analyze a proximity matrix.
- Write a proximity matrix to disk for later analysis.
- Specify any values for power and root in the customized (Power) distance measure.
- Specify names for saved variables.

See the *Command Syntax Reference* for complete syntax information.

K-Means Cluster Analysis

This procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that can handle large numbers of cases. However, the algorithm requires you to specify the number of clusters. You can specify initial cluster centers if you know this information. You can select one of two methods for classifying cases, either updating cluster centers iteratively or classifying only. You can save cluster membership, distance information, and final cluster centers. Optionally, you can specify a variable whose values are used to label casewise output. You can also request analysis of variance *F* statistics. While these statistics are opportunistic (the procedure tries to form groups that do differ), the relative size of the statistics provides information about each variable's contribution to the separation of the groups.

Example. What are some identifiable groups of television shows that attract similar audiences within each group? With *k*-means cluster analysis, you could cluster television shows (cases) into *k* homogeneous groups based on viewer characteristics. This process can be used to identify segments for marketing. Or you can cluster cities (cases) into homogeneous groups so that comparable cities can be selected to test various marketing strategies.

Statistics. Complete solution: initial cluster centers, ANOVA table. Each case: cluster information, distance from cluster center.

K-Means Cluster Analysis Data Considerations

Data. Variables should be quantitative at the interval or ratio level. If your variables are binary or counts, use the Hierarchical Cluster Analysis procedure.

Case and initial cluster center order. The default algorithm for choosing initial cluster centers is not invariant to case ordering. The **Use running means** option in the Iterate dialog box makes the resulting solution potentially dependent on case order, regardless of how initial cluster centers are chosen. If you are using either of these methods, you may want to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution. Specifying initial cluster centers and not using the **Use running means** option will avoid issues related to case order. However, ordering of the initial cluster centers may affect the solution if there are tied distances from cases to cluster centers. To assess the stability of a given solution, you can compare results from analyses with different permutations of the initial center values.

Assumptions. Distances are computed using simple Euclidean distance. If you want to use another distance or similarity measure, use the Hierarchical Cluster Analysis procedure. Scaling of variables is an important consideration. If your variables are measured on different scales (for example, one variable is expressed in dollars and another variable is expressed in years), your results may be misleading. In such cases, you should consider standardizing your variables before you perform the *k*-means cluster analysis (this task can be done in the Descriptives procedure). The procedure assumes that you have selected the appropriate number of clusters and that you have included all relevant variables. If you have chosen an inappropriate number of clusters or omitted important variables, your results may be misleading.

To Obtain a K-Means Cluster Analysis

1. From the menus choose:

   Analyze > **Classify** > **K-Means Cluster...**
2. Select the variables to be used in the cluster analysis.
3. Specify the number of clusters. (The number of clusters must be at least 2 and must not be greater than the number of cases in the data file.)
4. Select either **Iterate and classify** or **Classify only**.
5. Optionally, select an identification variable to label cases.

K-Means Cluster Analysis Efficiency

The *k*-means cluster analysis command is efficient primarily because it does not compute the distances between all pairs of cases, as do many clustering algorithms, including the algorithm that is used by the hierarchical clustering command.

For maximum efficiency, take a sample of cases and select the **Iterate and classify** method to determine cluster centers. Select **Write final as**. Then restore the entire data file and select **Classify only** as the method and select **Read initial from** to classify the entire file using the centers that are estimated from the sample. You can write to and read from a file or a dataset. Datasets are available for subsequent use in the same session but are not saved as files unless explicitly saved prior to the end of the session. Dataset names must conform to variable-naming rules. See the topic for more information.

K-Means Cluster Analysis Iterate

Note: These options are available only if you select the **Iterate and classify** method from the K-Means Cluster Analysis dialog box.

Maximum Iterations. Limits the number of iterations in the *k*-means algorithm. Iteration stops after this many iterations even if the convergence criterion is not satisfied. This number must be between 1 and 999.

To reproduce the algorithm used by the Quick Cluster command prior to version 5.0, set **Maximum Iterations** to 1.

Convergence Criterion. Determines when iteration ceases. It represents a proportion of the minimum distance between initial cluster centers, so it must be greater than 0 but not greater than 1. If the criterion equals 0.02, for example, iteration ceases when a complete iteration does not move any of the cluster centers by a distance of more than 2% of the smallest distance between any initial cluster centers.

Use running means. Allows you to request that cluster centers be updated after each case is assigned. If you do not select this option, new cluster centers are calculated after all cases have been assigned.

K-Means Cluster Analysis Save

You can save information about the solution as new variables to be used in subsequent analyses:

Cluster membership. Creates a new variable indicating the final cluster membership of each case. Values of the new variable range from 1 to the number of clusters.

Distance from cluster center. Creates a new variable indicating the Euclidean distance between each case and its classification center.

K-Means Cluster Analysis Options

Statistics. You can select the following statistics: initial cluster centers, ANOVA table, and cluster information for each case.

- *Initial cluster centers*. First estimate of the variable means for each of the clusters. By default, a number of well-spaced cases equal to the number of clusters is selected from the data. Initial cluster centers are used for a first round of classification and are then updated.
- *ANOVA table*. Displays an analysis-of-variance table which includes univariate F tests for each clustering variable. The F tests are only descriptive and the resulting probabilities should not be interpreted. The ANOVA table is not displayed if all cases are assigned to a single cluster.
- *Cluster information for each case*. Displays for each case the final cluster assignment and the Euclidean distance between the case and the cluster center used to classify the case. Also displays Euclidean distance between final cluster centers.

Missing Values. Available options are **Exclude cases listwise** or **Exclude cases pairwise**.

- **Exclude cases listwise.** Excludes cases with missing values for any clustering variable from the analysis.
- **Exclude cases pairwise.** Assigns cases to clusters based on distances that are computed from all variables with nonmissing values.

QUICK CLUSTER Command Additional Features

The K-Means Cluster procedure uses `QUICK CLUSTER` command syntax. The command syntax language also allows you to:

- Accept the first *k* cases as initial cluster centers, thereby avoiding the data pass that is normally used to estimate them.
- Specify initial cluster centers directly as a part of the command syntax.
- Specify names for saved variables.

See the *Command Syntax Reference* for complete syntax information.

Nonparametric Tests

Nonparametric tests make minimal assumptions about the underlying distribution of the data. The tests that are available in these dialogs can be grouped into three broad categories based on how the data are organized:

- A one-sample test analyzes one field.
- A test for related samples compares two or more fields for the same set of cases.
- An independent-samples test analyzes one field that is grouped by categories of another field.

One-Sample Nonparametric Tests

One-sample nonparametric tests identify differences in single fields using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective? The objectives allow you to quickly specify different but commonly used test settings.

- **Automatically compare observed data to hypothesized.** This objective applies the Binomial test to categorical fields with only two categories, the Chi-Square test to all other categorical fields, and the Kolmogorov-Smirnov test to continuous fields.

- **Test sequence for randomness.** This objective uses the Runs test to test the observed sequence of data values for randomness.
- **Custom analysis.** When you want to manually amend the test settings on the Settings tab, select this option. Note that this setting is automatically selected if you subsequently make changes to options on the Settings tab that are incompatible with the currently selected objective.

Obtaining One-Sample Nonparametric Tests

From the menus choose:

Analyze > **Nonparametric Tests** > **One Sample...**

1. Click **Run**.

Optionally, you can:

- Specify an objective on the Objective tab.
- Specify field assignments on the Fields tab.
- Specify expert settings on the Settings tab.

Fields Tab

The Fields tab specifies which fields should be tested.

Use predefined roles. This option uses existing field information. All fields with a predefined role as Input, Target, or Both will be used as test fields. At least one test field is required.

Use custom field assignments. This option allows you to override field roles. After selecting this option, specify the fields below:

- **Test Fields.** Select one or more fields.

Settings Tab

The Settings tab comprises several different groups of settings that you can modify to fine-tune how the algorithm processes your data. If you make any changes to the default settings that are incompatible with the currently selected objective, the Objective tab is automatically updated to select the **Customize analysis** option.

Choose Tests

These settings specify the tests to be performed on the fields specified on the Fields tab.

Automatically choose the tests based on the data. This setting applies the Binomial test to categorical fields with only two valid (non-missing) categories, the Chi-Square test to all other categorical fields, and the Kolmogorov-Smirnov test to continuous fields.

Customize tests. This setting allows you to choose specific tests to be performed.

- **Compare observed binary probability to hypothesized (Binomial test).** The Binomial test can be applied to all fields. This produces a one-sample test that tests whether the observed distribution of a flag field (a categorical field with only two categories) is the same as what is expected from a specified binomial distribution. In addition, you can request confidence intervals. See “Binomial Test Options ” on page 186 for details on the test settings.
- **Compare observed probabilities to hypothesized (Chi-Square test).** The Chi-Square test is applied to nominal and ordinal fields. This produces a one-sample test that computes a chi-square statistic based on the differences between the observed and expected frequencies of categories of a field. See “Chi-Square Test Options ” on page 186 for details on the test settings.
- **Test observed distribution against hypothesized (Kolmogorov-Smirnov test).** The Kolmogorov-Smirnov test is applied to continuous and ordinal fields. This produces a one-sample test of whether the sample cumulative distribution function for a field is homogenous with a uniform, normal, Poisson,

or exponential distribution. See “Kolmogorov-Smirnov Options ” on page 186 for details on the test settings.

- **Compare median to hypothesized (Wilcoxon signed-rank test).** The Wilcoxon signed-rank test is applied to continuous and ordinal fields. This produces a one-sample test of median value of a field. Specify a number as the hypothesized median.
- **Test sequence for randomness (Runs test).** The Runs test is applied to all fields. This produces a one-sample test of whether the sequence of values of a dichotomized field is random. See “Runs Test Options ” on page 187 for details on the test settings.

Binomial Test Options

The binomial test is intended for flag fields (categorical fields with only two categories), but is applied to all fields by using rules for defining "success".

Hypothesized proportion. This specifies the expected proportion of records defined as "successes", or *p*. Specify a value greater than 0 and less than 1. The default is 0.5.

Confidence Interval. The following methods for computing confidence intervals for binary data are available:

- **Clopper-Pearson (exact).** An exact interval based on the cumulative binomial distribution.
- **Jeffreys.** A Bayesian interval based on the posterior distribution of *p* using the Jeffreys prior.
- **Likelihood ratio.** An interval based on the likelihood function for *p*.

Define Success for Categorical Fields. This specifies how "success", the data value(s) tested against the hypothesized proportion, is defined for categorical fields.

- **Use first category found in data** performs the binomial test using the first value found in the sample to define "success". This option is only applicable to nominal or ordinal fields with only two values; all other categorical fields specified on the Fields tab where this option is used will not be tested. This is the default.
- **Specify success values** performs the binomial test using the specified list of values to define "success". Specify a list of string or numeric values. The values in the list do not need to be present in the sample.

Define Success for Continuous Fields. This specifies how "success", the data value(s) tested against the test value, is defined for continuous fields. Success is defined as values equal to or less than a cut point.

- **Sample midpoint** sets the cut point at the average of the minimum and maximum values.
- **Custom cutpoint** allows you to specify a value for the cut point.

Chi-Square Test Options

All categories have equal probability. This produces equal frequencies among all categories in the sample. This is the default.

Customize expected probability. This allows you to specify unequal frequencies for a specified list of categories. Specify a list of string or numeric values. The values in the list do not need to be present in the sample. In the **Category** column, specify category values. In the **Relative Frequency** column, specify a value greater than 0 for each category. Custom frequencies are treated as ratios so that, for example, specifying frequencies 1, 2, and 3 is equivalent to specifying frequencies 10, 20, and 30, and both specify that 1/6 of the records are expected to fall into the first category, 1/3 into the second, and 1/2 into the third. When custom expected probabilities are specified, the custom category values must include all the field values in the data; otherwise the test is not performed for that field.

Kolmogorov-Smirnov Options

This dialog specifies which distributions should be tested and the parameters of the hypothesized distributions.

When certain parameters of the distribution have to be estimated from the sample, the Kolmogorov-Smirnov test no longer applies. In these instances, the Lilliefors test statistic can be used to estimate the *p*-value by using the Monte Carlo sampling for testing normality with mean and variance unknown. The

Lilliefors test applies to the three continuous distributions (**Normal**, **Exponential**, and **Uniform**). Note that the test does not apply if the underlying distribution is discrete (**Poisson**). The test is only defined for one-sample inference when the corresponding distribution parameters are not specified.

Normal
: **Use sample data** uses the observed mean and standard deviation and provides options for selecting the existing **Asymptotic test** results, or use **Lilliefors test based on the Monte Carlo sampling**. **Custom** allows you to specify values.

Uniform
: **Use sample data** uses the observed minimum and maximum and uses Lilliefors test based on the Monte Carlo sampling. **Custom** allows you to specify minimum and maximum values.

Exponential
: **Sample mean** uses the observed mean and uses Lilliefors test based on the Monte Carlo sampling. **Custom** allows you to specify an observed mean value.

Poisson
: **Mean** allows you to specify an observed mean value.

Runs Test Options

The runs test is intended for flag fields (categorical fields with only two categories), but can be applied to all fields by using rules for defining the groups.

Define Groups for Categorical Fields. The following options are available:

- **There are only 2 categories in the sample** performs the runs test using the values found in the sample to define the groups. This option is only applicable to nominal or ordinal fields with only two values; all other categorical fields specified on the Fields tab where this option is used will not be tested.
- **Recode data into 2 categories** performs the runs test using the specified list of values to define one of the groups. All other values in the sample define the other group. The values in the list do not all need to be present in the sample, but at least one record must be in each group.

Define Cut Point for Continuous Fields. This specifies how groups are defined for continuous fields. The first group is defined as values equal to or less than a cut point.

- **Sample median** sets the cut point at the sample median.
- **Sample mean** sets the cut point at the sample mean.
- **Custom** allows you to specify a value for the cut point.

Test Options

Significance level
: This specifies the significance level (alpha) for all tests. Specify a numeric value between 0 and 1. 0.05 is the default.

Confidence interval (%)
: This specifies the confidence level for all confidence intervals produced. Specify a numeric value between 0 and 100. 95 is the default.

Excluded Cases
: This specifies how to determine the case basis for tests.

  Exclude cases test by test
  : Records with missing values for a field that is used for a specific test are omitted from that test. When several tests are specified in the analysis, each test is evaluated separately.

  Exclude cases listwise
  : Records with missing values for any field that is named on the Fields tab are excluded from all analyses.

Monte Carlo Sampling

: When certain parameters of the distribution have to be estimated from the sample, the Kolmogorov-Smirnov test no longer applies. In these instances, the Lilliefors test statistic can be used to estimate

the *p*-value by using the Monte Carlo sampling for testing normality with mean and variance unknown. The Lilliefors test applies to the three continuous distributions (**Normal**, **Exponential**, and **Uniform**). Note that the test does not apply if the underlying distribution is discrete (**Poisson**). The test is only defined for one-sample inference when the corresponding distribution parameters are not specified.

Set custom seed
When enabled, this setting provides the option of resetting the random **Seed** value that is used for Monte Carlo sampling. The value must be a single integer between 1 and 2,147,483,647 (the default value is 2,000,000). By default, the setting is not enabled, which produces a random seed value.

Number of samples
Resets the number of Monte Carlo sampling replicates that are used by the Lilliefors test. The value must be a single integer between 100 and the largest integer. The default value is 10,000.

Simulation confidence level (%)
Resets the Kolmogorov-Smirnov test's estimated confidence interval level. The value must be a single value between 0 and 100. The default value is 99.

User-Missing Values

User-Missing Values for Categorical Fields. Categorical fields must have valid values for a record to be included in the analysis. These controls allow you to decide whether user-missing values are treated as valid among categorical fields. System-missing values and missing values for continuous fields are always treated as invalid.

NPTESTS command additional features

The command syntax language also allows you to:

- Specify one-sample, independent-samples, and related-samples tests in a single run of the procedure.

See the *Command Syntax Reference* for complete syntax information.

One-Sample Wilcoxon Test

The Wilcoxon signed-rank test is a non-parametric statistical test that is used to test the location of a population based on sample data or to compare the locations of two populations. The one-sample Wilcoxon signed rank test is a non-parametric alternative to the sample t-test when the data does not follow a normal distribution. It helps to determine whether there is a significant difference between the median of a sample and a hypothesized median value. Here, the test statistic is computed based on the ranks of the difference between observed values and the hypothesized value.

The one sample Wilcoxon signed rank test makes the assumption that the parent population is symmetric about the median. As the random samples come from a population with a continuous cumulative distribution function and are symmetric about the median, under the null hypothesis, the differences between observed values and hypothesized value are symmetrically distributed about zero so that the positive and negative differences of equal absolute magnitude have the same probability of occurrence.

Example

Suppose that a new startup company is interested in hiring software engineers but wants to ensure that they are offering competitive salaries. Assume that they have collected salary data from a sample of 30 software engineers who are currently employed at various companies in the same industry. If the company wants to determine whether the median salary they are offering to their potential hires is significantly different from the median salary of software engineers in the industry, then one sample Wilcoxon sign rank can be used for this purpose. Here, we assume the null hypothesis that the median salary offered by the startup is not significantly different from the median salary in the industry, that is, the median salary difference is zero.

To run the one-sample Wilcoxon test

(This feature requires the Statistics Base option.)

From the menu choose:

1. **Analyze > Nonparametric Tests > One Sample...**
2. Click **'Settings'**.
3. Under **'Select an item'**, select **'Choose Tests'**.
4. Choose **'Customise Tests'**.
5. Select **'Compare median to hypothesized (Wilcoxon signed rank test)'**.
6. Enter the value of the **'Hypothesized median'** and click **'Run'**.

For more information, see https://www.ibm.com/support/pages/can-spss-statistics-produce-one-sample-wilcoxon-test

Independent-Samples Nonparametric Tests

Independent-samples nonparametric tests identify differences between two or more groups using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective? The objectives allow you to quickly specify different but commonly used test settings.

- **Automatically compare distributions across groups.** This objective applies the Mann-Whitney U test to data with 2 groups, or the Kruskal-Wallis 1-way ANOVA to data with *k* groups.
- **Compare medians across groups.** This objective uses the Median test to compare the observed medians across groups.
- **Custom analysis.** When you want to manually amend the test settings on the Settings tab, select this option. Note that this setting is automatically selected if you subsequently make changes to options on the Settings tab that are incompatible with the currently selected objective.

To Obtain Independent-Samples Nonparametric Tests

From the menus choose:

Analyze > **Nonparametric Tests** > **Independent Samples...**

1. Click **Run**.

Optionally, you can:

- Specify an objective on the Objective tab.
- Specify field assignments on the Fields tab.
- Specify expert settings on the Settings tab.

Fields Tab

The Fields tab specifies which fields should be tested and the field used to define groups.

Use predefined roles. This option uses existing field information. All continuous and ordinal fields with a predefined role as Target or Both will be used as test fields. If there is a single categorical field with a predefined role as Input, it will be used as a grouping field. Otherwise no grouping field is used by default and you must use custom field assignments. At least one test field and a grouping field is required.

Use custom field assignments. This option allows you to override field roles. After selecting this option, specify the fields below:

- **Test Fields.** Select one or more continuous or ordinal fields.
- **Groups.** Select a categorical field.

Settings Tab

The Settings tab comprises several different groups of settings that you can modify to fine tune how the algorithm processes your data. If you make any changes to the default settings that are incompatible with the currently selected objective, the Objective tab is automatically updated to select the **Customize analysis** option.

Choose Tests

These settings specify the tests to be performed on the fields specified on the Fields tab.

Automatically choose the tests based on the data. This setting applies the Mann-Whitney U test to data with 2 groups, or the Kruskal-Wallis 1-way ANOVA to data with *k* groups.

Customize tests. This setting allows you to choose specific tests to be performed.

- **Compare Distributions across Groups.** These produce independent-samples tests of whether the samples are from the same population.

  Mann-Whitney U (2 samples) uses the rank of each case to test whether the groups are drawn from the same population. The first value in ascending order of the grouping field defines the first group and the second defines the second group. If the grouping field has more than two values, this test is not produced.

  Kolmogorov-Smirnov (2 samples) is sensitive to any difference in median, dispersion, skewness, and so forth, between the two distributions. If the grouping field has more than two values, this test is not produced.

  Test sequence for randomness (Wald-Wolfowitz for 2 samples) produces a runs test with group membership as the criterion. If the grouping field has more than two values, this test is not produced.

  Kruskal-Wallis 1-way ANOVA (k samples) is an extension of the Mann-Whitney U test and the nonparametric analog of one-way analysis of variance. You can optionally request multiple comparisons of the *k* samples, either **all pairwise** multiple comparisons or **stepwise step-down** comparisons.

  Test for ordered alternatives (Jonckheere-Terpstra for k samples) is a more powerful alternative to Kruskal-Wallis when the *k* samples have a natural ordering. For example, the *k* populations might represent *k* increasing temperatures. The hypothesis that different temperatures produce the same response distribution is tested against the alternative that as the temperature increases, the magnitude of the response increases. Here, the alternative hypothesis is ordered; therefore, Jonckheere-Terpstra is the most appropriate test to use. **Smallest to largest** specifies the alternative hypothesis that the location parameter of the first group is less than or equal to the second, which is less than or equal to the third, and so on. **Largest to smallest** specifies the alternative hypothesis that the location parameter of the first group is greater than or equal to the second, which is greater than or equal to the third, and so on. For both options, the alternative hypothesis also assumes that the locations are not all equal. You can optionally request multiple comparisons of the *k* samples, either **All pairwise** multiple comparisons or **Stepwise step-down** comparisons.
- **Compare Ranges across Groups.** This produces an independent-samples tests of whether the samples have the same range. **Moses extreme reaction (2 samples)** tests a control group versus a comparison group. The first value in ascending order of the grouping field defines the control group and the second defines the comparison group. If the grouping field has more than two values, this test is not produced.
- **Compare Medians across Groups.** This produces an independent-samples tests of whether the samples have the same median. **Median test (k samples)** can use either the pooled sample median (calculated across all records in the dataset) or a custom value as the hypothesized median. You can optionally request multiple comparisons of the *k* samples, either **All pairwise** multiple comparisons or **Stepwise step-down** comparisons.
- **Estimate Confidence Intervals across Groups.** **Hodges-Lehman estimate (2 samples)** produces an independent samples estimate and confidence interval for the difference in the medians of two groups. If the grouping field has more than two values, this test is not produced.

Test Options

Significance level. This specifies the significance level (alpha) for all tests. Specify a numeric value between 0 and 1. 0.05 is the default.

Confidence interval (%). This specifies the confidence level for all confidence intervals produced. Specify a numeric value between 0 and 100. 95 is the default.

Excluded Cases. This specifies how to determine the case basis for tests. **Exclude cases listwise** means that records with missing values for any field that is named on any subcommand are excluded from all analyses. **Exclude cases test by test** means that records with missing values for a field that is used for a specific test are omitted from that test. When several tests are specified in the analysis, each test is evaluated separately.

User-Missing Values

User-Missing Values for Categorical Fields. Categorical fields must have valid values for a record to be included in the analysis. These controls allow you to decide whether user-missing values are treated as valid among categorical fields. System-missing values and missing values for continuous fields are always treated as invalid.

NPTESTS command additional features

The command syntax language also allows you to:

- Specify one-sample, independent-samples, and related-samples tests in a single run of the procedure.

See the *Command Syntax Reference* for complete syntax information.

Related-Samples Nonparametric Tests

Identifies differences between two or more related fields using one or more nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

Data Considerations. Each record corresponds to a given subject for which two or more related measurements are stored in separate fields in the dataset. For example, a study concerning the effectiveness of a dieting plan can be analyzed using related-samples nonparametric tests if each subject's weight is measured at regular intervals and stored in fields like *Pre-diet weight*, *Interim weight*, and *Post-diet weight*. These fields are "related".

What is your objective? The objectives allow you to quickly specify different but commonly used test settings.

- **Automatically compare observed data to hypothesized data.** This objective applies McNemar's Test to categorical data when 2 fields are specified, Cochran's Q to categorical data when more than 2 fields are specified, the Wilcoxon Matched-Pair Signed-Rank test to continuous data when 2 fields are specified, and Friedman's 2-Way ANOVA by Ranks to continuous data when more than 2 fields are specified.
- **Custom analysis.** When you want to manually amend the test settings on the Settings tab, select this option. Note that this setting is automatically selected if you subsequently make changes to options on the Settings tab that are incompatible with the currently selected objective.

When fields of differing measurement level are specified, they are first separated by measurement level and then the appropriate test is applied to each group. For example, if you choose **Automatically compare observed data to hypothesized data** as your objective and specify 3 continuous fields and 2 nominal fields, then Friedman's test is applied to the continuous fields and McNemar's test is applied to the nominal fields.

To Obtain Related-Samples Nonparametric Tests

From the menus choose:

Analyze > **Nonparametric Tests** > **Related Samples...**

1. Click **Run**.

Optionally, you can:

- Specify an objective on the Objective tab.
- Specify field assignments on the Fields tab.
- Specify expert settings on the Settings tab.

Fields Tab

The Fields tab specifies which fields should be tested.

Use predefined roles. This option uses existing field information. All fields with a predefined role as Target or Both will be used as test fields. At least two test fields are required.

Use custom field assignments. This option allows you to override field roles. After selecting this option, specify the fields below:

- **Test Fields.** Select two or more fields. Each field corresponds to a separate related sample.

Settings Tab

The Settings tab comprises several different groups of settings that you can modify to fine tune how the procedure processes your data. If you make any changes to the default settings that are incompatible with the other objectives, the Objective tab is automatically updated to select the **Customize analysis** option.

Choose Tests

These settings specify the tests to be performed on the fields specified on the Fields tab.

Automatically choose the tests based on the data. This setting applies McNemar's Test to categorical data when 2 fields are specified, Cochran's Q to categorical data when more than 2 fields are specified, the Wilcoxon Matched-Pair Signed-Rank test to continuous data when 2 fields are specified, and Friedman's 2-Way ANOVA by Ranks to continuous data when more than 2 fields are specified.

Customize tests. This setting allows you to choose specific tests to be performed.

- **Test for Change in Binary Data. McNemar's test (2 samples)** can be applied to categorical fields. This produces a related-samples test of whether combinations of values between two flag fields (categorical fields with only two values) are equally likely. If there are more than two fields specified on the Fields tab, this test is not performed. See “McNemar's Test: Define Success ” on page 193 for details on the test settings. **Cochran's Q (k samples)** can be applied to categorical fields. This produces a related-samples test of whether combinations of values between *k* flag fields (categorical fields with only two values) are equally likely. You can optionally request multiple comparisons of the *k* samples, either **all pairwise** multiple comparisons or **stepwise step-down** comparisons. See “Cochran's Q: Define Success ” on page 193 for details on the test settings.
- **Test for Changes in Multinomial Data. Marginal homogeneity test (2 samples)** produces a related samples test of whether combinations of values between two paired ordinal fields are equally likely. The marginal homogeneity test is typically used in repeated measures situations. This test is an extension of the McNemar test from binary response to multinomial response. If there are more than two fields specified on the Fields tab, this test is not performed.
- **Compare Median Difference to Hypothesized.** These tests each produce a related-samples test of whether the median difference between two fields is different from 0. The test applies to continuous and ordinal fields. If there are more than two fields specified on the Fields tab, these tests are not performed.
- **Estimate Confidence Interval.** This produces a related samples estimate and confidence interval for the median difference between two paired fields. The test applies to continuous and ordinal fields. If there are more than two fields specified on the Fields tab, this test is not performed.

- **Quantify Associations.** **Kendall's coefficient of concordance (k samples)** produces a measure of agreement among judges or raters, where each record is one judge's rating of several items (fields). You can optionally request multiple comparisons of the *k* samples, either **All pairwise** multiple comparisons or **Stepwise step-down** comparisons.
- **Compare Distributions.** **Friedman's 2-way ANOVA by ranks (k samples)** produces a related samples test of whether *k* related samples have been drawn from the same population. You can optionally request multiple comparisons of the *k* samples, either **All pairwise** multiple comparisons or **Stepwise step-down** comparisons.

McNemar's Test: Define Success

McNemar's test is intended for flag fields (categorical fields with only two categories), but is applied to all categorical fields by using rules for defining "success".

Define Success for Categorical Fields. This specifies how "success" is defined for categorical fields.

- **Use first category found in data** performs the test using the first value found in the sample to define "success". This option is only applicable to nominal or ordinal fields with only two values; all other categorical fields specified on the Fields tab where this option is used will not be tested. This is the default.
- **Specify success values** performs the test using the specified list of values to define "success". Specify a list of string or numeric values. The values in the list do not need to be present in the sample.

Cochran's Q: Define Success

Cochran's Q test is intended for flag fields (categorical fields with only two categories), but is applied to all categorical fields by using rules for defining "success".

Define Success for Categorical Fields. This specifies how "success" is defined for categorical fields.

- **Use first category found in data** performs the test using the first value found in the sample to define "success". This option is only applicable to nominal or ordinal fields with only two values; all other categorical fields specified on the Fields tab where this option is used will not be tested. This is the default.
- **Specify success values** performs the test using the specified list of values to define "success". Specify a list of string or numeric values. The values in the list do not need to be present in the sample.

Test Options

Significance level. This specifies the significance level (alpha) for all tests. Specify a numeric value between 0 and 1. 0.05 is the default.

Confidence interval (%). This specifies the confidence level for all confidence intervals produced. Specify a numeric value between 0 and 100. 95 is the default.

Excluded Cases. This specifies how to determine the case basis for tests.

- **Exclude cases listwise** means that records with missing values for any field that is named on any subcommand are excluded from all analyses.
- **Exclude cases test by test** means that records with missing values for a field that is used for a specific test are omitted from that test. When several tests are specified in the analysis, each test is evaluated separately.

User-Missing Values

User-Missing Values for Categorical Fields. Categorical fields must have valid values for a record to be included in the analysis. These controls allow you to decide whether user-missing values are treated as valid among categorical fields. System-missing values and missing values for continuous fields are always treated as invalid.

NPTESTS command additional features

The command syntax language also allows you to:

- Specify one-sample, independent-samples, and related-samples tests in a single run of the procedure.

See the *Command Syntax Reference* for complete syntax information.

NPTESTS command additional features

The command syntax language also allows you to:

- Specify one-sample, independent-samples, and related-samples tests in a single run of the procedure.

See the *Command Syntax Reference* for complete syntax information.

Legacy Dialogs

There are a number of "legacy" dialogs that also perform nonparametric tests. These dialogs support the functionality provided by the Exact Tests option.

Chi-Square Test. Tabulates a variable into categories and computes a chi-square statistic based on the differences between observed and expected frequencies.

Binomial Test. Compares the observed frequency in each category of a dichotomous variable with expected frequencies from the binomial distribution.

Runs Test. Tests whether the order of occurrence of two values of a variable is random.

One-Sample Kolmogorov-Smirnov Test. Compares the observed cumulative distribution function for a variable with a specified theoretical distribution, which may be normal, uniform, exponential, or Poisson.

Two-Independent-Samples Tests. Compares two groups of cases on one variable. The Mann-Whitney *U* test, two-sample Kolmogorov-Smirnov test, Moses test of extreme reactions, and Wald-Wolfowitz runs test are available.

Two-Related-Samples Tests. Compares the distributions of two variables. The Wilcoxon signed-rank test, the sign test, and the McNemar test are available.

Tests for Several Independent Samples. Compares two or more groups of cases on one variable. The Kruskal-Wallis test, the Median test, and the Jonckheere-Terpstra test are available.

Tests for Several Related Samples. Compares the distributions of two or more variables. Friedman's test, Kendall's *W*, and Cochran's *Q* are available.

Quartiles and the mean, standard deviation, minimum, maximum, and number of nonmissing cases are available for all of the above tests.

Chi-Square Test

The Chi-Square Test procedure tabulates a variable into categories and computes a chi-square statistic. This goodness-of-fit test compares the observed and expected frequencies in each category to test that all categories contain the same proportion of values or test that each category contains a user-specified proportion of values.

Examples. The chi-square test could be used to determine whether a bag of jelly beans contains equal proportions of blue, brown, green, orange, red, and yellow candies. You could also test to see whether a bag of jelly beans contains 5% blue, 30% brown, 10% green, 20% orange, 15% red, and 15% yellow candies.

Statistics. Mean, standard deviation, minimum, maximum, and quartiles. The number and the percentage of nonmissing and missing cases; the number of cases observed and expected for each category; residuals; and the chi-square statistic.

Chi-Square Test Data Considerations

Data. Use ordered or unordered numeric categorical variables (ordinal or nominal levels of measurement). To convert string variables to numeric variables, use the Automatic Recode procedure, which is available on the Transform menu.

Assumptions. Nonparametric tests do not require assumptions about the shape of the underlying distribution. The data are assumed to be a random sample. The expected frequencies for each category should be at least 1. No more than 20% of the categories should have expected frequencies of less than 5.

To Obtain a Chi-Square Test

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **Chi-Square...**
2. Select one or more test variables. Each variable produces a separate test.
3. Optionally, click **Options** for descriptive statistics, quartiles, and control of the treatment of missing data.

Chi-Square Test Expected Range and Expected Values

Expected Range. By default, each distinct value of the variable is defined as a category. To establish categories within a specific range, select **Use specified range** and enter integer values for lower and upper bounds. Categories are established for each integer value within the inclusive range, and cases with values outside of the bounds are excluded. For example, if you specify a value of 1 for Lower and a value of 4 for Upper, only the integer values of 1 through 4 are used for the chi-square test.

Expected Values. By default, all categories have equal expected values. Categories can have user-specified expected proportions. Select **Values**, enter a value that is greater than 0 for each category of the test variable, and then click **Add**. Each time you add a value, it appears at the bottom of the value list. The order of the values is important; it corresponds to the ascending order of the category values of the test variable. The first value of the list corresponds to the lowest group value of the test variable, and the last value corresponds to the highest value. Elements of the value list are summed, and then each value is divided by this sum to calculate the proportion of cases expected in the corresponding category. For example, a value list of 3, 4, 5, 4 specifies expected proportions of 3/16, 4/16, 5/16, and 4/16.

Chi-Square Test Options

Statistics. You can choose one or both summary statistics.

- **Descriptive.** Displays the mean, standard deviation, minimum, maximum, and number of nonmissing cases.
- **Quartiles.** Displays values corresponding to the 25th, 50th, and 75th percentiles.

Missing Values. Controls the treatment of missing values.

- **Exclude cases test-by-test.** When several tests are specified, each test is evaluated separately for missing values.
- **Exclude cases listwise.** Cases with missing values for any variable are excluded from all analyses.

NPAR TESTS Command Additional Features (Chi-Square Test)

The command syntax language also allows you to:

- Specify different minimum and maximum values or expected frequencies for different variables (with the `CHISQUARE` subcommand).
- Test the same variable against different expected frequencies or use different ranges (with the `EXPECTED` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Binomial Test

The Binomial Test procedure compares the observed frequencies of the two categories of a dichotomous variable to the frequencies that are expected under a binomial distribution with a specified probability parameter. By default, the probability parameter for both groups is 0.5. To change the probabilities, you can enter a test proportion for the first group. The probability for the second group will be 1 minus the specified probability for the first group.

Example. When you toss a dime, the probability of a head equals 1/2. Based on this hypothesis, a dime is tossed 40 times, and the outcomes are recorded (heads or tails). From the binomial test, you might find that 3/4 of the tosses were heads and that the observed significance level is small (0.0027). These results indicate that it is not likely that the probability of a head equals 1/2; the coin is probably biased.

Statistics. Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.

Binomial Test Data Considerations

Data. The variables that are tested should be numeric and dichotomous. To convert string variables to numeric variables, use the Automatic Recode procedure, which is available on the Transform menu. A **dichotomous variable** is a variable that can take only two possible values: *yes* or *no*, *true* or *false*, 0 or 1, and so on. The first value encountered in the dataset defines the first group, and the other value defines the second group. If the variables are not dichotomous, you must specify a cut point. The cut point assigns cases with values that are less than or equal to the cut point to the first group and assigns the rest of the cases to the second group.

Assumptions. Nonparametric tests do not require assumptions about the shape of the underlying distribution. The data are assumed to be a random sample.

To Obtain a Binomial Test

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **Binomial...**
2. Select one or more numeric test variables.
3. Optionally, click **Options** for descriptive statistics, quartiles, and control of the treatment of missing data.

Binomial Test Options

Statistics. You can choose one or both summary statistics.

- **Descriptive.** Displays the mean, standard deviation, minimum, maximum, and number of nonmissing cases.
- **Quartiles.** Displays values corresponding to the 25th, 50th, and 75th percentiles.

Missing Values. Controls the treatment of missing values.

- **Exclude cases test-by-test.** When several tests are specified, each test is evaluated separately for missing values.
- **Exclude cases listwise.** Cases with missing values for any variable that is tested are excluded from all analyses.

NPAR TESTS Command Additional Features (Binomial Test)

The command syntax language also allows you to:

- Select specific groups (and exclude other groups) when a variable has more than two categories (with the `BINOMIAL` subcommand).
- Specify different cut points or probabilities for different variables (with the `BINOMIAL` subcommand).
- Test the same variable against different cut points or probabilities (with the `EXPECTED` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Runs Test

The Runs Test procedure tests whether the order of occurrence of two values of a variable is random. A run is a sequence of like observations. A sample with too many or too few runs suggests that the sample is not random.

Examples. Suppose that 20 people are polled to find out whether they would purchase a product. The assumed randomness of the sample would be seriously questioned if all 20 people were of the same gender. The runs test can be used to determine whether the sample was drawn at random.

Statistics. Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles.

Runs Test Data Considerations

Data. The variables must be numeric. To convert string variables to numeric variables, use the Automatic Recode procedure, which is available on the Transform menu.

Assumptions. Nonparametric tests do not require assumptions about the shape of the underlying distribution. Use samples from continuous probability distributions.

To Obtain a Runs Test

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **Runs...**
2. Select one or more numeric test variables.
3. Optionally, click **Options** for descriptive statistics, quartiles, and control of the treatment of missing data.

Runs Test Cut Point

Cut Point. Specifies a cut point to dichotomize the variables that you have chosen. You can use the observed mean, median, or mode, or you can use a specified value as a cut point. Cases with values that are less than the cut point are assigned to one group, and cases with values that are greater than or equal to the cut point are assigned to another group. One test is performed for each chosen cut point.

Runs Test Options

Statistics. You can choose one or both summary statistics.

- **Descriptive.** Displays the mean, standard deviation, minimum, maximum, and number of nonmissing cases.
- **Quartiles.** Displays values corresponding to the 25th, 50th, and 75th percentiles.

Missing Values. Controls the treatment of missing values.

- **Exclude cases test-by-test.** When several tests are specified, each test is evaluated separately for missing values.
- **Exclude cases listwise.** Cases with missing values for any variable are excluded from all analyses.

NPAR TESTS Command Additional Features (Runs Test)

The command syntax language also allows you to:

- Specify different cut points for different variables (with the RUNS subcommand).
- Test the same variable against different custom cut points (with the RUNS subcommand).

See the *Command Syntax Reference* for complete syntax information.

One-Sample Kolmogorov-Smirnov Test

The One-Sample Kolmogorov-Smirnov Test procedure compares the observed cumulative distribution function for a variable with a specified theoretical distribution, which may be normal, uniform, Poisson, or exponential. The Kolmogorov-Smirnov *Z* is computed from the largest difference (in absolute value)

between the observed and theoretical cumulative distribution functions. This goodness-of-fit test tests whether the observations could reasonably have come from the specified distribution.

Starting with version 27.0, the Lilliefors test statistic can be used to estimate the *p*-value by using the Monte Carlo sampling for testing against a normal distribution with estimated parameters (this functionality was previously possible only through the Explore procedure).

Example
Many parametric tests require normally distributed variables. The one-sample Kolmogorov-Smirnov test can be used to test that a variable (for example, *income*) is normally distributed.

Statistics
Mean, standard deviation, minimum, maximum, number of non-missing cases, quartiles, Lilliefors test, and Monte Carlo simulation.

One-Sample Kolmogorov-Smirnov test data considerations

Data
Use quantitative variables (interval or ratio level of measurement).

Assumptions
The Kolmogorov-Smirnov test assumes that the parameters of the test distribution are specified in advance. This procedure estimates the parameters from the sample. The sample mean and sample standard deviation are the parameters for a normal distribution, the sample minimum and maximum values define the range of the uniform distribution, the sample mean is the parameter for the Poisson distribution, and the sample mean is the parameter for the exponential distribution. The power of the test to detect departures from the hypothesized distribution may be seriously diminished.

When certain parameters of the distribution have to be estimated from the sample, the Kolmogorov-Smirnov test no longer applies. In these instances, the Lilliefors test statistic can be used to estimate the *p*-value by using the Monte Carlo sampling for testing normality with mean and variance unknown. The Lilliefors test applies to the three continuous distributions (**Normal**, **Exponential**, and **Uniform**). Note that the test does not apply if the underlying distribution is discrete (**Poisson**). The test is only defined for one-sample inference when the corresponding distribution parameters are not specified.

Obtaining a One-Sample Kolmogorov-Smirnov test

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **1-Sample K-S...**
2. Select one or more numeric test variables. Each variable produces a separate test.
3. Optionally, select a test distribution method:

   Normal
   When selected, you can specify whether distribution parameter(s) are estimated from sample data (the default setting) or from custom settings. When **Use sample data** is selected, both the existing asymptotic results and Lilliefors significance correction based on the Monte Carlo sampling are used. When **Custom** is selected, provide values for both **Mean** and **Std Dev**.

   Uniform
   When selected, you can specify whether distribution parameter(s) are estimated from sample data (the default setting) or from custom settings. When **Use sample data** is selected, the Lilliefors test is used. When **Custom** is selected, provide values for both **Min** and **Max**.

   Poisson
   When selected, specify a **Mean** parameter value.

   Exponential
   When selected, you can specify whether distribution parameter(s) are estimated from the sample mean (the default setting) or from custom settings. When **Use sample data** is selected, the Lilliefors test is used. When **Custom** is selected, provide a **Mean** parameter value.

4. Optionally, click **Simulation** to specify Monte Carlo simulation parameters, click **Exact** to specify exact test parameters, or click **Options** for descriptive statistics, quartiles, and control of the treatment of missing data.

One-Sample Kolmogorov-Smirnov Test: Simulation

When certain parameters of the distribution have to be estimated from the sample, the Kolmogorov-Smirnov test no longer applies. In these instances, the Lilliefors test statistic can be used to estimate the *p*-value by using the Monte Carlo sampling for testing normality with mean and variance unknown. The Lilliefors test applies to the three continuous distributions (**Normal**, **Exponential**, and **Uniform**). Note that the test does not apply if the underlying distribution is discrete (**Poisson**). The test is only defined for one-sample inference when the corresponding distribution parameters are not specified.

Monte Carlo Simulation Parameters

Confidence level
This optional setting resets the confidence interval level that is estimated by the Kolmogorov-Smirnov test when using the Monte Carlo simulations. The value must be between 0 and 100. The default setting is 99.

Number of samples
This optional setting resets the number of replicates that the Lilliefors test uses for the Monte Carlo sampling. The value must be a single integer between 10000 and the largest number of samples value. The default value is 10000.

Suppress the Monte Carlo results for the normal distribution
This optional setting suppresses the Monte Carlo sampling for the normal distribution results. By default, the setting is not selected (which means both the existing asymptotic results and the Lilliefors test results, that are based on the Monte Carlo sampling, are presented).

One-Sample Kolmogorov-Smirnov Test: Options

Statistics
You can choose one or both summary statistics.

Descriptive
Displays the mean, standard deviation, minimum, maximum, and number of nonmissing cases.

Quartiles
Displays values corresponding to the 25th, 50th, and 75th percentiles.

Missing Values
Controls the treatment of missing values.

Exclude cases test-by-test
When several tests are specified, each test is evaluated separately for missing values.

Exclude cases listwise
Cases with missing values for any variable are excluded from all analyses.

NPAR TESTS Command Additional Features (One-Sample Kolmogorov-Smirnov Test)

The command syntax language also allows you to specify the parameters of the test distribution (with the K-S subcommand).

See the *Command Syntax Reference* for complete syntax information.

Two-Independent-Samples Tests

The Two-Independent-Samples Tests procedure compares two groups of cases on one variable.

Example. New dental braces have been developed that are intended to be more comfortable, to look better, and to provide more rapid progress in realigning teeth. To find out whether the new braces have to be worn as long as the old braces, 10 children are randomly chosen to wear the old braces, and another 10 children are chosen to wear the new braces. From the Mann-Whitney *U* test, you might find that, on

average, children with the new braces did not have to wear the braces as long as children with the old braces.

Statistics. Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Mann-Whitney *U*, Moses extreme reactions, Kolmogorov-Smirnov *Z*, Wald-Wolfowitz runs.

Two-Independent-Samples Tests Data Considerations

Data. Use numeric variables that can be ordered.

Assumptions. Use independent, random samples. The Mann-Whitney *U* test tests equality of two distributions. In order to use it to test for differences in location between two distributions, one must assume that the distributions have the same shape.

To Obtain Two-Independent-Samples Tests

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **2 Independent Samples...**
2. Select one or more numeric variables.
3. Select a grouping variable and click **Define Groups** to split the file into two groups or samples.

Two-Independent-Samples Test Types

Test Type. Four tests are available to test whether two independent samples (groups) come from the same population.

The **Mann-Whitney U test** is the most popular of the two-independent-samples tests. It is equivalent to the Wilcoxon rank sum test and the Kruskal-Wallis test for two groups. Mann-Whitney tests that two sampled populations are equivalent in location. The observations from both groups are combined and ranked, with the average rank assigned in the case of ties. The number of ties should be small relative to the total number of observations. If the populations are identical in location, the ranks should be randomly mixed between the two samples. The test calculates the number of times that a score from group 1 precedes a score from group 2 and the number of times that a score from group 2 precedes a score from group 1. The Mann-Whitney *U* statistic is the smaller of these two numbers. The Wilcoxon rank sum *W* statistic is also displayed. *W* is the sum of the ranks for the group with the smaller mean rank, unless the groups have the same mean rank, in which case it is the rank sum from the group that is named last in the Two-Independent-Samples Define Groups dialog box.

The **Kolmogorov-Smirnov Z test** and the **Wald-Wolfowitz runs test** are more general tests that detect differences in both the locations and shapes of the distributions. The Kolmogorov-Smirnov test is based on the maximum absolute difference between the observed cumulative distribution functions for both samples. When this difference is significantly large, the two distributions are considered different. The Wald-Wolfowitz runs test combines and ranks the observations from both groups. If the two samples are from the same population, the two groups should be randomly scattered throughout the ranking.

The **Moses extreme reactions test** assumes that the experimental variable will affect some subjects in one direction and other subjects in the opposite direction. The test tests for extreme responses compared to a control group. This test focuses on the span of the control group and is a measure of how much extreme values in the experimental group influence the span when combined with the control group. The control group is defined by the group 1 value in the Two-Independent-Samples Define Groups dialog box. Observations from both groups are combined and ranked. The span of the control group is computed as the difference between the ranks of the largest and smallest values in the control group plus 1. Because chance outliers can easily distort the range of the span, 5% of the control cases are trimmed automatically from each end.

Two-Independent-Samples Tests Define Groups

To split the file into two groups or samples, enter an integer value for Group 1 and another value for Group 2. Cases with other values are excluded from the analysis.

Two-Independent-Samples Tests Options

Statistics. You can choose one or both summary statistics.

- **Descriptive.** Displays the mean, standard deviation, minimum, maximum, and the number of nonmissing cases.
- **Quartiles.** Displays values corresponding to the 25th, 50th, and 75th percentiles.

Missing Values. Controls the treatment of missing values.

- **Exclude cases test-by-test.** When several tests are specified, each test is evaluated separately for missing values.
- **Exclude cases listwise.** Cases with missing values for any variable are excluded from all analyses.

NPAR TESTS Command Additional Features (Two-Independent-Samples Tests)

The command syntax language also allows you to specify the number of cases to be trimmed for the Moses test (with the MOSES subcommand).

See the *Command Syntax Reference* for complete syntax information.

Two-Related-Samples Tests

The Two-Related-Samples Tests procedure compares the distributions of two variables.

Example. In general, do families receive the asking price when they sell their homes? By applying the Wilcoxon signed-rank test to data for 10 homes, you might learn that seven families receive less than the asking price, one family receives more than the asking price, and two families receive the asking price.

Statistics. Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Wilcoxon signed-rank, sign, McNemar. If the Exact Tests option is installed (available only on Windows operating systems), the marginal homogeneity test is also available.

Two-Related-Samples Tests Data Considerations

Data. Use numeric variables that can be ordered.

Assumptions. Although no particular distributions are assumed for the two variables, the population distribution of the paired differences is assumed to be symmetric.

To Obtain Two-Related-Samples Tests

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **2 Related Samples...**
2. Select one or more pairs of variables.

Two-Related-Samples Test Types

The tests in this section compare the distributions of two related variables. The appropriate test to use depends on the type of data.

If your data are continuous, use the sign test or the Wilcoxon signed-rank test. The **sign test** computes the differences between the two variables for all cases and classifies the differences as positive, negative, or tied. If the two variables are similarly distributed, the number of positive and negative differences will not differ significantly. The **Wilcoxon signed-rank test** considers information about both the sign of the differences and the magnitude of the differences between pairs. Because the Wilcoxon signed-rank test incorporates more information about the data, it is more powerful than the sign test.

If your data are binary, use the **McNemar test**. This test is typically used in a repeated measures situation, in which each subject's response is elicited twice, once before and once after a specified event occurs. The McNemar test determines whether the initial response rate (before the event) equals the final response rate (after the event). This test is useful for detecting changes in responses due to experimental intervention in before-and-after designs.

If your data are categorical, use the **marginal homogeneity test**. This test is an extension of the McNemar test from binary response to multinomial response. It tests for changes in response (using the chi-square distribution) and is useful for detecting response changes due to experimental intervention in before-and-after designs. The marginal homogeneity test is available only if you have installed Exact Tests.

Two-Related-Samples Tests Options

Statistics. You can choose one or both summary statistics.

- **Descriptive.** Displays the mean, standard deviation, minimum, maximum, and the number of nonmissing cases.
- **Quartiles.** Displays values corresponding to the 25th, 50th, and 75th percentiles.

Missing Values. Controls the treatment of missing values.

- **Exclude cases test-by-test.** When several tests are specified, each test is evaluated separately for missing values.
- **Exclude cases listwise.** Cases with missing values for any variable are excluded from all analyses.

NPAR TESTS Command Additional Features (Two Related Samples)

The command syntax language also allows you to test a variable with each variable on a list.

See the *Command Syntax Reference* for complete syntax information.

Tests for Several Independent Samples

The Tests for Several Independent Samples procedure compares two or more groups of cases on one variable.

Example. Do three brands of 100-watt lightbulbs differ in the average time that the bulbs will burn? From the Kruskal-Wallis one-way analysis of variance, you might learn that the three brands do differ in average lifetime.

Statistics. Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Kruskal-Wallis *H*, median.

Tests for Several Independent Samples Data Considerations

Data. Use numeric variables that can be ordered.

Assumptions. Use independent, random samples. The Kruskal-Wallis *H* test requires that the tested samples be similar in shape.

To Obtain Tests for Several Independent Samples

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **K Independent Samples...**
2. Select one or more numeric variables.
3. Select a grouping variable and click **Define Range** to specify minimum and maximum integer values for the grouping variable.

Tests for Several Independent Samples Test Types

Three tests are available to determine if several independent samples come from the same population. The Kruskal-Wallis *H* test, the median test, and the Jonckheere-Terpstra test all test whether several independent samples are from the same population.

The **Kruskal-Wallis H test**, an extension of the Mann-Whitney *U* test, is the nonparametric analog of one-way analysis of variance and detects differences in distribution location. The **median test**, which is a more general test (but not as powerful), detects distributional differences in location and shape. The Kruskal-Wallis *H* test and the median test assume that there is no *a priori* ordering of the *k* populations from which the samples are drawn.

When there *is* a natural *a priori* ordering (ascending or descending) of the *k* populations, the **Jonckheere-Terpstra test** is more powerful. For example, the *k* populations might represent *k* increasing temperatures. The hypothesis that different temperatures produce the same response distribution is tested against the alternative that as the temperature increases, the magnitude of the response increases. Here, the alternative hypothesis is ordered; therefore, Jonckheere-Terpstra is the most appropriate test to use. The Jonckheere-Terpstra test is available only if you have installed the Exact Tests add-on module.

Tests for Several Independent Samples Define Range

To define the range, enter integer values for **Minimum** and **Maximum** that correspond to the lowest and highest categories of the grouping variable. Cases with values outside of the bounds are excluded. For example, if you specify a minimum value of 1 and a maximum value of 3, only the integer values of 1 through 3 are used. The minimum value must be less than the maximum value, and both values must be specified.

Tests for Several Independent Samples Options

Statistics. You can choose one or both summary statistics.

- **Descriptive.** Displays the mean, standard deviation, minimum, maximum, and the number of nonmissing cases.
- **Quartiles.** Displays values corresponding to the 25th, 50th, and 75th percentiles.

Missing Values. Controls the treatment of missing values.

- **Exclude cases test-by-test.** When several tests are specified, each test is evaluated separately for missing values.
- **Exclude cases listwise.** Cases with missing values for any variable are excluded from all analyses.

NPAR TESTS Command Additional Features (K Independent Samples)

The command syntax language also allows you to specify a value other than the observed median for the median test (with the `MEDIAN` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Tests for Several Related Samples

The Tests for Several Related Samples procedure compares the distributions of two or more variables.

Example. Does the public associate different amounts of prestige with a doctor, a lawyer, a police officer, and a teacher? Ten people are asked to rank these four occupations in order of prestige. Friedman's test indicates that the public does associate different amounts of prestige with these four professions.

Statistics. Mean, standard deviation, minimum, maximum, number of nonmissing cases, and quartiles. Tests: Friedman, Kendall's *W*, and Cochran's *Q*.

Tests for Several Related Samples Data Considerations

Data. Use numeric variables that can be ordered.

Assumptions. Nonparametric tests do not require assumptions about the shape of the underlying distribution. Use dependent, random samples.

To Obtain Tests for Several Related Samples

1. From the menus choose:

   Analyze > **Nonparametric Tests** > **Legacy Dialogs** > **K Related Samples...**
2. Select two or more numeric test variables.

Tests for Several Related Samples Test Types

Three tests are available to compare the distributions of several related variables.

The **Friedman test** is the nonparametric equivalent of a one-sample repeated measures design or a two-way analysis of variance with one observation per cell. Friedman tests the null hypothesis that *k* related variables come from the same population. For each case, the *k* variables are ranked from 1 to *k*. The test statistic is based on these ranks.

Kendall's W is a normalization of the Friedman statistic. Kendall's *W* is interpretable as the coefficient of concordance, which is a measure of agreement among raters. Each case is a judge or rater, and each variable is an item or person being judged. For each variable, the sum of ranks is computed. Kendall's *W* ranges between 0 (no agreement) and 1 (complete agreement).

Cochran's Q is identical to the Friedman test but is applicable when all responses are binary. This test is an extension of the McNemar test to the *k*-sample situation. Cochran's *Q* tests the hypothesis that several related dichotomous variables have the same mean. The variables are measured on the same individual or on matched individuals.

Tests for Several Related Samples Statistics

You can choose statistics.

- **Descriptive.** Displays the mean, standard deviation, minimum, maximum, and the number of nonmissing cases.
- **Quartiles.** Displays values corresponding to the 25th, 50th, and 75th percentiles.

NPAR TESTS Command Additional Features (K Related Samples)

See the *Command Syntax Reference* for complete syntax information.

Multiple Response Analysis

Multiple Response Analysis

Two procedures are available for analyzing multiple dichotomy and multiple category sets. The Multiple Response Frequencies procedure displays frequency tables. The Multiple Response Crosstabs procedure displays two- and three-dimensional crosstabulations. Before using either procedure, you must define multiple response sets.

Example. This example illustrates the use of multiple response items in a market research survey. The data are fictitious and should not be interpreted as real. An airline might survey passengers flying a particular route to evaluate competing carriers. In this example, American Airlines wants to know about its passengers' use of other airlines on the Chicago-New York route and the relative importance of schedule and service in selecting an airline. The flight attendant hands each passenger a brief questionnaire upon boarding. The first question reads: Circle all airlines you have flown at least once in the last six months on this route--American, United, TWA, USAir, Other. This is a multiple response question, since the passenger can circle more than one response. However, this question cannot be coded directly because a variable can have only one value for each case. You must use several variables to map responses to each question. There are two ways to do this. One is to define a variable corresponding to each of the choices (for example, American, United, TWA, USAir, and Other). If the passenger circles United, the variable *united* is assigned a code of 1, otherwise 0. This is a **multiple dichotomy method** of mapping variables. The other way to map responses is the **multiple category method**, in which you estimate the maximum number of possible responses to the question and set up the same number of variables, with codes used to specify the airline flown. By perusing a sample of questionnaires, you might discover that no user has flown more than three different airlines on this route in the last six months. Further, you find that due to the deregulation of airlines, 10 other airlines are named in the Other category. Using the multiple response method, you would define three variables, each coded as 1 = *american*, 2 = *united*, 3 = *twa*, 4 = *usair*, 5 = *delta*, and so on. If a given passenger circles American and TWA, the first variable has a code of 1, the second has a code of 3, and the third has a missing-value code. Another passenger might have circled American and entered Delta. Thus, the first variable has a code of 1, the second has a code of 5, and the third a missing-value code. If you use the multiple dichotomy method, on the other hand, you end up with 14 separate variables. Although either method of mapping is feasible for this survey, the method you choose depends on the distribution of responses.

Multiple Response Define Sets

The Define Multiple Response Sets procedure groups elementary variables into multiple dichotomy and multiple category sets, for which you can obtain frequency tables and crosstabulations. You can define up to 20 multiple response sets. Each set must have a unique name. To remove a set, highlight it on the list of multiple response sets and click **Remove**. To change a set, highlight it on the list, modify any set definition characteristics, and click **Change**.

You can code your elementary variables as dichotomies or categories. To use dichotomous variables, select **Dichotomies** to create a multiple dichotomy set. Enter an integer value for Counted value. Each variable having at least one occurrence of the counted value becomes a category of the multiple dichotomy set. Select **Categories** to create a multiple category set having the same range of values as the component variables. Enter integer values for the minimum and maximum values of the range for categories of the multiple category set. The procedure totals each distinct integer value in the inclusive range across all component variables. Empty categories are not tabulated.

Each multiple response set must be assigned a unique name of up to seven characters. The procedure prefixes a dollar sign ($) to the name you assign. You cannot use the following reserved names: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length*, and *width*. The name of the multiple response set exists only for use in multiple response procedures. You cannot refer to multiple response set names in other procedures. Optionally, you can enter a descriptive variable label for the multiple response set. The label can be up to 40 characters long.

To Define Multiple Response Sets

1. From the menus choose:

   Analyze > **Multiple Response** > **Define Variable Sets...**
2. Select two or more variables.
3. If your variables are coded as dichotomies, indicate which value you want to have counted. If your variables are coded as categories, define the range of the categories.
4. Enter a unique name for each multiple response set.
5. Click **Add** to add the multiple response set to the list of defined sets.

Multiple Response Frequencies

The Multiple Response Frequencies procedure produces frequency tables for multiple response sets. You must first define one or more multiple response sets (see "Multiple Response Define Sets").

For multiple dichotomy sets, category names shown in the output come from variable labels defined for elementary variables in the group. If the variable labels are not defined, variable names are used as labels. For multiple category sets, category labels come from the value labels of the first variable in the group. If categories missing for the first variable are present for other variables in the group, define a value label for the missing categories.

Missing Values. Cases with missing values are excluded on a table-by-table basis. Alternatively, you can choose one or both of the following:

- **Exclude cases listwise within dichotomies.** Excludes cases with missing values for any variable from the tabulation of the multiple dichotomy set. This applies only to multiple response sets defined as dichotomy sets. By default, a case is considered missing for a multiple dichotomy set if none of its component variables contains the counted value. Cases with missing values for some (but not all variables) are included in the tabulations of the group if at least one variable contains the counted value.
- **Exclude cases listwise within categories.** Excludes cases with missing values for any variable from tabulation of the multiple category set. This applies only to multiple response sets defined as category sets. By default, a case is considered missing for a multiple category set only if none of its components has valid values within the defined range.

Example. Each variable created from a survey question is an elementary variable. To analyze a multiple response item, you must combine the variables into one of two types of multiple response sets: a multiple dichotomy set or a multiple category set. For example, if an airline survey asked which of three airlines

(American, United, TWA) you have flown in the last six months and you used dichotomous variables and defined a **multiple dichotomy set**, each of the three variables in the set would become a category of the group variable. The counts and percentages for the three airlines are displayed in one frequency table. If you discover that no respondent mentioned more than two airlines, you could create two variables, each having three codes, one for each airline. If you define a **multiple category set**, the values are tabulated by adding the same codes in the elementary variables together. The resulting set of values is the same as those for each of the elementary variables. For example, 30 responses for United are the sum of the five United responses for airline 1 and the 25 United responses for airline 2. The counts and percentages for the three airlines are displayed in one frequency table.

Statistics. Frequency tables displaying counts, percentages of responses, percentages of cases, number of valid cases, and number of missing cases.

Multiple Response Frequencies Data Considerations

Data. Use multiple response sets.

Assumptions. The counts and percentages provide a useful description for data from any distribution.

Related procedures. The Multiple Response Define Sets procedure allows you to define multiple response sets.

To Obtain Multiple Response Frequencies

1. From the menus choose:

   Analyze > **Multiple Response** > **Frequencies...**
2. Select one or more multiple response sets.

Multiple Response Crosstabs

The Multiple Response Crosstabs procedure crosstabulates defined multiple response sets, elementary variables, or a combination. You can also obtain cell percentages based on cases or responses, modify the handling of missing values, or get paired crosstabulations. You must first define one or more multiple response sets (see "To Define Multiple Response Sets").

For multiple dichotomy sets, category names shown in the output come from variable labels defined for elementary variables in the group. If the variable labels are not defined, variable names are used as labels. For multiple category sets, category labels come from the value labels of the first variable in the group. If categories missing for the first variable are present for other variables in the group, define a value label for the missing categories. The procedure displays category labels for columns on three lines, with up to eight characters per line. To avoid splitting words, you can reverse row and column items or redefine labels.

Example. Both multiple dichotomy and multiple category sets can be crosstabulated with other variables in this procedure. An airline passenger survey asks passengers for the following information: Circle all of the following airlines you have flown at least once in the last six months (American, United, TWA). Which is more important in selecting a flight--schedule or service? Select only one. After entering the data as dichotomies or multiple categories and combining them into a set, you can crosstabulate the airline choices with the question involving service or schedule.

Statistics. Crosstabulation with cell, row, column, and total counts, and cell, row, column, and total percentages. The cell percentages can be based on cases or responses.

Multiple Response Crosstabs Data Considerations

Data. Use multiple response sets or numeric categorical variables.

Assumptions. The counts and percentages provide a useful description of data from any distribution.

Related procedures. The Multiple Response Define Sets procedure allows you to define multiple response sets.

To Obtain Multiple Response Crosstabs

1. From the menus choose:

Analyze > **Multiple Response** > **Crosstabs...**

2. Select one or more numeric variables or multiple response sets for each dimension of the crosstabulation.
3. Define the range of each elementary variable.

Optionally, you can obtain a two-way crosstabulation for each category of a control variable or multiple response set. Select one or more items for the Layer(s) list.

Multiple Response Crosstabs Define Ranges

Value ranges must be defined for any elementary variable in the crosstabulation. Enter the integer minimum and maximum category values that you want to tabulate. Categories outside the range are excluded from analysis. Values within the inclusive range are assumed to be integers (non-integers are truncated).

Multiple Response Crosstabs Options

Cell Percentages. Cell counts are always displayed. You can choose to display row percentages, column percentages, and two-way table (total) percentages.

Percentages Based on. You can base cell percentages on cases (or respondents). This is not available if you select matching of variables across multiple category sets. You can also base cell percentages on responses. For multiple dichotomy sets, the number of responses is equal to the number of counted values across cases. For multiple category sets, the number of responses is the number of values in the defined range.

Missing Values. You can choose one or both of the following:

- **Exclude cases listwise within dichotomies.** Excludes cases with missing values for any variable from the tabulation of the multiple dichotomy set. This applies only to multiple response sets defined as dichotomy sets. By default, a case is considered missing for a multiple dichotomy set if none of its component variables contains the counted value. Cases with missing values for some, but not all, variables are included in the tabulations of the group if at least one variable contains the counted value.
- **Exclude cases listwise within categories.** Excludes cases with missing values for any variable from tabulation of the multiple category set. This applies only to multiple response sets defined as category sets. By default, a case is considered missing for a multiple category set only if none of its components has valid values within the defined range.

By default, when crosstabulating two multiple category sets, the procedure tabulates each variable in the first group with each variable in the second group and sums the counts for each cell; therefore, some responses can appear more than once in a table. You can choose the following option:

Match variables across response sets. Pairs the first variable in the first group with the first variable in the second group, and so on. If you select this option, the procedure bases cell percentages on responses rather than respondents. Pairing is not available for multiple dichotomy sets or elementary variables.

MULT RESPONSE Command Additional Features

The command syntax language also allows you to:

- Obtain crosstabulation tables with up to five dimensions (with the BY subcommand).
- Change output formatting options, including suppression of value labels (with the FORMAT subcommand).

See the *Command Syntax Reference* for complete syntax information.

Reporting Results

Reporting Results

Case listings and descriptive statistics are basic tools for studying and presenting data. You can obtain case listings with the Data Editor or the Summarize procedure, frequency counts and descriptive statistics with the Frequencies procedure, and subpopulation statistics with the Means procedure. Each of these uses a format designed to make information clear. If you want to display the information in a different format, Report Summaries in Rows and Report Summaries in Columns give you the control you need over data presentation.

Report Summaries in Rows

Report Summaries in Rows produces reports in which different summary statistics are laid out in rows. Case listings are also available, with or without summary statistics.

Example. A company with a chain of retail stores keeps records of employee information, including salary, job tenure, and the store and division in which each employee works. You could generate a report that provides individual employee information (listing) broken down by store and division (break variables), with summary statistics (for example, mean salary) for each store, division, and division within each store.

Data Columns. Lists the report variables for which you want case listings or summary statistics and controls the display format of data columns.

Break Columns. Lists optional break variables that divide the report into groups and controls the summary statistics and display formats of break columns. For multiple break variables, there will be a separate group for each category of each break variable within categories of the preceding break variable in the list. Break variables should be discrete categorical variables that divide cases into a limited number of meaningful categories. Individual values of each break variable appear, sorted, in a separate column to the left of all data columns.

Report. Controls overall report characteristics, including overall summary statistics, display of missing values, page numbering, and titles.

Display cases. Displays the actual values (or value labels) of the data-column variables for every case. This produces a listing report, which can be much longer than a summary report.

Preview. Displays only the first page of the report. This option is useful for previewing the format of your report without processing the whole report.

Data are already sorted. For reports with break variables, the data file must be sorted by break variable values before generating the report. If your data file is already sorted by values of the break variables, you can save processing time by selecting this option. This option is particularly useful after running a preview report.

To Obtain a Summary Report: Summaries in Rows

1. From the menus choose:

   Analyze > **Reports** > **Report Summaries in Rows...**
2. Select one or more variables for Data Columns. One column in the report is generated for each variable selected.
3. For reports sorted and displayed by subgroups, select one or more variables for Break Columns.
4. For reports with summary statistics for subgroups defined by break variables, select the break variable in the Break Column Variables list and click **Summary** in the Break Columns group to specify the summary measure(s).
5. For reports with overall summary statistics, click **Summary** to specify the summary measure(s).

Report Data Column/Break Format

The Format dialog boxes control column titles, column width, text alignment, and the display of data values or value labels. Data Column Format controls the format of data columns on the right side of the report page. Break Format controls the format of break columns on the left side.

Column Title. For the selected variable, controls the column title. Long titles are automatically wrapped within the column. Use the Enter key to manually insert line breaks where you want titles to wrap.

Value Position within Column. For the selected variable, controls the alignment of data values or value labels within the column. Alignment of values or labels does not affect alignment of column headings. You can either indent the column contents by a specified number of characters or center the contents.

Column Content. For the selected variable, controls the display of either data values or defined value labels. Data values are always displayed for any values that do not have defined value labels. (Not available for data columns in column summary reports.)

Report Summary Lines for/Final Summary Lines

The two Summary Lines dialog boxes control the display of summary statistics for break groups and for the entire report. Summary Lines controls subgroup statistics for each category defined by the break variable(s). Final Summary Lines controls overall statistics, displayed at the end of the report.

Available summary statistics are sum, mean, minimum, maximum, number of cases, percentage of cases above or below a specified value, percentage of cases within a specified range of values, standard deviation, kurtosis, variance, and skewness.

Report Break Options

Break Options controls spacing and pagination of break category information.

Page Control. Controls spacing and pagination for categories of the selected break variable. You can specify a number of blank lines between break categories or start each break category on a new page.

Blank Lines before Summaries. Controls the number of blank lines between break category labels or data and summary statistics. This is particularly useful for combined reports that include both individual case listings and summary statistics for break categories; in these reports, you can insert space between the case listings and the summary statistics.

Report Options

Report Options controls the treatment and display of missing values and report page numbering.

Exclude cases with missing values listwise. Eliminates (from the report) any case with missing values for any of the report variables.

Missing Values Appear as. Allows you to specify the symbol that represents missing values in the data file. The symbol can be only one character and is used to represent both *system-missing* and *user-missing* values.

Number Pages from. Allows you to specify a page number for the first page of the report.

Report Layout

Report Layout controls the width and length of each report page, placement of the report on the page, and the insertion of blank lines and labels.

Page Layout. Controls the page margins expressed in lines (top and bottom) and characters (left and right) and reports alignment within the margins.

Page Titles and Footers. Controls the number of lines that separate page titles and footers from the body of the report.

Break Columns. Controls the display of break columns. If multiple break variables are specified, they can be in separate columns or in the first column. Placing all break variables in the first column produces a narrower report.

Column Titles. Controls the display of column titles, including title underlining, space between titles and the body of the report, and vertical alignment of column titles.

Data Column Rows and Break Labels. Controls the placement of data column information (data values and/or summary statistics) in relation to the break labels at the start of each break category. The first row of data column information can start either on the same line as the break category label or on a specified number of lines after the break category label. (Not available for column summary reports.)

Report Titles

Report Titles controls the content and placement of report titles and footers. You can specify up to 10 lines of page titles and up to 10 lines of page footers, with left-justified, centered, and right-justified components on each line.

If you insert variables into titles or footers, the current value label or value of the variable is displayed in the title or footer. In titles, the value label corresponding to the value of the variable at the beginning of the page is displayed. In footers, the value label corresponding to the value of the variable at the end of the page is displayed. If there is no value label, the actual value is displayed.

Special Variables. The special variables *DATE* and *PAGE* allow you to insert the current date or the page number into any line of a report header or footer. If your data file contains variables named *DATE* or *PAGE*, you cannot use these variables in report titles or footers.

Report Summaries in Columns

Report Summaries in Columns produces summary reports in which different summary statistics appear in separate columns.

Example. A company with a chain of retail stores keeps records of employee information, including salary, job tenure, and the division in which each employee works. You could generate a report that provides summary salary statistics (for example, mean, minimum, and maximum) for each division.

Data Columns. Lists the report variables for which you want summary statistics and controls the display format and summary statistics displayed for each variable.

Break Columns. Lists optional break variables that divide the report into groups and controls the display formats of break columns. For multiple break variables, there will be a separate group for each category of each break variable within categories of the preceding break variable in the list. Break variables should be discrete categorical variables that divide cases into a limited number of meaningful categories.

Report. Controls overall report characteristics, including display of missing values, page numbering, and titles.

Preview. Displays only the first page of the report. This option is useful for previewing the format of your report without processing the whole report.

Data are already sorted. For reports with break variables, the data file must be sorted by break variable values before generating the report. If your data file is already sorted by values of the break variables, you can save processing time by selecting this option. This option is particularly useful after running a preview report.

To Obtain a Summary Report: Summaries in Columns

1. From the menus choose:

   Analyze > **Reports** > **Report Summaries in Columns...**

2. Select one or more variables for Data Columns. One column in the report is generated for each variable selected.

3. To change the summary measure for a variable, select the variable in the Data Column Variables list and click **Summary**.
4. To obtain more than one summary measure for a variable, select the variable in the source list and move it into the Data Column Variables list multiple times, one for each summary measure you want.
5. To display a column containing the sum, mean, ratio, or other function of existing columns, click **Insert Total**. This places a variable called *total* into the Data Columns list.
6. For reports sorted and displayed by subgroups, select one or more variables for Break Columns.

Data Columns Summary Function

Summary Lines controls the summary statistic displayed for the selected data column variable.

Available summary statistics are sum, mean, minimum, maximum, number of cases, percentage of cases above or below a specified value, percentage of cases within a specified range of values, standard deviation, variance, kurtosis, and skewness.

Data Columns Summary for Total Column

Summary Column controls the total summary statistics that summarize two or more data columns.

Available total summary statistics are sum of columns, mean of columns, minimum, maximum, difference between values in two columns, quotient of values in one column divided by values in another column, and product of columns values multiplied together.

Sum of columns. The *total* column is the sum of the columns in the Summary Column list.

Mean of columns. The *total* column is the average of the columns in the Summary Column list.

Minimum of columns. The *total* column is the minimum of the columns in the Summary Column list.

Maximum of columns. The *total* column is the maximum of the columns in the Summary Column list.

1st column – 2nd column. The *total* column is the difference of the columns in the Summary Column list. The Summary Column list must contain exactly two columns.

1st column / 2nd column. The *total* column is the quotient of the columns in the Summary Column list. The Summary Column list must contain exactly two columns.

% 1st column / 2nd column. The *total* column is the first column's percentage of the second column in the Summary Column list. The Summary Column list must contain exactly two columns.

Product of columns. The *total* column is the product of the columns in the Summary Column list.

Report Column Format

Data and break column formatting options for Report Summaries in Columns are the same as those described for Report Summaries in Rows.

Report Summaries in Columns Break Options

Break Options controls subtotal display, spacing, and pagination for break categories.

Subtotal. Controls the display subtotals for break categories.

Page Control. Controls spacing and pagination for categories of the selected break variable. You can specify a number of blank lines between break categories or start each break category on a new page.

Blank Lines before Subtotal. Controls the number of blank lines between break category data and subtotals.

Report Summaries in Columns Options

Options controls the display of grand totals, the display of missing values, and pagination in column summary reports.

Grand Total. Displays and labels a grand total for each column; displayed at the bottom of the column.

Missing values. You can exclude missing values from the report or select a single character to indicate missing values in the report.

Report Layout for Summaries in Columns

Report layout options for Report Summaries in Columns are the same as those described for Report Summaries in Rows.

REPORT Command Additional Features

The command syntax language also allows you to:

- Display different summary functions in the columns of a single summary line.
- Insert summary lines into data columns for variables other than the data column variable or for various combinations (composite functions) of summary functions.
- Use Median, Mode, Frequency, and Percent as summary functions.
- Control more precisely the display format of summary statistics.
- Insert blank lines at various points in reports.
- Insert blank lines after every *n*th case in listing reports.

Because of the complexity of the REPORT syntax, you may find it useful, when building a new report with syntax, to approximate the report generated from the dialog boxes, copy and paste the corresponding syntax, and refine that syntax to yield the exact report that you want.

See the *Command Syntax Reference* for complete syntax information.

Reliability Analysis

Reliability analysis allows you to study the properties of measurement scales and the items that compose the scales. The Reliability Analysis procedure calculates a number of commonly used measures of scale reliability and also provides information about the relationships between individual items in the scale. Intra-class correlation coefficients can be used to compute inter-rater reliability estimates.

Reliability analysis also provides Fleiss' Multiple Rater Kappa statistics that assess the interrater agreement to determine the reliability among the various raters. A higher agreement provides more confidence in the ratings reflecting the true circumstance. The Fleiss' Multiple Rater Kappa options are available in the “Reliability Analysis: Statistics” on page 213 dialog.

Example

Does my questionnaire measure customer satisfaction in a useful way? Using reliability analysis, you can determine the extent to which the items in your questionnaire are related to each other, you can get an overall index of the repeatability or internal consistency of the scale as a whole, and you can identify problem items that should be excluded from the scale.

Statistics

Descriptives for each variable and for the scale, summary statistics across items, inter-item correlations and covariances, reliability estimates, ANOVA table, intraclass correlation coefficients, Hotelling's T^2, Tukey's test of additivity, and Fleiss' Multiple Rater Kappa.

Models

The following models of reliability are available:

Alpha (Cronbach)

This model is a measure of internal consistency based on the average inter-item correlation.

Omega (McDonald's)

This model assumes that the model is uni-dimensional including a single factor with no local item dependence in the form of error covariances. The model implies that the covariance of the two different items is the product of their loadings.

Split-half
This model splits the scale into two parts and examines the correlation between the parts.

Guttman
This model computes Guttman's lower bounds for true reliability.

Parallel
This model assumes that all items have equal variances and equal error variances across replications.

Strict parallel
This model makes the assumptions of the Parallel model and also assumes equal means across items.

Reliability Analysis data considerations

Data
Data can be dichotomous, ordinal, or interval, but the data should be coded numerically.

Assumptions
Observations should be independent, and errors should be uncorrelated between items. Each pair of items should have a bivariate normal distribution. Scales should be additive, so that each item is linearly related to the total score. The following assumptions apply for Fleiss' Multiple Rater Kappa statistics:

- At least two item variables must be selected to run any reliability statistic.
- When at least two ratings variables are selected, the Fleiss' Multiple Rater Kappa syntax is pasted.
- There is no connection between raters.
- The number of raters is a constant.
- Each subject is rated by the same group containing only a single rater.
- No weights can be assigned to the various disagreements.

Related procedures
If you want to explore the dimensionality of your scale items (to see whether more than one construct is needed to account for the pattern of item scores), use factor analysis or multidimensional scaling. To identify homogeneous groups of variables, use hierarchical cluster analysis to cluster variables.

To obtain a Reliability Analysis

1. From the menus choose:

   Analyze > **Scale** > **Reliability Analysis...**
2. Select two or more variables as potential components of an additive scale.
3. Choose a model from the **Model** drop-down list.
4. Optionally, click **Statistics** to select various statistics that describe your scale items or interrater agreement.

Reliability Analysis: Statistics

You can select various statistics that describe your scale, items and the interrater agreement to determine the reliability among the various raters. Statistics that are reported by default include the number of cases, the number of items, and reliability estimates as follows:

Alpha models
Coefficient alpha; for dichotomous data, this is equivalent to the Kuder-Richardson 20 (KR20) coefficient.

Omega models
Estimation of McDonald’s omega to evaluate reliability.

Split-half models
Correlation between forms, Guttman split-half reliability, Spearman-Brown reliability (equal and unequal length), and coefficient alpha for each half.

Guttman models
Reliability coefficients lambda 1 through lambda 6.

Parallel and Strict parallel models
Test for goodness of fit of model; estimates of error variance, common variance, and true variance; estimated common inter-item correlation; estimated reliability; and unbiased estimate of reliability.

Descriptives for
Produces descriptive statistics for scales or items across cases.

Item
Produces descriptive statistics for items across cases.

Scale
Produces descriptive statistics for scales.

Scale if item deleted
Displays summary statistics comparing each item to the scale that is composed of the other items. Statistics include scale mean and variance if the item were to be deleted from the scale, correlation between the item and the scale that is composed of other items, and Cronbach's alpha if the item were to be deleted from the scale.

Summaries
Provides descriptive statistics of item distributions across all items in the scale.

Means
Summary statistics for item means. The smallest, largest, and average item means, the range and variance of item means, and the ratio of the largest to the smallest item means are displayed.

Variances
Summary statistics for item variances. The smallest, largest, and average item variances, the range and variance of item variances, and the ratio of the largest to the smallest item variances are displayed.

Correlations
Summary statistics for inter-item correlations. The smallest, largest, and average inter-item correlations, the range and variance of inter-item correlations, and the ratio of the largest to the smallest inter-item correlations are displayed.

Covariances
Summary statistics for inter-item covariances. The smallest, largest, and average inter-item covariances, the range and variance of inter-item covariances, and the ratio of the largest to the smallest inter-item covariances are displayed.

Inter-Item
Produces matrices of correlations or covariances between items.

ANOVA Table
Produces tests of equal means.

F test
Displays a repeated measures analysis-of-variance table.

Friedman chi-square
Displays Friedman's chi-square and Kendall's coefficient of concordance. This option is appropriate for data that are in the form of ranks. The chi-square test replaces the usual F test in the ANOVA table.

Cochran chi-square
Displays Cochran's Q. This option is appropriate for data that are dichotomous. The Q statistic replaces the usual F statistic in the ANOVA table.

Interrater Agreement: Fleiss' Kappa

Assesses the interrater agreement to determine the reliability among the various raters. A higher agreement provides more confidence in the ratings reflecting the true circumstance. The generalized unweighted kappa statistic measures the agreement among any constant number of raters while assuming:

- At least two item variables must be specified to run any reliability statistic.
- At least two ratings variables must be specified.
- The variables selected as items can also be selected as ratings.
- There is no connection between raters.
- The number of raters is a constant.
- Each subject is rated by the same group containing only a single rater.
- No weights can be assigned to the various disagreements.

Display agreement on individual categories

Specifies whether or not to output the agreement on individual categories. By default, the output suppresses the estimation on any individual categories. When enabled, multiple tables display in the output.

Ignore string cases

Controls whether or not the string variables are case sensitive. By default, string rating values are case sensitive.

String category labels are displayed in uppercase

Controls whether the category labels in the output tables are displayed in uppercase or lowercase. The setting is enabled by default, which displays the string category labels in uppercase.

Asymptotic significance level (%)

Specifies the significance level for the asymptotic confidence intervals. 95 is the default setting.

Missing

Exclude both user-missing and system missing values

Controls the exclusion of user-missing and system-missing values. By default, user-missing and system-missing values are excluded.

User-missing values are treated as valid

When enabled, treats user-missing and system-missing values as valid data. The setting is disabled by default.

Hotelling's T-square

Produces a multivariate test of the null hypothesis that all items on the scale have the same mean.

Tukey's test of additivity

Produces a test of the assumption that there is no multiplicative interaction among the items.

Intraclass correlation coefficient

Produces measures of consistency or agreement of values within cases.

Model

Select the model for calculating the intraclass correlation coefficient. Available models are Two-Way Mixed, Two-Way Random, and One-Way Random. Select **Two-Way Mixed** when people effects are random and the item effects are fixed, select **Two-Way Random** when people effects and the item effects are random, or select **One-Way Random** when people effects are random.

Type

Select the type of index. Available types are Consistency and Absolute Agreement.

Confidence interval (%)

Specify the level for the confidence interval. The default is 95%.

Test value

Specify the hypothesized value of the coefficient for the hypothesis test. This value is the value to which the observed value is compared. The default value is 0.

RELIABILITY Command Additional Features

The command syntax language also allows you to:

- Read and analyze a correlation matrix.
- Write a correlation matrix for later analysis.
- Specify splits other than equal halves for the split-half method.

See the *Command Syntax Reference* for complete syntax information.

Weighted Kappa

Cohen's weighted kappa is broadly used in cross-classification as a measure of agreement between observed raters. It is an appropriate index of agreement when ratings are nominal scales with no order structure. The development of Cohen's weighted kappa was motivated by the factor that some assignments in a contingency table might be of greater gravity than the others. The statistic relies on the predefined cell weights reflecting either agreement or disagreement.

The Weighted Kappa procedure provides options for estimating Cohen's weighted kappa, an important generalization of the kappa statistic that measures the agreement of two ordinal subjects with identical categories.

Note: The Weighted Kappa procedure supersedes the functionality previously provided by the `STATS WEIGHTED KAPPA.spe` extension.

Example

There are situations where the differences between raters should not be treated as equally important. An example would be in the healthcare industry where multiple people collect research or clinical data. In such cases the reliability of the data can come into question given the variability among those collecting data.

Statistics

Cohen's weighted kappa, linear scale, quadratic scale, asymptotic confidence interval.

Weighted Kappa data considerations

Data

A two-way table that is based on an active data set is required in order to estimate the Cohen's weighted kappa statistic.

Rating variables must be of the same type (all string or all numeric).

The estimation of Cohen's weighted kappa makes sense only when the categories of the two rating variables, represented by the row and column in the table, are appropriately ordered (for a pair of numeric variables, numerical order is applied; for a pair of string variables, alphabetical order is applied).

Assumptions

When mixed variable pairs are selected, Cohen's weighted kappa is not estimated.

The rating variables are assumed to share the same set of categories.

To obtain a Weighted Kappa analysis

1. From the menus choose:

   Analyze > **Scale** > **Weighted Kappa...**

2. Select two or more string or numeric variables to specify as **Pairwise raters**.

   Note: You must select either all string variables or all numeric variables.

3. Optionally, enable the **Specify raters for rows and columns** setting to control the display of pairwise raters or rows/column raters.

- When enabled, pairwise raters are suppressed and row/column raters display. The user interface updates to provide **Row rater(s)** and **Column rater(s)** fields (effectively replacing the **Pairwise raters** field.
- When disabled, row/column raters are suppressed and pairwise raters display (the default setting)

When **Specify raters for rows and columns** is enabled, specify at least one variable for both **Row rater(s)** and **Column rater(s)**.

Note: If both **Row rater(s)** and **Column rater(s)** contain only one variable, the selected variables cannot be the same for both.

4. Optionally, click **Criteria** to specify the weighting scale and missing values settings, or **Print** to specify the display format and crosstabulation settings.

Weighted Kappa: Criteria

The Criteria dialog provides options for specifying the estimation of the Cohen's weighted kappa statistics.

Weighting Scale
Provides options for specifying either a linear or quadratic weighting for agreement. The use of linear (also known as Cicchetti-Allison) weighting is the default setting.

Missing Values
Provides options for removing cases with missing values on a pairwise basis and treating user-missing values as valid.

String rating variables are case sensitive
When selected, string variables are treated as case sensitive.

Asymptotic confidence interval (%)
This optional setting specifies the confidence level for the estimation of the asymptotic confidence intervals. Must be a single double value between 0 and 100 (95 is the default setting).

Weighted Kappa: Print

The Print dialog provides options for controlling the crosstabulation tables.

Display and Format
Provides options for controlling the crosstabulation table display and format.

Rating categories are displayed in ascending order
When selected, rating categories in the crosstabulation tables display in ascending order. When the setting is not selected, rating categories in the crosstabulation tables display in descending order. The setting is enabled by default.

String category labels are displayed in uppercase
When selected, crosstabulation tables display as uppercase. When the setting is not selected, crosstabulation tables display as lowercase. The setting is enabled by default.

Crosstabulation
Provides options for specifying the rating variables that are used in crosstabulation. By default, crosstabulation settings are not enabled, which suppresses the crosstabulation of any rating variables.

Display the crosstabulation of rating variables
When selected, this setting the enables the crosstabulation of all or user-specified rating variables.

Include all rating variables
When selected, crosstabulation tables are printed for all defined rating variable pairs.

Include user-specified rating variables
When selected, use the **Available variables**, **Row rater(s)**, and **Column rater(s)** fields to select which rating variables are included in the crosstabulation tables.

Multidimensional Scaling

Multidimensional scaling attempts to find the structure in a set of distance measures between objects or cases. This task is accomplished by assigning observations to specific locations in a conceptual space (usually two- or three-dimensional) such that the distances between points in the space match the given dissimilarities as closely as possible. In many cases, the dimensions of this conceptual space can be interpreted and used to further understand your data.

If you have objectively measured variables, you can use multidimensional scaling as a data reduction technique (the Multidimensional Scaling procedure will compute distances from multivariate data for you, if necessary). Multidimensional scaling can also be applied to subjective ratings of dissimilarity between objects or concepts. Additionally, the Multidimensional Scaling procedure can handle dissimilarity data from multiple sources, as you might have with multiple raters or questionnaire respondents.

Example. How do people perceive relationships between different cars? If you have data from respondents indicating similarity ratings between different makes and models of cars, multidimensional scaling can be used to identify dimensions that describe consumers' perceptions. You might find, for example, that the price and size of a vehicle define a two-dimensional space, which accounts for the similarities that are reported by your respondents.

Statistics. For each model: data matrix, optimally scaled data matrix, S-stress (Young's), stress (Kruskal's), RSQ, stimulus coordinates, average stress and RSQ for each stimulus (RMDS models). For individual difference (INDSCAL) models: subject weights and weirdness index for each subject. For each matrix in replicated multidimensional scaling models: stress and RSQ for each stimulus. Plots: stimulus coordinates (two- or three-dimensional), scatterplot of disparities versus distances.

Multidimensional Scaling Data Considerations

Data. If your data are dissimilarity data, all dissimilarities should be quantitative and should be measured in the same metric. If your data are multivariate data, variables can be quantitative, binary, or count data. Scaling of variables is an important issue--differences in scaling may affect your solution. If your variables have large differences in scaling (for example, one variable is measured in dollars and the other variable is measured in years), consider standardizing them (this process can be done automatically by the Multidimensional Scaling procedure).

Assumptions. The Multidimensional Scaling procedure is relatively free of distributional assumptions. Be sure to select the appropriate measurement level (ordinal, interval, or ratio) in the Multidimensional Scaling Options dialog box so that the results are computed correctly.

Related procedures. If your goal is data reduction, an alternative method to consider is factor analysis, particularly if your variables are quantitative. If you want to identify groups of similar cases, consider supplementing your multidimensional scaling analysis with a hierarchical or *k*-means cluster analysis.

To Obtain a Multidimensional Scaling Analysis

1. From the menus choose:

   Analyze > **Scale** > **Multidimensional Scaling...**
2. Select at least four numeric variables for analysis.
3. In the Distances group, select either **Data are distances** or **Create distances from data**.
4. If you select **Create distances from data**, you can also select a grouping variable for individual matrices. The grouping variable can be numeric or string.

Optionally, you can also:

- Specify the shape of the distance matrix when data are distances.
- Specify the distance measure to use when creating distances from data.

Multidimensional Scaling Shape of Data

If your active dataset represents distances among a set of objects or represents distances between two sets of objects, specify the shape of your data matrix in order to get the correct results.

Note: You cannot select **Square symmetric** if the Model dialog box specifies row conditionality.

Multidimensional Scaling Create Measure

Multidimensional scaling uses dissimilarity data to create a scaling solution. If your data are multivariate data (values of measured variables), you must create dissimilarity data in order to compute a multidimensional scaling solution. You can specify the details of creating dissimilarity measures from your data.

Measure. Allows you to specify the dissimilarity measure for your analysis. Select one alternative from the Measure group corresponding to your type of data, and then choose one of the measures from the drop-down list corresponding to that type of measure. Available alternatives are:

- **Interval**. Euclidean distance, Squared Euclidean distance, Chebychev, Block, Minkowski, or Customized.
- **Counts**. Chi-square measure or Phi-square measure.
- **Binary**. Euclidean distance, Squared Euclidean distance, Size difference, Pattern difference, Variance, or Lance and Williams.

Create Distance Matrix. Allows you to choose the unit of analysis. Alternatives are Between variables or Between cases.

Transform Values. In certain cases, such as when variables are measured on very different scales, you may want to standardize values before computing proximities (not applicable to binary data). Choose a standardization method from the Standardize drop-down list. If no standardization is required, choose **None**.

Multidimensional Scaling Model

Correct estimation of a multidimensional scaling model depends on aspects of the data and the model itself.

Level of Measurement. Allows you to specify the level of your data. Alternatives are Ordinal, Interval, or Ratio. If your variables are ordinal, selecting **Untie tied observations** requests that the variables be treated as continuous variables, so that ties (equal values for different cases) are resolved optimally.

Conditionality. Allows you to specify which comparisons are meaningful. Alternatives are Matrix, Row, or Unconditional.

Dimensions. Allows you to specify the dimensionality of the scaling solution(s). One solution is calculated for each number in the range. Specify integers between 1 and 6; a minimum of 1 is allowed only if you select **Euclidean distance** as the scaling model. For a single solution, specify the same number for minimum and maximum.

Scaling Model. Allows you to specify the assumptions by which the scaling is performed. Available alternatives are Euclidean distance or Individual differences Euclidean distance (also known as INDSCAL). For the Individual differences Euclidean distance model, you can select **Allow negative subject weights**, if appropriate for your data.

Multidimensional Scaling Options

You can specify options for your multidimensional scaling analysis.

Display. Allows you to select various types of output. Available options are Group plots, Individual subject plots, Data matrix, and Model and options summary.

Criteria. Allows you to determine when iteration should stop. To change the defaults, enter values for **S-stress convergence**, **Minimum s-stress value**, and **Maximum iterations**.

Treat distances less than n as missing. Distances that are less than this value are excluded from the analysis.

ALSCAL Command Additional Features

The command syntax language also allows you to:

- Use three additional model types, known as ASCAL, AINDS, and GEMSCAL in the literature about multidimensional scaling.
- Carry out polynomial transformations on interval and ratio data.
- Analyze similarities (rather than distances) with ordinal data.
- Analyze nominal data.
- Save various coordinate and weight matrices into files and read them back in for analysis.
- Constrain multidimensional unfolding.

See the *Command Syntax Reference* for complete syntax information.

Ratio Statistics

The Ratio Statistics procedure provides a comprehensive list of summary statistics for describing the ratio between two scale variables.

You can sort the output by values of a grouping variable in ascending or descending order. The ratio statistics report can be suppressed in the output, and the results can be saved to an external file.

Example
Is there good uniformity in the ratio between the appraisal price and sale price of homes in each of five counties? From the output, you might learn that the distribution of ratios varies considerably from county to county.

Statistics
Median, mean, weighted mean, confidence intervals, coefficient of dispersion (COD), coefficient of variation (COV), median-centered coefficient of variation, mean-centered coefficient of variation, price-related bias (PRB), price-related differential (PRD), standard deviation, average absolute deviation (AAD), range, minimum and maximum values, and the concentration index computed for a user-specified range or percentage within the median ratio.

Data considerations

Data
Use numeric codes or strings to code grouping variables (nominal or ordinal level measurements).

Assumptions
The variables that define the numerator and denominator of the ratio should be scale variables that take positive values.

Obtaining Ratio Statistics

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Ratio...**
2. Select a numerator variable.
3. Select a denominator variable.

Optionally:

- Select a grouping variable and specify the ordering of the groups in the results.
- Choose whether to display the results in the Viewer.
- Choose to add an **N** keyword to the **PRINT** subcommand. The setting adds the sample sizes to the statistics table in the output.
- Choose whether to save the results to an external file for later use, and specify the name of the file to which the results are saved.

Ratio Statistics: Statistics

Central Tendency

Measures of central tendency are statistics that describe the distribution of ratios.

Median

The value such that the number of ratios that are less than this value and the number of ratios that are greater than this value are the same.

Mean

The result of summing the ratios and dividing the result by the total number of ratios.

Weighted Mean

The result of dividing the mean of the numerator by the mean of the denominator. Weighted mean is also the mean of the ratios weighted by the denominator.

Confidence intervals for PRB and measures of central tendency

Displays confidence intervals for PRB, the mean, the median, and the weighted mean (if requested). Specify a value that is greater than or equal to 0 and less than 100 as the confidence level. The default setting is 95%.

Dispersion

These statistics measure the amount of variation, or spread, in the observed values.

AAD

The average absolute deviation is the result of summing the absolute deviations of the ratios about the median and dividing the result by the total number of ratios.

COD

The coefficient of dispersion is the result of expressing the average absolute deviation as a percentage of the median.

COV

The coefficient of variation. The median-centered coefficient of variation is the result of expressing the root mean squares of deviation from the median as a percentage of the median. The mean-centered coefficient of variation is the result of expressing the standard deviation as a percentage of the mean.

PRB

The price-related bias. An index of whether assessment to price ratios are systematically higher or lower for higher-priced properties. Regresses percentage differences in assessment ratios from the median ratio on the base 2 logarithms of value proxy measures calculated as the averages of sales prices and the ratios of assessed values to the median ratio. Gives the percentage change in assessment ratios for a 100 percent change in value.

PRD

The price-related differential, also known as the index of regressivity, is the result of dividing the mean by the weighted mean.

Standard deviation

The standard deviation is the result of summing the squared deviations of the ratios about the mean, dividing the result by the total number of ratios minus one, and taking the positive square root.

Range

The range is the result of subtracting the minimum ratio from the maximum ratio.

Minimum

The minimum is the smallest ratio.

Maximum

The maximum is the largest ratio.

Concentration Index

The coefficient of concentration measures the percentage of ratios that fall within an interval. It can be computed in two different ways:

Between Proportions
Here the interval is defined explicitly by specifying the low and high values of the interval. Enter values for the low proportion and high proportion, and click **Add** to obtain an interval.

Within Percentage of Median
Here the interval is defined implicitly by specifying the percentage of the median. Enter a value between 0 and 100, and click **Add**. The lower end of the interval is equal to (1 – 0.01 × value) × median, and the upper end is equal to (1 + 0.01 × value) × median.

P-P Plot

The P-P plots procedure produces probability plots of one or more sequence or time series variables. The variables can be standardized, differenced, and transformed before plotting.

Available test distributions include beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, normal, pareto, Student's t, Weibull, and uniform. Depending on the distribution selected, you can specify degrees of freedom and other parameters.

- You obtain probability plots for transformed values. Transformation options include natural log, standardize values, difference, and seasonally difference.
- You can specify the method for calculating expected distributions and for resolving "ties" (multiple observations with the same value).

Test distribution
Specify a distribution type for your data. The drop-down list provides the following options:

Beta
Beta distribution. The shape1 and shape2 parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation. All observations must be in the range 0 - 1, inclusive.

Chi-square
Chi-square distribution. You must specify the degrees of freedom (*df*). Negative observations are not allowed.

Exponential
Exponential distribution. The scale parameter *a* must be positive. If the parameter is not specified, `DISTRIBUTION` estimates it from the sample mean. Negative observations are not allowed.

Gamma
Gamma distribution. The shape and scale parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation. Negative observations are not allowed.

Half-normal
Half-normal distribution. Data are assumed to be location-free or centralized. (Location parameter=0.) You can specify the scale parameter *a* or let `DISTRIBUTION` estimate it by using the maximum likelihood method.

Laplace
Laplace or double exponential distribution. `LAPLACE` takes a location and a scale parameter (*a* and *b).* The scale parameter *(b*) must be positive. If the parameters are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation.

Logistic
Logistic distribution. `LOGISTIC` takes a location and a scale parameter (*a* and *b).* The scale parameter *(b*) must be positive. If the parameters are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation.

Lognormal
Lognormal distribution. The scale and shape parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the mean and standard deviation of the natural logarithm of the sample data. Negative observations are not allowed.

Normal
Normal distribution. The location parameter *a* can be any numeric value, while the scale parameter *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation.

Pareto
Pareto distribution. The threshold and shape parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` assumes *a* equals the minimum observation and estimates *b* by the maximum likelihood method. Negative observations are not allowed.

Student t
Student's t *distribution.* You must specify the degrees of freedom (*df).*

Uniform
Uniform distribution. `UNIFORM` takes a minimum and a maximum parameter *(a* and *b).* Parameter *a* must be equal to or greater than *b.* If the parameters are not specified, `DISTRIBUTION` assumes them from the sample data.

Weibull
Weibull distribution. The scale and shape parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them using the least square method. Negative observations are not allowed.

Distribution parameters
Provides distribution strategy and parameter options.

Estimate from data
When selected, this setting estimates the distribution parameters based on the data and selected distribution type.

Specify
When selected, you can specify the distribution parameters for the selected distribution type.

Note: The available parameters will vary based on the selected distribution type.

Transform
The provided options set the transform and periodicity settings.

Natural log transform
Transforms the data by using the natural logarithm (base *e*) to remove varying amplitude.

Standardize values
Transforms the sequence or time series variables into a sample with a mean of 0 and a standard deviation of 1.

Difference
Specifies the degree of differencing that is used before plotting to convert a non-stationary variable into a stationary variable with a constant mean and variance. Enter an appropriate value in the field.

Seasonally difference
If the variable exhibits a seasonal or periodic pattern, you can use this setting to seasonally difference the variable before plotting.

Note: This setting is enabled only when a sequence or time series variable, with a defined periodicity, is selected as one of the quantitative variables.

Proportion estimation formula
The provided options set the formula that is used to estimate proportions.

Blom's
Creates new ranking variable based on proportion estimates that uses the formula `(r-3/8) / (w+1/4)`, where `w` is the sum of the case weights and `r` is the rank.

Rankit
Uses the formula `(r-1/2) / w`, where `w` is the number of observations and `r` is the rank, ranging from 1 to `w`.

Tukey's
Uses the formula `(r-1/3) / (w+1/3)`, where `r` is the rank and `w` is the sum of the case weights.

Van der Waerden's
Van der Waerden's transformation, defined by the formula `r/(w+1)`, where `w` is the sum of the case weights and `r` is the rank, ranging from 1 to `w`.

Rank assigned to ties
The provided options control the method for determining how to handle tie values. The following table shows how the different methods assign ranks to tied values.

Table 3. Ranking methods and results

Value	Mean	Low	High	Break ties arbitrarily
10	1	1	1	1
15	3	2	4	2
15	3	2	4	2
15	3	2	4	2
16	5	5	5	3
20	6	6	6	4

Obtaining P-P probability plots

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **P-P Plots...**
2. Select one or more numeric variables and move them onto the Variables field.
3. Select a test distribution.

Optionally, you can select transformation options to obtain probability plots for transformed values and specify a method for calculating expected distributions.

Q-Q Plot

The Q-Q plots procedure produces probability plots for transformed values. Available test distributions include beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, normal, pareto, Student's t, Weibull, and uniform. Depending on the distribution selected, you can specify degrees of freedom and other parameters.

- You obtain probability plots for transformed values. Transformation options include natural log, standardize values, difference, and seasonally difference.
- You can specify the method for calculating expected distributions and for resolving "ties" (multiple observations with the same value).

Test distribution
Specify a distribution type for your data. The drop-down list provides the following options:

Beta
Beta distribution. The shape1 and shape2 parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation. All observations must be in the range 0 - 1, inclusive.

Chi-square
Chi-square distribution. You must specify the degrees of freedom (*df*). Negative observations are not allowed.

Exponential
Exponential distribution. The scale parameter *a* must be positive. If the parameter is not specified, `DISTRIBUTION` estimates it from the sample mean. Negative observations are not allowed.

Gamma
Gamma distribution. The shape and scale parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation. Negative observations are not allowed.

Half-normal
Half-normal distribution. Data are assumed to be location-free or centralized. (Location parameter=0.) You can specify the scale parameter *a* or let `DISTRIBUTION` estimate it by using the maximum likelihood method.

Laplace
Laplace or double exponential distribution. `LAPLACE` takes a location and a scale parameter (*a* and *b).* The scale parameter *(b*) must be positive. If the parameters are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation.

Logistic
Logistic distribution. `LOGISTIC` takes a location and a scale parameter (*a* and *b).* The scale parameter *(b*) must be positive. If the parameters are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation.

Lognormal
Lognormal distribution. The scale and shape parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the mean and standard deviation of the natural logarithm of the sample data. Negative observations are not allowed.

Normal
Normal distribution. The location parameter *a* can be any numeric value, while the scale parameter *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them from the sample mean and sample standard deviation.

Pareto
Pareto distribution. The threshold and shape parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` assumes *a* equals the minimum observation and estimates *b* by the maximum likelihood method. Negative observations are not allowed.

Student t
Student's t *distribution.* You must specify the degrees of freedom (*df).*

Uniform
Uniform distribution. `UNIFORM` takes a minimum and a maximum parameter *(a* and *b).* Parameter *a* must be equal to or greater than *b.* If the parameters are not specified, `DISTRIBUTION` assumes them from the sample data.

Weibull
Weibull distribution. The scale and shape parameters *a* and *b* must be positive. If they are not specified, `DISTRIBUTION` estimates them using the least square method. Negative observations are not allowed.

Distribution parameters
Provides distribution strategy and parameter options.

Estimate from data
When selected, this setting estimates the distribution parameters based on the data and selected distribution type.

Specify
When selected, you can specify the distribution parameters for the selected distribution type.

Note: The available parameters will vary based on the selected distribution type.

Transform
The provided options set the transform and periodicity settings.

Natural log transform
Transforms the data by using the natural logarithm (base *e*) to remove varying amplitude.

Standardize values
Transforms the sequence or time series variables into a sample with a mean of 0 and a standard deviation of 1.

Difference
Specifies the degree of differencing that is used before plotting to convert a non-stationary variable into a stationary variable with a constant mean and variance. Enter an appropriate value in the field.

Seasonally difference
If the variable exhibits a seasonal or periodic pattern, you can use this setting to seasonally difference the variable before plotting.

Note: This setting is enabled only when a sequence or time series variable, with a defined periodicity, is selected as one of the quantitative variables.

Proportion estimation formula
The provided options set the formula that is used to estimate proportions.

Blom's
Creates new ranking variable based on proportion estimates that uses the formula `(r-3/8) / (w+1/4)`, where `w` is the sum of the case weights and `r` is the rank.

Rankit
Uses the formula `(r-1/2) / w`, where `w` is the number of observations and `r` is the rank, ranging from 1 to `w`.

Tukey's
Uses the formula `(r-1/3) / (w+1/3)`, where `r` is the rank and `w` is the sum of the case weights.

Van der Waerden's
Van der Waerden's transformation, defined by the formula `r/(w+1)`, where `w` is the sum of the case weights and `r` is the rank, ranging from 1 to `w`.

Rank assigned to ties
The provided options control the method for determining how to handle tie values. The following table shows how the different methods assign ranks to tied values.

Table 4. Ranking methods and results

Value	Mean	Low	High	Break ties arbitrarily
10	1	1	1	1
15	3	2	4	2
15	3	2	4	2
15	3	2	4	2
16	5	5	5	3
20	6	6	6	4

Obtaining Q-Q probability plots

1. From the menus choose:

   Analyze > **Descriptive Statistics** > **Q-Q Plots...**
2. Select one or more numeric variables and move them onto the Variables field.
3. Select a test distribution.

Optionally, you can select transformation options to obtain probability plots for transformed values and specify a method for calculating expected distributions.

ROC Analysis

Receiver operating characteristic (ROC) Analysis is a useful way to assess the accuracy of model predictions by plotting sensitivity versus (1-specificity) of a classification test (as the threshold varies over an entire range of diagnostic test results). The full area under a given ROC curve, or AUC, formulates an important statistic that represents the probability that the prediction will be in the correct order when a test variable is observed (for one subject that is randomly selected from the case group, and the other randomly selected from the control group). ROC Analysis supports the inference regarding a single AUC, precision-recall (PR) curves, and provides options for comparing two ROC curves that are generated from either independent groups or paired subjects.

The old ROC Curve procedure supports the statistical inference about a single ROC curve. This may also be recovered by the new ROC Analysis procedure. Furthermore, the new ROC Analysis procedure can compare two ROC curves generated from either independent groups or paired subjects.

PR curves plot precision versus recall, tend to be more informative when the observed data samples are highly skewed, and provide an alternative to ROC curves for data with a large skew in the class distribution.

Example

It is in a bank's interest to correctly classify customers into those customers who will and will not default on their loans, so special models are developed for making these decisions. ROC Analysis can be used to evaluate and assess the accuracy of the model predictions.

Statistics

AUC, negative group, missing values, positive classification, cutoff value, strength of conviction, two-sided asymptotic confidence interval, distribution, standard error, independent-group design, paired-sample design, nonparametric assumption, bi-negative exponential distribution assumption, midpoint, cut point, PR curve, stepwise interpolation, asymptotic significance (2-tail), Sensitivity and (1-Specicity), Precision and Recall.

Methods

The areas under two ROC curves, that are generated from either independent groups or paired subjects, are compared. Comparing two ROC curves can provide more information in the accuracy resulted from two comparative diagnostic approaches.

ROC Analysis data considerations

Data

PR curves plot Precision versus Recall, and tend to be more informative when the observed data samples are highly skewed. A simple linear interpolation may mistakenly yield an overly-optimistic estimate of a PR curve.

Assumptions

The prediction will be in the correct order when a test variable is observed for one subject that is randomly selected from the case group and the other is randomly selected from the control group. Each defined group will contain at least one valid observation. Only a single grouping variable is used for a single procedure.

Obtaining an ROC Analysis

1. From the menus choose:

   Analyze > **Classify** > **ROC Analysis**
2. Select one or more **Test Variable** test probability variables.
3. Select one **State Variable**.
4. Identify the *positive* **Value of the State Variable**.
5. Optionally select the **Paired-sample design** option, or select a single grouping variable (you cannot select both options).

- Use the **Paired-sample design** setting to request the paired-sample design for the test variable(s). The paired-sample design compares two ROC curves in a paired-sample scenario when multiple test values are measured on the same subjects that are associated with a state variable.

  Note: When **Paired-sample design** is selected, the **Grouping Variable** and **Distribution Assumption** (in the Options dialog) options are disabled.
- When a numeric grouping variable is selected, you can click **Define Groups...** to request the independent group design for the test variable(s), and to specify two values, a midpoint, or a cut point.

6. Optionally, click Options to define the classification, test direction, standard error parameters, and missing values settings.
7. Optionally, click Display to define the plotting and print settings (which include ROC Curve, Precision-Recall Curve and model quality settings).
8. Click **OK**.

ROC Analysis: Options

You can specify the following options for your ROC analysis:

Classification
Allows you to specify whether the cutoff value should be included or excluded when making a *positive* classification. This setting currently has no effect on the output.

Test Direction
Provides options for specifying which direction of the test result variable indicates increasing strength of conviction that the subject is test positive.

Parameters for Standard Error of Area
Allows you to specify the method of estimating the standard error of the area under the curve. Available methods are nonparametric and bi-negative exponential. The default **Nonparametric** setting provides estimates under the nonparametric assumption. The **Bi-negative exponential** setting provides estimates under the bi-negative exponential distribution assumption.

The section also allows you to specify the confidence level for the two-sided asymptotic confidence interval of the AUC. The available range is 0.0% to 100.0% (the default value is 95%).

Note: The setting only applies to the independent-group design and has no effect in the paired-sample design.

Missing Values
Allows you to specify how missing values are handled. When the setting is not selected, both user-missing values and system-missing values are excluded. When the setting is selected, user-missing values as treated as valid, system-missing values are excluded. Cases with system-missing values, in either the test variable or the state variable, are always excluded from the analysis.

ROC Analysis: Display

You can specify the following display settings for your ROC analysis:

Plot
Provides options for plotting the ROC and Precision-Recall curves.

ROC Curve
When selected, a ROC Curve chart displays in the output. Select **With diagonal reference line** to draw a diagonal reference line on the ROC Curve chart.

Precision-Recall Curve
When selected, a Precision-Recall Curve chart displays in the output. Precision-Recall Curves tend to be more informative when the observed data samples are highly skewed and provide an alternative to ROC Curves for data with a large skew in the class distribution. The default **Interpolate along the true positives** setting makes the stepwise interpolation along the true

positives. The **Interpolate along the false positives** setting makes the stepwise interpolation along the false positives.

Overall model quality
The setting controls whether or not a bar chart is created to display the value of the lower bound of the confidence interval of the estimated AUC. By default, the setting is not selected, which suppresses the bar chart.

Print
Provides options for defining the output for the corresponding statistics.

Standard error and confidence interval
The setting controls which statistics display in the "Area Under the Curve" table. When the setting is not selected, only the estimated AUC displays. When the setting is selected, additional statistics display, including the standard error of the AUC, the asymptotic significance (2-tail), and the Asymptotic Confidence Interval bounds under the null hypothesis.

Coordinate points of ROC Curve
The setting controls the coordinate points of the ROC Curve, along with the cutoff values. When the setting is not selected, the output of coordinate points is suppressed. When the setting is selected, the pairs of Sensitivity and (1-Specificity) values are given with the cutoff values for each ROC Curve.

Youden's Index

Displays the value of Youden's Index for each cutoff value of the ROC curve.

Coordinate points of the Precision-Recall Curve
The setting controls the coordinate points of the Precision-Recall Curve, along with the cutoff values. When the setting is not selected, the output of coordinate points is suppressed. When the setting is selected, the pairs of Precision and Recall values are given with the cutoff values for each Precision-Recall Curve.

Classifier evaluation metrics
The setting controls the display of the Classifier Evaluation Metrics table in the output. The table shows how well a classification model fits the data compared to a random assignment and provides the following information:

- The user-specified test variables
- Group information
- Gini Index (the Gini index is 2*AUC - 1, where AUC is the area under the ROC curve)
- Max K-S and Cutoff values

ROC Analysis: Define Groups (string)

For string grouping variables, enter a string for Group 1 and another value for Group 2, such as *yes* and *no*. Cases with other strings are excluded from the analysis.

Note: The specified values must exist in the variable, otherwise an error message displays to indicate that at least one of the groups is empty.

ROC Analysis: Define Groups (numeric)

For numeric grouping variables, define the two groups for the *t* test by specifying two values, a midpoint, or a cut point.

Note: The specified values must exist in the variable, otherwise an error message displays to indicate that at least one of the groups is empty.

- **Use specified values.** Enter a value for Group 1 and another value for Group 2. Cases with any other values are excluded from the analysis. Numbers need not be integers (for example, 6.25 and 12.5 are valid).
- **Use midpoint value.** When selected, the groups are separated into < and ≥ midpoint values.

- **Use cut point.**
  - **Cutpoint.** Enter a number that splits the values of the grouping variable into two sets. All cases with values that are less than the cutpoint form one group, and cases with values that are greater than or equal to the cutpoint form the other group.

ROC Curve

This procedure is a useful way to evaluate the performance of classification schemes in which there is one variable with two categories by which subjects are classified.

Note: Version 29.0.1 onwards, the menus and dialogs for ROC command are removed and this command is available as syntax only.

Example. It is in a bank's interest to correctly classify customers into those customers who will and will not default on their loans, so special methods are developed for making these decisions. ROC curves can be used to evaluate how well these methods perform.

Statistics. Area under the ROC curve with confidence interval and coordinate points of the ROC curve. Plots: ROC curve.

Methods. The estimate of the area under the ROC curve can be computed either nonparametrically or parametrically using a binegative exponential model.

ROC Curve Data Considerations

The old ROC Curve procedure supports the statistical inference about a single ROC curve. This may also be recovered by the new ROC Analysis procedure. Furthermore, the new ROC Analysis procedure can compare two ROC curves generated from either independent groups or paired subjects.

Data. Test variables are quantitative. Test variables are often composed of probabilities from discriminant analysis or logistic regression or composed of scores on an arbitrary scale indicating a rater's "strength of conviction" that a subject falls into one category or another category. The state variable can be of any type and indicates the true category to which a subject belongs. The value of the state variable indicates which category should be considered *positive*.

Assumptions. It is assumed that increasing numbers on the rater scale represent the increasing belief that the subject belongs to one category, while decreasing numbers on the scale represent the increasing belief that the subject belongs to the other category. The user must choose which direction is *positive*. It is also assumed that the *true* category to which each subject belongs is known.

To obtain an ROC Curve

1. From the menus choose: **Analyze** > **Classify** > **ROC Analysis...**
2. Select one or more **Test Variable**.
3. Select one **State Variable**.
4. Identify the *positive* value for the state variable.
5. Click **Display**, select **ROC Curve** and **With diagonal reference line**, if needed.
6. Click **Continue**

ROC Curve Options

You can specify the following options for your ROC analysis:

Classification
Allows you to specify whether the cutoff value should be included or excluded when making a *positive* classification. This setting currently has no effect on the output.

Test Direction
Allows you to specify the direction of the scale in relation to the *positive* category.

Parameters for Standard Error of Area

Allows you to specify the method of estimating the standard error of the area under the curve. Available methods are 'Nonparametric' and 'Bi-negative' exponential. Also allows you to set the level for the confidence interval. The available range is 50.1% to 99.9%.

Missing Values

Allows you to specify how missing values are handled.

Simulation

Predictive models, such as linear regression, require a set of known inputs to predict an outcome or target value. In many real world applications, however, values of inputs are uncertain. Simulation allows you to account for uncertainty in the inputs to predictive models and evaluate the likelihood of various outcomes of the model in the presence of that uncertainty. For example, you have a profit model that includes the cost of materials as an input, but there is uncertainty in that cost due to market volatility. You can use simulation to model that uncertainty and determine the effect it has on profit.

Simulation in IBM SPSS Statistics uses the Monte Carlo method. Uncertain inputs are modeled with probability distributions (such as the triangular distribution), and simulated values for those inputs are generated by drawing from those distributions. Inputs whose values are known are held fixed at the known values. The predictive model is evaluated using a simulated value for each uncertain input and fixed values for the known inputs to calculate the target (or targets) of the model. The process is repeated many times (typically tens of thousands or hundreds of thousands of times), resulting in a distribution of target values that can be used to answer questions of a probabilistic nature. In the context of IBM SPSS Statistics, each repetition of the process generates a separate case (record) of data that consists of the set of simulated values for the uncertain inputs, the values of the fixed inputs, and the predicted target (or targets) of the model.

You can also simulate data in the absence of a predictive model by specifying probability distributions for variables that are to be simulated. Each generated case of data consists of the set of simulated values for the specified variables.

To run a simulation, you need to specify details such as the predictive model, the probability distributions for the uncertain inputs, correlations between those inputs and values for any fixed inputs. Once you've specified all of the details for a simulation, you can run it and optionally save the specifications to a **simulation plan** file. You can share the simulation plan with other users, who can then run the simulation without needing to understand the details of how it was created.

Two interfaces are available for working with simulations. The Simulation Builder is an advanced interface for users who are designing and running simulations. It provides the full set of capabilities for designing a simulation, saving the specifications to a simulation plan file, specifying output and running the simulation. You can build a simulation based on an IBM SPSS model file, or on a set of custom equations that you define in the Simulation Builder. You can also load an existing simulation plan into the Simulation Builder, modify any of the settings and run the simulation, optionally saving the updated plan. For users who have a simulation plan and primarily want to run the simulation, a simpler interface is available. It allows you to modify settings that enable you to run the simulation under different conditions, but does not provide the full capabilities of the Simulation Builder for designing simulations.

To design a simulation based on a model file

1. From the menus choose:

   Analyze > **Simulation...**

2. Click **Select SPSS Model File** and click **Continue**.
3. Open the model file.

   The model file is an XML file that contains model PMML created from IBM SPSS Statistics or IBM SPSS Modeler. See the topic “Model tab ” on page 234 for more information.

4. On the Simulation tab (in the Simulation Builder), specify probability distributions for simulated inputs and values for fixed inputs. If the active dataset contains historical data for simulated inputs, click

Fit All to automatically determine the distribution that most closely fits the data for each such input as well as determining correlations between them. For each simulated input that is not being fit to historical data, you must explicitly specify a distribution by selecting a distribution type and entering the required parameters.

5. Click **Run** to run the simulation. By default, the simulation plan, specifying the details of the simulation, is saved to the location specified on the Save settings.

The following options are available:

- Modify the location for the saved simulation plan.
- Specify known correlations between simulated inputs.
- Automatically compute a contingency table of associations between categorical inputs and use those associations when data are generated for those inputs.
- Specify sensitivity analysis to investigate the effect of varying the value of a fixed input or varying a distribution parameter for a simulated input.
- Specify advanced options such as setting the maximum number of cases to generate or requesting tail sampling.
- Customize output.
- Save the simulated data to a data file.

To design a simulation based on custom equations

1. From the menus choose:

   Analyze > **Simulation...**

2. Click **Type in the Equations** and click **Continue**.
3. Click **New Equation** on the Model tab (in the Simulation Builder) to define each equation in your predictive model.
4. Click the Simulation tab and specify probability distributions for simulated inputs and values for fixed inputs. If the active dataset contains historical data for simulated inputs, click **Fit All** to automatically determine the distribution that most closely fits the data for each such input as well as determining correlations between them. For each simulated input that is not being fit to historical data, you must explicitly specify a distribution by selecting a distribution type and entering the required parameters.
5. Click **Run** to run the simulation. By default, the simulation plan, specifying the details of the simulation, is saved to the location specified on the Save settings.

The following options are available:

- Modify the location for the saved simulation plan.
- Specify known correlations between simulated inputs.
- Automatically compute a contingency table of associations between categorical inputs and use those associations when data are generated for those inputs.
- Specify sensitivity analysis to investigate the effect of varying the value of a fixed input or varying a distribution parameter for a simulated input.
- Specify advanced options such as setting the maximum number of cases to generate or requesting tail sampling.
- Customize output.
- Save the simulated data to a data file.

To design a simulation without a predictive model

1. From the menus, choose:

   Analyze > **Simulation...**

2. Click **Create Simulated Data** and click **Continue**.

3. On the Model tab (in the Simulation Builder), select the fields that you want to simulate. You can select fields from the active dataset or you can define new fields by clicking **New**.
4. Click the Simulation tab and specify probability distributions for the fields that are to be simulated. If the active dataset contains historical data for any of those fields, click **Fit All** to automatically determine the distribution that most closely fits the data and to determine correlations between the fields. For fields that are not fit to historical data, you must explicitly specify a distribution by selecting a distribution type and entering the required parameters.
5. Click **Run** to run the simulation. By default, the simulated data are saved to the new dataset specified on the Save settings. In addition, the simulation plan, which specifies the details of the simulation, is saved to the location specified on the Save settings.

The following options are available:

- Modify the location for the simulated data or the saved simulation plan.
- Specify known correlations between simulated fields.
- Automatically compute a contingency table of associations between categorical fields and use those associations when data are generated for those fields.
- Specify sensitivity analysis to investigate the effect of varying a distribution parameter for a simulated field.
- Specify advanced options such as setting the number of cases to generate.

To run a simulation from a simulation plan

Two options are available for running a simulation from a simulation plan. You can use the Run Simulation dialog, which is primarily designed for running from a simulation plan, or you can use the Simulation Builder.

To use the Run Simulation dialog:

1. From the menus choose:

   Analyze > **Simulation...**
2. Click **Open an Existing Simulation Plan**.
3. Make sure the **Open in Simulation Builder** check box is not checked and click **Continue**.
4. Open the simulation plan.
5. Click **Run** in the Run Simulation dialog.

To run the simulation from the Simulation Builder:

1. From the menus choose:

   Analyze > **Simulation...**
2. Click **Open an Existing Simulation Plan**.
3. Select the **Open in Simulation Builder** check box and click **Continue**.
4. Open the simulation plan.
5. Modify any settings you want to modify on the Simulation tab.
6. Click **Run** to run the simulation.

Optionally, you can do the following:

- Set up or modify sensitivity analysis to investigate the effect of varying the value of a fixed input or varying a distribution parameter for a simulated input.
- Refit distributions and correlations for simulated inputs to new data.
- Change the distribution for a simulated input.
- Customize output.
- Save the simulated data to a data file.

Simulation Builder

The Simulation Builder provides the full set of capabilities for designing and running simulations. It allows you to perform the following general tasks:

- Design and run a simulation for an IBM SPSS model defined in a PMML model file.
- Design and run a simulation for a predictive model defined by a set of custom equations that you specify.
- Design and run a simulation that generates data in the absence of a predictive model.
- Run a simulation based on an existing simulation plan, optionally modifying any plan settings.

Model tab

For simulations based on a predictive model, the Model tab specifies the source of the model. For simulations that do not include a predictive model, the Model tab specifies the fields that are to be simulated.

Select an SPSS model file. This option specifies that the predictive model is defined in an IBM SPSS model file. An IBM SPSS model file is an XML file or a compressed file archive (.zip file) that contains model PMML created from IBM SPSS Statistics or IBM SPSS Modeler. Predictive models are created by procedures, such as Linear Regression and Decision Trees within IBM SPSS Statistics, and can be exported to a model file. You can use a different model file by clicking **Browse** and navigating to the file you want.

PMML models supported by Simulation

- Linear Regression
- Automatic Linear Model
- Generalized Linear Model
- Generalized Linear Mixed Model
- General Linear Model
- Binary Logistic Regression
- Multinomial Logistic Regression
- Ordinal Multnomial Regression
- Cox Regression
- Tree
- Boosted Tree (C5)
- Discriminant
- Two-step Cluster
- K-Means Cluster
- Neural Net
- Ruleset (Decision List)

Note:

- PMML models that have multiple target fields (variables) or splits are not supported for use in Simulation.
- Values of string inputs to binary logistic regression models are limited to 8 bytes in the model. If you are fitting such string inputs to the active dataset, make sure that the values in the data do not exceed 8 bytes in length. Data values that exceed 8 bytes are excluded from the associated categorical distribution for the input, and are displayed as unmatched in the Unmatched Categories output table.

Type in the equations for the model. This option specifies that the predictive model consists of one or more custom equations to be created by you. Create equations by clicking **New Equation**. This opens the

Equation Editor. You can modify existing equations, copy them to use as templates for new equations, reorder them and delete them.

- The Simulation Builder does not support systems of simultaneous equations or equations that are non-linear in the target variable.
- Custom equations are evaluated in the order in which they are specified. If the equation for a given target depends on another target, then the other target must be defined by a preceding equation.

  For example, given the set of three equations below, the equation for *profit* depends on the values of *revenue* and *expenses*, so the equations for *revenue* and *expenses* must precede the equation for *profit*.

  ```
  revenue = price*volume
  ```

  ```
  expenses = fixed + volume*(unit_cost_materials + unit_cost_labor)
  ```

  ```
  profit = revenue - expenses
  ```

Create simulated data without a model. Select this option to simulate data without a predictive model. Specify the fields that are to be simulated by selecting fields from the active dataset or by clicking **New** to define new fields.

Equation Editor

The Equation Editor allows you to create or modify a custom equation for your predictive model.

- The expression for the equation can contain fields from the active dataset or new input fields that you define in the Equation Editor.
- You can specify properties of the target such as its measurement level, value labels and whether output is generated for the target.
- You can use targets from previously defined equations as inputs to the current equation, allowing you to create coupled equations.
- You can attach a descriptive comment to the equation. Comments are displayed along with the equation on the Model tab.

1. Enter the name of the target. Optionally, click **Edit** under the Target text box to open the Defined Inputs dialog, allowing you to change the default properties of the target.
2. To build an expression, either paste components into the Numeric Expression field or type directly in the Numeric Expression field.

- You can build your expression using fields from the active dataset or you can define new inputs by clicking the **New** button. This opens the Define Inputs dialog.
- You can paste functions by selecting a group from the Function group list and double-clicking the function in the Functions list (or select the function and click the arrow adjacent to the Function group list). Enter any parameters indicated by question marks. The function group labeled **All** provides a listing of all available functions. A brief description of the currently selected function is displayed in a reserved area in the dialog box.
- String constants must be enclosed in quotation marks.
- If values contain decimals, a period (.) must be used as the decimal indicator.

Note: Simulation does not support custom equations with string targets.

Defined Inputs

The Defined Inputs dialog allows you to define new inputs and set properties for targets.

- If an input to be used in an equation does not exist in the active dataset, you must define it before it can be used in the equation.
- If you are simulating data without a predictive model, you must define all simulated inputs that do not exist in the active dataset.

Name. Specify the name for a target or input.

Target. You can specify the measurement level of a target. The default measurement level is continuous. You can also specify whether output will be created for this target. For example, for a set of coupled equations you may only be interested in output from the target for the final equation, so you would suppress output from the other targets.

Input to be simulated. This specifies that the values of the input will be simulated according to a specified probability distribution (the probability distribution is specified on the Simulation tab). The measurement level determines the default set of distributions that are considered when finding the distribution that most closely fits the data for the input (by clicking **Fit** or **Fit All** on the Simulation tab). For example, if the measurement level is continuous, then the normal distribution (appropriate for continuous data) would be considered but the binomial distribution would not.

Note: Select a measurement level of String for string inputs. String inputs that are to be simulated are restricted to the Categorical distribution.

Fixed value input. This specifies that the value of the input is known and will be fixed at the known value. Fixed inputs can be numeric or string. Specify a value for the fixed input. String values should not be enclosed in quotation marks.

Value labels. You can specify value labels for targets, simulated inputs and fixed inputs. Value labels are used in output charts and tables.

Simulation tab

The Simulation tab specifies all properties of the simulation other than the predictive model. You can perform the following general tasks on the Simulation tab:

- Specify probability distributions for simulated inputs and values for fixed inputs.
- Specify correlations between simulated inputs. For categorical inputs, you can specify that associations that exist between those inputs in the active dataset are used when data are generated for those inputs.
- Specify advanced options such as tail sampling and criteria for fitting distributions to historical data.
- Customize output.
- Specify where to save the simulation plan and optionally save the simulated data.

Simulated Fields

To run a simulation, each input field must be specified as fixed or simulated. Simulated inputs are those whose values are uncertain and will be generated by drawing from a specified probability distribution. When historical data are available for the inputs to be simulated, the distributions that most closely fit the data can be automatically determined, along with any correlations between those inputs. You can also manually specify distributions or correlations if historical data are not available or you require specific distributions or correlations.

Fixed inputs are those whose values are known and remain constant for each case generated in the simulation. For example, you have a linear regression model for sales as a function of a number of inputs including price, and you want to hold the price fixed at the current market price. You would then specify price as a fixed input.

For simulations based on a predictive model, each predictor in the model is an input field for the simulation. For simulations that do not include a predictive model, the fields that are specified on the Model tab are the inputs for the simulation.

Automatically fitting distributions and calculating correlations for simulated inputs. If the active dataset contains historical data for the inputs that you want to simulate, then you can automatically find the distributions that most closely fit the data for those inputs as well as determine any correlations between them. The steps are as follows:

1. Verify that each of the inputs that you want to simulate is matched up with the correct field in the active dataset. Inputs are listed in the Input column and the Fit to column displays the matched field in the active dataset. You can match an input to a different field in the active dataset by selecting a different item from the Fit to dropdown list.

A value of -*None*- in the Fit to column indicates that the input could not be automatically matched to a field in the active dataset. By default, inputs are matched to dataset fields on name, measurement level and type (numeric or string). If the active dataset does not contain historical data for the input, then manually specify the distribution for the input or specify the input as a fixed input, as described below.

2. Click **Fit All**.

The closest fitting distribution and its associated parameters are displayed in the Distribution column along with a plot of the distribution superimposed on a histogram (or bar chart) of the historical data. Correlations between simulated inputs are displayed on the Correlations settings. You can examine the fit results and customize automatic distribution fitting for a particular input by selecting the row for the input and clicking **Fit Details**. See the topic "Fit Details " on page 238 for more information.

You can run automatic distribution fitting for a particular input by selecting the row for the input and clicking **Fit**. Correlations for all simulated inputs that match fields in the active dataset are also automatically calculated.

Note:

- Cases with missing values for any simulated input are excluded from distribution fitting, computation of correlations, and computation of the optional contingency table (for inputs with a Categorical distribution). You can optionally specify whether user-missing values of inputs with a Categorical distribution are treated as valid. By default, they are treated as missing. For more information, see the topic "Advanced Options " on page 240.
- For continuous and ordinal inputs, if an acceptable fit cannot be found for any of the tested distributions, then the Empirical distribution is suggested as the closest fit. For continuous inputs, the Empirical distribution is the cumulative distribution function of the historical data. For ordinal inputs, the Empirical distribution is the categorical distribution of the historical data.

Manually specifying distributions. You can manually specify the probability distribution for any simulated input by selecting the distribution from the **Type** dropdown list and entering the distribution parameters in the Parameters grid. Once you have entered the parameters for a distribution, a sample plot of the distribution, based on the specified parameters, will be displayed adjacent to the Parameters grid. Following are some notes on particular distributions:

- **Categorical.** The categorical distribution describes an input field that has a fixed number of values, referred to as categories. Each category has an associated probability such that the sum of the probabilities over all categories equals one. To enter a category, click the left-hand column in the Parameters grid and specify the category value. Enter the probability associated with the category in the right-hand column.

  Note: Categorical inputs from a PMML model have categories that are determined from the model and cannot be modified.
- **Negative Binomial - Failures.** Describes the distribution of the number of failures in a sequence of trials before a specified number of successes are observed. The parameter *thresh* is the specified number of successes and the parameter *prob* is the probability of success in any given trial.
- **Negative Binomial - Trials.** Describes the distribution of the number of trials required before a specified number of successes are observed. The parameter *thresh* is the specified number of successes and the parameter *prob* is the probability of success in any given trial.
- **Range.** This distribution consists of a set of intervals with a probability assigned to each interval such that the sum of the probabilities over all intervals equals 1. Values within a given interval are drawn from a uniform distribution defined on that interval. Intervals are specified by entering a minimum value, a maximum value and an associated probability.

  For example, you believe that the cost of a raw material has a 40% chance of falling in the range of $10 - $15 per unit and a 60% chance of falling in the range of $15 - $20 per unit. You would model the cost with a Range distribution consisting of the two intervals [10 - 15] and [15 - 20], setting the probability associated with the first interval to 0.4 and the probability associated with the second interval to 0.6. The intervals do not have to be contiguous and they can even be overlapping. For example, you could have specified the intervals $10 - $15 and $20 - $25 or $10 - $15 and $13 - $16.

- **Weibull.** The parameter *c* is an optional location parameter, which specifies where the origin of the distribution is located.

Parameters for the following distributions have the same meaning as in the associated random variable functions available in the Compute Variable dialog box: Bernoulli, beta, binomial, exponential, gamma, lognormal, negative binomial (trials and failures), normal, Poisson and uniform.

Specifying fixed inputs. Specify a fixed input by selecting Fixed from the **Type** dropdown list in the Distribution column and entering the fixed value. The value can be numeric or string depending on whether the input is numeric or string. String values should not be enclosed in quotation marks.

Specifying bounds on simulated values. Most distributions support specifying upper and lower bounds on the simulated values. You can specify a lower bound by entering a value into the **Min** text box and you can specify an upper bound by entering a value into the **Max** text box.

Locking inputs. Locking an input, by selecting the check box in the column with the lock icon, excludes the input from automatic distribution fitting. This is most useful when you manually specify a distribution or a fixed value and want to ensure that it will not be affected by automatic distribution fitting. Locking is also useful if you intend to share your simulation plan with users who will be running it in the Run Simulation dialog and you want to prevent any changes to certain inputs. In that regard, specifications for locked inputs cannot be modified in the Run Simulation dialog.

Sensitivity Analysis. Sensitivity analysis allows you to investigate the effect of systematic changes in a fixed input or in a distribution parameter for a simulated input by generating an independent set of simulated cases—effectively, a separate simulation—for each specified value. To specify sensitivity analysis, select a fixed or simulated input and click **Sensitivity Analysis**. Sensitivity analysis is limited to a single fixed input or a single distribution parameter for a simulated input. See the topic "Sensitivity Analysis " on page 239 for more information.

Fit status icons

Icons in the Fit to column indicate the fit status for each input field.

Table 5. Status icons

Icon	Description
	No distribution has been specified for the input and the input has not been specified as fixed. In order to run the simulation, you must either specify a distribution for this input or define it to be fixed and specify the fixed value.
	The input was previously fit to a field that does not exist in the active dataset. No action is necessary unless you want to refit the distribution for the input to the active dataset.
	The closest fitting distribution has been replaced with an alternate distribution from the Fit Details dialog.
	The input is set to the closest fitting distribution.
	The distribution has been manually specified or sensitivity analysis iterations have been specified for this input.

Fit Details

The Fit Details dialog displays the results of automatic distribution fitting for a particular input. Distributions are ordered by goodness of fit, with the closest fitting distribution listed first. You can override the closest fitting distribution by selecting the radio button for the distribution you want in the Use column. Selecting a radio button in the Use column also displays a plot of the distribution superimposed on a histogram (or bar chart) of the historical data for that input.

Fit statistics. By default and for continuous fields, the Anderson-Darling test is used for determining goodness of fit. Alternatively, and for continuous fields only, you can specify the Kolmogorov-Smirnoff test for goodness of fit by selecting that choice on the Advanced Options settings. For continuous inputs, results of both tests are shown in the Fit Statistics column (A for Anderson-Darling and K for Kolmogorov-Smirnoff), with the chosen test used to order the distributions. For ordinal and nominal inputs the chi-square test is used. The p-values associated with the tests are also shown.

Parameters. The distribution parameters associated with each fitted distribution are displayed in the Parameters column. Parameters for the following distributions have the same meaning as in the associated random variable functions available in the Compute Variable dialog box: Bernoulli, beta, binomial, exponential, gamma, lognormal, negative binomial (trials and failures), normal, Poisson and uniform. For the categorical distribution, the parameter names are the categories and the parameter values are the associated probabilities.

Refitting with a customized distribution set. By default, the measurement level of the input is used to determine the set of distributions considered for automatic distribution fitting. For example, continuous distributions such as lognormal and gamma are considered when fitting a continuous input but discrete distributions such as Poisson and binomial are not. You can choose a subset of the default distributions by selecting the distributions in the Refit column. You can also override the default set of distributions by selecting a different measurement level from the **Treat as (Measure)** dropdown list and selecting the distributions in the Refit column. Click **Run Refit** to refit with the custom distribution set.

Note:

- Cases with missing values for any simulated input are excluded from distribution fitting, computation of correlations, and computation of the optional contingency table (for inputs with a Categorical distribution). You can optionally specify whether user-missing values of inputs with a Categorical distribution are treated as valid. By default, they are treated as missing. For more information, see the topic "Advanced Options " on page 240.
- For continuous and ordinal inputs, if an acceptable fit cannot be found for any of the tested distributions, then the Empirical distribution is suggested as the closest fit. For continuous inputs, the Empirical distribution is the cumulative distribution function of the historical data. For ordinal inputs, the Empirical distribution is the categorical distribution of the historical data.

Sensitivity Analysis

Sensitivity analysis allows you to investigate the effect of varying a fixed input or a distribution parameter for a simulated input over a specified set of values. An independent set of simulated cases--effectively, a separate simulation--is generated for each specified value, allowing you to investigate the effect of varying the input. Each set of simulated cases is referred to as an **iteration**.

Iterate. This choice allows you to specify the set of values over which the input will be varied.

- If you are varying the value of a distribution parameter, then select the parameter from the drop-down list. Enter the set of values in the Parameter value by iteration grid. Clicking **Continue** will add the specified values to the Parameters grid for the associated input, with an index specifying the iteration number of the value.
- For the Categorical and Range distributions, the probabilities of the categories or intervals respectively can be varied but the values of the categories and the endpoints of the intervals cannot be varied. Select a category or interval from the drop-down list and specify the set of probabilities in the Parameter value by iteration grid. The probabilities for the other categories or intervals will be automatically adjusted accordingly.

No iterations. Use this option to cancel iterations for an input. Clicking **Continue** will remove the iterations.

Correlations

Input fields to be simulated are often known to be correlated--for example, height and weight. Correlations between inputs that will be simulated must be accounted for in order to ensure that the simulated values preserve those correlations.

Recalculate correlations when fitting. This choice specifies that correlations between simulated inputs are automatically calculated when fitting distributions to the active dataset through the **Fit All** or **Fit** actions on the Simulated Fields settings.

Do not recalculate correlations when fitting. Select this option if you want to manually specify correlations and prevent them from being overwritten when automatically fitting distributions to the active dataset. Values that are entered in the Correlations grid must be between -1 and 1. A value of 0 specifies that there is no correlation between the associated pair of inputs.

Reset. This resets all correlations to 0.

Use fitted multiway contingency table for inputs with a categorical distribution. For inputs with a categorical distribution, you can automatically compute a multiway contingency table from the active dataset that describes the associations between those inputs. The contingency table is then used when data are generated for those inputs. If you choose to save the simulation plan, the contingency table is saved in the plan file and is used when you run the plan.

- **Compute contingency table from the active dataset.** If you are working with an existing simulation plan that contains a contingency table, you can recompute the contingency table from the active dataset. This action overrides the contingency table from the loaded plan file.
- **Use contingency table from loaded simulation plan.** By default, when you load a simulation plan that contains a contingency table, the table from the plan is used. You can recompute the contingency table from the active dataset by selecting **Compute contingency table from the active dataset**.

Advanced Options

Maximum Number of Cases. This specifies the maximum number of cases of simulated data, and associated target values, to generate. When sensitivity analysis is specified, this is the maximum number of cases for each iteration.

Target for stopping criteria. If your predictive model contains more than one target, then you can select the target to which stopping criteria are applied.

Stopping criteria. These choices specify criteria for stopping the simulation, potentially before the maximum number of allowable cases has been generated.

- **Continue until maximum is reached.** This specifies that simulated cases will be generated until the maximum number of cases is reached.
- **Stop when the tails have been sampled.** Use this option when you want to ensure that one of the tails of a specified target distribution has been adequately sampled. Simulated cases will be generated until the specified tail sampling is complete or the maximum number of cases is reached. If your predictive model contains multiple targets then select the target, to which this criteria will be applied, from the **Target for stopping criteria** dropdown list.

  Type. You can define the boundary of the tail region by specifying a value of the target such as 10,000,000 or a percentile such as the 99th percentile. If you choose Value in the **Type** dropdown list, then enter the value of the boundary in the Value text box and use the **Side** dropdown list to specify whether this is the boundary of the Left tail region or the Right tail region. If you choose Percentile in the **Type** dropdown list, then enter a value in the Percentile text box.

  Frequency. Specify the number of values of the target that must lie in the tail region in order to ensure that the tail has been adequately sampled. Generation of cases will stop when this number has been reached.
- **Stop when the confidence interval of the mean is within the specified threshold.** Use this option when you want to ensure that the mean of a given target is known with a specified degree of accuracy. Simulated cases will be generated until the specified degree of accuracy has been achieved or the maximum number of cases is reached. To use this option, you specify a confidence level and a threshold. Simulated cases will be generated until the confidence interval associated the specified level is within the threshold. For example, you can use this option to specify that cases are generated until the confidence interval of the mean at the 95% confidence level is within 5% of the mean value. If your

predictive model contains multiple targets then select the target, to which this criteria will be applied, from the **Target for stopping criteria** dropdown list.

Threshold Type. You can specify the threshold as a numeric value or as a percent of the mean. If you choose Value in the **Threshold Type** dropdown list, then enter the threshold in the Threshold as Value text box. If you choose Percent in the **Threshold Type** dropdown list, then enter a value in the Threshold as Percent text box.

Number of cases to sample. This specifies the number of cases to use when automatically fitting distributions for simulated inputs to the active dataset. If your dataset is very large you might want to consider limiting the number of cases used for distribution fitting. If you select **Limit to N cases**, the first N cases will be used.

Goodness of fit criteria (Continuous). For continuous inputs, you can use the Anderson-Darling test or the Kolmogorov-Smirnoff test of goodness of fit to rank distributions when fitting distributions for simulated inputs to the active dataset. The Anderson-Darling test is selected by default and is especially recommended when you want to ensure the best possible fit in the tail regions.

Empirical Distribution. For continuous inputs, the Empirical distribution is the cumulative distribution function of the historical data. You can specify the number of bins used for calculating the Empirical distribution for continuous inputs. The default is 100 and the maximum is 1000.

Replicate results. Setting a random seed allows you to replicate your simulation. Specify an integer or click **Generate**, which will create a pseudo-random integer between 1 and 2147483647, inclusive. The default is 629111597.

Note: For a particular random seed, results are replicated unless the number of threads is changed. On a particular computer, the number of threads is constant unless you change it by running SET THREADS command syntax. The number of threads might change if you run the simulation on a different computer because an internal algorithm is used to determine the number of threads for each computer.

User-missing values for inputs with a Categorical distribution. These controls specify whether user-missing values of inputs with a Categorical distribution are treated as valid. System-missing values and user-missing values for all other types of inputs are always treated as invalid. All inputs must have valid values for a case to be included in distribution fitting, computation of correlations, and computation of the optional contingency table.

Density Functions

These settings allow you to customize output for probability density functions and cumulative distribution functions for continuous targets, as well as bar charts of predicted values for categorical targets.

Probability Density Function (PDF). The probability density function displays the distribution of target values. For continuous targets, it allows you to determine the probability that the target is within a given region. For categorical targets (targets with a measurement level of nominal or ordinal), a bar chart is generated that displays the percentage of cases that fall in each category of the target. Additional options for categorical targets of PMML models are available with the Category values to report setting described below.

For Two-Step cluster models and K-Means cluster models, a bar chart of cluster membership is produced.

Cumulative Distribution Function (CDF). The cumulative distribution function displays the probability that the value of the target is less than or equal to a specified value. It is only available for continuous targets.

Slider positions. You can specify the initial positions of the moveable reference lines on PDF and CDF charts. Values that are specified for the lower and upper lines refer to positions along the horizontal axis, not percentiles. You can remove the lower line by selecting **-Infinity** or you can remove the upper line by selecting **Infinity**. By default, the lines are positioned at the 5-th and 95-th percentiles. When multiple distribution functions are displayed on a single chart (because of multiple targets or results from sensitivity analysis iterations), the default refers to the distribution for the first iteration or first target.

Reference Lines (Continuous). You can request various vertical reference lines to be added to probability density functions and cumulative distribution functions for continuous targets.

- **Sigmas.** You can add reference lines at plus and minus a specified number of standard deviations from the mean of a target.
- **Percentiles.** You can add reference lines at one or two percentile values of the distribution of a target by entering values into the Bottom and Top text boxes. For example, a value of 95 in the Top text box represents the 95th percentile, which is the value below which 95% of the observations fall. Likewise, a value of 5 in the Bottom text box represents the 5th percentile, which is the value below which 5% of the observations fall.
- **Custom reference lines.** You can add reference lines at specified values of the target.

Note: When multiple distribution functions are displayed on a single chart (because of multiple targets or results from sensitivity analysis iterations), reference lines are only applied to the distribution for the first iteration or first target. You can add reference lines to the other distributions from the Chart Options dialog, which is accessed from the PDF or CDF chart.

Overlay results from separate continuous targets. In the case of multiple continuous targets, this specifies whether distribution functions for all such targets are displayed on a single chart, with one chart for probability density functions and another for cumulative distribution functions. When this option is not selected, results for each target will be displayed on a separate chart.

Category values to report. For PMML models with categorical targets, the result of the model is a set of predicted probabilities, one for each category, that the target value falls in each category. The category with the highest probability is taken to be the predicted category and used in generating the bar chart described for the **Probability Density Function** setting above. Selecting **Predicted category** will generate the bar chart. Selecting **Predicted probabilities** will generate histograms of the distribution of predicted probabilities for each of the categories of the target.

Grouping for sensitivity analysis. Simulations that include sensitivity analysis generate an independent set of predicted target values for each iteration defined by the analysis (one iteration for each value of the input that is being varied). When iterations are present, the bar chart of the predicted category for a categorical target is displayed as a clustered bar chart that includes the results for all iterations. You can choose to group categories together or you can group iterations together.

Output

Tornado charts. Tornado charts are bar charts that display relationships between targets and simulated inputs using a variety of metrics.

- **Correlation of target with input.** This option creates a tornado chart of the correlation coefficients between a given target and each of its simulated inputs. This type of tornado chart does not support targets with a nominal or ordinal measurement level or simulated inputs with a categorical distribution.
- **Contribution to variance.** This option creates a tornado chart that displays the contribution to the variance of a target from each of its simulated inputs, allowing you to assess the degree to which each input contributes to the overall uncertainty in the target. This type of tornado chart does not support targets with ordinal or nominal measurement levels, or simulated inputs with any of the following distributions: categorical, Bernoulli, binomial, Poisson, or negative binomial.
- **Sensitivity of target to change.** This option creates a tornado chart that displays the effect on the target of modulating each simulated input by plus or minus a specified number of standard deviations of the distribution associated with the input. This type of tornado chart does not support targets with ordinal or nominal measurement levels, or simulated inputs with any of the following distributions: categorical, Bernoulli, binomial, Poisson, or negative binomial.

Box plots of target distributions. Box plots are available for continuous targets. Select **Overlay results from separate targets** if your predictive model has multiple continuous targets and you want to display the box plots for all targets on a single chart.

Scatterplots of targets versus inputs. Scatterplots of targets versus simulated inputs are available for both continuous and categorical targets and include scatters of the target with both continuous and categorical inputs. Scatters involving a categorical target or a categorical input are displayed as a heat map.

Create a table of percentile values. For continuous targets, you can obtain a table of specified percentiles of the target distributions. Quartiles (the 25th, 50th, and 75th percentiles) divide the observations into four groups of equal size. If you want an equal number of groups other than four, select **Intervals** and specify the number. Select **Custom percentiles** to specify individual percentiles--for example, the 99th percentile.

Descriptive statistics of target distributions. This option creates tables of descriptive statistics for continuous and categorical targets as well as for continuous inputs. For continuous targets the table includes the mean, standard deviation, median, minimum and maximum, confidence interval of the mean at the specified level, and the 5th and 95th percentiles of the target distribution. For categorical targets the table includes the percentage of cases that fall in each category of the target. For categorical targets of PMML models, the table also includes the mean probability of each category of the target. For continuous inputs, the table includes the mean, standard deviation, minimum and maximum.

Correlations and contingency table for inputs. This option displays a table of correlation coefficients between simulated inputs. When inputs with categorical distributions are generated from a contingency table, the contingency table of the data that are generated for those inputs is also displayed.

Simulated inputs to include in the output. By default, all simulated inputs are included in the output. You can exclude selected simulated inputs from output. This will exclude them from tornado charts, scatterplots and tabular output.

Limit ranges for continuous targets. You can specify the range of valid values for one or more continuous targets. Values outside of the specified range are excluded from all output and analyses associated with the targets. To set a lower limit, select **Lower** in the Limit column and enter a value in the Minimum column. To set an upper limit, select **Upper** in the Limit column and enter a value in the Maximum column. To set both a lower and an upper limit, select **Both** in the Limit column and enter values in the Minimum and Maximum columns.

Display Formats. You can set the format used when displaying values of targets and inputs (both fixed inputs and simulated inputs).

Save

Save the plan for this simulation. You can save the current specifications for your simulation to a simulation plan file. Simulation plan files have the extension *.splan*. You can re-open the plan in the Simulation Builder, optionally make modifications and run the simulation. You can share the simulation plan with other users, who can then run it in the Run Simulation dialog. Simulation plans include all specifications except the following: settings for Density Functions; Output settings for charts and tables; Advanced Options settings for Fitting, Empirical Distribution and Random Seed.

Save the simulated data as a new data file. You can save simulated inputs, fixed inputs and predicted target values to an SPSS Statistics data file, a new dataset in the current session, or an Excel file. Each case (or row) of the data file consists of the predicted values of the targets along with the simulated inputs and fixed inputs that generate the target values. When sensitivity analysis is specified, each iteration gives rise to a contiguous set of cases that are labeled with the iteration number.

Run Simulation dialog

The Run Simulation dialog is designed for users who have a simulation plan and primarily want to run the simulation. It also provides the features you need to run the simulation under different conditions. It allows you to perform the following general tasks:

- Set up or modify sensitivity analysis to investigate the effect of varying the value of a fixed input or varying a distribution parameter for a simulated input.
- Refit probability distributions for uncertain inputs (and correlations between those inputs) to new data.
- Modify the distribution for a simulated input.
- Customize output.
- Run the simulation.

Simulation tab

The Simulation tab allows you to specify sensitivity analysis, refit probability distributions for simulated inputs and correlations between simulated inputs to new data, and modify the probability distribution associated with a simulated input.

The Simulated inputs grid contains an entry for each input field that is defined in the simulation plan. Each entry displays the name of the input and the probability distribution type associated with the input, along with a sample plot of the associated distribution curve. Each input also has an associated status icon (a colored circle with a check mark) that is useful when you are refitting distributions to new data. In addition, inputs may include a lock icon which indicates that the input is locked and cannot be modified or refit to new data in the Run Simulation dialog. To modify a locked input you will need to open the simulation plan in the Simulation Builder.

Each input is either simulated or fixed. Simulated inputs are those whose values are uncertain and will be generated by drawing from a specified probability distribution. Fixed inputs are those whose values are known and remain constant for each case generated in the simulation. To work with a particular input, select the entry for the input in the Simulated inputs grid.

Specifying sensitivity analysis

Sensitivity analysis allows you to investigate the effect of systematic changes in a fixed input or in a distribution parameter for a simulated input by generating an independent set of simulated cases—effectively, a separate simulation—for each specified value. To specify sensitivity analysis, select a fixed or simulated input and click **Sensitivity Analysis**. Sensitivity analysis is limited to a single fixed input or a single distribution parameter for a simulated input. See the topic "Sensitivity Analysis " on page 239 for more information.

Refitting distributions to new data

To automatically refit probability distributions for simulated inputs (and correlations between simulated inputs) to data in the active dataset:

1. Verify that each of the model inputs is matched up with the correct field in the active dataset. Each simulated input is fit to the field in the active dataset specified in the **Field** dropdown list associated with that input. You can easily identify inputs that are unmatched by looking for inputs with a status icon that includes a check mark with a question mark, as shown below.

2. Modify any necessary field matching by selecting **Fit to a field in the dataset** and selecting the field from the list.
3. Click **Fit All**.

For each input that was fit, the distribution that most closely fits the data is displayed along with a plot of the distribution superimposed on a histogram (or bar chart) of the historical data. If an acceptable fit cannot be found then the Empirical distribution is used. For inputs that are fit to the Empirical distribution, you will only see a histogram of the historical data because the Empirical distribution is in fact represented by that histogram.

Note: For a complete list of status icons, see the topic "Simulated Fields" on page 236.

Modifying probability distributions

You can modify the probability distribution for a simulated input and optionally change a simulated input to a fixed input or vice versa.

1. Select the input and select **Manually set the distribution**.
2. Select the distribution type and specify the distribution parameters. To change a simulated input to a fixed input, select Fixed in the **Type** dropdown list.

Once you have entered the parameters for a distribution, the sample plot of the distribution (displayed in the entry for the input) will be updated to reflect your changes. For more information on manually specifying probability distributions, see the topic "Simulated Fields" on page 236.

Include user-missing values of categorical inputs when fitting. This specifies whether user-missing values of inputs with a Categorical distribution are treated as valid when you are refitting to data in the active dataset. System-missing values and user-missing values for all other types of inputs are always treated as invalid. All inputs must have valid values for a case to be included in distribution fitting and computation of correlations.

Output tab

The Output tab allows you to customize the output generated by the simulation.

Density Functions. Density functions are the primary means of probing the set of outcomes from your simulation.

- **Probability Density Function.** The probability density function displays the distribution of target values, allowing you to determine the probability that the target is within a given region. For targets with a fixed set outcomes--such as "poor service", "fair service", "good service" and "excellent service"--a bar chart is generated that displays the percentage of cases that fall in each category of the target.
- **Cumulative Distribution Function.** The cumulative distribution function displays the probability that the value of the target is less than or equal to a specified value.

Tornado Charts. Tornado charts are bar charts that display relationships between targets and simulated inputs using a variety of metrics.

- **Correlation of target with input.** This option creates a tornado chart of the correlation coefficients between a given target and each of its simulated inputs.
- **Contribution to variance.** This option creates a tornado chart that displays the contribution to the variance of a target from each of its simulated inputs, allowing you to assess the degree to which each input contributes to the overall uncertainty in the target.
- **Sensitivity of target to change.** This option creates a tornado chart that displays the effect on the target of modulating each simulated input by plus or minus one standard deviation of the distribution associated with the input.

Scatterplots of targets versus inputs. This option generates scatterplots of targets versus simulated inputs.

Box plots of target distributions. This option generates box plots of the target distributions.

Quartiles table. This option generates a table of the quartiles of the target distributions. The quartiles of a distribution are the 25th, 50th, and 75th percentiles of the distribution, and divide the observations into four groups of equal size.

Correlations and contingency table for inputs. This option displays a table of correlation coefficients between simulated inputs. A contingency table of associations between inputs with a categorical distribution is displayed when the simulation plan specifies generating categorical data from a contingency table.

Overlay results from separate targets. If the predictive model you are simulating contains multiple targets, you can specify whether results from separate targets are displayed on a single chart. This setting applies to charts for probability density functions, cumulative distribution functions and box plots. For example, if you select this option then the probability density functions for all targets will be displayed on a single chart.

Save the plan for this simulation. You can save any modifications to your simulation to a simulation plan file. Simulation plan files have the extension *.splan*. You can re-open the plan in the Run Simulation dialog or in the Simulation Builder. Simulation plans include all specifications except output settings.

Save the simulated data as a new data file. You can save simulated inputs, fixed inputs and predicted target values to an SPSS Statistics data file, a new dataset in the current session, or an Excel file. Each case (or row) of the data file consists of the predicted values of the targets along with the simulated

inputs and fixed inputs that generate the target values. When sensitivity analysis is specified, each iteration gives rise to a contiguous set of cases that are labeled with the iteration number.

If you require more customization of output than is available here, then consider running your simulation from the Simulation Builder. See the topic "To run a simulation from a simulation plan" on page 233 for more information.

Working with chart output from Simulation

A number of the charts generated from a simulation have interactive features that allow you to customize the display. Interactive features are available by activating (double-clicking) the chart object in the Output Viewer. All simulation charts are graphboard visualizations.

Probability density function charts for continuous targets. This chart has two sliding vertical reference lines that divide the chart into separate regions. The table below the chart displays the probability that the target is in each of the regions. If multiple density functions are displayed on the same chart, the table has a separate row for the probabilities associated with each density function. Each of the reference lines has a slider (inverted triangle) that allows you to easily move the line. A number of additional features are available by clicking the **Chart Options** button on the chart. In particular, you can explicitly set the positions of the sliders, add fixed reference lines and change the chart view from a continuous curve to a histogram or vice versa. See the topic "Chart Options " on page 246 for more information.

Cumulative distribution function charts for continuous targets. This chart has the same two moveable vertical reference lines and associated table described for the probability density function chart above. It also provides access to the Chart Options dialog, which allows you to explicitly set the positions of the sliders, add fixed reference lines and specify whether the cumulative distribution function is displayed as an increasing function (the default) or a decreasing function. See the topic "Chart Options " on page 246 for more information.

Bar charts for categorical targets with sensitivity analysis iterations. For categorical targets with sensitivity analysis iterations, results for the predicted target category are displayed as a clustered bar chart that includes the results for all iterations. The chart includes a dropdown list that allows you to cluster on category or on iteration. For Two-Step cluster models and K-Means cluster models, you can choose to cluster on cluster number or iteration.

Box plots for multiple targets with sensitivity analysis iterations. For predictive models with multiple continuous targets and sensitivity analysis iterations, choosing to display box plots for all targets on a single chart produces a clustered box plot. The chart includes a dropdown list that allows you to cluster on target or on iteration.

Chart Options

The Chart Options dialog allows you to customize the display of activated charts of probability density functions and cumulative distribution functions generated from a simulation.

View. The **View** dropdown list only applies to the probability density function chart. It allows you to toggle the chart view from a continuous curve to a histogram. This feature is not available when multiple density functions are displayed on the same chart. In that case, the density functions can only be viewed as continuous curves.

Order. The **Order** dropdown list only applies to the cumulative distribution function chart. It specifies whether the cumulative distribution function is displayed as an ascending function (the default) or a descending function. When displayed as a descending function, the value of the function at a given point on the horizontal axis is the probability that the target lies to the right of that point.

Slider positions. You can explicitly set the positions of the sliding reference lines by entering values in the Upper and Lower text boxes. You can remove the left-hand line by selecting **-Infinity**, effectively setting the position to negative infinity, and you can remove the right-hand line by selecting **Infinity**, effectively setting its position to infinity.

Reference lines. You can add various fixed vertical reference lines to probability density functions and cumulative distribution functions. When multiple functions are displayed on a single chart (because of

multiple targets or results from sensitivity analysis iterations), you can specify the particular functions to which the lines are applied.

- **Sigmas.** You can add reference lines at plus and minus a specified number of standard deviations from the mean of a target.
- **Percentiles.** You can add reference lines at one or two percentile values of the distribution of a target by entering values into the Bottom and Top text boxes. For example, a value of 95 in the Top text box represents the 95th percentile, which is the value below which 95% of the observations fall. Likewise, a value of 5 in the Bottom text box represents the 5th percentile, which is the value below which 5% of the observations fall.
- **Custom positions.** You can add reference lines at specified values along the horizontal axis.

Label reference lines. This option controls whether labels are applied to the selected reference lines.

Reference lines are removed by clearing the associated choice in the Chart Options dialog and clicking **Continue**.

Geospatial Modeling

Geospatial modeling techniques are designed to discover patterns in data that include a geospatial (map) component. The Geospatial Modeling Wizard provides methods for analyzing geospatial data with and without a time component.

Find associations based on event and geospatial data (Geospatial Associations Rules)
Using geospatial association rules, you can find patterns in data based on both the spatial and non-spatial properties. For example, you might identify patterns in crime data by location and demographic attributes. From these patterns, you can build rules that predict where certain types of crimes are likely to occur.

Make predictions using time series and geospatial data (Spatio-Temporal Prediction)
Spatial temporal prediction uses data that contains location data, input fields for prediction (predictors), one or more time fields, and a target field. Each location has numerous rows in the data that represent the values of each predictor and the target at each time interval.

Using the Geospatial Modeling Wizard

1. From the menus choose:

   Analyze > **Spatial and Temporal Modeling** > **Spatial Modeling**
2. Follow the steps in the wizard.

Examples

Detailed examples are available in the help system.

- Geospatial association rules: **Help** > **Topics** > **Case Studies** > **Statistics Base** > **Spatial association rules**
- Spatial temporal prediction: **Help** > **Topics** > **Case Studies** > **Statistics Base** > **Spatial temporal prediction**

Selecting Maps

Geospatial modeling can use one or more map data sources. Map data sources contain information that defines geographic areas and other geographic features, such as roads or rivers. Many map sources also contain demographic or other descriptive data and event data such as crime reports or unemployment rates. You can use a previously defined map specification file or define map specifications here and save those specifications for subsequent use.

Load a Map Specification
Loads a previously defined map specification (.mplan) file. Map data sources that you define here can be saved in a map specification file. For spatial temporal prediction, if you select a map specification file that identifies more than one map, you are prompted to select one map from the file.

Add Map File
Add an ESRI shape (.shp) file or .zip archive that contains an ESRI shape file.

- There must be a corresponding .dbf file in the same location as the .shp file, and that file must have the same root name as the .shp file.
- If the file is a .zip archive, the .shp and .dbf files must have the same root name as the .zip archive.
- If there is no corresponding projection (.prj) file, you are prompted to select a projection system.

Relationship
For geospatial association rules, this column defines how events relate to the features in the map. This setting is not available for spatial temporal prediction.

Move up, Move down
The layer order of the map elements is determined by the order in which they appear in the list. The first map in the list is the bottom layer.

Selecting a Map

For spatial temporal prediction, if you select a map specification file that identifies more than one map, your are prompted to select one map from the file. Spatial temporal prediction does not support multiple maps.

Geospatial Relationship

For geospatial association rules, the Geospatial Relationship dialog defines how events relate to the features in the map.

- This setting applies only to geospatial association rules.
- This setting only affects data sources associated with maps that are specified as context data on the step for selecting data sources.

Relationship

Close
The event occurs close to a specified point or area on the map.

Within
The event occurs within a specified area on the map.

Contains
The event area contains a map context object.

Intersects
Locations where lines or regions from the different maps intersect each other.

Cross
For multiple maps, locations where lines (for roads, rivers, railroads) from different lines cross each other.

North of, Sourth of, East of, West of
The event occurs within an area north, south, east, or west of a specified point on the map.

Set Coordinate System

If there is no projection (.prj) file with the map or you define two fields from a data source as a set of coordinates, you must set the coordinate system.

Default geographic (longitude and latitude)
The coordinate system is longitude and latitude.

Simple Cartesian (X and Y)
The coordinate system is simple X and Y coordinates.

Use a Well Known ID (WKID)
"Well known ID" for common projections.

Use a Coordinate System Name
The coordinate system is based on the named projection. The name is enclosed in parentheses.

Setting the Projection

If the projection system cannot be determined from the information provided with the map, you need to specify the projection system. The most common cause of this condition is the absence of a projection (.prj) file associated with the map or a projection file that cannot be used.

- **A city, region or country (Mercator)**
- **A large country, several countries, or continents (Winkel Tripel)**
- **An area very close to the equator (Mercator)**
- **An area close to one of the poles (Stereographic)**

The Mercator projection is a common projection used in many maps. This projection treats the globe as cylinder that is rolled out onto a flat surface. The Mercator projection distorts the size and shape of large objects. This distortion increases as you move farther from the equator and closer to the poles. The Winkel Tripel and Stereographic projections make adjustments for the fact that a map represents a portion of a three-dimensional sphere displayed in two dimensions.

Projection and Coordinate System

If you select more than one map and the maps have different projection and coordinate systems, you must select the map with the projection system that you want to use. That projection system will be used for all maps when they are combined together in the output.

Data Sources

A data source can be a dBase file that is provided with the shape file, an IBM SPSS Statistics data file, or an open dataset in the current session.

Context Data. A context data identifies features on the map. A context data can also contain fields that can be used as inputs for the model. To use a context dBase (.dbf) file that is associated with a map shape (.shp) file, the context dBase file must be in the same location as the shape file and must have the same root name. For example, if the shape file is `geodata.shp`, the dBase file must be named `geodata.dbf`

Event Data. Event data contains information on events that occur, such as crimes or accidents. This option is available only for geospatial association rules.

Point Density. Time interval and coordinate data for kernel density estimates. This option is available only for spatial temporal prediction.

Add. Opens a dialog for adding data sources. A data source can be a dBase file that is provided with the shape file, an IBM SPSS Statistics data file, or an open dataset in the current session.

Associate. Opens a dialog for specifying the identifiers (coordinates or keys) used to associate data with maps. Each data source must contain one or more identifiers that associate the data with the map. dBase files that come with a shape file typically contain a field that is automatically used as the default identifier. For other data sources, you must specify the fields that are used as identifiers.

Validate Key. Opens a dialog to validate key matching between the map and the data source.

Geospatial association rules

- At least one data source must be an event data source.
- All event data sources must use the same form of map association identifiers: coordinates or key values.

- If the event data sources are associated with the maps with key values, all event sources must use the same map feature type (for example, polygons, points, lines).

Spatial temporal prediction

- There must be a context data source.
- If there is only one data source (a data file with no associated map), it must include coordinate values.
- If you have two data sources, one data source must be context data, and the other data source must be point density data.
- You cannot include more than two data sources.

Add a Data Source

A data source can be a dBase file that is provided with the shape file and context file, an IBM SPSS Statistics data file, or an open dataset in the current session.

You can add the same data source multiple times if you want to use a different spatial association with each one.

Data and Map Association

Each data source must contain one or more identifiers that associate the data with the map.

Coordinates
: The data source contains fields that represent Cartesian coordinates, select the fields that represent that X and Y coordinates. For geospatial association rules, there can also be a Z coordinate.

Key values
: Key values in fields in the data source correspond to selected map keys. For example, a map of regions might have a name identifier (map key) labeling each region. That identifier corresponds with a field in the data that also contains the names of the regions (data key). Fields are matched to map keys based on the order they are displayed in the two lists.

Validate Keys

The Validate Keys dialog provides a summary of record matching between the map and the data source, based on the selected identifier keys. If there are unmatched data key values, you can manually match them to map key values.

Geospatial Association Rules

For geospatial association rules, after defining maps and data sources, the remaining steps in the wizard are:

- If there are multiple event data sources, define how event data sources are merged.
- Select fields to use as conditions and predictions in the analysis.

Optionally, you can also:

- Select different output options.
- Save a scoring model file.
- Create new fields for predicted values and rules in the data sources used in the model.
- Customize settings for building association rules.
- Customize binning and aggregation settings.

Define Event Data Fields

For geospatial association rules, if there is more than one event data source, the event data sources are merged.

- By default, only fields that are common to all event data sources are included.
- You can display a list of common fields, fields for a specific data source, or fields from all data sources and select the fields that you want to include.
- For common fields, the **Type** and **Measurement** must be the same for all data sources. If there are conflicts, you can specify the type and measurement level to use for each common field.

Select Fields

The list of available fields includes fields from the event data sources and fields from the context data sources.

- You can control the list of displayed fields by selecting a data source from the **Data Sources** list.
- You must select at least two fields. At least one must be a condition, and at least one must be a prediction. There are a number of ways to meet this requirement, including selecting two fields for the **Both (Condition and Prediction)** list.
- Association rules predict values of the prediction fields that are based on values of the condition fields. For example, in the rule "If x=1 and y=2, then z=3", the values of x and y are conditions, and the value of z is the prediction.

Output

Rules Tables
: Each rules table displays the top rules and values for confidence, rule support, lift, condition support, and deployability. Each table is sorted by values of the selected criterion. You can display all rules or the top **Number** of rules, based on the selected criterion.

Sortable Word Cloud
: A list of the top rules, based on the values of the selected criterion. The size of the text indicates the relative importance of the rule. The interactive output object contains the top rules for confidence, rule support, lift, condition support, and deployability. The selected criterion determines which list of rules is displayed by default. You can select a different criterion interactively in the output. **Max rules to display** determines the number of rules that are displayed in the output.

Map
: Interactive bar chart and map of the top rules, based on the selected criterion. Each interactive output object contains the top rules for confidence, rule support, lift, condition support, and deployability. The selected criterion determines which list of rules is displayed by default. You can select a different criterion interactively in the output. **Max rules to display** determines the number of rules that are displayed in the output.

Model Information Tables

Field Transformations.
: Describes the transformations that are applied to fields used in the analysis.

Record Summary.
: Number and percentage of included and excluded records.

Rule Statistics.
: Summary statistics for condition support, confidence, rule support, lift, and deployability. The statistics include mean, minimum, maximum, and standard deviation.

Most Frequent Items.
: Items that occur most frequently. An item is included in a condition or a prediction in a rule. For example, age < 18 or gender=female.

Most Frequent Fields.
: Fields that occur most frequently in the rules.

Excluded Inputs.
: Fields that are excluded from the analysis and the reason each field was excluded.

Criterion for Rules Tables, Word Cloud, and Maps

Confidence.
The percentage of correct rule predictions.

Rule Support.
The percentage of cases for which the rule is true. For example, if the rule is "If x=1 and y=2, then z=3," rule support is the actual percentage of cases in the data for which x=1, y=2, and z=3.

Lift.
Lift is a measure of how much the rule improves prediction compared to random chance. It is the ratio of correct predictions to the overall occurrence of the predicted value. The value must be greater than 1. For example, if the predicted value occurs 20% of the time and the confidence in the prediction is 80%, then the lift value is 4.

Condition Support.
The percentage of cases for which the rule condition exists. For example, if the rule is "If x=1 and y=2, then z=3," condition support is the proportion of cases in the data for which x=1 and y=2.

Deployability.
The percentage of incorrect predictions when the conditions are true. Deployability is equal to (1-confidence) multiplied by condition support or condition support minus rule support.

Save

Save the map and context data as a map specification
Save the map specifications to an external file (.mplan). You can load this map specification file into the wizard for subsequent analyses. You can also use the map specification file with the `SPATIAL ASSOCIATION RULES` command.

Copy any map and data files into the specification
Data from map shape files, external data files, and datasets used in the map specification are saved in the map specification file.

Scoring
Saves best rule values, confidence values for the rules, and numeric ID values for the rules as new fields in the specified data source.

Data Source to Score
The data source or sources where the new fields are created. If the data source is not open in the current session, it is opened in the current session. You must explicitly save the modified file to save the new fields.

Target Values
Create new fields for the selected target (prediction) fields.

- Two new fields are created for each target field: predicted value and confidence value.
- For continuous (scale) target fields, the predicted value is a string that describes a value range. A value of the form "(value1, value2]" means "greater than value1 and less than or equal to value2."

Number of best rules
Create new fields for the number of best rules specified. Three new fields are created for each rule: rule value, confidence value, and a numeric ID value for the rule.

Name Prefix
Prefix to use for the new field names.

Rule Building

Rule building parameters set the criteria for the generated association rules.

Items Per Rule
Number of field values that can be included in rule conditions and predictions. The total number of items cannot exceed 10. For example, in the rule "If x=1 and y=2, then z=3", there are two condition items and one prediction item.

Maximum predictions.
Maximum number of field values that can occur in the predictions for a rule.

Maximum conditions.
Maximum number of field values that can occur in the conditions for a rule.

Exclude Pair
Excludes the specified pairs of fields from being included in the same rule.

Rule Criteria

Confidence.
Minimum confidence a rule must have to be included in the output. Confidence is the percentage of correct predictions.

Rule Support.
Minimum rule support a rule must have to be included in the output. The value represents the percentage of cases for which the rule is true in the observed data. For example, if the rule is "If x=1 and y=2, then z=3," rule support is the actual percentage of cases in the data for which x=1, y=2, and z=3.

Condition Support.
Minimum condition support a rule must have to be included in the output. The value represents the percentage of cases for which the condition exists. For example, if the rule is "If x=1 and y=2, then z=3," condition support is the percentage of cases in the data for which x=1 and y=2.

Lift.
Minimum lift a rule must have to be included in the output. Lift is a measure of how much the rule improves prediction over random chance. It is the ratio of correct predictions to the overall occurrence of the predicted value. For example, if the predicted value occurs 20% of the time and the confidence in the prediction is 80%, then the lift value is 4.

Treat as same
Identifies pairs of fields that should be treated as the same field.

Binning and Aggregation

- Aggregation is necessary when there are more records in the data than there are features in the map. For example, you have data records for individual counties but you have a map of states.
- You can specify the aggregate summary measure method for continuous and ordinal fields. Nominal fields are aggregated based on the modal value.

Continuous
For continuous (scale) fields, the summary measure can be mean, median, or sum.

Ordinal
For ordinal fields, the summary measure can be median, mode, highest, or lowest.

Number of bins
Sets the maximum number of bins for continuous (scale) fields. Continuous fields are always grouped or "binned" into ranges of values. For example: less than or equal to 5, greater than 5 and less than or equal to 10, or greater than 10.

Aggregate the map
Apply aggregation to both data and maps.

Custom settings for specific fields
You can override the default summary measure and number of bins for specific fields.

- Click the icon to open the **Field Chooser** dialog and select a field to add to the list.
- In the **Aggregation** column, select a summary measure.

- For continuous fields, click the button in the **Bins** column to specify a custom number of bins for the field in the **Bins** dialog.

Spatial Temporal Prediction

For spatial temporal prediction, after defining maps and data sources, the remaining steps in the wizard are:

- Specify the target field, time fields, and optional predictors.
- Define time intervals or cyclic periods for time fields.

Optionally, you can also:

- Select different output options.
- Customize model building parameters.
- Customize aggregation settings.
- Save predicted values to a dataset in the current session or to an IBM SPSS Statistics format data file.

Select Fields

The list of available fields includes fields from the selected data sources. You can control the list of displayed fields by selecting a data source from the **Data Sources** list.

Target
A target field is required. The target is the field for which values are predicted.

- The target field must be a continuous (scale), numeric field.
- If there are two data sources, the target is kernel density estimates, and "Density" is displayed as the target name. You cannot change this selection.

Predictors
One or more predictor fields can be specified. This setting is optional.

Time Fields
You must select one or more fields that represent time periods or select **Cyclic Periods**.

- If there are two data sources, you must select time fields from both data sources. Both time fields must represent the same interval.
- For cyclic periods, you must specify the fields that define periodicity cycles in the **Time Intervals panel** of the wizard.

Time Intervals

The options in this panel are based on the choice of **Time Fields** or **Cyclic period** in the step for selecting fields.

Time fields

Selected Time Fields. If you select one or more time fields in the step for selecting fields, those fields are displayed in this list.

Time Interval. Select the appropriate time interval from the list. Depending on the time interval, you can also specify other settings, such as interval between observations (increment) and starting value. This time interval is used for all selected time fields.

- The procedure assumes that all cases (records) represent equally spaced intervals.
- Based on the selected time interval, the procedure can detect missing observations or multiple observations in the same time interval that need to be aggregated together. For example, if the time interval is days and the date 2014-10-27 is followed by 2014-10-29, then there is a missing observation for 2014-10-28. If the time interval is month, then multiple dates in the same month are aggregated together.

- For some time intervals, the additional setting can define breaks in the normal equally spaced intervals. For example, if the time interval is days, but only weekdays are valid, you can specify that there are five days in a week, and the week begins on Monday.
- If the selected time field is not a date format or time format field, the time interval is automatically set to **Periods** and cannot be changed.

Cycle fields

If you select **Cyclic period** on the step for selecting fields, you must specify the fields that define the cyclic periods. A cyclic period identifies repetitive cyclical variation, such as the number of months in a year or the number of days in a week.

- You can specify up to three fields that define cyclic periods.
- The first cycle field represents the highest level of the cycle. For example, if there is cyclic variation by year, quarter, and month, the field that represents year is the first cycle field.
- The cycle length for the first and second cycle fields is the periodicity at the subsequent level. For example, if the cycle fields are year, quarter, and month, the first cycle length is 4 and the second cycle length is 3.
- The starting value for the second and third cycle fields is the first value in each of those cyclic periods.
- Cycle length and starting values must be positive integers.

Aggregation

- If you select any **Predictors** in the step for selecting fields, you can select the aggregation summary method for the predictors.
- Aggregation is necessary when there is more than one record in a defined time interval. For example, if the time interval is month, then multiple dates in the same month are aggregated together.
- You can specify the aggregation summary measure method for continuous and ordinal fields. Nominal fields are aggregated based on the modal value.

Continuous
For continuous (scale) fields, the summary measure can be mean, median, or sum.

Ordinal
For ordinal fields, the summary measure can be median, mode, highest, or lowest.

Custom settings for specific fields
You can override the default aggregation summary measure for specific predictors.

- Click the icon to open the **Field Chooser** dialog and select a field to add to the list.
- In the **Aggregation** column, select a summary measure.

Output

Maps

Target values.
Map of values for the selected target field.

Correlation
Map of correlations.

Clusters
Map that highlights clusters of locations that are similar to each other. Maps of clusters are available only for empirical models.

Location similarity threshold.
The similarity that is required to create clusters. The value must be a number greater than zero and less than 1.

Specify maximum number of clusters.
The maximum number of clusters to display.

Model Evaluation Tables

Model Specifications.
Summary of specifications that are used to run the analysis, including target, input, and location fields.

Temporal Information Summary.
Identifies the time fields and time intervals that are used in the model.

Test of Effects in Mean Structure.
The output includes test statistics value, degrees of freedom, and significance level for the model and each effect.

Mean Structure of Model Coefficients.
The output includes the coefficient value, standard error, test statistics value, significance level, and confidence intervals for each model term.

Autoregressive Coefficients.
The output includes the coefficient value, standard error, test statistics value, significance level, and confidence intervals for each lag.

Tests of Spatial Covariance.
For variogram-based parametric models, displays the goodness of fit test results for spatial covariance structure. The test results can determine whether to model the spatial covariance structure parametrically or to use a nonparametric model.

Parametric Spatial Covariance.
For variogram-based parametric models, displays parameter estimates for parametric spatial covariance.

Model Options

Model Settings

Automatically include an intercept
Include the intercept in the model.

Maximum autoregression lag
The maximum autoregression lag. The value must be an integer between 1 and 5.

Spatial Covariance
Specifies the estimation method for spatial covariance.

Parametric
The estimation method is parametric. The method can be **Gaussian**, **Exponential** or **Power Exponential**. For Power Exponential, you can specify the **Power** value.

Nonparametric
The estimation method is nonparametric.

Save

Save the map and context data as a map specification
Save the map specifications to an external file (.mplan). You can load this map specification file into the wizard for subsequent analysis. You can also use the map specification file with the `SPATIAL TEMPORAL PREDICTION` command.

Copy any map and data files into the specification
Data from map shape files, external data files, and datasets that are used in the map specification are saved in the map specification file.

Scoring
Saves predicted values, variance, and upper and lower confidence bounds for the target field in the selected data file.

- You can save predicted values to an open dataset in the current session or an IBM SPSS Statistics format data file.
- The data file cannot be a data source that is used in the model.
- The data file must contain all the time fields and predictors that are used in the model.
- The time values must be greater than the time values used in the model.

Advanced

Maximum cases with missing values (%)
The maximum percentage of cases with missing values.

Significance level
The significance level for determining whether a variogram-based parametric model is appropriate. The value must be greater than 0 and less than 1. The default value is 0.05. The significance level is used in the goodness of fit test for spatial covariance structure. The goodness of fit statistic is used to determine whether to use a parametric or non-parametric model.

Uncertainty factor (%)
The uncertainty factor is a percentage value that represents the growth in uncertainty for future forecasts. The upper and lower limits of forecast uncertainty increase by the specified percentage for each step into the future.

Finish

In the last step of the Geospatial Modeling Wizard you can either run the model or paste the generated command syntax to a syntax window. You can modify and save the generated syntax for subsequent use.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© Copyright IBM Corp. 2021. Portions of this code are derived from IBM Corp. Sample Programs.

© Copyright IBM Corp. 1989 - 2021. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

A

adjusted R 2
in Linear Regression 140
adjusted R-square
in linear models 132
Agresti-Caffo
in Independent-Samples Proportions 105
Agresti-Coull
in One-Sample Proportions 100
Agresti-Min
in Paired-Samples Proportions 102
Akaike information criterion
in linear models 132
alpha coefficient
in Reliability Analysis 212, 213
alpha factoring 170
Analysis dialog
Meta-Analysis Binary 53
Meta-Analysis Binary Effect Size 61
Meta-Analysis Continuous 38
Meta-Analysis Continuous Effect Size 45
analysis of variance
in Curve Estimation 154
in Linear Regression 140
in Means 94
in One-Way ANOVA 111
Anderson-Rubin factor scores 172
Andrews' wave estimator
in Explore 85
ANOVA
in GLM Univariate 115
in linear models 135
in Means 94
in One-Way ANOVA 111
model 116
Anscombe
in One-Sample Proportions 100
Asymptotic significance level 213
automatic data preparation
in linear models 134
automatic distribution fitting
in simulation 236
auxiliary regression model
in GLM 125
average absolute deviation (AAD)
in Ratio Statistics 221

B

backward elimination
backward elimination *(continued)*
in Linear Regression 137
bagging
in linear models 131
bar charts
in Frequencies 82
Bartlett factor scores 172
Bartlett's test of sphericity
in Factor Analysis 170
best subsets
in linear models 132
beta coefficients
in Linear Regression 140
Bias dialog
Meta-Analysis Binary 56
Meta-Analysis Binary Effect Size 64
Meta-Analysis Continuous 40
Meta-Analysis Continuous Effect Size 47
binary outcomes
pre-calculated effect size data 35
raw effect size data 35
binomial test
One-Sample Nonparametric Tests 185, 186
Binomial Test
command additional features 196
dichotomies 196
missing values 196
options 196
statistics 196
Bivariate Correlations
command additional features 128
confidence interval 126
confidence intervals 127
correlation coefficients 126
missing values 127
options 127
significance level 126
statistics 127
block distance
in Distances 130
Bonett-Price
in Paired-Samples Proportions 102
Bonferroni
in GLM 120
in One-Way ANOVA 112
boosting
in linear models 131
Box's M test
in Discriminant Analysis 167
boxplots
comparing factor levels 86
comparing variables 86
in Explore 86

boxplots *(continued)*
in simulation 242
Brown-Forsythe statistic
in One-Way ANOVA 114
Brown-Li-Jeffreys
in Independent-Samples Proportions 105
Bubble Plot
Meta-Analysis Regression 74
Bubble Plot tab
Meta-Analysis Binary 74
Meta-Analysis Binary Effect Size 74
Meta-Analysis Continuous 74
Meta-Analysis Continuous Effect Size 74
build terms 116, 144, 145

C

case-control study
Paired-Samples T Test 108
casewise diagnostic information
in Linear Regression 140
charts
case labels 154
in ROC Analysis 227
in ROC Curve 230
probability plots 222, 224
Chebychev distance
in Distances 130
chi-square
expected range 195
expected values 195
Fisher's exact test 89
for independence 89
in Crosstabs 89
likelihood-ratio 89
linear-by-linear association 89
missing values 195
one-sample test 194
options 195
Pearson 89
statistics 195
Yates' correction for continuity 89
chi-square distance
in Distances 130
chi-square test
One-Sample Nonparametric Tests 185, 186
city-block distance
in Nearest Neighbor Analysis 160
classification
in ROC Analysis 227
in ROC Curve 230
classification table
in Nearest Neighbor Analysis 165
Clopper-Pearson (Exact)
in One-Sample Proportions 100
Clopper-Pearson intervals
Clopper-Pearson intervals *(continued)*
One-Sample Nonparametric Tests 186
cluster analysis
efficiency 183
Hierarchical Cluster Analysis 180
K-Means Cluster Analysis 182
cluster frequencies
in TwoStep Cluster Analysis 175
cluster viewer
about cluster models 176
basic view 178
cell content display 178
cell distribution view 179
cluster centers view 177
cluster comparison view 179
cluster display sort 178
cluster predictor importance view 178
cluster sizes view 178
clusters view 177
comparison of clusters 179
distribution of cells 179
feature display sort 177
filtering records 180
flip clusters and features 177
model summary 176
overview 176
predictor importance 178
size of clusters 178
sort cell contents 178
sort clusters 178
sort features 177
summary view 176
transpose clusters and features 177
using 179
clustering
choosing a procedure 172
overall display 176
viewing clusters 176
Cochran's Q
in Tests for Several Related Samples 203
Cochran's Q test
Related-Samples Nonparametric Tests 192, 193
Cochran's statistic
in Crosstabs 89
Codebook
output 78
statistics 79
coefficient of dispersion (COD)
in Ratio Statistics 221
coefficient of variation (COV)
in Ratio Statistics 221
Cohen's kappa
in Crosstabs 89
Cohen's Weighted Kappa 216, 217
collinearity diagnostic information
in Linear Regression 140
column percentages
in Crosstabs 90
column proportions statistics
in Crosstabs 90
column summary reports 210
combining rules

combining rules *(continued)*
in linear models 133
comparing groups
in OLAP Cubes 98
comparing variables
in OLAP Cubes 98
compound model
in Curve Estimation 155
concentration index
in Ratio Statistics 221
confidence interval
in Bivariate Correlations 126
confidence intervals
in Bivariate Correlations 127
in Explore 85
in GLM 117
in Independent-Samples T Test 108
in Linear Regression 140
in One-Sample T Test 110
in One-Way ANOVA 114
in Paired-Samples T Test 109
in ROC Analysis 228
in ROC Curve 230
saving in Linear Regression 139
contingency coefficient
in Crosstabs 89
contingency tables 87
continuous outcomes
Meta Analysis 35
pre-calculated effect size data 35
raw effect size data 35
Contrast dialog
Meta-Analysis Binary 55
Meta-Analysis Binary Effect Size 63
Meta-Analysis Continuous 40
Meta-Analysis Continuous Effect Size 47
contrasts
in GLM 117, 118
in One-Way ANOVA 112
control variables
in Crosstabs 88
convergence
in Factor Analysis 170, 171
in K-Means Cluster Analysis 183
Cook's distance
in GLM 123
in Linear Regression 139
correlation matrix
in Discriminant Analysis 167
in Factor Analysis 169, 170
in Ordinal Regression 143
correlations
in Bivariate Correlations 126
in Crosstabs 89
in Partial Correlations 128
in simulation 239
zero-order 129
covariance matrix
in Discriminant Analysis 167, 168
in GLM 123
in Linear Regression 140
covariance matrix *(continued)*
in Ordinal Regression 143
covariance ratio
in Linear Regression 139
Cox and Snell R2
in Ordinal Regression 143
Cramér's V
in Crosstabs 89
Criteria dialog
Meta-Analysis Binary 52
Meta-Analysis Binary Effect Size 60
Meta-Analysis Continuous 37, 44
Meta-Analysis Regression 68
Cronbach's alpha
in Reliability Analysis 212, 213
Crosstabs
cell display 90
clustered bar charts 88
control variables 88
formats 91
layers 88
statistics 89
suppressing tables 87
crosstabulation
in Crosstabs 87
multiple response 206
cubic model
in Curve Estimation 155
cumulative distribution functions
in simulation 241
Cumulative Forest Plot 71
Cumulative Forest Plot tab
Meta-Analysis Binary 73
Meta-Analysis Binary Effect Size 73
Meta-Analysis Continuous 73
Meta-Analysis Continuous Effect Size 73
cumulative frequencies
in Ordinal Regression 143
Curve Estimation
analysis of variance 154
forecast 155
including constant 154
models 155
saving predicted values 155
saving prediction intervals 155
saving residuals 155
custom models
in GLM 116

D

d
in Crosstabs 89
Define Multiple Response Sets
categories 205
dichotomies 205
set labels 205
set names 205
deleted residuals
in GLM 123
in Linear Regression 139
dendrograms

dendrograms *(continued)*
in Hierarchical Cluster Analysis 181
dependent t test
in Paired-Samples T Test 108
descriptive statistics
in Descriptives 82
in Explore 85
in Frequencies 81
in Ratio Statistics 221
in Summarize 92
in TwoStep Cluster Analysis 175
Descriptives
command additional features 84
display order 83
saving z scores 82
statistics 83
designs for heteroskedasticity tests
in GLM 125
detrended normal plots
in Explore 86
deviation contrasts
in GLM 117, 118
DfBeta
in Linear Regression 139
DfFit
in Linear Regression 139
dictionary
Codebook 77
difference contrasts
in GLM 117, 118
differences between groups
in OLAP Cubes 98
differences between variables
in OLAP Cubes 98
direct oblimin rotation
in Factor Analysis 171
Discriminant Analysis
command additional features 169
covariance matrix 168
criteria 167
defining ranges 166
descriptive statistics 167
discriminant methods 167
display options 167, 168
example 165
exporting model information 168
function coefficients 167
grouping variables 165
independent variables 165
Mahalanobis distance 167
matrices 167
missing values 168
plots 168
prior probabilities 168
Rao's V 167
saving classification variables 168
selecting cases 166
statistics 165, 167
stepwise methods 165
Wilks' lambda 167
distance measures
in Distances 130
in Hierarchical Cluster Analysis 181
distance measures *(continued)*
in Nearest Neighbor Analysis 160
Distances
command additional features 131
computing distances between cases 129
computing distances between variables 129
dissimilarity measures 130
example 129
similarity measures 130
statistics 129
transforming measures 130
transforming values 130
distribution fitting
in simulation 236
division
dividing across report columns 211
Duncan's multiple range test
in GLM 120
in One-Way ANOVA 112
Dunnett's C
in GLM 120
in One-Way ANOVA 112
Dunnett's t test
in GLM 120
in One-Way ANOVA 112
Dunnett's T3
in GLM 120
in One-Way ANOVA 112
Durbin-Watson statistic
in Linear Regression 140

E

effect size
in Independent-Samples T Test 107
in Paired-Samples T-Test 108
eigenvalues
in Factor Analysis 170
in Linear Regression 140
ensembles
in linear models 133
equamax rotation
in Factor Analysis 171
error summary
in Nearest Neighbor Analysis 165
eta
in Crosstabs 89
in Means 94
eta-squared
in Means 94
Euclidean distance
in Distances 130
in Nearest Neighbor Analysis 160
Exact Binomial
in Paired-Samples Proportions 103
expected count
in Crosstabs 90
expected frequencies
in Ordinal Regression 143

Explore
command additional features 87
missing values 87
options 87
plots 86
power transformations 86
statistics 85
exponential model
in Curve Estimation 155
extreme values
in Explore 85

F

F statistic
in linear models 132
Factor Analysis
coefficient display format 172
command additional features 172
convergence 170, 171
descriptives 170
example 169
extraction methods 170
factor scores 172
loading plots 171
missing values 172
overview 169
rotation methods 171
selecting cases 170
statistics 169, 170
factor scores 172
feature selection
in Nearest Neighbor Analysis 165
feature space chart
in Nearest Neighbor Analysis 163
first
in Means 94
in OLAP Cubes 96
in Summarize 92
Fisher's exact test
in Crosstabs 89
Fisher's LSD
in GLM 120
Fleiss' Multiple Rater Kappa 212, 213
forecast
in Curve Estimation 155
Forest Plot 71
Forest Plot tab
Meta-Analysis Binary 71
Meta-Analysis Binary Effect Size 71
Meta-Analysis Continuous 71
Meta-Analysis Continuous Effect Size 71
formatting
columns in reports 209
forward selection
in Linear Regression 137
in Nearest Neighbor Analysis 160
forward stepwise
in linear models 132
Frequencies
charts 82
display order 82
formats 82
Frequencies *(continued)*
statistics 81
suppressing tables 82
frequency tables
in Explore 85
in Frequencies 80
Friedman test
in Tests for Several Related Samples 203
Related-Samples Nonparametric Tests 192
full factorial models
in GLM 116
Funnel Plot 71
Funnel Plot tab
Meta-Analysis Binary 75
Meta-Analysis Binary Effect Size 75
Meta-Analysis Continuous 75
Meta-Analysis Continuous Effect Size 75

G

Gabriel's pairwise comparisons test
in GLM 120
in One-Way ANOVA 112
Galbraith Plot 71
Galbraith Plot tab
Meta-Analysis Binary 76
Meta-Analysis Binary Effect Size 76
Meta-Analysis Continuous 76
Meta-Analysis Continuous Effect Size 76
Games and Howell's pairwise comparisons test
in GLM 120
in One-Way ANOVA 112
gamma
in Crosstabs 89
generalized least squares
in Factor Analysis 170
geometric mean
in Means 94
in OLAP Cubes 96
in Summarize 92
geospatial modeling 247–257
GLM
model 116
post hoc tests 120
profile plots 118
saving matrices 123
saving variables 123
sum of squares 116
GLM Univariate
contrasts 117, 118
Goodman and Kruskal's gamma
in Crosstabs 89
Goodman and Kruskal's lambda
in Crosstabs 89
Goodman and Kruskal's tau
in Crosstabs 89
goodness of fit
in Ordinal Regression 143
grand totals
in column summary reports 211
group means 93, 96

grouped median
in Means 94
in OLAP Cubes 96
in Summarize 92
growth model
in Curve Estimation 155
Guttman model
in Reliability Analysis 212, 213

H

Hampel's redescending M-estimator
in Explore 85
harmonic mean
in Means 94
in OLAP Cubes 96
in Summarize 92
Hauck-Anderson
in Independent-Samples Proportions 105
Helmert contrasts
in GLM 117, 118
Hierarchical Cluster Analysis
agglomeration schedules 181
cluster membership 181, 182
clustering cases 180
clustering methods 181
clustering variables 180
command additional features 182
dendrograms 181
distance matrices 181
distance measures 181
example 180
icicle plots 181
plot orientation 181
saving new variables 182
similarity measures 181
statistics 180, 181
transforming measures 181
transforming values 181
hierarchical decomposition 117
histograms
in Explore 86
in Frequencies 82
in Linear Regression 138
Hochberg's GT2
in GLM 120
in One-Way ANOVA 112
Hodges-Lehman estimates
Related-Samples Nonparametric Tests 192
holdout sample
in Nearest Neighbor Analysis 161
homogeneity-of-variance tests
in One-Way ANOVA 114
Hotelling's T 2
in Reliability Analysis 212, 213
Huber's M-estimator
in Explore 85

I

ICC. See intraclass correlation coefficient 213
icicle plots
in Hierarchical Cluster Analysis 181
image factoring 170
Independent-Samples Nonparametric Tests
Fields tab 189
Independent-Samples Proportions 104, 105
Independent-Samples T Test
confidence intervals 108
missing values 108
options 108
Independent-Samples T-Test
defining groups 107
grouping variables 107
string variables 107
Inference dialog
Meta-Analysis Binary 54
Meta-Analysis Binary Effect Size 62
Meta-Analysis Continuous 39
Meta-Analysis Continuous Effect Size 46
Meta-Analysis Regression 69
information criteria
in linear models 132
initial threshold
in TwoStep Cluster Analysis 174
interaction terms 116, 144, 145
Interrater Agreement 213
intraclass correlation coefficient (ICC)
in Reliability Analysis 213
inverse model
in Curve Estimation 155
iteration history
in Ordinal Regression 143
iterations
in Factor Analysis 170, 171
in K-Means Cluster Analysis 183

J

Jeffreys
in One-Sample Proportions 100
Jeffreys intervals
One-Sample Nonparametric Tests 186

K

k and feature selection
in Nearest Neighbor Analysis 165
k selection
in Nearest Neighbor Analysis 165
K-Means Cluster Analysis
cluster distances 183
cluster membership 183
command additional features 184
convergence criteria 183

K-Means Cluster Analysis *(continued)*
efficiency 183
examples 182
iterations 183
methods 182
missing values 184
overview 182
saving cluster information 183
statistics 182, 184
Kaiser normalization
in Factor Analysis 171
kappa
in Crosstabs 89
Kendall's coefficient of concordance (W)
Related-Samples Nonparametric Tests 192
Kendall's tau-b
in Bivariate Correlations 126
in Crosstabs 89
Kendall's tau-c
in Crosstabs 89
Kendall's W
in Tests for Several Related Samples 203
Kolmogorov-Smirnov test
One-Sample Nonparametric Tests 185, 186
Kolmogorov-Smirnov Test
Lilliefors Test 186
Kolmogorov-Smirnov Z
in One-Sample Kolmogorov-Smirnov Test 197
in Two-Independent-Samples Tests 200
KR20
in Reliability Analysis 213
Kruskal-Wallis H
in Two-Independent-Samples Tests 202
Kruskal's tau
in Crosstabs 89
Kuder-Richardson 20 (KR20)
in Reliability Analysis 213
kurtosis
in Descriptives 83
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Report Summaries in Columns 211
in Report Summaries in Rows 209
in Summarize 92

L

L'Abb´e Plot 71
L'Abb´e Plot tab
Meta-Analysis Binary 76
Meta-Analysis Binary Effect Size 76
Meta-Analysis Continuous 76
Meta-Analysis Continuous Effect Size 76
lambda
in Crosstabs 89
Lance and Williams dissimilarity measure
in Distances 130
last
in Means 94
in OLAP Cubes 96
in Summarize 92
layers
in Crosstabs 88
least significant difference
in GLM 120
in One-Way ANOVA 112
Levene test
in Explore 86
in One-Way ANOVA 114
leverage values
in GLM 123
in Linear Regression 139
likelihood ratio intervals
One-Sample Nonparametric Tests 186
likelihood-ratio chi-square
in Crosstabs 89
in Ordinal Regression 143
Lilliefors test
in Explore 86
Lilliefors Test 186, 197, 199
Linear Elastic Net Regression
options 146
partition 145
variables 145
Linear Lasso Regression
options 149
partition 148
variables 148
linear model
in Curve Estimation 155
linear models
ANOVA table 135
automatic data preparation 132, 134
coefficients 135
combining rules 133
confidence level 132
ensembles 133
estimated means 136
information criterion 134
model building summary 136
model options 134
model selection 132
model summary 134
objectives 131
outliers 135
predicted by observed 135
predictor importance 134
R-square statistic 134
replicating results 133
residuals 135
Linear Regression
blocks 137
command additional features 141
exporting model information 139
missing values 141
plots 138
residuals 139
saving new variables 139

Linear Regression *(continued)*
selection variable 138
statistics 140
variable selection methods 137, 141
weights 137
Linear Ridge Regression
options 152
partition 151
variables 151
linear-by-linear association
in Crosstabs 89
link
in Ordinal Regression 142
listing cases 91
loading plots
in Factor Analysis 171
location model
in Ordinal Regression 144
logarithmic model
in Curve Estimation 155
logistic model
in Curve Estimation 155
Logit
in One-Sample Proportions 100

M

M-estimators
in Explore 85
Mahalanobis distance
in Discriminant Analysis 167
in Linear Regression 139
Manhattan distance
in Nearest Neighbor Analysis 160
Mann-Whitney U
in Two-Independent-Samples Tests 200
Mantel-Haenszel statistic
in Crosstabs 89
marginal homogeneity test
in Two-Related-Samples Tests 201
Related-Samples Nonparametric Tests 192
matched-pairs study
in Paired-Samples T Test 108
maximum
comparing report columns 211
in Descriptives 83
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Ratio Statistics 221
in Summarize 92
maximum branches
in TwoStep Cluster Analysis 174
maximum likelihood
in Factor Analysis 170
McDonald's Omega
in Reliability Analysis 212, 213
McFadden R2
in Ordinal Regression 143
McNemar
in Paired-Samples Proportions 103
McNemar (continuity corrected)
in Paired-Samples Proportions 103
McNemar test
in Crosstabs 89
in Two-Related-Samples Tests 201
Related-Samples Nonparametric Tests 192, 193
mean
in Descriptives 83
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in One-Way ANOVA 114
in Ratio Statistics 221
in Report Summaries in Columns 211
in Report Summaries in Rows 209
in Summarize 92
of multiple report columns 211
subgroup 93, 96
Means
options 94
statistics 94
measures of central tendency
in Explore 85
in Frequencies 81
in Ratio Statistics 221
measures of dispersion
in Descriptives 83
in Explore 85
in Frequencies 81
in Ratio Statistics 221
measures of distribution
in Descriptives 83
in Frequencies 81
median
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Ratio Statistics 221
in Summarize 92
median test
in Two-Independent-Samples Tests 202
memory allocation
in TwoStep Cluster Analysis 174
Meta Analysis
binary outcomes 35
continuous outcomes 35
regression 35
Meta-Analysis
Plots 71
Meta-Analysis Binary
Analysis dialog 53
Bias dialog 56
Bubble Plot tab 74
Contrast dialog 55
Criteria dialog 52
Cumulative Forest Plot tab 73

Meta-Analysis Binary *(continued)*
Forest Plot tab 71
Funnel Plot tab 75
Galbraith Plot tab 76
Inference dialog 54
L'Abb´e Plot tab 76
Plot dialog 59
Print dialog 58
Save dialog 58
Trim-and-Fill dialog 56
Meta-Analysis Binary Effect Size
Analysis dialog 61
Bias dialog 64
Bubble Plot tab 74
Contrast dialog 63
Criteria dialog 60
Cumulative Forest Plot tab 73
Forest Plot tab 71
Funnel Plot tab 75
Galbraith Plot tab 76
Inference dialog 62
L'Abb´e Plot tab 76
Plot dialog 67
Print dialog 66
Save dialog 66
Trim-and-Fill dialog 64
Meta-Analysis Continuous
Analysis dialog 38
Bias dialog 40
Bubble Plot tab 74
Contrast dialog 40
Criteria dialog 37, 44
Cumulative Forest Plot tab 73
Forest Plot tab 71
Funnel Plot tab 75
Galbraith Plot tab 76
Inference dialog 39
L'Abb´e Plot tab 76
Plot dialog 43
Print dialog 42
Save dialog 43
Trim-and-Fill dialog 41
Meta-Analysis Continuous Effect Size
Analysis dialog 45
Bias dialog 47
Bubble Plot tab 74
Contrast dialog 47
Cumulative Forest Plot tab 73
Forest Plot tab 71
Funnel Plot tab 75
Galbraith Plot tab 76
Inference dialog 46
L'Abb´e Plot tab 76
Plot dialog 51
Print dialog 49
Save dialog 50
Trim-and-Fill dialog 48
Meta-Analysis Regression
Bubble Plot 74
Criteria dialog 68
Inference dialog 69
Plot dialog 71
Print dialog 70
Meta-Analysis Regression *(continued)*
Save dialog 70
Mid-p Adjusted Binomial
in Paired-Samples Proportions 103
minimum
comparing report columns 211
in Descriptives 83
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Ratio Statistics 221
in Summarize 92
Minkowski distance
in Distances 130
missing values
in Binomial Test 196
in Bivariate Correlations 127
in Chi-Square Test 195
in column summary reports 211
in Explore 87
in Factor Analysis 172
in Independent-Samples Proportions 106
in Independent-Samples T Test 108
in Linear Regression 141
in Multiple Response Crosstabs 207
in Multiple Response Frequencies 205
in Nearest Neighbor Analysis 162
in One-Sample Kolmogorov-Smirnov Test 199
in One-Sample Proportions 101
in One-Sample T Test 110
in One-Way ANOVA 114
in Paired-Samples Proportions 103
in Paired-Samples T Test 109
in Partial Correlations 129
in Report Summaries in Rows 209
in ROC Analysis 228
in ROC Curve 230
in Runs Test 197
in Tests for Several Independent Samples 203
in Two-Independent-Samples Tests 201
in Two-Related-Samples Tests 202
Missing Values
in Percentiles 84
mode
in Frequencies 81
model view
in Nearest Neighbor Analysis 162
Monte Carlo simulation 231
Moses extreme reaction test
in Two-Independent-Samples Tests 200
Multidimensional Scaling
command additional features 220
conditionality 219
creating distance matrices 219
criteria 219
defining data shape 218
dimensions 219
display options 219
distance measures 219
example 218
levels of measurement 219

Multidimensional Scaling *(continued)*
scaling models 219
statistics 218
transforming values 219
multiple comparisons
in One-Way ANOVA
112
multiple R
in Linear Regression 140
multiple regression
in Linear Regression 137
Multiple Response
command additional features 207
multiple response analysis
crosstabulation 206
frequency tables 205
Multiple Response Crosstabs 206
Multiple Response Frequencies 205
Multiple Response Crosstabs
cell percentages 207
defining value ranges 207
matching variables across response sets 207
missing values 207
percentages based on cases 207
percentages based on responses 207
Multiple Response Frequencies
missing values 205
multiple response sets
Codebook 77
multiplication
multiplying across report columns 211

N

Nagelkerke R2
in Ordinal Regression 143
Nearest Neighbor Analysis
feature selection 160
model view 162
neighbors 160
options 162
output 162
partitions 161
saving variables 162
nearest neighbor distances
in Nearest Neighbor Analysis 164
Newcombe
in Independent-Samples Proportions
105
in Paired-Samples Proportions 102
Newcombe (continuity corrected)
in Independent-Samples Proportions
105
Newman-Keuls
in GLM 120
noise handling
in TwoStep Cluster Analysis 174
nonparametric tests
chi-square 194
One-Sample Kolmogorov-Smirnov Test 197
Runs Test 197
Tests for Several Independent Samples 202
Tests for Several Related Samples 203
Two-Independent-Samples Tests 199
nonparametric tests *(continued)*
Two-Related-Samples Tests 201
normal probability plots
in Explore 86
in Linear Regression 138
P-P 222
Q-Q 224
normality tests
in Explore 86
number of cases
in Means 94
in OLAP Cubes 96
in Summarize 92

O

observed count
in Crosstabs 90
observed frequencies
in Ordinal Regression 143
OLAP Cubes
statistics 96
titles 98
One-Sample Kolmogorov-Smirnov
Test
command additional features 199
Lilliefors Test 197, 199
missing values 199
options 199
statistics 199
test distribution 197
One-Sample Nonparametric
Tests
binomial test 186
chi-square test 186
fields 185
Kolmogorov-Smirnov test 186
runs test 187
One-Sample Proportions 99, 100
One-Sample T Test
command additional features 109, 110
confidence intervals 110
missing values 110
options 110
One-Sample Wilcoxon Tests 188
One-Samples Proportions 102
One-Way ANOVA
command additional features 114
contrasts 112
factor variables 111
missing values 114
multiple comparisons 112
options 114
polynomial contrasts 112
post hoc tests 112
statistics 114
Ordinal Regression
command additional features 145
link 142
location model 144
options 142
scale model 144
statistics 142
outliers

outliers *(continued)*
in Explore 85
in Linear Regression 138
in TwoStep Cluster Analysis 174
overfit prevention criterion
in linear models 132

P

P-P probability plots 222
page control
in column summary reports 211
in row summary reports 209
page numbering
in column summary reports 211
in row summary reports 209
Paired-Samples Proportions 101
Paired-Samples T Test
missing values 109
options 109
selecting paired variables 108
Paired-Samples *t*-Test 108
parallel model
in Reliability Analysis 212, 213
parameter estimates
in Ordinal Regression 143
Partial Correlations
command additional features 129
in Linear Regression 140
missing values 129
options 129
statistics 129
zero-order correlations 129
Partial Least Squares Regression
export variables 158
model 157
partial plots
in Linear Regression 138
pattern difference measure
in Distances 130
pattern matrix
in Factor Analysis 169
Pearson chi-square
in Crosstabs 89
in Ordinal Regression 143
Pearson correlation
in Bivariate Correlations 126
in Crosstabs 89
Pearson residuals
in Ordinal Regression 143
peers
in Nearest Neighbor Analysis 164
percentages
in Crosstabs 90
percentiles
in Explore 85
in Frequencies 81
in simulation 242
Percentiles 84
phi
in Crosstabs 89
phi-square distance measure
in Distances 130
pie charts
pie charts *(continued)*
in Frequencies 82
Plot dialog
Meta-Analysis Binary 59
Meta-Analysis Binary Effect Size 67
Meta-Analysis Continuous 43
Meta-Analysis Continuous Effect Size 51
Meta-Analysis Regression 71
Plots
Meta-Analysis 71
PLUM
in Ordinal Regression 142
polynomial contrasts
in GLM 117, 118
in One-Way ANOVA 112
post hoc multiple comparisons 112
Power Analysis
statistics 1
power model
in Curve Estimation 155
predicted values
saving in Curve Estimation 155
saving in Linear Regression 139
prediction intervals
saving in Curve Estimation 155
saving in Linear Regression 139
predictor importance
linear models 134
price-related differential (PRD)
in Ratio Statistics 221
principal axis factoring 170
principal components analysis 169, 170
Print dialog
Meta-Analysis Binary 58
Meta-Analysis Binary Effect Size 66
Meta-Analysis Continuous 42
Meta-Analysis Continuous Effect Size 49
Meta-Analysis Regression 70
probability density functions
in simulation 241
probability plots
P-P 222
Q-Q 224
profile plots
in GLM 118
Proximities
in Hierarchical Cluster Analysis 180

Q

Q-Q probability plots 224
quadrant map
in Nearest Neighbor Analysis 165
quadratic model
in Curve Estimation 155
quartiles
in Frequencies 81
quartimax rotation
in Factor Analysis 171

R

R 2
in Linear Regression 140
in Means 94
R 2 change 140
r correlation coefficient
in Bivariate Correlations 126
in Crosstabs 89
R statistic
in Linear Regression 140
in Means 94
R-E-G-W F
in GLM 120
in One-Way ANOVA 112
R-E-G-W Q
in GLM 120
in One-Way ANOVA 112
R-square
in linear models 134
range
in Descriptives 83
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Ratio Statistics 221
in Summarize 92
rank correlation coefficient
in Bivariate Correlations 126
Rao's V
in Discriminant Analysis 167
Ratio Statistics
statistics 221
reference category
in GLM 117, 118
regression
Linear Regression 137
Meta Analysis 35
multiple regression 137
plots 138
regression coefficients
in Linear Regression 140
related samples 201, 203
Related-Samples Nonparametric Tests
Cochran's Q test 193
fields 192
McNemar test 193
relative risk
in Crosstabs 89
Reliability Analysis
ANOVA table 213
command additional features 216
descriptives 213
example 212
Hotelling's T 2 213
inter-item correlations and covariances 213
intraclass correlation coefficient 213
Kuder-Richardson 20 213
statistics 212, 213
Tukey's test of additivity 213
repeated contrasts
repeated contrasts *(continued)*
in GLM 117, 118
Report Summaries in Columns
column format 209
command additional features 212
grand total 211
missing values 211
page control 211
page layout 209
page numbering 211
subtotals 211
total columns 211
Report Summaries in Rows
break columns 208
break spacing 209
column format 209
command additional features 212
data columns 208
footers 210
missing values 209
page control 209
page layout 209
page numbering 209
sorting sequences 208
titles 210
variables in titles 210
reports
column summary reports 210
comparing columns 211
composite totals 211
dividing column values 211
multiplying column values 211
row summary reports 208
total columns 211
residuals
in Crosstabs 90
saving in Curve Estimation 155
saving in Linear Regression 139
rho
in Bivariate Correlations 126
in Crosstabs 89
risk
in Crosstabs 89
ROC Analysis
statistics and plots 228
ROC Curve
statistics and plots 230
row percentages
in Crosstabs 90
runs test
One-Sample Nonparametric Tests 185, 187
Runs Test
command additional features 197
cut points 197
missing values 197
options 197
statistics 197
Ryan-Einot-Gabriel-Welsch multiple F
in GLM 120
in One-Way ANOVA 112
Ryan-Einot-Gabriel-Welsch multiple range
in GLM 120
in One-Way ANOVA 112

S

S model
in Curve Estimation 155
S-stress
in Multidimensional Scaling 218
Save dialog
Meta-Analysis Binary 58
Meta-Analysis Binary Effect Size 66
Meta-Analysis Continuous 43
Meta-Analysis Continuous Effect Size 50
Meta-Analysis Regression 70
scale
in Multidimensional Scaling 218
in Reliability Analysis 212
in Weighted Kappa 216
scale model
in Ordinal Regression 144
scatterplot
in simulation 242
scatterplots
in Linear Regression 138
Scheffé test
in GLM 120
in One-Way ANOVA 112
Score 100
Score (Continuity Corrected) 100
selection criteria
in Linear Regression 140
selection variable
in Linear Regression 138
sensitivity analysis
in simulation 239
Shapiro-Wilk's test
in Explore 86
Sidak's t test
in GLM 120
in One-Way ANOVA 112
sign test
in Two-Related-Samples Tests 201
Related-Samples Nonparametric Tests 192
similarity measures
in Distances 130
in Hierarchical Cluster Analysis 181
simple contrasts
in GLM 117, 118
simulation
box plots 242
chart options 246
correlations between inputs 239
creating a simulation plan 231, 232
creating new inputs 235
cumulative distribution function 241
customizing distribution fitting 238
display formats for targets and inputs 242
distribution fitting 236
distribution fitting results 238
equation editor 235
interactive charts 246
model specification 234
simulation *(continued)*
output 241, 242
percentiles of target distributions 242
probability density function 241
refitting distributions to new data 244
running a simulation plan 233, 243
save simulated data 243
save simulation plan 243
scatter plots 242
sensitivity analysis 239
Simulation Builder 234
stopping criteria 240
supported models 234
tail sampling 240
tornado charts 242
what-if analysis 239
Simulation Builder 234
size difference measure
in Distances 130
skewness
in Descriptives 83
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Report Summaries in Columns 211
in Report Summaries in Rows 209
in Summarize 92
Somers' d
in Crosstabs 89
spatial modeling 247
Spearman correlation coefficient
in Bivariate Correlations 126
in Crosstabs 89
Spearman-Brown reliability
in Reliability Analysis 213
split-half reliability
in Reliability Analysis 212, 213
spread-versus-level plots
in Explore 86
squared Euclidean distance
in Distances 130
standard deviation
in Descriptives 83
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Ratio Statistics 221
in Report Summaries in Columns 211
in Report Summaries in Rows 209
in Summarize 92
standard error
in Descriptives 83
in Explore 85
in Frequencies 81
in GLM 123
in ROC Analysis 228
in ROC Curve 230
standard error of kurtosis
in Means 94
in OLAP Cubes 96
in Summarize 92
standard error of skewness

standard error of skewness *(continued)*
in Means 94
in OLAP Cubes 96
in Summarize 92
standard error of the mean
in Means 94
in OLAP Cubes 96
in Summarize 92
standardization
in TwoStep Cluster Analysis 174
standardized residuals
in GLM 123
in Linear Regression 139
standardized values
in Descriptives 82
stem-and-leaf plots
in Explore 86
stepwise selection
in Linear Regression 137
stress
in Multidimensional Scaling 218
strictly parallel model
in Reliability Analysis 212, 213
Student-Newman-Keuls
in GLM 120
in One-Way ANOVA
112
Student's t test 107
Studentized residuals
in Linear Regression 139
subgroup means 93, 96
subtotals
in column summary reports 211
sum
in Descriptives 83
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Summarize 92
sum of squares
in GLM 116
Summarize
options 92
statistics 92

T

t test
in Independent-Samples T Test
107
in One-Sample T Test 109
in Paired-Samples T Test 108
Tamhane's T2
in GLM 120
in One-Way ANOVA
112
tau-b
in Crosstabs 89
tau-c
in Crosstabs 89
test of parallel lines
in Ordinal Regression 143
tests for independence
chi-square 89
tests for linearity
in Means 94
Tests for Several Independent Samples
command additional features 203
defining range 203
grouping variables 203
missing values 203
options 203
statistics 203
test types 202
Tests for Several Related Samples
command additional features 204
statistics 204
test types 203
time series analysis
forecast 155
predicting cases 155
titles
in OLAP Cubes 98
tolerance
in Linear Regression 140
Tolerance
Linear Regression 141
tornado charts
in simulation 242
total column
in reports 211
total percentages
in Crosstabs 90
training sample
in Nearest Neighbor Analysis 161
transformation matrix
in Factor Analysis 169
tree depth
in TwoStep Cluster Analysis 174
Trim-and-Fill dialog
Meta-Analysis Binary 56
Meta-Analysis Binary Effect Size 64
Meta-Analysis Continuous 41
Meta-Analysis Continuous Effect Size
48
trimmed mean
in Explore 85
Tukey's b test
in GLM 120
in One-Way ANOVA
112
Tukey's biweight estimator
in Explore 85
Tukey's honestly significant difference
in GLM 120
in One-Way ANOVA 112
Tukey's test of additivity
in Reliability Analysis 212, 213
Two-Independent-Samples Tests
command additional features 201
defining groups 200
grouping variables 200
missing values 201
options 201
statistics 201
test types 200
Two-Related-Samples Tests
command additional features 202

Two-Related-Samples Tests *(continued)*
missing values 202
options 202
statistics 202
test types 201
two-sample t test
in Independent-Samples T Test 107
TwoStep Cluster Analysis
options 174
save to external file 175
save to working file 175
statistics 175

U

uncertainty coefficient
in Crosstabs 89
unstandardized residuals
in GLM 123
unweighted least squares
in Factor Analysis 170

V

V
in Crosstabs 89
variable importance
in Nearest Neighbor Analysis 164
variance
in Descriptives 83
in Explore 85
in Frequencies 81
in Means 94
in OLAP Cubes 96
in Report Summaries in Columns 211
in Report Summaries in Rows 209
in Summarize 92
variance inflation factor
in Linear Regression 140
varimax rotation
in Factor Analysis 171
visualization
clustering models 176

W

Wald
in Independent-Samples Proportions 105
in One-Sample Proportions 100
in Paired-Samples Proportions 102, 103
Wald (continuity corrected)
in Independent-Samples Proportions 105
in One-Sample Proportions 100
in Paired-Samples Proportions 102
Wald (Continuity Corrected)
in Independent-Samples Proportions 105
in Paired-Samples Proportions 103
Wald H0
Wald H0 *(continued)*
in Independent-Samples Proportions 105
Wald H0 (Continuity Corrected)
in Independent-Samples Proportions 105
Wald-Wolfowitz runs
in Two-Independent-Samples Tests 200
Waller-Duncan t test
in GLM 120
in One-Way ANOVA 112
Weighted Kappa
criteria 217
crosstabulation 217
example 216
print 217
statistics 216, 217
weighted least squares
in Linear Regression 137
weighted mean
in Ratio Statistics 221
weighted predicted values
in GLM 123
Welch statistic
in One-Way ANOVA 114
what-if analysis
in simulation 239
Wilcoxon signed-rank test
in Two-Related-Samples Tests 201
One-Sample Nonparametric Tests 185
Related-Samples Nonparametric Tests 192
Wilks' lambda
in Discriminant Analysis 167
Wilson Score
in One-Sample Proportions 100
Wilson Score (continuity corrected)
in One-Sample Proportions 100

Y

Yates' correction for continuity
in Crosstabs 89

Z

z scores
in Descriptives 82
saving as variables 82
zero-order correlations
in Partial Correlations 129

IBM®