JMP STATISTICAL DISCOVERY

Version 17

Basic Analysis

> *"The real voyage of discovery consists not in seeking new landscapes,*
> *but in having new eyes."*
> Marcel Proust

JMP Statistical Discovery LLC
SAS Campus Drive
Cary, North Carolina 27513-2414

17.0

The correct bibliographic citation for this manual is as follows: JMP Statistical Discovery LLC 2022. *JMP® 17 Basic Analysis*. Cary, NC: JMP Statistical Discovery LLC

JMP® 17 Basic Analysis

JMP Statistical Discovery LLC, SAS Campus Drive, Cary, North Carolina 27513-2414.

October 2022

Get the Most from JMP

Whether you are a first-time or a long-time user, there is always something to learn about JMP.

Visit JMP.com to find the following:

- live and recorded webcasts about how to get started with JMP
- video demos and webcasts of new features and advanced techniques
- details on registering for JMP training
- schedules for seminars being held in your area
- success stories showing how others use JMP
- the JMP user community, resources for users including examples of add-ins and scripts, a forum, blogs, conference information, and so on

jmp.com/getstarted

Contents

Basic Analysis

4 Introduction to Fit Y by X 101

Examine Relationships between Two Variables

5 Bivariate Analysis 103

Examine Relationships between Two Continuous Variables

6 Oneway Analysis 149

Examine Relationships between a Continuous Y and a Categorical X Variable

7 Contingency Analysis 233

Examine Relationships between Two Categorical Variables

8 Logistic Analysis 271

Examine Relationships between a Categorical Y and a Continuous X Variable

9 Tabulate

10 Simulate

11 Bootstrapping

12 Text Explorer 375

Explore Unstructured Text in Your Data

Chapter 1

Learn about JMP

Documentation and Additional Resources

Learn about JMP documentation, such as book conventions, descriptions of each JMP document, the Help system, and where to find additional support.

Contents

Formatting Conventions in JMP Documentation

These conventions help you relate written material to information that you see on your screen:

- Sample data table names, column names, pathnames, filenames, file extensions, and folders appear in Helvetica (or sans-serif online) font.
- Code appears in `Lucida Sans Typewriter` (or monospace online) font.
- Code output appears in *`Lucida Sans Typewriter`* italic (or monospace italic online) font and is indented farther than the preceding code.
- **Helvetica bold** formatting (or bold sans-serif online) indicates items that you select to complete a task:
  - buttons
  - check boxes
  - commands
  - list names that are selectable
  - menus
  - options
  - tab names
  - text boxes
- The following items appear in italics:
  - words or phrases that are important or have definitions specific to JMP
  - book titles
  - variables
- Features that are for JMP Pro only are noted with the JMP Pro icon JMP PRO. For an overview of JMP Pro features, visit jmp.com/software/pro.

Note: Special information and limitations appear within a Note.

Tip: Helpful information appears within a Tip.

JMP Help

JMP Help in the Help menu enables you to search for information about JMP features, statistical methods, and the JMP Scripting Language (or *JSL*). You can open JMP Help in several ways:

- Search and view JMP Help on Windows by selecting **Help > JMP Help**.
- On Windows, press the F1 key to open the Help system in the default browser.
- Get help on a specific part of a data table or report window. Select the Help tool ? from the **Tools** menu and then click anywhere in a data table or report window to see the Help for that area.
- Within a JMP window, click the **Help** button.

Note: The JMP Help is available for users with Internet connections. Users without an Internet connection can search all books in a PDF file by selecting **Help > JMP Documentation Library**. See "JMP Documentation Library" for more information.

JMP Documentation Library

The Help system content is also available in one PDF file called *JMP Documentation Library*. Select **Help > JMP Documentation Library** to open the file. You can also download the Documentation PDF Files add-in if you prefer searching individual PDF files of each document in the JMP library. Download the available add-ins from https://community.jmp.com.

The following table describes the purpose and content of each document in the JMP library.

Document Title	Document Purpose	Document Content
Discovering JMP	If you are not familiar with JMP, start here.	Introduces you to JMP and gets you started creating and analyzing data. Also learn how to share your results.
Using JMP	Learn about JMP data tables and how to perform basic operations.	Covers general JMP concepts and features that span across all of JMP, including importing data, modifying columns properties, sorting data, and connecting to SAS.

Document Title	Document Purpose	Document Content
Basic Analysis	Perform basic analysis using this document.	Describes these Analyze menu platforms: • Distribution • Fit Y by X • Tabulate • Text Explorer Covers how to perform bivariate, one-way ANOVA, and contingency analyses through Analyze > Fit Y by X. How to approximate sampling distributions using bootstrapping and how to perform parametric resampling with the Simulate platform are also included.
Essential Graphing	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Scatterplot Matrix • Ternary Plot • Treemap • Chart • Overlay Plot The book also covers how to create background and custom maps.
Profilers	Learn how to use interactive profiling tools, which enable you to view cross-sections of any response surface.	Covers all profilers listed in the Graph menu. Analyzing noise factors is included along with running simulations using random inputs.

Document Title	**Document Purpose**	**Document Content**
Design of Experiments Guide	Learn how to design experiments and determine appropriate sample sizes.	Covers all topics in the DOE menu.
Fitting Linear Models	Learn about Fit Model platform and many of its personalities.	Describes these personalities, all available within the Analyze menu Fit Model platform: • Standard Least Squares • Stepwise • Generalized Regression • Mixed Model • Generalized Linear Mixed Model • MANOVA • Loglinear Variance • Nominal Logistic • Ordinal Logistic • Generalized Linear Model

Document Title	**Document Purpose**	**Document Content**
Predictive and Specialized Modeling	Learn about additional modeling techniques.	Describes these Analyze > Predictive Modeling menu platforms: • Neural • Partition • Bootstrap Forest • Boosted Tree • K Nearest Neighbors • Naive Bayes • Support Vector Machines • Model Comparison • Model Screening • Make Validation Column • Formula Depot Describes these Analyze > Specialized Modeling menu platforms: • Fit Curve • Nonlinear • Functional Data Explorer • Gaussian Process • Time Series • Matched Pairs Describes these Analyze > Screening menu platforms: • Explore Outliers • Explore Missing Values • Explore Patterns • Response Screening • Predictor Screening • Association Analysis • Process History Explorer

Document Title	Document Purpose	Document Content
Multivariate Methods	Read about techniques for analyzing several variables simultaneously.	Describes these Analyze > Multivariate Methods menu platforms: • Multivariate • Principal Components • Discriminant • Partial Least Squares • Multiple Correspondence Analysis • Structural Equation Models • Factor Analysis • Multidimensional Scaling • Multivariate Embedding • Item Analysis Describes these Analyze > Clustering menu platforms: • Hierarchical Cluster • K Means Cluster • Normal Mixtures • Latent Class Analysis • Cluster Variables

Document Title	**Document Purpose**	**Document Content**
Quality and Process Methods	Read about tools for evaluating and improving processes.	Describes these Analyze > Quality and Process menu platforms: • Control Chart Builder and individual control charts • Measurement Systems Analysis (EMP and Type 1 Gauge) • Variability / Attribute Gauge Charts • Process Screening • Process Capability • Model Driven Multivariate Control Chart • Legacy Control Charts • Pareto Plot • Diagram • Manage Spec Limits • OC Curves
Reliability and Survival Methods	Learn to evaluate and improve reliability in a product or system and analyze survival data for people and products.	Describes these Analyze > Reliability and Survival menu platforms: • Life Distribution • Fit Life by X • Cumulative Damage • Recurrence Analysis • Degradation • Repeated Measures Degradation • Destructive Degradation • Reliability Forecast • Reliability Growth • Reliability Block Diagram • Repairable Systems Simulation • Survival • Fit Parametric Survival • Fit Proportional Hazards

Document Title	Document Purpose	Document Content
Consumer Research	Learn about methods for studying consumer preferences and using that insight to create better products and services.	Describes these Analyze > Consumer Research menu platforms: • Categorical • Choice • MaxDiff • Uplift • Multiple Factor Analysis
Genetics	Learn about methods that are available in JMP to help you analyze your genetic data and use that data to simulate a breeding program to predict the optimum genetic crosses to make.	Describes these Analyze > Genetics menu platforms: • Marker Statistics • Marker Simulation
Scripting Guide	Learn about taking advantage of the powerful JMP Scripting Language (JSL).	Covers a variety of topics, such as writing and debugging scripts, manipulating data tables, constructing display boxes, and creating JMP applications.
JSL Syntax Reference	Read about many JSL functions on functions and their arguments, and messages that you send to objects and display boxes.	Includes syntax, examples, and notes for JSL commands.

Additional Resources for Learning JMP

In addition to reading JMP help, you can also learn about JMP using the following resources:

- "Search JMP"
- "JMP Tutorials"
- "Sample Data Tables"
- "Learn about Statistical and JSL Terms"
- "Learn JMP Tips and Tricks"
- "JMP Tooltips"
- "JMP User Community"
- "Free Online Statistical Thinking Course"
- "JMP New User Welcome Kit"
- "Statistics Knowledge Portal"
- "JMP Training"
- "JMP Books by Users"
- "The JMP Starter Window"

Search JMP

If you are not sure where to find a statistical procedure, do a search across JMP. Results are tailored to the window that you launch the search from, such as a data table or report.

1. Click **Help** > **Search JMP**. Or, press Ctrl+comma.
2. Enter your search text.
3. Click the result that contains the procedure that you want.

   On the right, you can see a description and the location of the procedure.
4. Click the corresponding button to open or go to a result.

JMP Tutorials

You can access JMP tutorials by selecting **Help > Tutorials**. The first item on the **Tutorials** menu is **Tutorials Directory**. This opens a new window with all the tutorials grouped by category.

If you are not familiar with JMP, start with the **Beginners Tutorial**. It steps you through the JMP interface and explains the basics of using JMP.

The rest of the tutorials help you with specific aspects of JMP, such as designing an experiment and comparing a sample mean to a constant.

Sample Data Tables

All of the examples in the JMP documentation suite use sample data. Select **Help > Sample Data Folder** to open the sample data directory.

To view an alphabetized list of sample data tables or view sample data within categories, select **Help > Sample Index**.

Sample data tables are installed in the following directory:

On Windows: C:\Program Files\SAS\JMP\17\Samples\Data

On macOS: \Library\Application Support\JMP\17\Samples\Data

In JMP Pro, sample data is installed in the JMPPRO (rather than JMP) directory.

To view examples using sample data, select **Help > Sample Index** and navigate to the Teaching Resources section. To learn more about the teaching resources, visit jmp.com/tools.

Learn about Statistical and JSL Terms

For help with statistical terms, select Help > Statistics Index. For help with JSL scripting and examples, select **Help > Scripting Index**.

Statistics Index Provides definitions of statistical terms.

Scripting Index Lets you search for information about JSL functions, objects, and display boxes. You can also edit and run sample scripts from the Scripting Index and get help on the commands.

Learn JMP Tips and Tricks

When you first start JMP, you see the Tip of the Day window. This window provides tips for using JMP.

To turn off the Tip of the Day, clear the **Show tips at startup** check box. To view it again, select **Help > Tip of the Day**. Or, you can turn it off using the Preferences window.

JMP Tooltips

JMP provides descriptive tooltips (or *hover labels*) when you hover over items, such as the following:

- Menu or toolbar options
- Labels in graphs
- Text results in the report window (move your cursor in a circle to reveal)

- Files or windows in the Home Window
- Code in the Script Editor

Tip: On Windows, you can hide tooltips in the JMP Preferences. Select **File > Preferences > General** and then deselect **Show menu tips**. This option is not available on macOS.

JMP User Community

The JMP User Community provides a range of options to help you learn more about JMP and connect with other JMP users. The learning library of one-page guides, tutorials, and demos is a good place to start. And you can continue your education by registering for a variety of JMP training courses.

Other resources include a discussion forum, sample data and script file exchange, webcasts, and social networking groups.

To access JMP resources on the website, select **Help > JMP on the Web > JMP User Community** or visit https://community.jmp.com.

Free Online Statistical Thinking Course

Learn practical statistical skills in this free online course on topics such as exploratory data analysis, quality methods, and correlation and regression. The course consists of short videos, demonstrations, exercises, and more. Visit jmp.com/statisticalthinking.

JMP New User Welcome Kit

The JMP New User Welcome Kit is designed to help you quickly get comfortable with the basics of JMP. You’ll complete its thirty short demo videos and activities, build your confidence in using the software, and connect with the largest online community of JMP users in the world. Visit jmp.com/welcome.

Statistics Knowledge Portal

The Statistics Knowledge Portal combines concise statistical explanations with illuminating examples and graphics to help visitors establish a firm foundation upon which to build statistical skills. Visit jmp.com/skp.

JMP Training

SAS offers training on a variety of topics led by a seasoned team of JMP experts. Public courses, live web courses, and on-site courses are available. You might also choose the online e-learning subscription to learn at your convenience. Visit jmp.com/training.

JMP Books by Users

Additional books about using JMP that are written by JMP users are available on the JMP website. Visit jmp.com/books.

The JMP Starter Window

The JMP Starter window is a good place to begin if you are not familiar with JMP or data analysis. Options are categorized and described, and you launch them by clicking a button. The JMP Starter window covers many of the options found in the Analyze, Graph, Tables, and File menus. The window also lists JMP Pro features and platforms.

- To open the JMP Starter window, select **View** (**Window** on macOS) **> JMP Starter**.
- To display the JMP Starter automatically when you open JMP on Windows, select **File > Preferences > General**, and then select **JMP Starter** from the Initial JMP Window list. On macOS, select **JMP > Preferences > Initial JMP Starter Window**.

JMP Technical Support

JMP technical support is provided by statisticians and engineers educated in SAS and JMP, many of whom have graduate degrees in statistics or other technical disciplines.

Many technical support options are provided at jmp.com/support, including the technical support phone number.

Chapter **2**

Introduction to Basic Analysis

Overview of Fundamental Analysis Methods

Basic Analysis describes the initial types of analyses that you often perform in JMP:

- The Distribution platform illustrates the distribution of a single variable using histograms, additional graphs, and reports. Once you know how your data is distributed, you can plan the appropriate type of analysis going forward. See "Distributions".
- The Fit Y by X platform analyzes the pair of X and Y variables that you specify, by context, based on modeling type. See "Introduction to Fit Y by X". The four types of analyses include:
  - The Bivariate platform, which analyzes the relationship between two continuous X variables. See "Bivariate Analysis".
  - The Oneway platform, which analyzes how the distribution of a continuous Y variable differs across groups defined by a categorical X variable. See "Oneway Analysis".
  - The Contingency platform, which analyzes the distribution of a categorical response variable Y as conditioned by the values of a categorical X factor. See "Contingency Analysis".
  - The Logistic platform, which fits the probabilities for response categories (Y) to a continuous X predictor. See "Logistic Analysis".
- The Tabulate platform interactively constructs tables of descriptive statistics. See "Tabulate".
- The Simulate feature provides parametric and nonparametric simulation capability. See "Simulate".
- Bootstrap analysis approximates the sampling distribution of a statistic. The data is re-sampled with replacement and the statistic is computed. This process is repeated to produce a distribution of values for the statistic. See "Bootstrapping".
- The Text Explorer platform enables you to categorize and analyze unformatted text data. You can use regular expressions to clean up the data before you proceed to analysis. See "Text Explorer".

Chapter 3

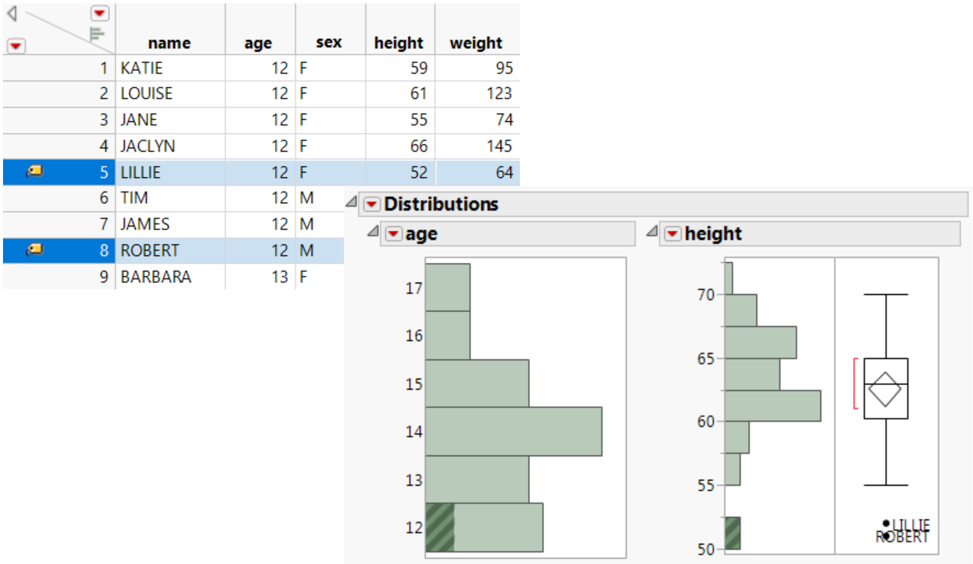
Distributions

Using the Distribution Platform

Use the Distribution platform to explore the distribution of a single variable using histograms, box plots, and summary statistics. You can perform several types of hypothesis tests, including *t* tests, *z* tests, chi-squared tests, and equivalence tests. You can also construct confidence, tolerance, and prediction intervals, and estimate process capability. The word *univariate* simply means involving one variable instead of two (bivariate) or many (multivariate). However, you can examine the distribution of several individual variables within a report. The report content for each variable depends on whether the variable is categorical (nominal or ordinal) or continuous.

The Distribution report window is interactive. Clicking on a histogram bar highlights the corresponding data in any other histograms and in the data table.

Figure 3.1 Example of the Distribution Platform



	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M		
7	JAMES	12	M		
8	ROBERT	12	M		
9	BARBARA	13	F		

Contents

Overview of the Distribution Platform

You can use the Distribution platform to analyze both categorical and continuous variables. The treatment of variables in the Distribution platform is different, depending on the modeling type of the variable, which can be categorical (nominal or ordinal) or continuous.

Categorical Variables

For categorical variables, the initial graph that appears is a histogram. The histogram shows a bar for each level of the ordinal or nominal variable. You can also add a divided (mosaic) bar chart.

The Frequencies report show counts and proportions. You can add confidence intervals and test the probabilities from the options in the red triangle menu.

Continuous Variables

For numeric continuous variables, the initial graphs show a histogram and an outlier box plot. The histogram shows a bar for grouped values of the continuous variable. The following options are also available:

- normal quantile plot
- quantile box plot
- stem and leaf plot
- CDF plot

The reports show selected quantiles and summary statistics. Additional report options are available in the red triangle menu for the following:

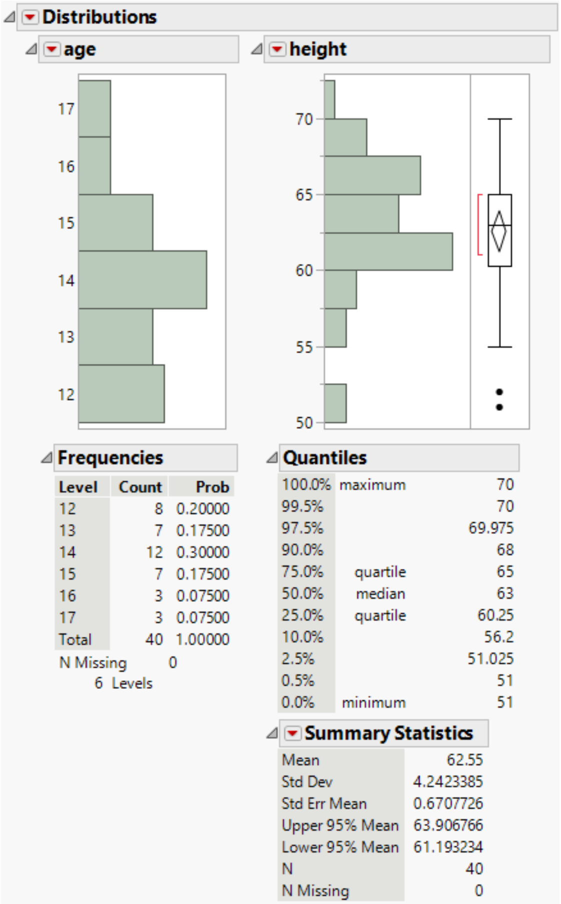
- saving ranks, probability scores, normal quantile values, and so on, as new columns in the data table (Save options)
- testing the mean and standard deviation of the column against a constant you specify (Test Mean and Test Std Dev options)
- fitting various distributions and nonparametric smoothing curves (Continuous Fit and Discrete Fit options)
- performing a process capability analysis for a quality control application
- confidence intervals, prediction intervals, and tolerance intervals

Example of the Distribution Platform

In this example, you have data on the ages and heights of 40 students. You want to visualize the distribution of the variables and find outliers in those distributions.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Distribution**.
3. Select age and height and click **Y, Columns**.
4. Click **OK**.

Figure 3.2 Example of the Distribution Platform



Frequencies

Level	Count	Prob
12	8	0.20000
13	7	0.17500
14	12	0.30000
15	7	0.17500
16	3	0.07500
17	3	0.07500
Total	40	1.00000

N Missing 0
6 Levels

Quantiles

100.0%	maximum	70
99.5%		70
97.5%		69.975
90.0%		68
75.0%	quartile	65
50.0%	median	63
25.0%	quartile	60.25
10.0%		56.2
2.5%		51.025
0.5%		51
0.0%	minimum	51

Summary Statistics

Mean	62.55
Std Dev	4.2423385
Std Err Mean	0.6707726
Upper 95% Mean	63.906766
Lower 95% Mean	61.193234
N	40
N Missing	0

From the histograms, you notice the following:

- The ages are not uniformly distributed.

- For height, there are two points with extreme values (that might be outliers).

Click the bar for 50 in the height histogram to take a closer look at the potential outliers.

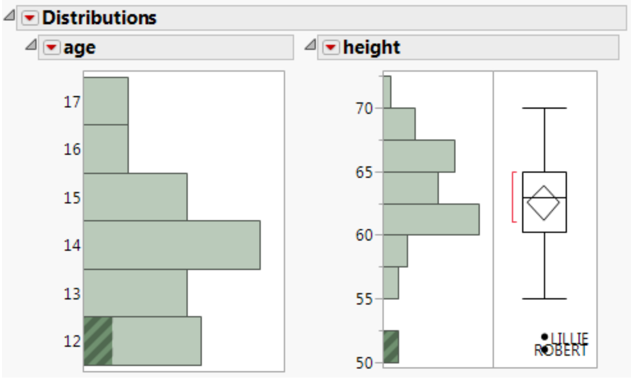
- The corresponding ages are highlighted in the age histogram. The potential outliers are age 12.
- The corresponding rows are highlighted in the data table. The names of the potential outliers are Lillie and Robert.

Add labels to the potential outliers in the height histogram.

1. Select both outliers.
2. Right-click one of the outliers and select **Row Label**.

   Label icons are added to these rows in the data table.
3. Resize the box plot wider to see the full labels.

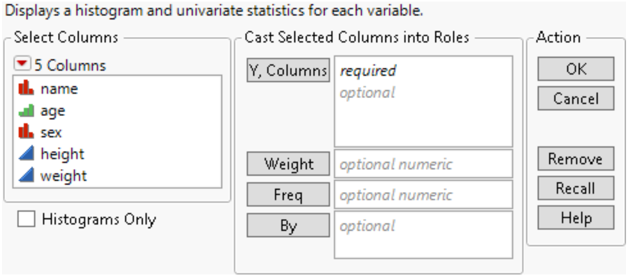
Figure 3.3 Potential Outliers Labeled



Launch the Distribution Platform

Launch the Distribution platform by selecting **Analyze > Distribution**.

Figure 3.4 The Distribution Launch Window



Displays a histogram and univariate statistics for each variable.

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns Assigns the variables that you want to analyze. A histogram and associated reports appear for each variable.

Weight Assigns a variable that specifies weights for observations on continuous Ys. For categorical Ys, the Weight column is ignored. Any statistic that is based on the sum of the weights is affected by weights.

Freq Assigns a frequency variable to this role. This is useful if your data are summarized. In this instance, you have one column for the Y values and another column for the frequency of occurrence of the Y values. The sum of this variable is included in the overall count appearing in the Summary Statistics report (represented by N). All other moment statistics (mean, standard deviation, and so on) are also affected by the **Freq** variable.

By Produces a separate report for each level of the **By** variable. If more than one **By** variable is assigned, a separate report is produced for each possible combination of the levels of the **By** variables.

Create Process Capability (Appears only if a column contains a Spec Limits column property.) Adds a Process Capability report for the analysis columns that contain a Spec Limits column property.

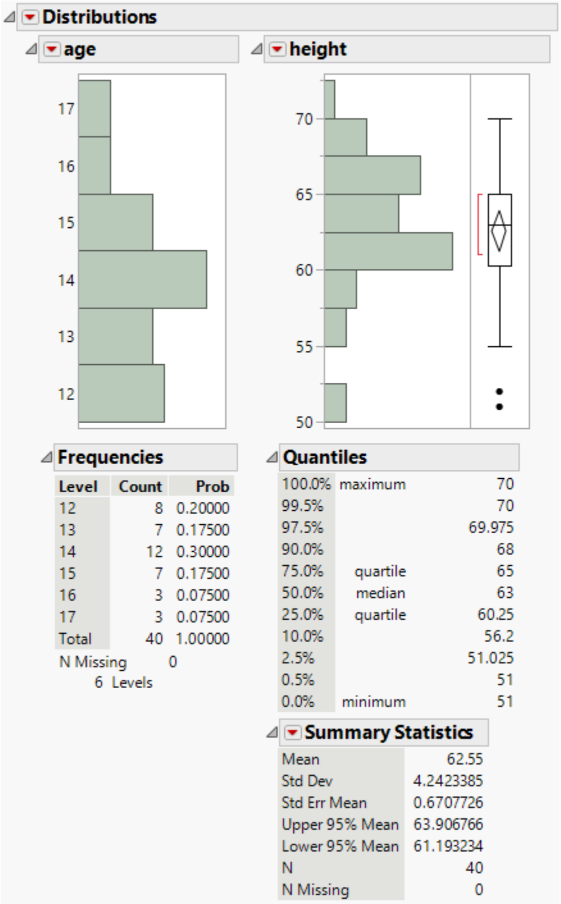
Histograms Only Removes everything except the histograms from the report window.

For more information about launch windows, see *Using JMP*.

The Distribution Report

The initial Distribution report contains a histogram and reports for each variable.

Figure 3.5 The Initial Distribution Report Window



Frequencies

Level	Count	Prob
12	8	0.20000
13	7	0.17500
14	12	0.30000
15	7	0.17500
16	3	0.07500
17	3	0.07500
Total	40	1.00000

N Missing 0
6 Levels

Quantiles

100.0%	maximum	70
99.5%		70
97.5%		69.975
90.0%		68
75.0%	quartile	65
50.0%	median	63
25.0%	quartile	60.25
10.0%		56.2
2.5%		51.025
0.5%		51
0.0%	minimum	51

Summary Statistics

Mean	62.55
Std Dev	4.2423385
Std Err Mean	0.6707726
Upper 95% Mean	63.906766
Lower 95% Mean	61.193234
N	40
N Missing	0

Note: Follow the instructions in “Example of the Distribution Platform” to produce the report shown in Figure 3.5.

Red Triangle Menus

The red triangle menu next to Distributions contains options that affect all of the variables. See “Distribution Platform Options”.

The red triangle menu next to each variable contains options that affect only that variable. See “Options for Categorical Variables” or “Options for Continuous Variables”.

Tip: If you hold down the Ctrl key and select a variable option, the option applies to all of the variables in the report that have the same modeling type.

Histograms and Reports

Histograms visually display your data. See “Histograms”.

The initial report for a categorical variable contains a Frequencies report. See “The Frequencies Report”.

The initial report for a continuous variable contains a Quantiles and a Summary Statistics report. See “The Quantiles Report” and “The Summary Statistics Report”.

Replace, Add, or Remove Variables in Histograms

To replace a variable in a histogram, from the Columns panel of the associated data table, drag and drop the variable into the axis of the histogram.

To create a new histogram in a report using a new variable, drag and drop the variable outside of an existing histogram. The new variable and histogram can be added before, between, or after the existing histograms.

To remove a variable from a histogram, select **Remove** from the red triangle menu.

Hidden and Excluded Rows

If you apply the Hidden row state to rows in the data table, the corresponding points do not appear in plots that show points. However, histograms are still constructed using the hidden rows.

If you want to exclude rows from the construction of the histograms and from analysis results, apply the Exclude row state. Then select **Redo > Redo Analysis** from the red triangle menu next to Distributions. Any rows that are excluded in the data table are also hidden in plots that show points.

Histograms

In the Distribution platform, you can use histograms to visually display your data. For categorical (nominal or ordinal) variables, the histogram shows a bar for each level of the ordinal or nominal variable. For continuous variables, the histogram shows a bar for grouped values of the continuous variable.

Highlighting data Click a histogram bar or an outlying point in the graph. The corresponding rows are highlighted in the data table, and corresponding sections of other histograms are also highlighted, if applicable. See “Highlight Bars and Select Rows”.

Creating a subset Double-click a histogram bar, or right-click a histogram bar and select **Subset**. A new data table that contains only the selected data is created.

Resizing the entire histogram Hover over the histogram borders until you see a double-sided arrow. Then click and drag the borders.

Rescaling the axis Click and drag on an axis to rescale it.

Alternatively, hover over the axis until you see a hand. Then, double-click the axis and set the parameters in the Axis Settings window.

Resizing histogram bars (Available only for continuous variables.) There are multiple options to resize histogram bars. See “Resize Histogram Bars for Continuous Variables”.

Specifying your selection Specify the data that you select in multiple histograms. See “Specify Your Selection in Multiple Histograms”.

To see additional options for the histogram or the associated data table:

- Right-click a histogram. See *Using JMP*.
- Right-click an axis. You can add a label or modify the axis. See *Using JMP*.
- Click the red triangle next to the variable, and select **Histogram Options**. Options are slightly different depending on the variable modeling type. See “Options for Categorical Variables” or “Options for Continuous Variables”.

Resize Histogram Bars for Continuous Variables

Resize histogram bars for continuous variables by using the following:

- the Grabber (hand) tool
- the **Set Bin Width** option
- the **Increment** option

Use the Grabber Tool

The Grabber (hand) tool is a quick way to explore your data.

1. Select **Tools > Grabber**.
2. Place the grabber tool anywhere in the histogram.
3. Click and drag the histogram bars.

Recall that histograms are based on binning the data. The height of each histogram bar is proportional to the number of observations that fall within the bin represented by the bar. For histograms in the default vertical orientation:

- Moving the hand to the left increases the bin width, which decreases the number of bins. The number of bars decreases as the bin width increases.
- Moving the hand to the right decreases the bin width, which increases the number of bins. The number of bars increases as the bin width decreases.
- Moving the hand up or down shifts the starting value for each bin, which shifts the data points that fall within each bin.

Note: If you have changed the histogram orientation to horizontal, reverse these directions. Move the hand down to increase bin width, up to decrease bin width, and left or right to shift bin starting values.

Use the Set Bin Width Option

The **Set Bin Width** option is a more precise way to set the width for all bars in a histogram. To use the Set Bin Width option, from the red triangle menu for the variable, select **Histogram Options > Set Bin Width**. Change the bin width value.

Use the Increment Option

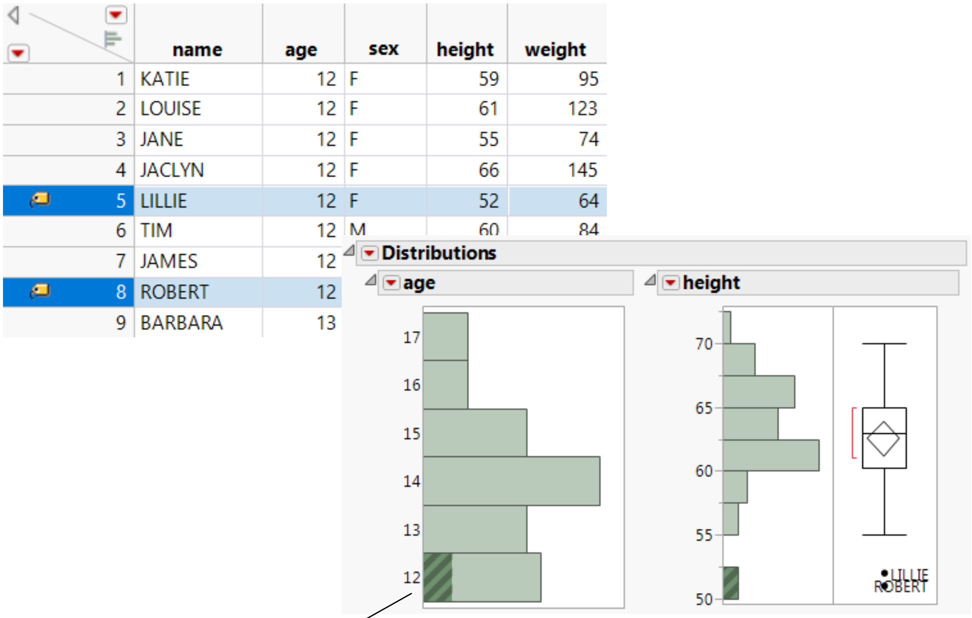
The **Increment** option is another precise way to set the bar width. To use the **Increment** option, double-click the axis, and change the Increment value.

Highlight Bars and Select Rows

Clicking on a histogram bar highlights the bar and selects the corresponding rows in the data table. The appropriate portions of all other graphical displays also highlight the selection. Figure 3.6 shows the results of highlighting a bar in the height histogram. The corresponding rows are selected in the data table.

Tip: To deselect specific histogram bars, press Ctrl and click the highlighted bars.

Figure 3.6 Highlighting Bars and Rows



	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84
7	JAMES	12			
8	ROBERT	12			
9	BARBARA	13			

Select a bar to highlight rows and parts of other output.

Specify Your Selection in Multiple Histograms

Extend or narrow your selection in histograms:

- To extend your selection, hold down the Shift key and select another bar. This is the equivalent of using an *or* operator.
- To narrow your selection, hold down the Ctrl and Alt keys (Windows) or Command and Option keys (macOS) and select another bar. This is the equivalent of using an *and* operator.

For an example, see “Example of Selecting Data in Multiple Histograms”.

The Frequencies Report

For nominal and ordinal variables, the Frequencies report in the Distribution platform lists the levels of the variables, along with the associated frequency of occurrence and probabilities.

For each level of a categorical (nominal or ordinal) variable, the Frequencies report contains the information described in the following list. Missing values are omitted from the analysis.

Tip: Click a value in the Frequencies report to select the corresponding data in the histogram and data table.

Level Lists each value found for a response variable.

Count Lists the number of rows found for each level of a response variable. If you use a Freq variable, the Count is the sum of the Freq variables for each level of the response variable.

Prob Lists the probability (or proportion) of occurrence for each level of a response variable. The probability is computed as the count divided by the total frequency of the variable, shown at the bottom of the table.

StdErr Prob Lists the standard error of the probabilities. This column might be hidden. To show the column, right-click in the table and select **Columns > StdErr Prob**.

Cum Prob Contains the cumulative sum of the column of probabilities. This column might be hidden. To show the column, right-click in the table and select **Columns > Cum Prob**.

The Quantiles Report

For continuous variables, the Quantiles report in the Distribution platform lists the values of selected quantiles (sometimes called *percentiles*). For statistical details, see “Statistical Details for Quantiles”.

The Summary Statistics Report

For continuous variables, the Summary Statistics report in the Distribution platform shows the mean, standard deviation, and other summary statistics. You can control which statistics appear in this report by selecting **Customize Summary Statistics** from the red triangle menu next to Summary Statistics.

Tip: You can specify which summary statistics show in the report each time you run a Distribution analysis for a continuous variable. Select **File > Preferences > Platforms > Distribution Summary Statistics**, and select the ones that you want to appear.

- “Description of the Summary Statistics Report” describes the statistics that appear by default.
- “Additional Summary Statistics” describes additional statistics that you can add to the report using the **Customize Summary Statistics** window.

Description of the Summary Statistics Report

Mean Estimates the expected value of the underlying distribution for the response variable, which is the arithmetic average of the column's values. It is the sum of the nonmissing values divided by the number of nonmissing values.

Std Dev The normal distribution is mainly defined by the mean and standard deviation. These parameters provide an easy way to summarize data as the sample becomes large:

- 68% of the values are within one standard deviation of the mean
- 95% of the values are within two standard deviations of the mean
- 99.7% of the values are within three standard deviations of the mean

Std Err Mean The standard error of the mean, which estimates the standard deviation of the distribution of the mean.

Upper and Lower Mean Confidence Limits The 95% confidence limits about the mean. They define an interval that is very likely to contain the true population mean.

N The total number of nonmissing values.

N Missing The number of missing observations.

Additional Summary Statistics

Sum Weight The sum of a column assigned to the role of Weight (in the launch window). Sum Wgt is used in the denominator for computations of the mean instead of N.

Sum The sum of the response values.

Variance The sample variance, and the square of the sample standard deviation.

Skewness Measures sidedness or symmetry.

Kurtosis Measures peakedness or heaviness of tails. See "Kurtosis" for formula details.

CV The percent coefficient of variation. It is computed as the standard deviation divided by the mean and multiplied by 100. The coefficient of variation can be used to assess relative variation. For example, it can be used when comparing the variation in data measured in different units or with different magnitudes.

N Zero The number of zero values.

N Unique The number of unique values.

Uncorrected SS The uncorrected sum of squares or sum of values squared.

Corrected SS The corrected sum of squares or sum of squares of deviations from the mean.

Autocorrelation (Appears only if you have not specified a Frequency variable.) First autocorrelation that tests if the residuals are correlated across the rows. This test helps detect non-randomness in the data.

Minimum Represents the 0 percentile of the data.

Maximum Represents the 100 percentile of the data.

Median Represents the 50th percentile of the data.

Mode The value that occurs most often in the data. If there are multiple modes, the smallest mode appears.

Trimmed Mean The mean calculated after removing the smallest p% and the largest p% of the data. The value of p is entered in the **Enter trimmed mean percent** text box at the bottom of the window. The Trimmed Mean option is not available if you have specified a Weight variable.

Geometric Mean The *n*th root of the product of the data. For example, geometric means are often used to calculate interest rates. The statistic is also helpful when the data contains a large value in a skewed distribution.

Note: Negative values result in missing numbers, and zero values (with no negative values) result in zero.

Range The difference between the maximum and minimum of the data.

Interquartile Range The difference between the 3rd and 1st quartiles.

Median Absolute Deviation (Does not appear if you have specified a Weight variable.) The median of the absolute deviations from the median.

Proportion Zero The proportion of nonmissing values that are equal to zero.

Proportion Nonzero The proportion of nonmissing values that are not equal to zero.

3*StdDev The value of 3 times the standard deviation. You can set the value of the multiplier using the K Sigma Multiplier platform preference under Preferences > Platforms > Distribution.

3*StdDev Above Mean The value of the mean plus 3 times the standard deviation. You can set the value of the multiplier using the K Sigma Multiplier platform preference under Preferences > Platforms > Distribution.

3*StdDev Below Mean The value of the mean minus 3 times the standard deviation. You can set the value of the multiplier using the K Sigma Multiplier platform preference under Preferences > Platforms > Distribution.

Robust Mean The robust mean, calculated in a way that is resistant to outliers, using Huber's M-estimation. See Huber and Ronchetti (2009).

Robust Std Dev The robust standard deviation, calculated in a way that is resistant to outliers, using Huber's M-estimation. See Huber and Ronchetti (2009).

Enter (1-alpha) for mean confidence interval Specify the alpha level for the mean confidence interval.

Enter trimmed mean percent Specify the trimmed mean percentage. The percentage is trimmed off each side of the data.

Summary Statistics Options

The red triangle menu next to Summary Statistics contains these options:

Customize Summary Statistics Select which statistics you want to appear from the list. You can select or deselect all summary statistics.

Show All Modes Shows all of the modes if there are multiple modes.

For statistical details, see "Statistical Details for Summary Statistics".

Distribution Platform Options

The Distributions red triangle menu contains options that affect all of the reports and graphs in the Distribution platform.

Uniform Scaling Scales all axes with the same minimum, maximum, and intervals so that the distributions can be easily compared.

Stack Changes the orientation of the histogram and the reports to horizontal and stacks the individual distribution reports vertically. Deselect this option to return the report window to its original layout.

Arrange in Rows Enter the number of plots that appear in a row. This option helps you view plots vertically rather than in one wide row.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Options for Categorical Variables

The red triangle menus next to each variable in the Distributions report contain additional options that apply to the variable. This section describes the options that are available for categorical (nominal or ordinal) variables.

Display Options Contains the following options to change the display of the report.

Frequencies Shows or hides the Frequencies report. See "The Frequencies Report".

Horizontal Layout Changes the orientation of the histogram and the reports to vertical or horizontal.

Axes on Left (Available only when the Horizontal Layout option is selected.) Moves the Count, Prob, and Density axes to the left side of the histogram instead of the right side.

Histogram Options See "Histogram Options for Categorical Variables".

Mosaic Plot Displays a mosaic bar chart for each nominal or ordinal response variable. A mosaic plot is a stacked bar chart where each segment is proportional to its group's frequency count.

Order By Reorders the histogram, mosaic plot, and Frequencies report in ascending or descending order, by count. To save the new order as a column property, use the **Save > Value Ordering** option.

Test Probabilities Shows a report that tests hypothesized probabilities. See "Example of Testing Probabilities for Two Levels" and "Example of Testing Probabilities for More Than Two Levels".

Confidence Interval This menu contains confidence levels. Select a value that is listed, or select **Other** to enter your own. JMP computes score confidence intervals.

Save See "Save Options for Categorical Variables".

Remove Permanently removes the variable and all its reports from the Distribution report.

Histogram Options for Categorical Variables

Histogram Shows or hides the histogram. See “Histograms”.

Vertical Changes the orientation of the histogram from a vertical to a horizontal orientation.

Std Error Bars Draws the standard error bar on each level of the histogram.

Separate Bars Separates the histogram bars.

Histogram Color Changes the color of the histogram bars.

Count Axis Adds an axis that shows the frequency of column values represented by the histogram bars.

Prob Axis Adds an axis that shows the proportion of column values represented by histogram bars.

Density Axis Adds an axis that shows the length of the bars in the histogram.

The count and probability axes are based on the following calculations:

prob = (bar width)*density

count = (bar width)*density*(total count)

Show Percents Labels the percent of column values represented by each histogram bar.

Note: To specify the number of decimal places, right-click the histogram and select **Customize > Histogram**.

Show Counts Labels the frequency of column values represented by each histogram bar.

Save Options for Categorical Variables

Level Numbers Creates a new column in the data table called Level <colname>. The level number of each observation corresponds to the histogram bar that contains the observation.

Value Ordering (Use with the **Order By** option) Creates a new value ordering column property in the data table, reflecting the new order.

Script to Log Displays the script commands to generate the current report in the log window. Select **View > Log** to see the log window.

Options for Continuous Variables

The red triangle menus next to each variable in the report window contain additional options that apply to the variable. This section describes the options that are available for continuous variables.

Display Options See "Display Options for Continuous Variables".

Histogram Options See "Histogram Options for Continuous Variables".

Normal Quantile Plot Helps you visualize the extent to which the variable is normally distributed. See "Normal Quantile Plot".

Outlier Box Plot Shows the distribution and helps you identify possible outliers. See "Outlier Box Plot".

Quantile Box Plot Shows specific quantiles from the Quantiles report. See "Quantile Box Plot".

Stem and Leaf See "Stem and Leaf".

CDF Plot Creates a plot of the empirical cumulative distribution function. See "CDF Plot".

Test Mean Perform a one-sample test for the mean. See "Test Mean".

Test Std Dev Perform a one-sample test for the standard deviation. See "Test Std Dev".

Test Equivalence Assesses whether a population mean is equivalent to a hypothesized value. See "Test Equivalence".

Confidence Interval Choose confidence intervals for the mean and standard deviation. See "Confidence Intervals".

Prediction Interval Choose prediction intervals for a single observation, or for the mean and standard deviation of the next randomly selected sample. See "Prediction Intervals".

Tolerance Interval Computes an interval to contain at least a specified proportion of the population. See "Tolerance Intervals".

Process Capability (Not available when the Y column has a Detection Limits column property.) Measures the conformance of a process to given specification limits. See "Process Capability".

Note: If the Y column has a Detection Limits column property, process capability analyses are available only from the fitted distribution red triangle menus. See "Process Capability".

Continuous Fit Fits distributions to continuous variables. If the Y column has a Detection Limits column property, the Continuous Fit options fit a censored distribution and only a subset of distributions are available. See “Fit Distributions”.

Discrete Fit (Available when all data values are integers.) Fits distributions to discrete variables. See “Fit Distributions”.

Save Saves information about continuous or categorical variables. See “Save Options for Continuous Variables”.

Remove Permanently removes the variable and all its reports from the Distribution report.

Display Options for Continuous Variables

Quantiles Shows or hides the Quantiles report. See “The Quantiles Report”.

Set Quantile Increment Changes the quantile increment or revert to the default quantile increment.

Custom Quantiles Sets custom quantiles by values or by increments. You can specify the confidence level and choose whether to compute smoothed empirical likelihood quantiles (for large data sets, this can take some time).

- For more information about how the weighted average quantiles are estimated, see “Statistical Details for Quantiles”.
- For more information about distribution-free confidence limits for the weighted average quantiles, see section 5.2 in Meeker et al. (2017).
- Smoothed empirical likelihood quantiles are based on a kernel density estimate. For more information about how these quantiles and their confidence limits are estimated, see Chen and Hall (1993).
- Confidence intervals and smoothed empirical likelihood quantiles are not available when fractional frequencies are used.

Summary Statistics Shows or hides the Summary Statistics report. See “The Summary Statistics Report”.

Customize Summary Statistics Adds or removes statistics from the Summary Statistics report. See “The Summary Statistics Report”.

Horizontal Layout Changes the orientation of the histogram and the reports to vertical or horizontal.

Axes on Left Moves the **Count**, **Prob**, **Density**, and **Normal Quantile Plot** axes to the left instead of the right.

This option is applicable only if **Horizontal Layout** is selected.

Histogram Options for Continuous Variables

Histogram Shows or hides the histogram. See "Histograms".

Shadowgram Replaces the histogram with a shadowgram. To understand a shadowgram, consider that if the bin width of a histogram is changed, the appearance of the histogram changes. A shadowgram overlays histograms with different bin widths. Dominant features of a distribution are less transparent on the shadowgram.

Note that the following options are not available for shadowgrams:

– Std Error Bars

– Show Counts

– Show Percents

Vertical Changes the orientation of the histogram from a vertical to a horizontal orientation.

Std Error Bars Draws the standard error bar on each level of the histogram using the standard error. The standard error bar adjusts automatically when you adjust the number of bars with the hand tool. See "Resize Histogram Bars for Continuous Variables" and "Statistical Details for Standard Error Bars".

Set Bin Width Changes the bin width of the histogram bars. See "Resize Histogram Bars for Continuous Variables".

Histogram Color Changes the color of the histogram bars.

Count Axis Adds an axis that shows the frequency of column values represented by the histogram bars.

Note: If you resize the histogram bars, the count axis also resizes.

Prob Axis Adds an axis that shows the proportion of column values represented by histogram bars.

Note: If you resize the histogram bars, the probability axis also resizes.

Density Axis The density is the length of the bars in the histogram. Both the count and probability are based on the following calculations:

prob = (bar width)*density

count = (bar width)*density*(total count)

When looking at density curves that are added by the Fit Distribution option, the density axis shows the point estimates of the curves.

Note: If you resize the histogram bars, the density axis also resizes.

Show Percents Labels the proportion of column values represented by each histogram bar.

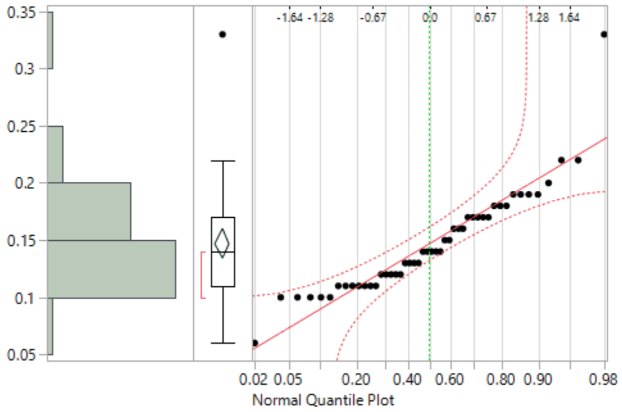
Show Counts Labels the frequency of column values represented by each histogram bar.

Normal Quantile Plot

Use the Normal Quantile Plot option in the Distribution platform to visualize the extent to which the variable is normally distributed. If a variable is normally distributed, the normal quantile plot approximates a diagonal straight line. This type of plot is also called a quantile-quantile plot, or Q-Q plot.

The normal quantile plot also shows Lilliefors confidence bounds (Conover 1980) and probability and normal quantile scales.

Figure 3.7 Normal Quantile Plot



Note the following information:

- The vertical axis shows the column values.
- The upper horizontal axis shows the normal quantile scale.
- The lower horizontal axis shows the empirical cumulative probability for each value.
- The dashed red line shows the Lilliefors confidence bounds.

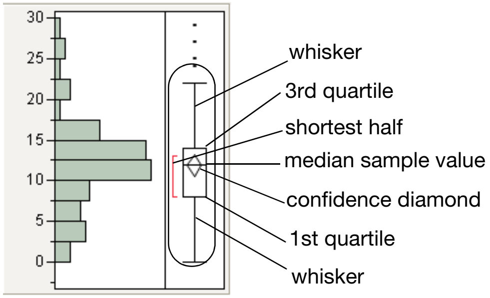
For statistical details, see "Statistical Details for the Normal Quantile Plot".

Outlier Box Plot

Use the Outlier Box Plot option in the Distribution platform to see the distribution and identify possible outliers. Generally, box plots show selected quantiles of continuous distributions.

Note: Outlier box plots are also called Tukey outlier box plots.

Figure 3.8 Outlier Box Plot



Note the following aspects about outlier box plots:

- The horizontal line within the box represents the median sample value.
- The confidence diamond contains the mean and the upper and lower 95% of the mean. If you drew a line through the middle of the diamond, you would have the mean. The top and bottom points of the diamond represent the upper and lower 95% of the mean.
- The ends of the box represent the 25th and 75th quantiles, also expressed as the 1st and 3rd *quartile*, respectively.
- The difference between the 1st and 3rd quartiles is called the *interquartile range*.
- The box has lines that extend from each end, sometimes called *whiskers*. The whiskers extend from the ends of the box to the outermost data point that falls within these distances:

  1st quartile - 1.5*(interquartile range)

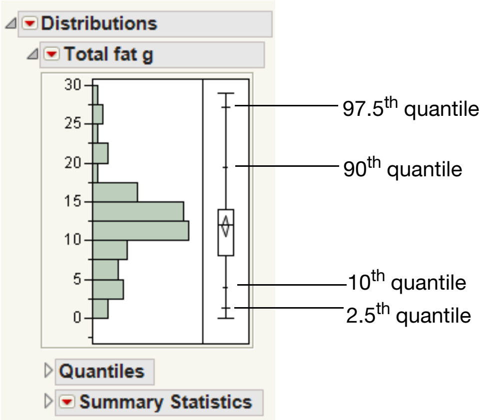
  3rd quartile + 1.5*(interquartile range)

  If the data points do not reach the computed ranges, then the whiskers are determined by the upper and lower data point values (not including outliers).
- The bracket outside of the box identifies the *shortest half*, which is the most dense 50% of the observations (Rousseeuw and Leroy 1987).
- To remove objects from outlier box plots, see “Remove Objects from the Outlier or Quantile Box Plot”.

Quantile Box Plot

Use the Quantile Box Plot option in the Distribution platform to display specific quantiles from the Quantiles report. At a glance, you can see whether the distribution is symmetric. If the distribution is symmetric, the quantiles in the box plot are approximately equidistant from each other. For example, if the quantile marks are grouped closely at one end, but have greater spacing at the other end, the distribution is skewed toward the end with more spacing.

Figure 3.9 Quantile Box Plot



Quantiles are values where the p^{th} quantile is larger than p% of the values. For example, 10% of the data lies below the 10^{th} quantile, and 90% of the data lies below the 90^{th} quantile.

Remove Objects from the Outlier or Quantile Box Plot

You can remove the confidence diamond and the shortest half from outlier or quantile box plots. You can remove them for a single graph, or remove them for all future graphs.

To remove them from an individual graph

1. Right-click the outlier box plot and select **Customize**.
2. Click **Box Plot**.
3. Deselect the check box next to **Confidence Diamond** or **Shortest Half**.

For more information about the Customize Graph window, see *Using JMP*.

To remove them for all future graphs

1. Select **File** > **Preferences** > **Platforms** > **Distribution**.
2. Deselect these options:
   - **Show Box Plot Confidence Diamond**

  - **Show Outlier Box Plot Shortest Half**
3. Click **OK**.

   Any box plots you now add in Distribution will not have the confidence diamond or shortest half.

Stem and Leaf

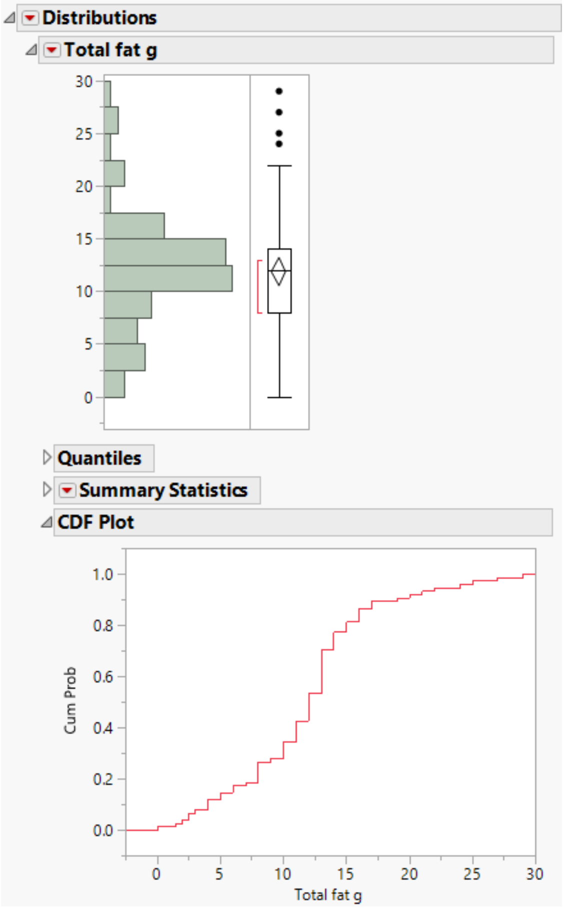
Use the Stem and Leaf option in the Distribution platform to show an interactive text plot of the data. Each line of the plot has a Stem value that is the leading digit of a range of column values. The Leaf values are made from the next-in-line digits of the values. You can see the data point by joining the stem and leaf. In some cases, the numbers on the stem and leaf plot are rounded versions of the actual data in the table. The stem-and-leaf plot actively responds to clicking and the brush tool.

Note: The stem-and-leaf plot converts fractional frequencies to the smallest integer greater than or equal to the specified frequency.

CDF Plot

Use the CDF Plot option in the Distribution platform to create a plot of the empirical cumulative distribution function. Use the CDF plot to determine the percent of data that is at or below a given value on the horizontal axis. When a distribution is fit to the data, the cumulative distribution function for the fitted distribution is overlaid on the CDF plot.

Figure 3.10 CDF Plot



For example, in this CDF plot, approximately 34% of the data are less than a total fat value of 10 grams.

Test Mean

Use the Test Mean option in the Distribution platform to specify options for and perform a one-sample test for the mean. If you specify a value for the standard deviation, a *z* test is performed. Otherwise, the sample standard deviation is used to perform a *t* test. You can also request the nonparametric Wilcoxon Signed-Rank test.

You can use the Test Mean option repeatedly to test different values. Each time you test the mean, a new Test Mean report appears.

Description of the Test Mean Report

Statistics That Are Calculated for Test Mean

t Test (or z Test) Lists the value of the test statistic and the *p*-values for the two-sided and one-sided alternatives.

Signed-Rank (Appears only if the Wilcoxon Signed-Rank test is selected.) Lists the value of the Wilcoxon signed-rank statistic followed by *p*-values for the two-sided and one-sided alternatives. The test uses the Pratt method to address zero values. This is a nonparametric test whose null hypothesis is that the median equals the postulated value. It assumes that the distribution is symmetric. See "Statistical Details for the Wilcoxon Signed Rank Test".

Probability Values

Prob > |t| The probability of obtaining an absolute *t* value that is greater than the observed *t* value when the population mean is equal to the hypothesized value. This is the *p*-value for observed significance of the two-tailed *t* test.

Prob > t The probability of obtaining a *t* value greater than the computed sample *t* ratio when the population mean is not different from the hypothesized value. This is the *p*-value for an upper-tailed test.

Prob < t The probability of obtaining a *t* value less than the computed sample *t* ratio when the population mean is not different from the hypothesized value. This is the *p*-value for a lower-tailed test.

Descriptions of the Test Mean Options

PValue animation Starts an interactive visual representation of the *p*-value. Enables you to change the hypothesized mean value while watching how the change affects the *p*-value.

Power animation Starts an interactive visual representation of power and beta. You can change the hypothesized mean and sample mean while watching how the changes affect power and beta.

Remove Test Removes the mean test.

Test Std Dev

Use the Test Std Dev option in the Distribution platform to perform a one-sample test for the standard deviation. You can use this option repeatedly to test different values. Each time you test the standard deviation, a new Test Standard Deviation report appears.

Test Statistic Provides the value of the chi-square test statistic. See "Statistical Details for the Standard Deviation Test".

Min PValue The probability of obtaining a more extreme chi-square value when the population standard deviation does not differ from the hypothesized value. See "Statistical Details for the Standard Deviation Test".

Prob>ChiSq The probability of obtaining a chi-square value greater than the computed sample chi-square when the population standard deviation is not different from the hypothesized value. This is the *p*-value for observed significance of a one-tailed *t* test.

Prob<ChiSq The probability of obtaining a chi-square value less than the computed sample chi-square when the population standard deviation is not different from the hypothesized value. This is the *p*-value for observed significance of a one-tailed *t* test.

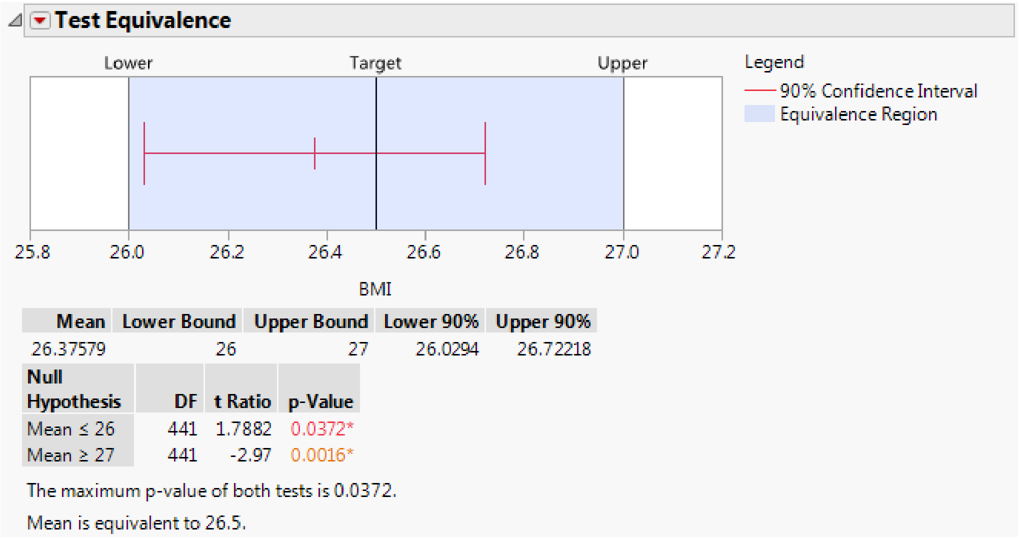
Test Equivalence

The Test Equivalence option in the Distribution platform enables you to assess whether a population mean is equivalent to a hypothesized value. You must define a threshold difference that is considered equivalent to no difference. The Test Equivalence option uses the Two One-Sided Tests (TOST) approach. Two one-sided *t* tests are constructed for the null hypotheses that the difference between the true mean and the hypothesized value exceeds the threshold. If both null hypotheses are rejected, this implies that the true difference does not exceed the threshold. You conclude that the mean can be considered practically equivalent to the hypothesized value.

When you select the Test Equivalence option, you specify the Hypothesized Mean, the threshold difference (Difference Considered Practically Zero), and the Confidence Level. The Confidence Level is 1 - alpha, where alpha is the significance level for each one-sided test.

The Test Equivalence report in Figure 3.11 is for the variable BMI in the Diabetes.jmp sample data table. The Hypothesized Mean is 26.5 and the Difference Considered Practically Zero is specified as 0.5.

Figure 3.11 Test Equivalence Report



Mean	Lower Bound	Upper Bound	Lower 90%	Upper 90%
26.37579	26	27	26.0294	26.72218

Null Hypothesis	DF	t Ratio	p-Value
Mean ≤ 26	441	1.7882	0.0372*
Mean ≥ 27	441	-2.97	0.0016*

The maximum p-value of both tests is 0.0372.

Mean is equivalent to 26.5.

The report shows the following:

- A plot of your defined equivalence region that shows the Target and boundaries, defined by vertical lines labeled Lower and Upper.
- A confidence interval for the calculated mean. This confidence interval is a 1 - 2*alpha level interval.
- A table that shows the calculated mean, the specified lower and upper bounds, and a (1 - 2*alpha) level confidence interval for the mean.
- A table that shows the results of the two one-sided tests.
- A note that summarizes the results, and states whether the mean can be considered equivalent to the Target value.

Confidence Intervals

Use the Confidence Interval option in the Distribution platform to show confidence intervals for the mean and standard deviation. The **0.90**, **0.95**, and **0.99** options compute two-sided confidence intervals for the mean and standard deviation. Use the **Confidence Interval > Other** option to select a confidence level, and select one-sided or two-sided confidence intervals. You can also enter a known sigma. If you use a known sigma, the confidence interval for the mean is based on *z*-values rather than *t*-values.

The Confidence Intervals report shows the mean and standard deviation parameter estimates with upper and lower confidence limits for 1 - α.

Prediction Intervals

Use the Prediction Interval option in the Distribution platform to show intervals for a single observation, or the mean and standard deviation of the next randomly selected sample. The calculations assume that the given sample is selected randomly from a normal distribution. Select one-sided or two-sided prediction intervals.

When you select the **Prediction Interval** option for a variable, the Prediction Intervals window appears. Use the window to specify the confidence level, the number of future samples, and either a one-sided or two-sided limit.

Related Information

- For statistical details, see "Statistical Details for Prediction Intervals".
- For an example, see "Example of Prediction Intervals".

Tolerance Intervals

Use the Tolerance Interval option in the Distribution platform to show intervals that contain at least a specified proportion of the population. It is an interval estimate for a specified proportion of the population. Complete discussions of tolerance intervals are found in Meeker et al. (2017) and in Tamhane and Dunlop (2000).

When you select the **Tolerance Interval** option for a variable, the Tolerance Intervals window appears. Use the window to specify the confidence level, the proportion to cover, a one-sided or two-sided limit, and the method. The two available methods are Assume Normal Distribution and Nonparametric. The Assume Normal Distribution option computes tolerance intervals that are based on the assumption that the sample was randomly selected from a normal distribution. The Nonparametric option computes distribution-free tolerance intervals.

Note: When you specify the Assume Normal Distribution option, the value of *k* used for the intervals is contained in the report table. To view the value, in the Tolerance Intervals report, right-click the table and select Columns > K-factor.

Related Information

- For statistical details, see "Statistical Details for Tolerance Intervals".
- For an example, see "Example of Tolerance Intervals".

Process Capability

Use the Process Capability option in the Distribution platform to show an analysis that measures how well a process is performing compared to given specification limits. A good process is one that is stable and consistently produces product that is well within specification limits. A capability index is a measure that relates process performance, summarized by process centering and variability, to specification limits.

Specification Limits

If a column contains a Spec Limits column property and the Create Process Capability option on the launch window is selected, a Process Capability report is automatically created. This report is based on the normal distribution, unless the column also contains a Distribution column property. If the column contains a Distribution column property, the Process Capability report is based on the distribution specified in the column property.

Tip: To add specification limits to several columns at once, see *Quality and Process Methods*.

If a column does not contain specification limits, select **Process Capability** from the red triangle next to the name of the analysis variable and set specification limits in the Process Capability Analysis window.

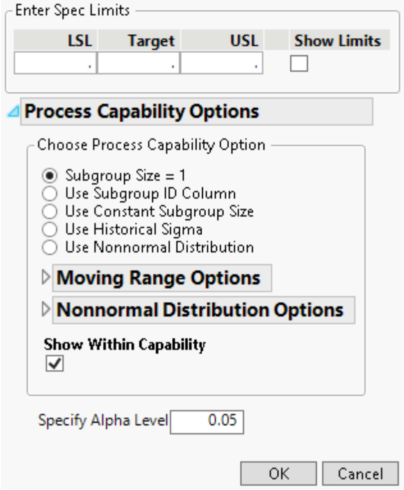
To save specification limits from a report to the data table as a column property, select **Save Spec Limits as Column Properties** from the Process Capability red triangle. When you repeat the process capability analysis, the saved specification limits are automatically retrieved.

Process Capability Analysis Window

The Process Capability Analysis window appears when you select the Process Capability option from the red triangle next to the name of the analysis variable.

Use the Process Capability Analysis window to specify options for the capability analysis, including specification limits, the underlying distribution for the analysis, and the estimation method for sigma. Process capability requires you to choose how to estimate sigma, the within-group (short-term) variation. Different suboptions appear depending on which process capability option you choose.

Figure 3.12 Process Capability Analysis Window



Enter Spec Limits

LSL	Target	USL	Show Limits
.	.	.	☐

Choose Process Capability Option

● Subgroup Size = 1
○ Use Subgroup ID Column
○ Use Constant Subgroup Size
○ Use Historical Sigma
○ Use Nonnormal Distribution

Show Within Capability
☑

Specify Alpha Level 0.05

Enter Spec Limits Specifies the Lower Spec Limit, the Target, and the Upper Spec Limit for the process capability analysis. At least one of these must be a nonmissing value. If you select the Show Limits option, the specification limits appear on the histogram in the Distribution platform report.

Process Capability Options Depending on which option you choose, different additional options appear. Choose one of the following options:

Subgroup Size = 1 Sets the subgroup size to 1 and provides additional Moving Range options. See *Quality and Process Methods*.

Use Subgroup ID Column Enables you to select a subgroup ID column and provides additional Subgrouping and Moving Range options. See *Quality and Process Methods*.

Use Constant Subgroup Size Enables you to set a constant subgroup size and provides additional Subgrouping and Moving Range options. See *Quality and Process Methods*.

Use Historical Sigma Assigns a historically accepted value for sigma. See *Quality and Process Methods*.

Use Nonnormal Distribution Enables you to select a nonnormal distribution and provides additional Nonnormal Distribution Options. See *Quality and Process Methods*.

Show Within Capability (Available only when Subgroup Size = 1 or Use Nonnormal Distribution is selected in Process Capability Options.) Specifies if estimates of within sigma are shown in the report.

Specify Alpha Level Specifies the significance level for confidence limits.

Process Capability Analysis Report

After you click OK in the Process Capability Analysis window, a Process Capability report appears that contains a capability report for the selected variable. For more information about this report, see *Quality and Process Methods*.

- For statistical details, see *Quality and Process Methods*.
- For an example, see "Example of Process Capability".
- For the Process Capability platform, see *Quality and Process Methods*.

Note: You can set preferences for many of the options in the Process Capability report in Distribution at **File > Preferences > Platforms > Process Capability**.

Fit Distributions

Use the options in the Continuous Fit or Discrete Fit submenus to fit a distribution to a continuous variable. When you fit a distribution to a continuous variable, a curve is overlaid on the histogram and a Compare Distributions report and a Fitted Distribution report are added to the report window. A red triangle menu in the Fitted Distribution report contains additional options. See "Fit Distribution Options". If a column contains a Distribution column property, the distribution in that column property is fit by default in the Distribution report.

Note: The Life Distribution platform also contains options for distribution fitting that might use different parameterizations and allow for censored observations. See *Reliability and Survival Methods*.

Continuous Fit

The Continuous Fit submenu contains options for fitting continuous distributions. For more information about the parameterization of these distributions, see "Statistical Details for Continuous Fit Distributions".

Fit Normal Fits a normal distribution to the data. The normal distribution is often used to model symmetric data with most of the values falling in the middle of the curve. The parameter estimation for the normal distribution uses the unbiased estimate.

Fit Cauchy Fits a Cauchy distribution to the data. The Cauchy distribution has an undefined mean and standard deviation. Although most data do not inherently follow a Cauchy distribution, it can be useful for estimating a robust location and scale for data that contain a large proportion of outliers (up to 50%).

Fit Student's t Fits a Student's *t* distribution to the data. The Student's *t* distribution is a robust option that spans the space between a normal distribution and a Cauchy distribution. As the degrees of freedom in the Student's *t* distribution approach infinity, the distribution is equivalent to the normal. When the degrees of freedom in the Student's *t* distribution equals 1, the distribution is equivalent to the Cauchy. The Distribution platform estimates the degrees of freedom value.

Fit SHASH Fits a sinh-arcsinh (SHASH) distribution to the data. The SHASH distribution is similar to Johnson distributions in that it is a transformation to normality, but the SHASH distribution includes the normal distribution as a special case. This distribution can be symmetric or asymmetric.

Fit Exponential (Available only when all observations are nonnegative.) Fits an exponential distribution to the data. The exponential distribution is right-skewed and is often used to model lifetimes or the time between successive events.

Fit Gamma (Available only when all observations are positive.) Fits a gamma distribution to the data. The gamma distribution is a flexible distribution for modeling positive values.

Fit Lognormal (Available only when all observations are positive.) Fits a lognormal distribution to the data. The lognormal distribution is right-skewed and is often used to model lifetimes or the time until an event. The parameter estimation for the lognormal distribution uses the maximum likelihood estimate.

Fit Weibull (Available only when all observations are positive.) Fits a Weibull distribution to the data. The Weibull distribution is a flexible distribution and is often used to model lifetimes or the time until an event.

Fit Normal 2 Mixture Fits a mixture of two normal distributions. This flexible distribution is capable of fitting bimodal data.

Fit Normal 3 Mixture Fits a mixture of three normal distributions. This flexible distribution is capable of fitting multi-modal data.

Fit Smooth Curve Fits a smooth curve using nonparametric density estimation ("Kernel Smoother Report"). Control the amount of smoothing by changing the bandwidth with the slider that appears in the Nonparametric Density report.

Fit Johnson Fits a Johnson distribution to the data. The most appropriate of the three types of Johnson distribution (Su, Sb, and Sl) is fit and reported. The Johnson family of distributions is useful for its data-fitting capabilities because it supports every possible combination of skewness and kurtosis. Information about selection procedures and parameter estimation for the Johnson distributions can be found in Slifker and Shapiro (1980).

Fit Beta (Available only when all observations are between 0 and 1.) Fits a beta distribution to the data. The beta distribution is useful for modeling data that are between 0 and 1 (not inclusive) and is often used to model proportions or rates.

Fit All Fits all available continuous distributions to a variable. The Compare Distributions report contains statistics about each fitted distribution. By default, the best fit distribution is selected and displayed on the histogram. Use the check boxes to show or hide a fit report and overlay curve for the selected distribution. Initially, the Compare Distributions list is sorted by AICc in ascending order.

Tip: You can quickly remove distributions from the Compare Distributions list by double-clicking the name of the distribution in the Distribution column. This action also removes the corresponding Fitted Distribution report.

Enable Legacy Fitters Shows or hides the Legacy Fitters submenu. Some features of distribution fitting were updated in JMP 15. This option enables you to use the older features from previous JMP releases that have been retained for compatibility purposes. For documentation on these legacy fitters, see the Details for the Legacy Distribution Fitters section of the JMP 16.1 Help.

Discrete Fit

The Discrete Fit submenu is available when all of the data values are integers. The Discrete Fit submenu contains options for fitting discrete distributions. For more information about the parameterization of these distributions, see “Statistical Details for Discrete Fit Distributions”.

Fit Poisson Fits a Poisson distribution to the data. The Poisson distribution is useful for modeling the number of events in a given interval and is often expressed as count data.

Fit Negative Binomial Fits a negative binomial distribution to the data. The negative binomial distribution is useful for modeling the number of successes before a specified number of failures. The negative binomial distribution is also equivalent to the Gamma Poisson distribution.

Fit ZI Poisson (Available only when there are values of zero in the data.) Fits a zero-inflated Poisson distribution to the data. The zero-inflated Poisson assumes a greater proportion of the data are zero values than would occur in a standard Poisson distribution.

Fit ZI Negative Binomial (Available only when there are values of zero in the data.) Fits a zero-inflated negative binomial distribution to the data. The zero-inflated negative binomial assumes a greater proportion of the data are zero values than would occur in a standard negative binomial distribution.

Fit Binomial Fits a binomial distribution to the data. The binomial distribution is useful for modeling the total number of successes in n independent trials that all have a fixed probability, p, of success. The sample size can be specified as a fixed sample size for all observations, or it can be specified as another column in the data table that contains sample sizes for each row.

Note: When a non-constant sample size is specified, density curves, diagnostic plots, and profilers are not available.

Fit Beta Binomial Fits a beta binomial distribution to the data. The beta binomial distribution is an overdispersed version of the binomial distribution. It requires a sample size greater than one for each observation. The sample size can be specified as a fixed sample size for all observations, or it can be specified as another column in the data table that contains sample sizes for each row.

Note: When a non-constant sample size is specified, density curves, diagnostic plots, and profilers are not available.

Fit Distribution Options

Each fitted distribution report has a red triangle menu that contains additional options.

Density Curve Uses the estimated parameters of the distribution to overlay a density curve on the histogram.

Diagnostic Plots Contains the following options:

QQ Plot Shows or hides a quantile-quantile (QQ) plot. This plot shows the relationship between the observations and the quantiles obtained using the estimated parameters.

PP Plot Shows or hides a percentile-percentile (PP) plot. This plot shows the relationship between the empirical cumulative distribution function (CDF) and the fitted CDF obtained using the estimated parameters.

Profilers Contains the following options:

Distribution Profiler Shows or hides a prediction profiler of the cumulative distribution function (CDF).

Quantile Profiler Shows or hides a prediction profiler of the quantile function.

Save Columns Contains the following options:

Save Density Formula Saves a column to the data table that contains the density formula computed using the estimated parameter values.

Save Distribution Formula Saves a column to the data table that contains the cumulative distribution function (CDF) formula computed using the estimated parameter values.

Save Simulation Formula Saves a column to the data table that contains a formula that generates simulated values using the estimated parameters. This column can be used in the Simulate utility as a Column to Switch In. See “Simulate”.

Save Transformed (Available only for Johnson and SHASH distribution fits.) Saves a column to the data table that contains a transform formula. The formula can be used to transform the analysis column to normality using the fitted distribution.

Goodness of Fit (Not available for Johnson, Smooth Curve, Normal Mixture, Binomial, or Beta Binomial distributions.) Shows or hides a Goodness-of-Fit Test report that contains a goodness-of-fit test for the fitted distribution.

For continuous fits, the goodness-of-fit test is the Anderson-Darling test. The *p*-value for the test is simulated using a parametric bootstrap, similar to the procedure described in Section 4.1 of Stephens (1974). For Normal distributions, the Shapiro-Wilk test for normality is also reported when the sample size is less than 2000 and there are no fixed parameters.

For discrete fits, the goodness-of-fit test is a Pearson chi-squared test.

Fix Parameters (Not available for Johnson distribution or smooth curve fits.) Enables you to fix parameters and re-estimate the non-fixed parameters. An Adequacy LR (likelihood ratio) Test report also appears, which tests your new parameters to determine whether they fit the data.

Process Capability (Not available for Cauchy, Student's *t*, or discrete distribution fits.) Enables you to create a Process Capability analysis using the fitted distribution, which is a measure of how well process performs with respect to the specification limits. When you select the Process Capability option from a Fitted Distribution red triangle menu, a window appears with the following options:

Enter Spec Limits Enables you to manually enter specification limits. To use the fitted distribution to calculate specification limits, leave this section blank and use the options under Calculate Quantile Spec Limits Options.

Calculate Quantile Spec Limits Options Enables you to calculate specification limits based on the fitted distribution. There are two methods available.

In the first method, you enter probabilities associated with the quantiles of the fitted distribution to calculate specification limits.

In the second method, you enter a K-Sigma Multiplier value that is used to calculate specification limits. This method has options for creating two-sided or one-sided limits.

After entering probabilities or a value for sigma multiplier, click **Calculate Spec Limits** to calculate the specification limits. These limits are entered into the Enter Spec Limits panel. Click **OK** to accept these limits and generate the Process Capability report.

Process Capability Options Contains the following options:

The Moving Range Options outline contains options that enable you to select the type of moving range statistic. See *Quality and Process Methods*.

The Nonnormal Distribution Options outline contains options that enable you to select methods used for nonnormal process capability calculations. See *Quality and Process Methods*.

For more information about the Process Capability options and report, see *Quality and Process Methods*.

Note: You can set preferences for many of the options in the Process Capability report in Distribution at **File > Preferences > Platforms > Process Capability**.

Remove Fit Removes the distribution fit from the report window.

Save Options for Continuous Variables

Use the Save menu options in the Distribution platform to save information about continuous variables. Each Save option generates a new column in the current data table. The new column is named by appending the variable name (denoted <colname> in the following definitions) to the Save command name (Table 3.1).

Select the Save options repeatedly to save the same information multiple times under different circumstances, such as before and after combining histogram bars. If you use a Save option multiple times, the column name is numbered (name1, name2, and so on) to ensure unique column names.

Table 3.1 Descriptions of Save Options

Option	Column Added to Data Table	Description
Level Numbers	Level <colname>	The level number of each observation corresponds to the histogram bar that contains the observation. The histogram bars are numbered from low to high, beginning with 1. **Note:** To maintain source information, value labels are added to the new column, but they are turned off by default.

Table 3.1 Descriptions of Save Options *(Continued)*

Option	Column Added to Data Table	Description
Level Midpoints	Midpoint <colname>	The midpoint value for each observation is computed by adding half the level width to the lower level bound. **Note:** To maintain source information, value labels are added to the new column, but they are turned off by default.
Ranks	Ranked <colname>	Provides a ranking for each of the corresponding column's values starting at 1. Duplicate response values are assigned consecutive ranks in order of their occurrence in the data table.
Ranks averaged	RankAvgd <colname>	If a value is unique, then the averaged rank is the same as the rank. If a value occurs k times, the average rank is computed as the sum of the value's ranks divided by k.
Prob Scores	Prob <colname>	For N nonmissing scores, the probability score of a value is computed as the averaged rank of that value divided by $N + 1$. This column is similar to the empirical cumulative distribution function.
Normal Quantiles	N-Quantile <colname>	Saves the Normal quantiles. See "Statistical Details for the Normal Quantile Plot".
Standardized	Std <colname>	Saves standardized values. See "Statistical Details for Saving Standardized Data".
Centered	Centered <colname>	Saves values for centering on zero.
Robust Standardized	Robust Std <colname>	Saves a column that contains the response value centered around the robust mean and standardized using the robust standard deviation.
Robust Centered	Robust Centered <colname>	Saves a column that contains the response value centered around the robust mean.
Script to Log	(none)	Prints the script to the log window. Run the script to re-create the analysis.

Additional Examples of the Distribution Platform

This section contains examples using the Distribution platform.

- "Example of Selecting Data in Multiple Histograms"
- "Example of a By Variable"
- "Example of Testing Probabilities for Two Levels"
- "Example of Testing Probabilities for More Than Two Levels"
- "Example of Prediction Intervals"
- "Example of Tolerance Intervals"
- "Example of Process Capability"

Example of Selecting Data in Multiple Histograms

This example shows you how to select and highlight data in several histograms at once.

1. Select **Help > Sample Data Folder** and open Companies.jmp.
2. Select **Analyze > Distribution**.
3. Select Type and Size Co and click **Y, Columns.**
4. Click **OK**.

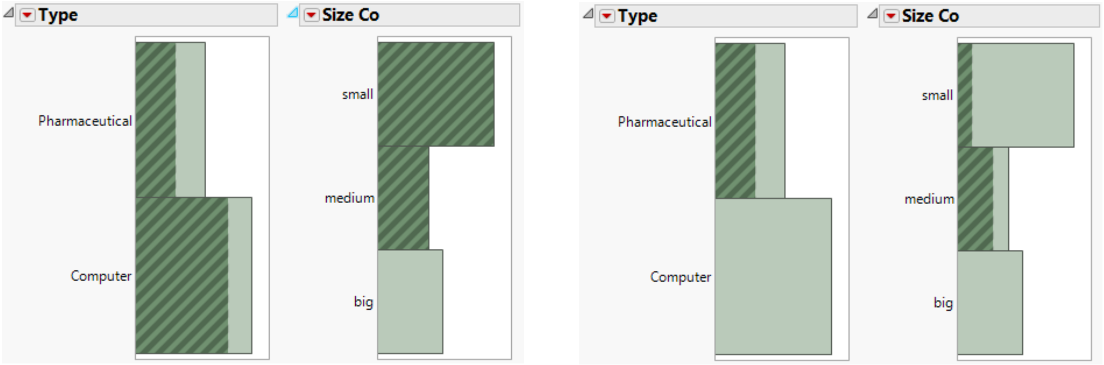
   You want to see the type distribution of companies that are small.
5. Click the bar next to small.

   You can see that there are more small computer companies than there are pharmaceutical companies. To broaden your selection, add medium companies.
6. Hold down the Shift key. In the Size Co histogram, click the bar next to medium.

   You can see the type distribution of small and medium sized companies. See Figure 3.13 at left. To narrow your selection, you want to see the small and medium pharmaceutical companies only.
7. Hold down the Ctrl and Shift keys (on Windows) or the Command and Shift keys (on macOS). In the Type histogram, click in the Computer bar to deselect it.

   You can see how many of the small and medium companies are pharmaceutical companies. See Figure 3.13 at right.

Figure 3.13 Selecting Data in Multiple Histograms



Broaden the selection using the Shift key.

Narrow the selection using Ctrl + Shift (Windows) or Command + Shift (macOS).

Example of a By Variable

This example shows you how to use a By variable to create a separate analysis for each level of a categorical variable.

1. Select **Help > Sample Data Folder** and open Lipid Data.jmp.
2. Select **Analyze > Distribution**.
3. Select Cholesterol and click **Y, Columns**.
4. Select Gender and click **By**.

   This results in a separate analysis for each level of Gender (female and male).
5. Click **OK**.

Change the orientation of the histograms and the reports.

6. Click the Distributions red triangle and select **Stack**.

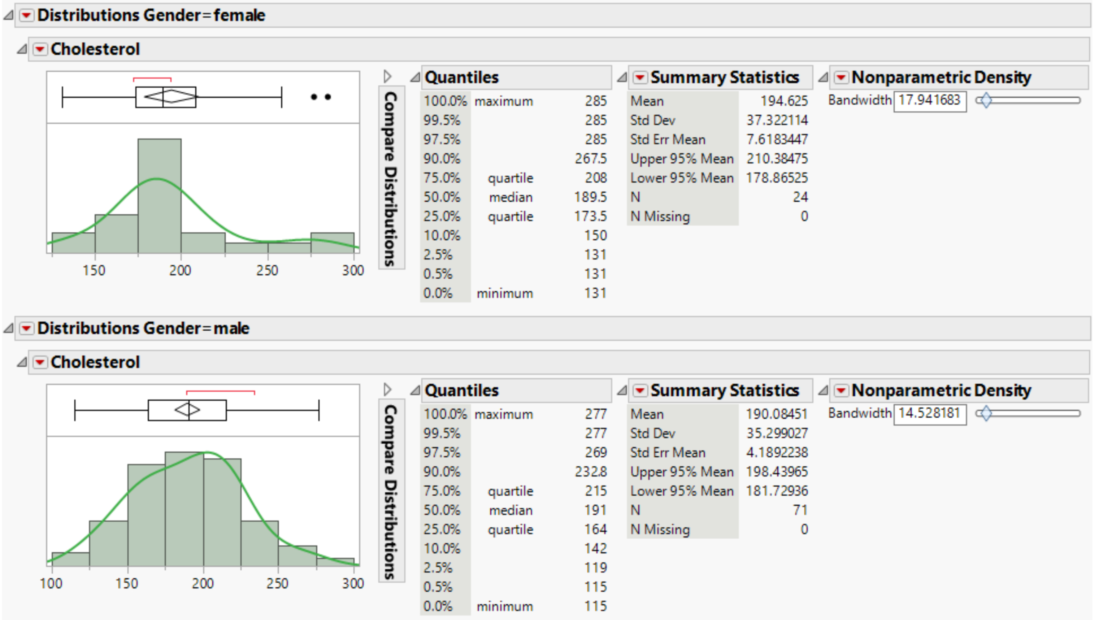
Add a smooth curve to both histograms.

7. Hold down the Ctrl key. Click the Cholesterol red triangle and select **Continuous Fit > Fit Smooth Curve**.

Hide the Compare Distributions report.

8. Hold down the Ctrl key. Click the gray disclosure icon next to **Compare Distributions**.

Figure 3.14 Separate Distributions by Gender



Distributions Gender=female

Cholesterol

Quantiles

100.0%	maximum	285
99.5%		285
97.5%		285
90.0%		267.5
75.0%	quartile	208
50.0%	median	189.5
25.0%	quartile	173.5
10.0%		150
2.5%		131
0.5%		131
0.0%	minimum	131

Summary Statistics

Mean	194.625
Std Dev	37.322114
Std Err Mean	7.6183447
Upper 95% Mean	210.38475
Lower 95% Mean	178.86525
N	24
N Missing	0

Nonparametric Density

Bandwidth 17.941683

Distributions Gender=male

Cholesterol

Quantiles

100.0%	maximum	277
99.5%		277
97.5%		269
90.0%		232.8
75.0%	quartile	215
50.0%	median	191
25.0%	quartile	164
10.0%		142
2.5%		119
0.5%		115
0.0%	minimum	115

Summary Statistics

Mean	190.08451
Std Dev	35.299027
Std Err Mean	4.1892238
Upper 95% Mean	198.43965
Lower 95% Mean	181.72936
N	71
N Missing	0

Nonparametric Density

Bandwidth 14.528181

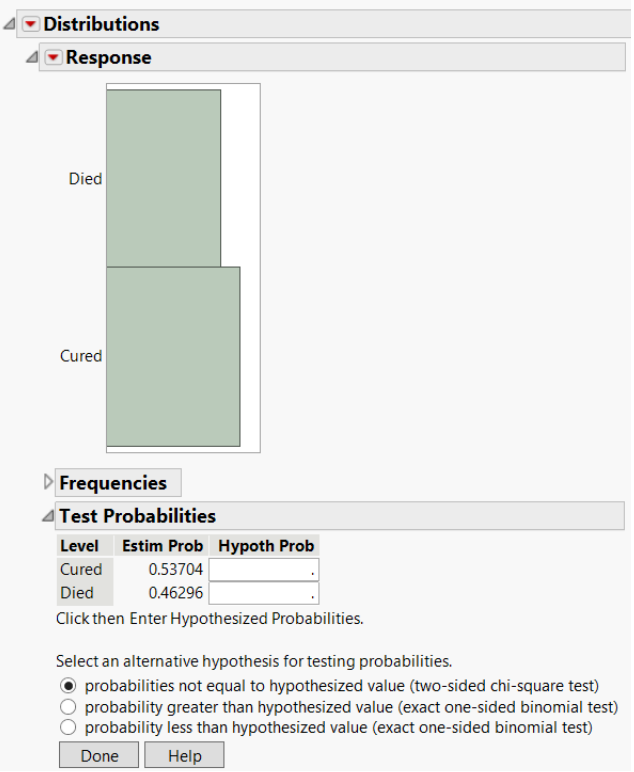
Example of Testing Probabilities for Two Levels

This example shows you how to create a Test Probabilities report in the Distribution platform for a variable with exactly two levels.

Initiate the Test Probabilities Report

1. Select **Help > Sample Data Folder** and open Penicillin.jmp.
2. Select **Analyze > Distribution**.
3. Select Response and click **Y, Columns**.
4. Click **OK**.
5. Click the Response red triangle and select **Test Probabilities**.

Figure 3.15 Test Probabilities Report Options for a Variable with Exactly Two Levels



Level	Estim Prob	Hypoth Prob
Cured	0.53704	.
Died	0.46296	.

Click then Enter Hypothesized Probabilities.

Select an alternative hypothesis for testing probabilities.

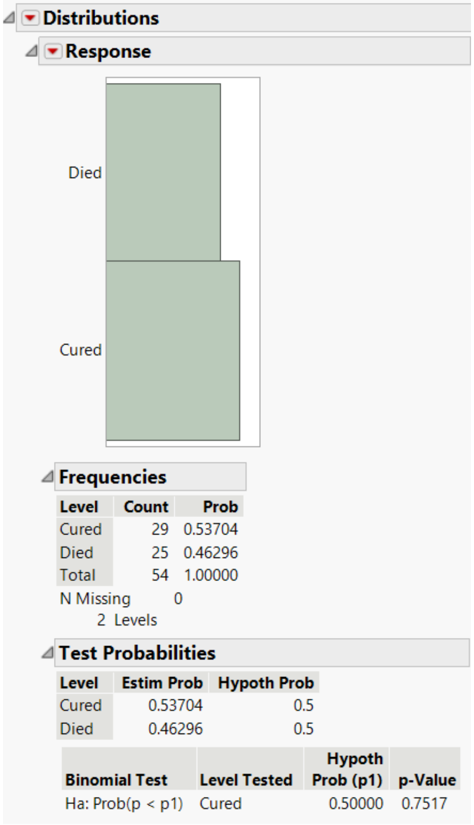
● probabilities not equal to hypothesized value (two-sided chi-square test)
○ probability greater than hypothesized value (exact one-sided binomial test)
○ probability less than hypothesized value (exact one-sided binomial test)

Generate the Test Probabilities Report

1. Type 0.5 in both Hypoth Prob fields.
2. Click the **probability less than hypothesized value** button.
3. Click **Done**.

   Exact probabilities are calculated for the binomial test.

Figure 3.16 Examples of Test Probabilities Reports for a Variable with Exactly Two Levels



Frequencies

Level	Count	Prob
Cured	29	0.53704
Died	25	0.46296
Total	54	1.00000

N Missing 0
2 Levels

Test Probabilities

Level	Estim Prob	Hypoth Prob
Cured	0.53704	0.5
Died	0.46296	0.5

Binomial Test	Level Tested	Hypoth Prob (p1)	p-Value
Ha: Prob(p < p1)	Cured	0.50000	0.7517

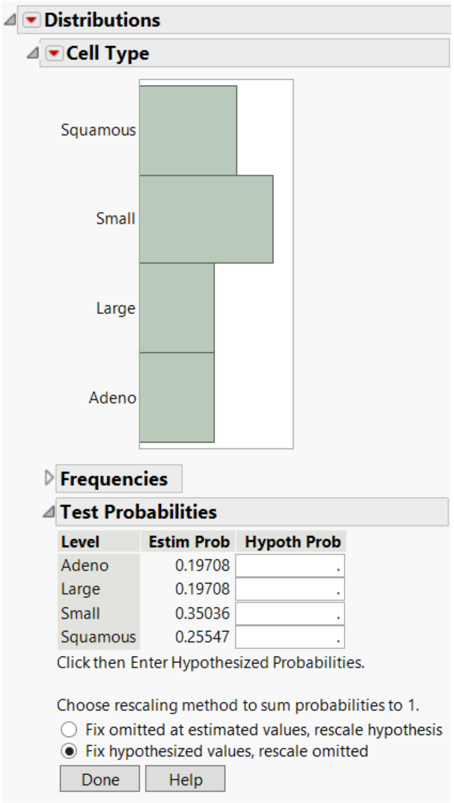
Example of Testing Probabilities for More Than Two Levels

This example shows you how to create a Test Probabilities report in the Distribution platform for a variable with more than two levels.

Initiate the Test Probabilities Report

1. Select **Help > Sample Data Folder** and open VA Lung Cancer.jmp.
2. Select **Analyze > Distribution**.
3. Select Cell Type and click **Y, Columns**.
4. Click **OK**.
5. Click the Cell Type red triangle and select **Test Probabilities**.

Figure 3.17 Test Probabilities Report Options for a Variable with More Than Two Levels



Level	Estim Prob	Hypoth Prob
Adeno	0.19708	.
Large	0.19708	.
Small	0.35036	.
Squamous	0.25547	.

Click then Enter Hypothesized Probabilities.

Choose rescaling method to sum probabilities to 1.

○ Fix omitted at estimated values, rescale hypothesis
● Fix hypothesized values, rescale omitted

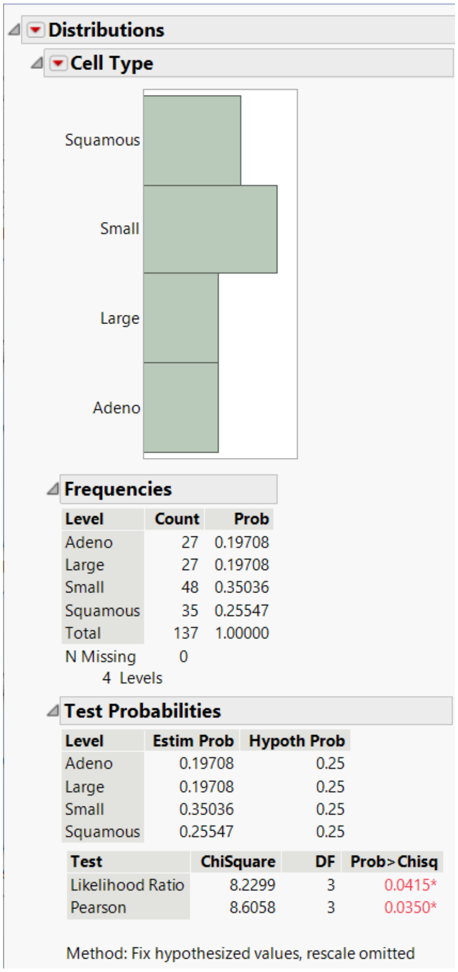
Generate the Test Probabilities Report

To generate a test probabilities report for a variable with more than two levels:

1. Type 0.25 in all four Hypoth Prob fields.
2. Select **Fix hypothesized values, rescale omitted**.
3. Click **Done**.

   Likelihood Ratio and Pearson Chi-square tests are calculated.

Figure 3.18 Test Probabilities Reports for a Variable with More Than Two Levels



Frequencies

Level	Count	Prob
Adeno	27	0.19708
Large	27	0.19708
Small	48	0.35036
Squamous	35	0.25547
Total	137	1.00000
N Missing	0	

4 Levels

Test Probabilities

Level	Estim Prob	Hypoth Prob
Adeno	0.19708	0.25
Large	0.19708	0.25
Small	0.35036	0.25
Squamous	0.25547	0.25

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	8.2299	3	0.0415*
Pearson	8.6058	3	0.0350*

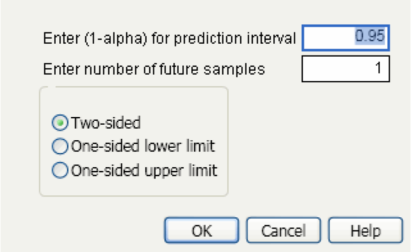
Method: Fix hypothesized values, rescale omitted

Example of Prediction Intervals

This example shows you how to add prediction intervals to the Distribution report. Suppose you are interested in computing prediction intervals for the next 10 observations of ozone level.

1. Select **Help > Sample Data Folder** and open Cities.jmp.
2. Select **Analyze > Distribution**.
3. Select OZONE and click **Y, Columns**.
4. Click **OK**.
5. Click the OZONE red triangle and select **Prediction Interval**.

Figure 3.19 The Prediction Intervals Window



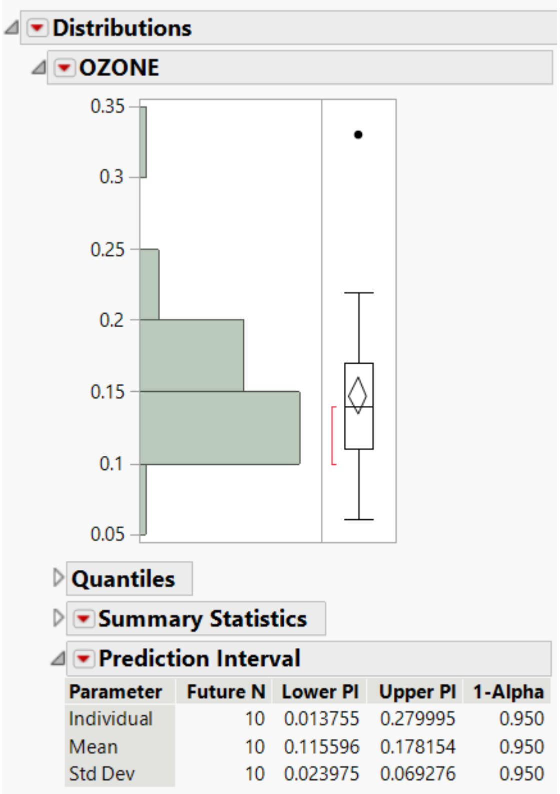
Enter (1-alpha) for prediction interval 0.95

Enter number of future samples 1

● Two-sided
○ One-sided lower limit
○ One-sided upper limit

6. In the Prediction Intervals window, type 10 next to **Enter number of future samples**.
7. Click **OK**.

Figure 3.20 Example of a Prediction Interval Report



Parameter	Future N	Lower PI	Upper PI	1-Alpha
Individual	10	0.013755	0.279995	0.950
Mean	10	0.115596	0.178154	0.950
Std Dev	10	0.023975	0.069276	0.950

In this example, you can be 95% confident about the following:

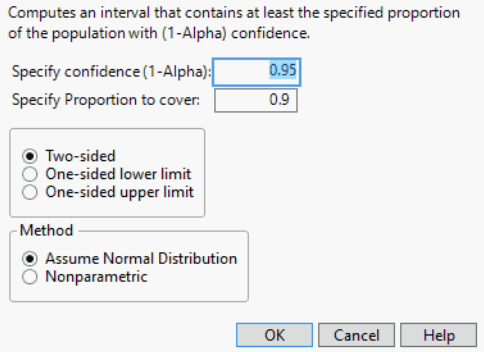
- Each of the next 10 observations will be between 0.013755 and 0.279995.
- The mean of the next 10 observations will be between 0.115596 and 0.178154.
- The standard deviation of the next 10 observations will be between 0.023975 and 0.069276.

Example of Tolerance Intervals

This example shows you how to add tolerance intervals to the Distribution report. Suppose you want to estimate an interval that contains 90% of ozone level measurements.

1. Select **Help > Sample Data Folder** and open Cities.jmp.
2. Select **Analyze > Distribution**.
3. Select OZONE and click **Y, Columns**.
4. Click **OK**.
5. Click the OZONE red triangle and select **Tolerance Interval**.

Figure 3.21 The Tolerance Intervals Window



Computes an interval that contains at least the specified proportion of the population with (1-Alpha) confidence.

Specify confidence (1-Alpha): 0.95

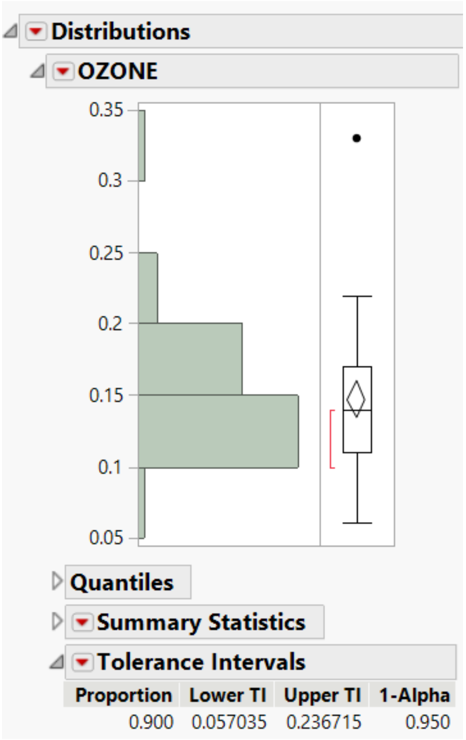
Specify Proportion to cover: 0.9

● Two-sided
○ One-sided lower limit
○ One-sided upper limit

Method
● Assume Normal Distribution
○ Nonparametric

6. Keep the default selections, and click **OK**.

Figure 3.22 Example of a Tolerance Interval Report



Proportion	Lower TI	Upper TI	1-Alpha
0.900	0.057035	0.236715	0.950

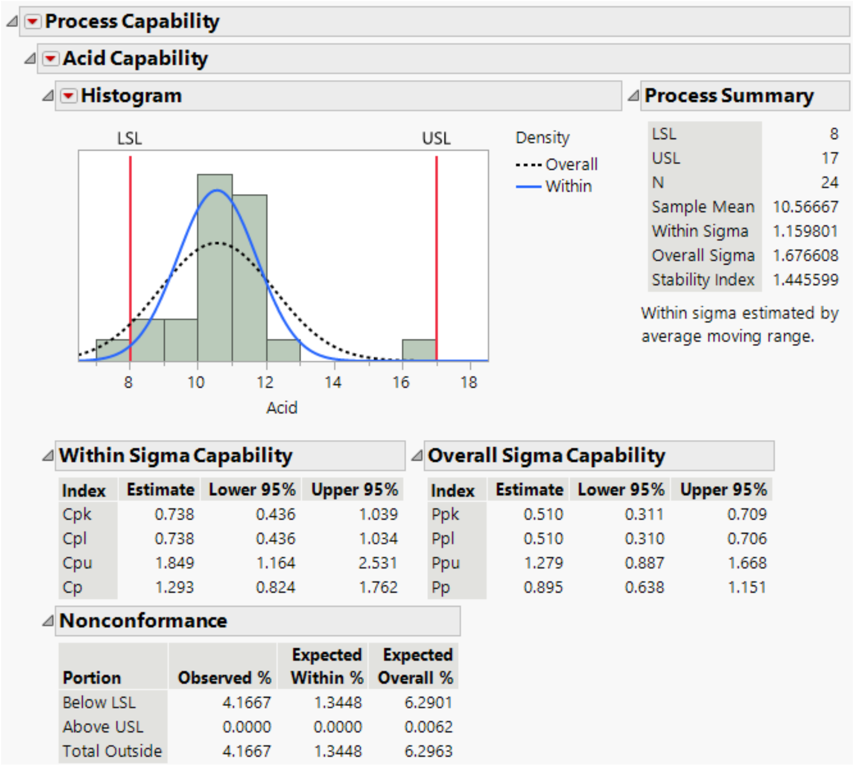
In this example, you can be 95% confident that at least 90% of the population lie between 0.057035 and 0.236715, based on the Lower TI (tolerance interval) and Upper TI values.

Example of Process Capability

This example shows you how to add process capability results to the Distribution report. Suppose you want to characterize the acidity of pickles. The lower and upper specification limits are 8 and 17, respectively.

1. Select **Help > Sample Data Folder** and open Quality Control/Pickles.jmp.
2. Select **Analyze > Distribution**.
3. Select Acid and click **Y, Columns**.
4. Click **OK**.
5. Click the Acid red triangle and select **Process Capability**.
6. Type 8 for the **LSL** (lower specification limit).
7. Type 17 for the **USL** (upper specification limit).
8. Click **OK**.

Figure 3.23 Example of the Process Capability Report



Process Capability

Acid Capability

Histogram

Process Summary

LSL	8
USL	17
N	24
Sample Mean	10.56667
Within Sigma	1.159801
Overall Sigma	1.676608
Stability Index	1.445599

Within sigma estimated by average moving range.

Within Sigma Capability

Index	Estimate	Lower 95%	Upper 95%
Cpk	0.738	0.436	1.039
Cpl	0.738	0.436	1.034
Cpu	1.849	1.164	2.531
Cp	1.293	0.824	1.762

Overall Sigma Capability

Index	Estimate	Lower 95%	Upper 95%
Ppk	0.510	0.311	0.709
Ppl	0.510	0.310	0.706
Ppu	1.279	0.887	1.668
Pp	0.895	0.638	1.151

Nonconformance

Portion	Observed %	Expected Within %	Expected Overall %
Below LSL	4.1667	1.3448	6.2901
Above USL	0.0000	0.0000	0.0062
Total Outside	4.1667	1.3448	6.2963

The Process Capability results are added to the report. The specification limits appear on the histogram in the Process Capability report so that the data can be visually compared to the limits. As you can see, some of the acidity levels are below the lower specification limit, and some are very close to the upper specification limit. The Ppk value is 0.510, indicating a process that is not capable, relative to the given specification limits.

Statistical Details for the Distribution Platform

This section contains statistical details for the Distribution platform.

- "Statistical Details for Standard Error Bars"
- "Statistical Details for Quantiles"
- "Statistical Details for Summary Statistics"
- "Statistical Details for the Normal Quantile Plot"
- "Statistical Details for the Wilcoxon Signed Rank Test"
- "Statistical Details for the Standard Deviation Test"
- "Statistical Details for Normal Quantiles"
- "Statistical Details for Saving Standardized Data"
- "Statistical Details for Prediction Intervals"
- "Statistical Details for Tolerance Intervals"
- "Statistical Details for Continuous Fit Distributions"
- "Statistical Details for Discrete Fit Distributions"

Statistical Details for Standard Error Bars

Standard error bars in the Distribution platform are calculated using the standard error $\sqrt{np_i(1-p_i)}$ where $p_i = n_i / n$.

Statistical Details for Quantiles

This section describes how quantiles are computed in the Distribution platform.

To compute the pth quantile of n nonmissing values in a column, arrange the n values in ascending order and call these column values $y_1, y_2, ..., y_n$. Compute the rank number for the pth quantile as $p / 100(n + 1)$.

- If the result is an integer, the pth quantile is that rank's corresponding value.
- If the result is not an integer, the pth quantile is found by interpolation. The pth quantile, denoted q_p, is defined as follows:

$$q_p = (1-f)y_i + (f)y_{i+1}$$

where:

- n is the number of nonmissing values for a variable
- $y_1, y_2, ..., y_n$ represents the ordered values of the variable

- y_{n+1} is taken to be y_n
- i is the integer part and f is the fractional part of $(n+1)p$.
- $(n + 1)p = i + f$

For example, suppose a data table has 15 rows and you want to find the 75th and 90th quantile values of a continuous column. After the column is arranged in ascending order, the ranks that contain these quantiles are computed as follows:

$$\frac{75}{100}(15+1) = 12 \text{ and } \frac{90}{100}(15+1) = 14.4$$

The value y_{12} is the 75th quantile. The 90th quantile is interpolated by computing a weighted average of the 14th and 15th ranked values as $y_{90} = 0.6y_{14} + 0.4y_{15}$.

Statistical Details for Summary Statistics

This section contains statistical details for specific statistics in the Distribution Summary Statistics report.

Mean

The mean is the sum of the nonmissing values divided by the number of nonmissing values. If you assigned a **Weight** or **Freq** variable, the mean is computed as follows:

1. Each column value is multiplied by its corresponding weight or frequency.
2. These values are added and divided by the sum of the weights or frequencies.

Std Dev

The standard deviation measures the spread of a distribution around the mean. It is often denoted as s and is the square root of the sample variance, denoted s^2.

$$s = \sqrt{s^2}$$

where

$$s^2 = \sum_{i=1}^{N} \frac{w_i(y_i - \bar{y}_w)^2}{N-1}$$

$\bar{y}_w$ = weighted mean

Std Err Mean

The standard error mean is computed by dividing the sample standard deviation, *s*, by the square root of *N*. In the launch window, if you specified a column for Weight or Freq, then the denominator is the square root of the sum of the weights or frequencies.

Skewness

Skewness is based on the third moment about the mean and is computed as follows:

$$\sum w_i^{\frac{3}{2}} z_i^3 \frac{N}{(N-1)(N-2)} \quad \text{where} \quad z_i = \frac{x_i - \bar{x}}{s}$$

and w_i is a weight term (= 1 for equally weighted items).

Kurtosis

Kurtosis is based on the fourth moment about the mean and is computed as follows:

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

where w_i is a weight term (= 1 for equally weighted items). Using this formula, the Normal distribution has a kurtosis of 0. This formula is often referred to as the excess kurtosis.

Statistical Details for the Normal Quantile Plot

The empirical cumulative probability for each normal quantile plot value in the Distribution platform is computed as follows:

$$\frac{r_i}{N+1}$$

where r_i is the rank of the *i*th observation, and *N* is the number of nonmissing (and nonexcluded) observations.

The normal quantile values are computed as follows:

$$\Phi^{-1}\left(\frac{r_i}{N+1}\right)$$

where Φ is the cumulative probability distribution function for the normal distribution.

These normal quantile values are Van Der Waerden approximations to the order statistics that are expected for the normal distribution.

Statistical Details for the Wilcoxon Signed Rank Test

The Wilcoxon signed-rank test in the Distribution platform can be used to test for the median of a single population or to test matched-pairs data for a common median. In the case of matched pairs, the test reduces to testing the single population of paired differences for a median of 0. The test assumes that the underlying population is symmetric.

The Wilcoxon test accommodates tied values. The test statistic is adjusted for differences of zero using a method suggested by Pratt. See Lehmann and D'Abrera (2006), Pratt (1959), and Cureton (1967).

Testing for the Median of a Single Population

- There are N observations:

  $X_1, X_2, \ldots, X_N$
- The null hypothesis is:

  H_0: distribution of X is symmetric around m
- The differences between observations and the hypothesized value m are calculated as follows:

  $D_j = X_j - m$

Testing for the Equality of Two Population Medians with Matched Pairs Data

A special case of the Wilcoxon signed-rank test is applied to matched-pairs data.

- There are N pairs of observations from two populations:

  $X_1, X_2, \ldots, X_N$ and $Y_1, Y_2, \ldots, Y_N$
- The null hypothesis is:

  H_0: distribution of $X - Y$ is symmetric around 0
- The differences between pairs of observations are calculated as follows:

  $D_j = X_j - Y_j$

Wilcoxon Signed-Rank Test Statistic

The test statistic is based on the sum of the signed ranks. Signed ranks are defined as follows:

- The absolute values of the differences, $|D_j|$, are ranked from smallest to largest.
- The ranks start with the value 1, even if there are differences of zero.
- When there are tied absolute differences, they are assigned the average, or *midrank*, of the ranks of the observations.

Denote the rank or midrank for a difference D_j by R_j. Define the signed rank for D_j as follows:

- If the difference D_j is positive, the signed rank is R_j.
- If the difference D_j is zero, the signed rank is 0.
- If the difference D_j is negative, the signed rank is $-R_j$.

The signed-rank statistic is computed as follows:

$$S = \frac{1}{2}\sum_{j=1}^{N} \text{signed ranks}$$

Define the following:

d_0 is the number of signed ranks that equal zero

R^+ is the sum of the positive signed ranks

Then the following holds:

$$S = R^+ - \frac{1}{4}[N(N+1) - d_0(d_0+1)]$$

Wilcoxon Signed-Rank Test P-Values

For $N \le 20$, exact p-values are calculated.

For $N > 20$, a Student's t approximation to the statistic defined below is used. Note that a correction for ties is applied. See Iman (1974) and Lehmann and D'Abrera (2006).

Under the null hypothesis, the mean of S is zero. The variance of S is given by the following:

$$Var(S) = \frac{1}{24}\left[N(N+1)(2N+1) - d_0(d_0+1)(2d_0+1) - \frac{1}{2}\sum_{i>0} d_i(d_i+1)(d_i-1)\right]$$

The last summation in the expression for $Var(S)$ is a correction for ties. The notation d_i for i > 0 represents the number of values in the i^{th} group of nonzero signed ranks. (If there are no ties for a given signed rank, then $d_i = 1$ and the summand is 0.)

The statistic t given by the following has an approximate t distribution with $N - 1$ degrees of freedom:

$$t = \frac{S}{\sqrt{\dfrac{N \cdot Var(S) - S^2}{N-1}}}$$

Statistical Details for the Standard Deviation Test

The test statistic in the standard deviation test in the Distribution platform is computed as follows:

$$\frac{(n-1)s^2}{\sigma^2}$$

The test statistic is distributed as a chi-square variable with n - 1 degrees of freedom when the population is normal.

The Min PValue is the *p*-value of the two-tailed test, and is calculated as follows:

2*min(p1,p2)

where p1 is the lower one-tail *p*-value and p2 is the upper one-tail *p*-value.

Statistical Details for Normal Quantiles

The normal quantile values in the Distribution platform are computed as follows:

$$\Phi^{-1}\left(\frac{r_i}{N+1}\right)$$

where:

Φ is the cumulative probability distribution function for the normal distribution.

r_i is the rank of the ith observation.

N is the number of nonmissing observations.

Statistical Details for Saving Standardized Data

The standardized values in the Distribution platform are computed as follows:

$$\frac{X-\overline{X}}{S_X}$$

where:

X is the original column

$\overline{X}$ is the mean of column X

S_X is the standard deviation of column X

Statistical Details for Prediction Intervals

The prediction intervals in the Distribution platform are defined as follows:

- For m future observations:

$$[\underset{\sim}{y}_m, \tilde{y}_m] = \overline{X} \pm t_{(1-\alpha/2m;n-1)} \times \sqrt{1+\frac{1}{n}} \times s \quad \text{for} \quad m \geq 1$$

- For the mean of m future observations:

$$[Y_l, Y_u] = \overline{X} \pm t_{(1-\alpha/2,\, n-1)} \times \sqrt{\frac{1}{m}+\frac{1}{n}} \times s \quad \text{for} \quad m \geq 1 \cdot$$

- For the standard deviation of m future observations:

$$[s_l, s_u] = \left[s \times \sqrt{\frac{1}{F_{(1-\alpha/2;(n-1,\, m-1))}}},\ s \times \sqrt{F_{(1-\alpha/2;(m-1,\, n-1))}} \right] \quad \text{for} \quad m \geq 2$$

where m = number of future observations, and n = number of points in current analysis sample.

- The one-sided intervals are formed by using 1-α in the quantile functions.

See Meeker et al. (2017, ch. 4).

Statistical Details for Tolerance Intervals

This section contains statistical details for one-sided and two-sided tolerance intervals in the Distribution platform.

Normal Distribution-Based Intervals

One-Sided Interval

The one-sided interval is computed as follows:

Lower Limit = $\bar{x} - k's$

Upper Limit = $\bar{x} + k's$

where

$$k' = t(1-\alpha, n-1, \Phi^{-1}(p) \cdot \sqrt{n}) / \sqrt{n}$$

s is the standard deviation

t is the quantile from the non-central t-distribution

Φ^{-1} is the standard normal quantile

Two-Sided Interval

The two-sided interval is computed as follows:

$$[T_{p_L}, T_{p_U}] = [\bar{x} - k_{(1-\alpha/2;p,n)}s, \bar{x} + k_{(1-\alpha/2;p,n)}s]$$

where s is the standard deviation and $k_{(1\text{-}\alpha/2;\, p,n)}$ is a constant.

To determine k, consider the fraction of the population captured by the tolerance interval. Tamhane and Dunlop (2000) define this fraction as follows:

$$\Phi\left(\frac{\bar{x} + ks - \mu}{\sigma}\right) - \Phi\left(\frac{\bar{x} - ks - \mu}{\sigma}\right)$$

where Φ denotes the standard normal cdf (cumulative distribution function).

Therefore, k solves the following equation:

$$P\left\{\Phi\left(\frac{\overline{X} + ks - \mu}{\sigma}\right) - \Phi\left(\frac{\overline{X} - ks - \mu}{\sigma}\right) \geq 1 - \gamma\right\} = 1 - \alpha$$

where $1 - \gamma$ is the fraction of all future observations contained in the tolerance interval.

For more information about normal distribution-based tolerance intervals, see Tables J.1a, J.1b, J.6a, and J.6b of Meeker et al. (2017).

Nonparametric Intervals

One-Sided Lower Limit

The lower $100(1 - \alpha)\%$ one-sided tolerance limit to contain at least a proportion β of the sampled distribution from a sample of size n is the order statistic $x_{(l)}$. The index l is computed as follows:

$$l = n - \Phi_{bin}^{-1}(1 - \alpha, n, \beta)$$

where $\Phi^{-1}_{\text{bin}}(1\text{-}\alpha, n, \beta)$ is the $(1 - \alpha)^{\text{th}}$ quantile of the binomial distribution with n trials and probability of success β.

The actual confidence level is computed as $\Phi_{\text{bin}}(n\text{-}l, n, \beta)$, where $\Phi_{\text{bin}}(x, n, \beta)$ is the probability of a binomially distributed random variable with n trials and probability of success β being less than or equal to x.

Note that to compute a lower one-sided distribution-free tolerance interval, the sample size n must be at least as large as $(\log\alpha)/(\log\beta)$.

One-Sided Upper Limit

The upper 100(1 - α)% one-sided tolerance limit to contain at least a proportion β of the sampled distribution from a sample of size n is the order statistic $x_{(u)}$. The index u is computed as follows:

$$u = 1 + \Phi_{bin}^{-1}(1 - \alpha, n, \beta)$$

where $\Phi^{-1}_{\text{bin}}(1-\alpha, n, \beta)$ is the (1 - α)th quantile of the binomial distribution with n trials and probability of success β.

The actual confidence level is computed as $\Phi_{\text{bin}}(u\text{-}1, n, \beta)$, where $\Phi_{\text{bin}}(x, n, \beta)$ is the probability of a binomially distributed random variable with n trials and probability of success β being less than or equal to x.

Note that to compute an upper one-sided distribution-free tolerance interval, the sample size n must be at least as large as $(\log\alpha)/(\log\beta)$.

Two-Sided Interval

The 100(1 - α)% two-sided tolerance interval to contain at least a proportion β of the sampled distribution from a sample of size n is computed as follows:

$$[\tilde{T}_{p_L}, \tilde{T}_{p_U}] = [x_{(l)}, x_{(u)}]$$

where $x_{(i)}$ is the ith order statistic and l and u are computed as follows:

Let $\nu = n - \Phi^{-1}_{\text{bin}}(1-\alpha, n, \beta)$, where $\Phi^{-1}_{\text{bin}}(1-\alpha, n, \beta)$ is the (1 - α)th quantile of the binomial distribution with n trials and probability of success β. If ν is less than 2, a two-sided distribution-free tolerance interval cannot be computed. If ν is greater than or equal to 2, $l = \text{floor}(\nu/2)$ and $u = \text{floor}(n + 1 - \nu/2)$.

The actual confidence level is computed as $\Phi_{\text{bin}}(u\text{-}l\text{-}1, n, \beta)$, where $\Phi_{\text{bin}}(x, n, \beta)$ is the probability of a binomially distributed random variable with n trials and probability of success β being less than or equal to x.

Note that to compute a two-sided distribution-free tolerance interval, the sample size n must be at least as large as the n in the following equation:

$$1 - \alpha = 1 - n\beta^{n-1} + (n-1)\beta^{n}$$

For more information about distribution-free tolerance intervals, see Meeker et al. (2017, sec. 5.3).

Statistical Details for Continuous Fit Distributions

This section contains statistical details for the options in the Continuous Fit menu in the Distribution platform. Unless otherwise specified, confidence intervals for parameter estimates use likelihood-based calculations. If the Y column has a Detection Limits column property, the Continuous Fit options fit a censored distribution and only a subset of distributions are available. For more information about fitting distributions to censored data, see Meeker and Escobar (1998).

Fit Normal

The Fit Normal option estimates the two parameters of the normal distribution:

- μ (the mean) defines the location of the distribution on the *x*-axis
- σ (standard deviation) defines the dispersion or spread of the distribution

The standard normal distribution occurs when $\mu = 0$ and $\sigma = 1$.

$$\text{pdf:} \quad \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{for} \quad -\infty < x < \infty; \quad -\infty < \mu < \infty; \quad 0 < \sigma$$

$E(x) = \mu$

$\mathrm{Var}(x) = \sigma^2$

Note: Confidence intervals for the mean estimate are based on the *t* distribution. Confidence intervals for the scale parameter are based on the χ^2 distribution.

Fit Cauchy

The Fit Cauchy option fits a Cauchy distribution with location μ and scale σ.

$$\text{pdf:} \quad \left\{\pi\sigma\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]\right\}^{-1} \quad \text{for} \quad -\infty < x < \infty; \quad -\infty < \mu < \infty; \quad 0 < \sigma$$

$E(x)$ = undefined

$\mathrm{Var}(x)$ = undefined

Fit Student's t

The Fit Student's *t* option fits a Student's *t* distribution with location μ, scale σ, and degrees of freedom ν.

$$\text{pdf:}\quad \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\nu\pi\sigma^2}}\left[1+\frac{(x-\mu)^2}{\nu\sigma^2}\right]^{-\left(\frac{\nu+1}{2}\right)} \quad \text{for} \quad -\infty < x < \infty\,;\quad -\infty < \mu < \infty\,;\quad 0 < \sigma;\quad 1 \le \nu$$

Fit SHASH

The Fit SHASH option fits a sinh-arcsinh (SHASH) distribution. The SHASH distribution is based on a transformation of the normal distribution and includes the normal distribution as a special case. It can be symmetric or asymmetric. The shape is determined by the two shape parameters, γ and δ. For more information about the SHASH distribution, see Jones and Pewsey (2009).

$$\text{pdf:}\quad f(x) = \frac{\delta\cosh(w)}{\sqrt{\sigma^2+(x-\theta)^2}}\phi[\sinh(w)] \quad \text{for} \quad -\infty < \gamma, x, \theta < \infty\,;\quad 0 < \delta, \sigma$$

where

$\phi(\cdot)$ is the standard normal pdf

$$w = \gamma + \delta\sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)$$

- When $\gamma = 0$ and $\delta = 1$, the SHASH distribution is equivalent to the normal distribution with location θ and scale σ.
- The transformation $\sinh(w)$ is normally distributed with $\mu = 0$ and $\sigma = 1$.

Fit Exponential

The exponential distribution is especially useful for describing events that randomly occur over time, such as survival data. The exponential distribution might also be useful for modeling elapsed time between the occurrence of non-overlapping events. Examples of non-overlapping events include the following: the time between a user's computer query and response of the server, the arrival of customers at a service desk, or calls coming in at a switchboard.

The Exponential distribution is a special case of the two-parameter Weibull when $\beta = 1$ and $\alpha = \sigma$, and also a special case of the Gamma distribution when $\alpha = 1$.

pdf: $$\frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right) \quad \text{for} \quad 0 < \sigma; \quad 0 \le x$$

$E(x) = \sigma$

$\text{Var}(x) = \sigma^2$

Devore (1995) notes that an exponential distribution is *memoryless*. Memoryless means that if you check a component after t hours and it is still working, the distribution of additional lifetime (the conditional probability of additional life given that the component has lived until t) is the same as the original distribution.

Fit Gamma

The Fit Gamma option estimates the gamma distribution parameters, $\alpha > 0$ and $\sigma > 0$. The parameter α, called alpha in the fitted gamma report, describes shape or curvature. The parameter σ, called sigma, is the scale parameter of the distribution. The data must be greater than zero.

pdf: $$\frac{1}{\Gamma(\alpha)\sigma^{\alpha}}x^{\alpha - 1}\exp(-x/\sigma) \quad \text{for} \quad 0 < x; \quad 0 < \alpha,\sigma$$

$E(x) = \alpha\sigma$

$\text{Var}(x) = \alpha\sigma^2$

- The *standard* gamma distribution has $\sigma = 1$. Sigma is called the scale parameter because values other than 1 stretch or compress the distribution along the horizontal axis.
- The chi-square $\chi^2_{(\nu)}$ distribution occurs when $\sigma = 2$ and $\alpha = \nu/2$.
- The exponential distribution occurs when $\alpha = 1$.

The standard gamma density function is strictly decreasing when $\alpha \le 1$. When $\alpha > 1$, the density function begins at zero, increases to a maximum, and then decreases.

Fit Lognormal

The Fit Lognormal option estimates the parameters μ (scale) and σ (shape) for the two-parameter lognormal distribution. A variable Y is lognormal if and only if $X = \ln(Y)$ is normal. The data must be greater than zero.

pdf: $$\frac{1}{\sigma\sqrt{2\pi}}\frac{\exp\left[\frac{-(\log(x) - \mu)^2}{2\sigma^2}\right]}{x} \quad \text{for} \quad 0 \le x; \quad -\infty < \mu < \infty; \quad 0 < \sigma$$

$$E(x) = \exp(\mu + \sigma^2/2)$$

$$\mathrm{Var}(x) = \exp(2(\mu + \sigma^2)) - \exp(2\mu + \sigma^2)$$

Fit Weibull

The Weibull distribution has different shapes depending on the values of α (scale) and β (shape). It often provides a good model for estimating the length of life, especially for mechanical devices and in biology.

The pdf for the Weibull distribution is defined as follows:

$$\text{pdf:}\quad \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right]\quad \text{for}\quad \alpha,\beta > 0;\quad 0 < x$$

$$E(x) = \alpha\Gamma\left(1 + \frac{1}{\beta}\right)$$

$$\mathrm{Var}(x) = \alpha^2\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right\}$$

where $\Gamma(\cdot)$ is the Gamma function.

Fit Normal 2 Mixture and Fit Normal 3 Mixture

The Fit Normal 2 Mixture and Fit Normal 3 Mixture options fit a mixture of two or three normal distributions. These flexible distributions are capable of fitting bimodal or multi-modal data. A separate mean, standard deviation, and proportion of the whole is estimated for each group. In the following equations, *k* equals the number of normal distributions in the mixture.

$$\text{pdf:}\quad \sum_{i=1}^{k}\frac{\pi_i}{\sigma_i}\phi\left(\frac{x-\mu_i}{\sigma_i}\right)$$

$$E(x) = \sum_{i=1}^{k}\pi_i\mu_i$$

$$\mathrm{Var}(x) = \sum_{i=1}^{k}\pi_i(\mu_i^2 + \sigma_i^2) - \left(\sum_{i=1}^{k}\pi_i\mu_i\right)^2$$

where μ_i, σ_i, and π_i are the respective mean, standard deviation, and proportion for the i^{th}

group, and $\phi(\cdot)$ is the standard normal pdf.

Note: Confidence intervals for normal mixture distribution parameter estimates use Wald-based calculations.

Fit Johnson

The Fit Johnson option selects and fits the best-fitting distribution from the Johnson system of distributions, which contains three distributions that are all based on a transformed normal distribution. These three distributions are the following:

- Johnson Su, which is unbounded.
- Johnson Sb, which has bounds on both tails. The bounds are defined by parameters that can be estimated.
- Johnson Sl, which is bounded in one tail. The bound is defined by a parameter that can be estimated. The Johnson Sl family contains the family of lognormal distributions.

Only the fit for the selected distribution is reported. Information about selection procedures and parameter estimation for the Johnson distributions can be found in Slifker and Shapiro (1980). The parameter estimation does not use maximum likelihood.

Johnson distributions are popular because of their flexibility. In particular, the Johnson distribution system is noted for its data-fitting capabilities because it supports every possible combination of skewness and kurtosis. However, the SHASH distribution is also very flexible and is recommended over the Johnson distributions.

If Z is a standard normal variate, then the system is defined as follows:

$$Z = \gamma + \delta f(Y)$$

where, for the Johnson Su:

$$f(Y) = \ln\left(Y + \sqrt{1 + Y^2}\right) = \sinh^{-1} Y$$

$$Y = \frac{X - \theta}{\sigma} \qquad -\infty < X < \infty$$

where, for the Johnson Sb:

$$f(Y) = \ln\left(\frac{Y}{1 - Y}\right)$$

$$Y = \frac{X - \theta}{\sigma} \qquad \theta < X < \theta + \sigma$$

and for the Johnson Sl, where $\sigma = \pm 1$.

$$f(Y) = \ln(Y)$$

$$Y = \frac{X-\theta}{\sigma} \qquad \begin{array}{ll} \theta < X < \infty & \text{if } \sigma = 1 \\ -\infty < X < \theta & \text{if } \sigma = -1 \end{array}$$

Johnson Su

pdf: $\frac{\delta}{\sigma}\left[1+\left(\frac{x-\theta}{\sigma}\right)^2\right]^{-1/2}\phi\left[\gamma+\delta\sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)\right]$ for $-\infty < x, \theta, \gamma < \infty$; $0 < \theta,\delta$

Johnson Sb

pdf: $\phi\left[\gamma+\delta\ln\left(\frac{x-\theta}{\sigma-(x-\theta)}\right)\right]\left(\frac{\delta\sigma}{(x-\theta)(\sigma-(x-\theta))}\right)$ for $\theta < x < \theta+\sigma$; $0 < \sigma$

Johnson Sl

pdf: $\frac{\delta}{|x-\theta|}\phi\left[\gamma+\delta\ln\left(\frac{x-\theta}{\sigma}\right)\right]$ for $\theta < x$ if $\sigma = 1$; $\theta > x$ if $\sigma = -1$

where $\phi(\cdot)$is the standard normal pdf.

Note: Confidence intervals for Johnson distribution parameter estimates use Wald-based calculations.

Fit Beta

The beta distribution is useful for modeling the behavior of random variables that are constrained to fall in the interval 0,1. For example, proportions always fall between 0 and 1. The Fit Beta option estimates two shape parameters, $\alpha > 0$ and $\beta > 0$. The beta distribution has values only in the interval 0,1.

pdf: $\frac{1}{B(\alpha,\beta)\sigma^{\alpha+\beta-1}}x^{\alpha-1}x^{\beta-1}$ for $0 < x < 1$; $0 < \sigma,\alpha,\beta$

$$E(x) = \sigma\frac{\alpha}{\alpha+\beta}$$

$$\mathrm{Var}(x) = \frac{\sigma^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

where $B(\cdot)$ is the Beta function.

Fit All

In the Compare Distributions report, the Distribution list is sorted by AICc in ascending order. Use the check boxes to show or hide a fit report and overlay curve for the selected distribution.

The formulas for AICc and BIC are defined as follows:

$$\text{AICc} = \text{-2logL} + 2k + \frac{2k(k+1)}{n-(k+1)}$$

$$\text{BIC} = \text{-2logL} + k\ln(n)$$

where:

- logL is the log-likelihood.
- n is the sample size.
- k is the number of parameters.

The AICc Weight column shows normalized AICc values that sum to one. The AICc weight can be interpreted as the probability that a particular distribution is the true distribution given that one of the fitted distributions is the truth. Therefore, the distribution with the AICc weight closest to one is the better fit. The AICc weights are calculated using only nonmissing AICc values:

AICcWeight = exp[-0.5(AICc-min(AICc))] / sum(exp[-0.5(AICc-min(AICc))])

where min(AICc) is the smallest AICc value among the fitted distributions.

For more information about the measures in the Compare Distributions report, see *Fitting Linear Models*.

Statistical Details for Discrete Fit Distributions

This section contains statistical details for the options in the Discrete Fit menu in the Distribution platform.

Fit Poisson

The Poisson distribution has a single scale parameter $\lambda > 0$.

$$\text{pmf:} \quad \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad 0 \le \lambda < \infty; \quad x = 0,1,2,...$$

$\mathrm{E}(x) = \lambda$

$\mathrm{Var}(x) = \lambda$

Since the Poisson distribution is a discrete distribution, the overlaid curve is a step function, with jumps that occur at every integer.

Fit Negative Binomial

The negative binomial distribution is useful for modeling the number of successes before a specified number of failures. The following parameterization contains mean parameter λ and dispersion parameter σ.

$$\text{pmf:}\quad \frac{\Gamma[x+(1/\sigma)]}{\Gamma[x+1]\Gamma[1/\sigma]}\left[\frac{(\lambda\sigma)^{x}}{(1+\lambda\sigma)^{x+(1/\sigma)}}\right],\ x = 0, 1, 2, \ldots$$

$E(x) = \lambda$

$\mathrm{Var}(x) = \lambda + \sigma\lambda^2$

where $\Gamma(\cdot)$ is the Gamma function.

Relationship between Negative Binomial and Gamma Poisson Distributions

The negative binomial distribution is equivalent to the Gamma Poisson distribution. The Gamma Poisson distribution is useful when the data are a combination of several Poisson(μ) distributions and each Poisson(μ) distribution has a different μ.

The Gamma Poisson distribution results from assuming that $x|\mu$ follows a Poisson distribution and μ follows a Gamma(α,τ). The Gamma Poisson has parameters $\lambda = \alpha\tau$ and $\sigma = \tau+1$. The parameter σ is a dispersion parameter. If $\sigma > 1$, there is over dispersion, meaning there is more variation in *x* than explained by the Poisson alone. If $\sigma = 1$, *x* reduces to Poisson(λ).

$$\text{pmf:}\quad \frac{\Gamma\left(x+\frac{\lambda}{\sigma-1}\right)}{\Gamma(x+1)\Gamma\left(\frac{\lambda}{\sigma-1}\right)}\left(\frac{\sigma-1}{\sigma}\right)^{x}\sigma^{-\frac{\lambda}{\sigma-1}}\quad \text{for}\quad 0<\lambda;\quad 1\le\sigma;\quad x=0,1,2,\ldots$$

$E(x) = \lambda$

$\mathrm{Var}(x) = \lambda\sigma$

where $\Gamma(\cdot)$ is the Gamma function.

The Gamma Poisson is equivalent to a Negative Binomial with $\sigma_{negbin} = (\sigma_{gp} - 1) / \lambda_{gp}$.

Run demoGammaPoisson.jsl in the JMP Samples/Scripts folder to compare a Gamma Poisson distribution with parameters λ and σ to a Poisson distribution with parameter λ.

Fit ZI Poisson

The zero-inflated (ZI) Poisson distribution has scale parameter λ > 0 and zero-inflation parameter π.

$$\text{pmf:}\quad \begin{cases} \pi + (1-\pi)\exp[-\lambda], \text{ for } x = 0 \\ (1-\pi)\dfrac{\lambda^x}{x!}\exp[-\lambda], \text{ for } x = 1, 2, \ldots \end{cases}$$

$E(x) = (1 - \pi)\lambda$

$\mathrm{Var}(x) = \lambda(1 - \pi)(1 + \lambda\pi)$

Fit ZI Negative Binomial

The zero-inflated (ZI) negative binomial distribution has scale parameter λ > 0, dispersion parameter σ > 0, and zero-inflation parameter π.

$$\text{pmf:}\quad \begin{cases} \pi + (1-\pi)(1+\lambda\sigma)^{-(1/\sigma)}, \text{ for } x = 0 \\ (1-\pi)\dfrac{\Gamma[x+(1/\sigma)]}{\Gamma[x+1]\Gamma[1/\sigma]}\left[\dfrac{(\lambda\sigma)^x}{(1+\lambda\sigma)^{x+(1/\sigma)}}\right], \text{ for } x = 1, 2, \ldots \end{cases}$$

$E(x) = (1 - \pi)\lambda$

$\mathrm{Var}(x) = \lambda(1 - \pi)[1 + \lambda(\sigma + \pi)]$

Fit Binomial

The Fit Binomial option accepts data in two formats: a constant sample size, or a column containing sample sizes.

$$\text{pmf:}\quad \binom{n}{x}p^x(1-p)^{n-x} \qquad \text{for} \quad 0 \le p \le 1; \quad x = 0,1,2,\ldots,n$$

$E(x) = np$

$\mathrm{Var}(x) = np(1-p)$

where n is the number of independent trials.

Note: The confidence interval for the binomial parameter is a Score interval. See Agresti and Coull (1998).

Fit Beta Binomial

The beta binomial distribution is useful when the data are a combination of several Binomial(p) distributions and each Binomial(p) distribution has a different p. One example is the overall number of defects combined from multiple manufacturing lines, when the mean number of defects (p) varies between the lines.

The beta binomial distribution results from assuming that $x|\pi$ follows a Binomial(n,π) distribution and π follows a Beta(α,β). The beta binomial has parameters p = α/(α+β) and δ = 1/(α+β+1). The parameter δ is a dispersion parameter. When δ > 0, there is over dispersion, meaning there is more variation in x than explained by the Binomial alone. When δ < 0, there is under dispersion. When δ = 0, x is distributed as Binomial(n,p). The beta binomial exists only when $n \geq 2$.

pmf: $$\binom{n}{x}\frac{\Gamma\left(\frac{1}{\delta}-1\right)\Gamma\left[x+p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n-x+(1-p)\left(\frac{1}{\delta}-1\right)\right]}{\Gamma\left[p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[(1-p)\left(\frac{1}{\delta}-1\right)\right]\Gamma\left(n+\frac{1}{\delta}-1\right)}$$

for $0 \leq p \leq 1$; $\max(-\frac{p}{n-p-1}, -\frac{1-p}{n-2+p}) \leq \delta \leq 1$; $x = 0,1,2,...,n$

$E(x) = np$

$Var(x) = np(1-p)[1+(n-1)\delta]$

where Γ(·) is the Gamma function.

Remember that $x|\pi \sim$ Binomial(n,π), while π ~ Beta(α,β). The parameters p = α/(α+β) and δ = 1/(α+β+1) are estimated by the platform. To obtain estimates of α and β, use the following formulas:

$$\hat{\alpha} = \hat{p}\left(\frac{1-\hat{\delta}}{\hat{\delta}}\right)$$

$$\hat{\beta} = (1-\hat{p})\left(\frac{1-\hat{\delta}}{\hat{\delta}}\right)$$

If the estimate of δ is 0, the formulas do not work. In that case, the beta binomial has reduced to the Binomial(n,p), and $\hat{p}$ is the estimate of p.

The confidence intervals for the beta binomial parameters are profile likelihood intervals.

Run demoBetaBinomial.jsl in the JMP Samples/Scripts folder to compare a beta binomial distribution with dispersion parameter δ to a Binomial distribution with parameters p and n = 20.

Chapter 4

Introduction to Fit Y by X

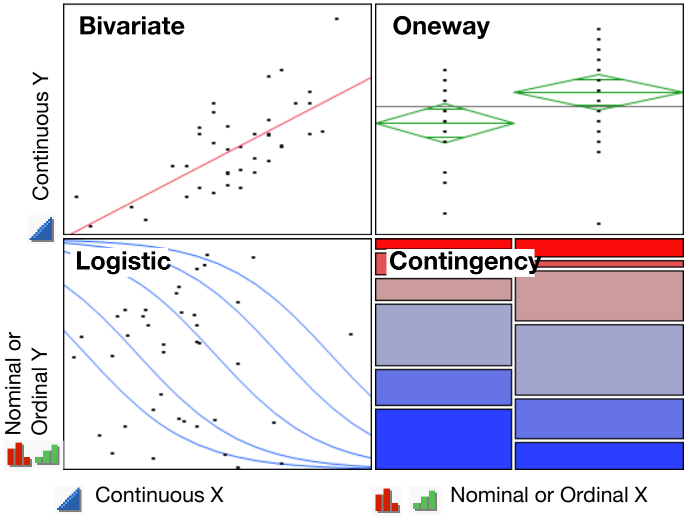
Examine Relationships between Two Variables

Use the Fit Y by X platform to analyze pairs of variables. You can do this using scatter plots, linear regression, ANOVA, multiple comparisons, logistic regression, contingency tables and much more. Specific analyses depend on the modeling types of the variables.

Fit Y by X launches one of four platforms:

- Use Bivariate to analyze a continuous Y and a continuous X
- Use One-way analysis, including ANOVA, to analyze a continuous Y and a categorical X
- Use Logistic regression to analyze a categorical Y and a continuous X
- Use Contingency to analyze a categorical Y and a categorical X

Figure 4.1 Examples of Four Types of Analyses



The Fit Y by X platform is a collection of four platforms (or types of analyses): Bivariate, Oneway, Logistic, and Contingency.

Specific Platform	Modeling Types	Description
Bivariate	Continuous Y by continuous X	Analyzes the relationship between two continuous variables. See “Bivariate Analysis”.
Oneway	Continuous Y by nominal or ordinal X	Analyzes how the distribution of a continuous Y variable differs across groups defined by a categorical X variable. See “Oneway Analysis”.
Contingency	Nominal or ordinal Y by nominal or ordinal X	Analyzes the distribution of a categorical response variable Y as conditioned by the values of a categorical X factor. See “Contingency Analysis”.
Logistic	Nominal or ordinal Y by continuous X	Fits the probabilities for response categories to a continuous X predictor. See “Logistic Analysis”.

Chapter 5

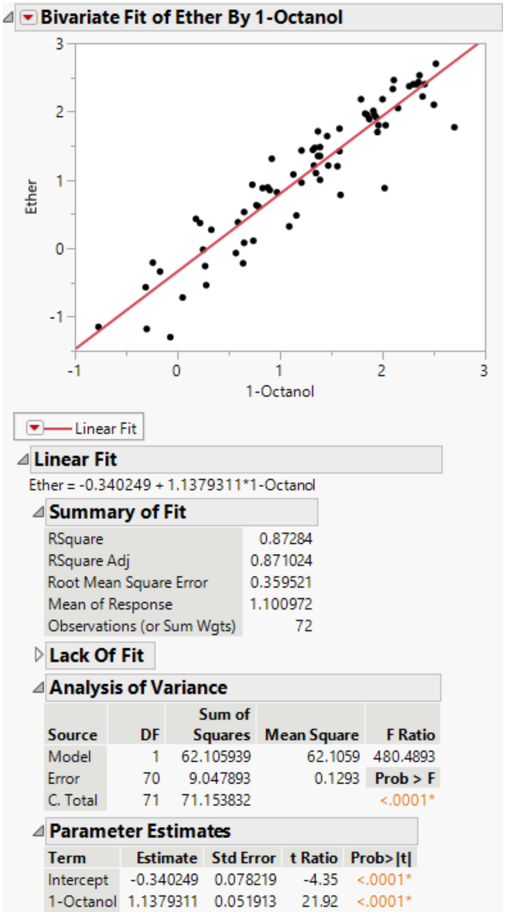
Bivariate Analysis

Examine Relationships between Two Continuous Variables

Use the Bivariate platform to investigate the relationship between two continuous variables. The scatterplot provides a graphical view of the data. Options enable you to fit simple linear regression lines, polynomial regression curves, or smoothers to the data.

The Bivariate platform is the *continuous by continuous* personality of the Fit Y by X platform. The word bivariate simply means involving two variables instead of one (univariate) or many (multivariate).

Figure 5.1 Example of Bivariate Analysis



Linear Fit

Ether = -0.340249 + 1.1379311*1-Octanol

Summary of Fit

RSquare	0.87284
RSquare Adj	0.871024
Root Mean Square Error	0.359521
Mean of Response	1.100972
Observations (or Sum Wgts)	72

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	62.105939	62.1059	480.4893
Error	70	9.047893	0.1293	**Prob > F**
C. Total	71	71.153832		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.340249	0.078219	-4.35	<.0001*
1-Octanol	1.1379311	0.051913	21.92	<.0001*

Contents

Overview of the Bivariate Platform

The Bivariate platform enables you to interactively fit models and view those fits on a scatterplot. You can compare multiple model fits on the same plot.

The Bivariate platform is launched from Fit Y by X when you have one or more continuous Y variables and one or more continuous X variables. The Bivariate Analysis platform initially shows scatterplots for each combination of X and Y variables. Use the scatterplots and red triangle options to interactively explore models for two continuous variables. You can fit a simple linear regression line to the data or you can fit more complex regression models. You can also explore density estimates.

Fitting option categories in the Bivariate platform include regression fits and density estimation.

Table 5.1 Bivariate Platform Fit Categories

Category	Description	Fitting Options
Regression Fits	Regression methods fit a model to the observed data points. The fitting methods include least squares fits as well as spline fits, kernel smoothing, orthogonal fits, transformations, and robust fits.	**Fit Mean** **Fit Line** **Fit Polynomial** **Fit Special** **Flexible** **Fit Orthogonal** **Fit Passing** **Bablok** **Robust**
Density Estimation	Density estimation fits a bivariate distribution to the points. You can either select a bivariate normal density, characterized by elliptical contours, or a general nonparametric density.	**Density Ellipse** **Nonpar Density**

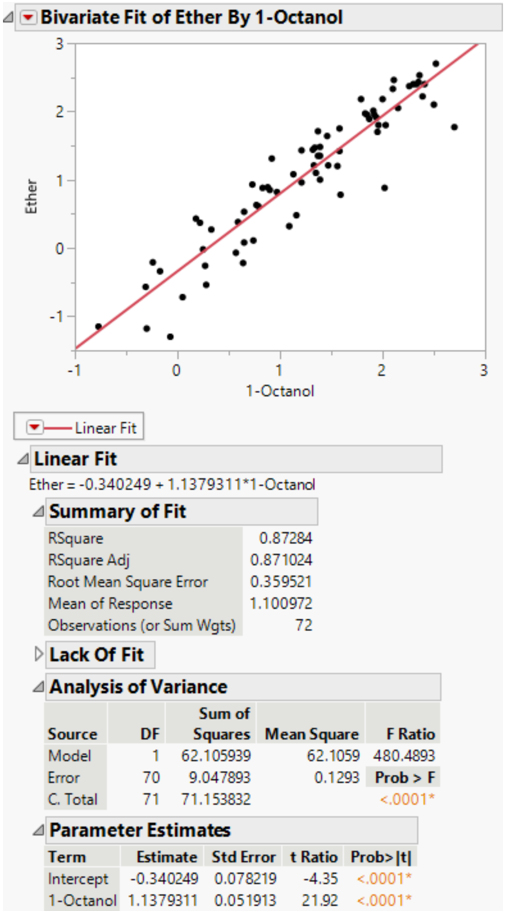
Example of Bivariate Analysis

Use the Bivariate platform to construct a scatterplot and fit a regression line for two continuous variables. In this example, solubility levels for chemical compounds were measured in different solvents. You want to compare the results of two solvents.

1. Select **Help > Sample Data Folder** and open Solubility.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Ether and click **Y, Response**.

4. Select 1-Octanol and click **X, Factor**.
5. Click **OK**.
6. Click the Bivariate Fit red triangle and select Fit Line.

Figure 5.2 Bivariate Fit Report



Linear Fit

Ether = -0.340249 + 1.1379311*1-Octanol

Summary of Fit

RSquare	0.87284
RSquare Adj	0.871024
Root Mean Square Error	0.359521
Mean of Response	1.100972
Observations (or Sum Wgts)	72

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	62.105939	62.1059	480.4893
Error	70	9.047893	0.1293	Prob > F
C. Total	71	71.153832		<.0001*

Parameter Estimates

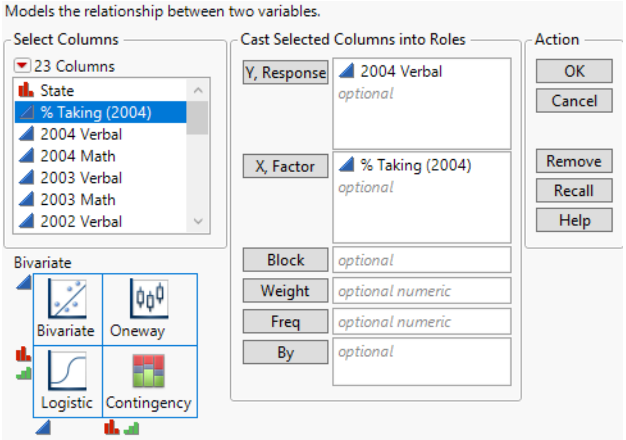
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.340249	0.078219	-4.35	<.0001*
1-Octanol	1.1379311	0.051913	21.92	<.0001*

You can see that the two solvents have similar results. The linear model has an RSquare value of 0.87, indicating that the model explains 87% of the variability in the response. Since both the X and Y variables are measurements, one might consider the Fit Orthogonal or Fit Passing Bablok options.

Launch the Bivariate Platform

Launch the Bivariate platform by selecting **Analyze > Fit Y by X**. The Fit Y by X launch window is used for four different types of analyses. When you enter a continuous Y variable and a continuous X variable, the Bivariate platform is launched.

Figure 5.3 The Bivariate Launch Window



Models the relationship between two variables.

Y, Response: 2004 Verbal

X, Factor: % Taking (2004)

For more information about the options in the Select Columns red triangle menu, see *Using JMP*. The Bivariate launch window contains the following options:

Y, Response The response variable or variables that you want to analyze. The response variable is often called the dependent variable. These variables must have a continuous modeling type.

X, Factor The predictor variable or variables that you want to analyze. The factor variables are often called independent variables. These variables must have a continuous modeling type.

Block (Not applicable for a bivariate analysis.) A column that specifies a Blocking variable.

Weight A column containing a weight for each observation in the data table. A row is included in the analyses only when its value is greater than zero.

Freq Assigns a frequency to each row in the analysis. Assigning a frequency is useful when your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Data Format

In the Bivariate platform, your data can consist of columns of unsummarized or summarized data:

Unsummarized data There is one row for each observation and columns for X and Y values.

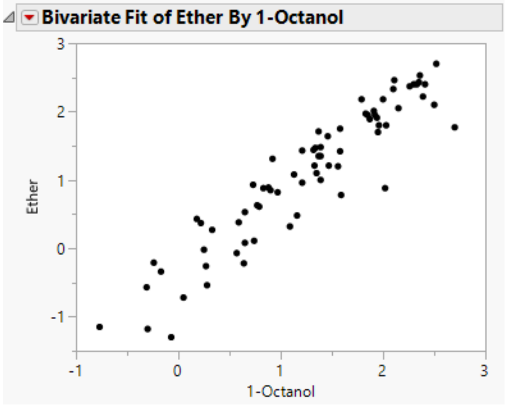
Summarized data Each row represents a set of observations with common X and Y values. The data table must contain a frequency column that contains the counts of observations for each row. Enter this column as Freq in the launch window.

Note: The Fit Y by X launch window accommodates columns with continuous, ordinal, and nominal modeling types. The Bivariate platform is launched for all pairs of Y, Response and X, Factor columns that have the continuous modeling type. The Fit Y by X launch window launches the Oneway, Contingency, or Logistic platforms for other column type combinations.

The Bivariate Report

The Bivariate report initially contains a scatterplot for each pair of X and Y variables. You can fit models to the data and view statistical reports using the red triangle options. See “Bivariate Platform Options”.

Figure 5.4 The Bivariate Plot



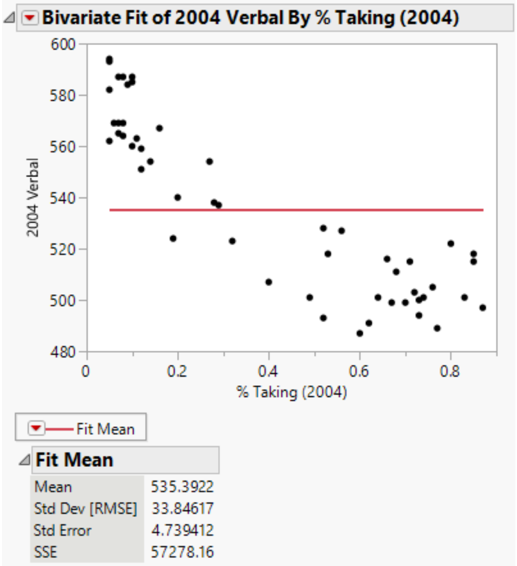
Replace Variables Interactively

You can interactively replace variables in the plot by dragging and dropping a variable from one axis to the other axis. You can also replace a variable by selecting a variable in the Columns panel of the associated data table and dragging it onto an axis.

Bivariate Platform Options

The Bivariate Fit red triangle menu contains display, fitting, and control options. Each fitting option adds a line, curve, or distribution to the scatterplot, a red triangle menu for the fit below the plot, and a fit-specific report to the report window.

Figure 5.5 Example of the Fit Mean Option



Fit Mean

Mean	535.3922
Std Dev [RMSE]	33.84617
Std Error	4.739412
SSE	57278.16

Note: The Fit Group menu appears only if you have specified multiple Y or multiple X variables. Use the Fit Group menu options to arrange reports or order them by RSquare. See *Fitting Linear Models*.

The Bivariate Fit red triangle menu contains the following options:

Show Points Shows or hides the points in the scatterplot.

Histogram Borders Shows or hides histograms on the horizontal and vertical axes of the scatterplot.

Note: Data points for hidden rows are hidden in the scatterplot but not in the histograms. To exclude rows from the histograms and from analysis results, apply the Hide and Exclude row state and select **Redo > Redo Analysis** from the Bivariate red triangle menu.

Summary Statistics Shows or hides the summary statistics for the variables that are plotted. Metrics include the correlation and covariance between the two variables as well as the univariate mean and standard deviation for each variable.

Fit Mean Fits the mean of the Y response variable. This is the simple regression model with slope constrained to zero. See “Fit Mean Report”.

Fit Line Fits a least squares regression line. The fit is shown on the plot and a fit report is provided. See “Fit Line, Fit Polynomial, and Fit Special Reports”.

Fit Polynomial Fits polynomial curves of a selected degree using least squares regression. See “Fit Line, Fit Polynomial, and Fit Special Reports”.

Fit Special Enables you to fit regression models that contain transformations for the Y and X variables. Transformations include: log, square root, square, reciprocal, and exponential. You can also turn off **center polynomials**, constrain the intercept and the slope, and fit polynomial models using the transformed variables. See “Fit Special Window”.

Flexible Enables you to fit flexible models. Models include splines, kernel smoothers, and pointwise fits.

Fit Spline Fits a penalized least squares model to the data. Use the smoothing parameter λ to adjust the degree of smoothness of the model fit. See “Fit Spline Report”.

Kernel Smoother Fits a locally weighted least squares model to the data. This model is also known as a LOWESS (locally weighted scatterplot smoothing) model. Use α, the smoothing parameter, to control the smoothness of the model. See “Kernel Smoother Report”.

Fit Each Value Connects the mean response for each X value. See “Fit Each Value Report”.

Fit Orthogonal Enables you to fit orthogonal regression models, which can be used when both X and Y variables are measured with error. See “Fit Orthogonal Report”.

Univariate Variances, Prin Comp Fits the standardized first principal component to the data.

Equal Variances Fits an orthogonal regression model with a variance ratio of one, which assumes that the error variances for X and Y are equal. This method is also known as *Deming regression*.

Fit X to Y Fits an orthogonal regression model with a variance ratio of zero, which indicates that Y has no variance.

Specified Variance Ratio Enables you to enter a specified variance ratio for the orthogonal regression model.

Fit Passing Bablok Fits a regression model using the Passing-Bablok procedure. This option includes the option for a Bland-Altman analysis. Use when both X and Y variables are measured with error. See “Passing-Bablok Fit Report” and “Statistical Details for the Fit Passing Bablok Option”.

Robust Enables you to fit robust regression models. Use robust models to reduce the influence of outliers on the model fit. See “Robust Fit Report”.

Fit Robust Fits a robust regression model using the Huber M-estimation method.

Fit Cauchy Fits a robust regression model where the parameters are estimated by maximum likelihood with a Cauchy link function.

Density Ellipse Enables you to add bivariate normal density ellipsoids of a specified percent to the plot. The contours contain the specified percent of the data points. You can use this option to estimate correlation. See “Density Ellipse Report”.

Nonpar Density Enables you to add nonparametric density contours to the plot. The contours describe the density of data points. See “Nonpar Density Report”.

Group By Enables you to specify a grouping variable. After specifying a grouping variable, subsequent fits are computed for each level of the grouping variable. Lines, curves, or ellipses are overlaid on the scatterplot by group. This enables you to visually compare the fits across groups. See “Example of Group By Using Density Ellipses” and “Example of Group By Using Regression Lines”.

Note: The Group By option allows for the group fits to be visualized on the one plot with the results in one report. Alternatively, the By option in the launch window results in grouped fits, each in its own report.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Bivariate Fit Options

The Bivariate platform model-specific red triangle options depend on the specified fit. These menus are found next to the model fit label under the scatterplot.

- "Options That Apply to Most Bivariate Fits"
- "Options That Apply to Normal Ellipses"
- "Options That Apply to Quantile Density Contours"

Options That Apply to Most Bivariate Fits

The Bivariate platform model-specific red triangle options depend on the specified fit. Not all options are available for all fits.

Line of Fit Shows or hides the line, curve, or contours that describe the model fit. Not applicable for Quantile Density Contours.

Confid Curves Fit Shows or hides the confidence limits (curves) for the expected value (mean).

Confid Curves Indiv Shows or hides the confidence limits for an individual predicted value. The confidence limits reflect variation in the error and variation in the parameter estimates.

Line Color Enables you to select a color for the fit curve.

Line Style Enables you select a line style for each fit.

Line Width Enables you to select a line width for each fit.

Report Shows or hides the report for each fit.

Note: This option does not modify the Bivariate plot.

Profiler Shows or hides a prediction trace for the X variable. Enables you to find optimum settings for one or more responses and to explore response distributions using simulation. See *Profilers*.

Save Predicteds Creates a new column in the current data table called Predicted <colname> where colname is the name of the Y variable. This column contains the prediction formula and the predicted values.

Tip: The prediction formula computes values automatically for new rows that you add to the table.

Save Residuals Creates a new column in the current data table called Residuals <colname> where colname is the name of the Y variable. This column contains the observed response values minus their predicted values.

Note: You can use the **Save Predicteds** and **Save Residuals** options for each fit. If you use these options multiple times or with a grouping variable, it is best to rename the resulting columns in the data table to reflect each fit.

Save Studentized Residuals Creates a new column in the current data table called Studentized Residuals <colname> where colname is the name of the Y variable. This column contains the residuals divided by their standard errors.

Mean Confidence Limit Formula Creates two new columns in the data table called Lower 95% Mean <colname> and Upper 95% Mean <colname> where colname is the name of the Y variable. These columns contain both the formulas and the values for lower and upper 95% confidence limits for the mean response.

Indiv Confidence Limit Formula Creates two new columns in the data table called Lower 95% Indiv <colname> and Upper 95% Indiv <colname> where colname is the name of the Y variable. These columns contain both the formulas and the values for lower and upper 95% confidence limits for an individual prediction.

Plot Residuals (Available for Linear, Polynomial, Passing Bablok, and Fit Special Only.) Shows or hides five diagnostic plots: residual by predicted, actual by predicted, residual by row, residual by X, and a normal quantile plot of the residuals.

Bland Altman Analysis (Available only for Passing Bablok.) Shows a matched pairs and Bland-Altman analysis in a new report window.

Set α Level Enables you to specify the alpha level used in computing confidence curves.

Confid Shaded Fit (Not available for all fits.) Shows or hides a shaded confidence region for the expected response (mean).

Confid Shaded Indiv (Not available for all fits.) Shows or hides a shaded confidence region for an individual prediction.

Save Coefficients (Available only for flexible spline fits.) Saves the spline coefficients as a new data table that contains columns named X, A, B, C, and D. The X column contains the knot points. A, B, C, and D are the intercept, linear, quadratic, and cubic coefficients of the third-degree polynomial. These coefficients span from the corresponding value in the X column to the next highest value.

Remove Fit Removes the fit from the graph and removes its report.

Options That Apply to Normal Ellipses

The Normal Ellipse red triangle options are available for Density Ellipse fits in the Bivariate platform.

Shaded Contour Shades or unshades the area inside the density ellipse.

Select Points Inside Selects the points inside the ellipse.

Select Points Outside Selects the points outside the ellipse.

Options That Apply to Quantile Density Contours

The Quantile Density Contours red triangle options are available for Nonpar Ellipse fits in the Bivariate platform.

Kernel Control Shows or hides a slider to control the standard deviation for each variable. The standard deviation defines the range of X and Y values for determining the density of contour lines.

5% Contours Shows or hides the 5% contours.

Contour Lines Shows or hides the contours.

Contour Fill Shows or hides filled contours.

Color Theme Enables you to change the color theme of the contours.

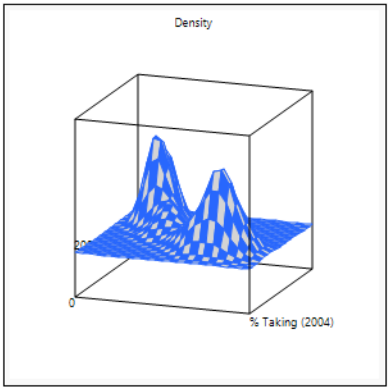
Select Points by Density Enables you to select points that fall in a user-specified quantile range.

Color by Density Quantile Colors the points according to density.

Save Density Quantile Creates a new column in the current data table that contains the density quantile for each point.

Mesh Plot Shows or hides a three-dimensional plot of the density over a grid of the two analysis variables.

Figure 5.6 Example of a Mesh Plot



Modal Clustering Shows or hides the results for a modal clustering of the data. Creates a new column in the current data table that contains the cluster number for each data pair. Modal clustering is based on the density estimates. JMP generates a grid of 10,404 density estimates. The number of modes in the distribution of the densities determines the number of clusters. The modes are the *cluster centers*. The remaining points are iteratively clustered to each mode based on distance and density estimates.

Note: For more information about additional clustering methods in JMP, see *Multivariate Methods*.

Save Density Grid Creates a new data table that contains the density estimates and the quantiles associated with them.

Note: If you save the modal clustering values first and then save the density grid, the grid table also contains the cluster values.

Bivariate Fit Reports

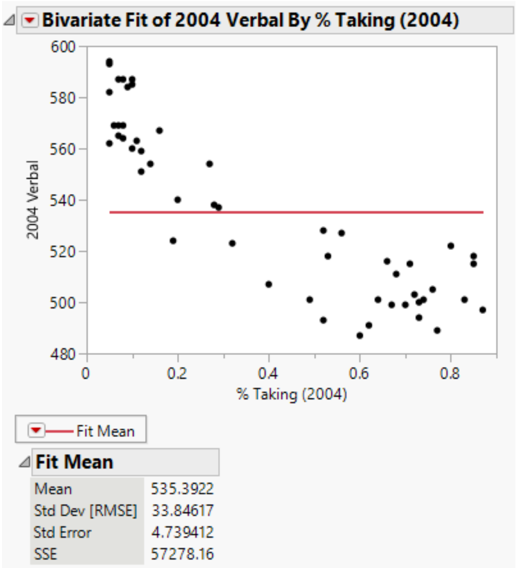
The Bivariate platform enables you to fit a wide variety of models to predict a Y variable based on a single X variable. This section contains information about the reports that are generated for specific model fits.

- "Fit Mean Report"
- "Fit Line, Fit Polynomial, and Fit Special Reports"
- "Fit Special Window"
- "Fit Spline Report"
- "Kernel Smoother Report"
- "Fit Each Value Report"
- "Fit Orthogonal Report"
- "Robust Fit Report"
- "Passing-Bablok Fit Report"
- "Density Ellipse Report"
- "Nonpar Density Report"

Fit Mean Report

Use the Fit Mean option in the Bivariate platform to fit the mean of the Y response variable. You can use the mean line as a reference to compare with other fits.

Figure 5.7 Example of Fit Mean



Fit Mean

Mean	535.3922
Std Dev [RMSE]	33.84617
Std Error	4.739412
SSE	57278.16

In the Bivariate platform, the Fit Mean report contains a table of summary statistics.

Mean Mean of the response variable. The predicted response when there are no specified effects in the model.

Std Dev [RMSE] Standard deviation of the response variable. Square root of the mean square error, also called the root mean square error (RMSE).

Std Error Standard deviation of the response mean. Calculated by dividing the RMSE by the square root of the number of values.

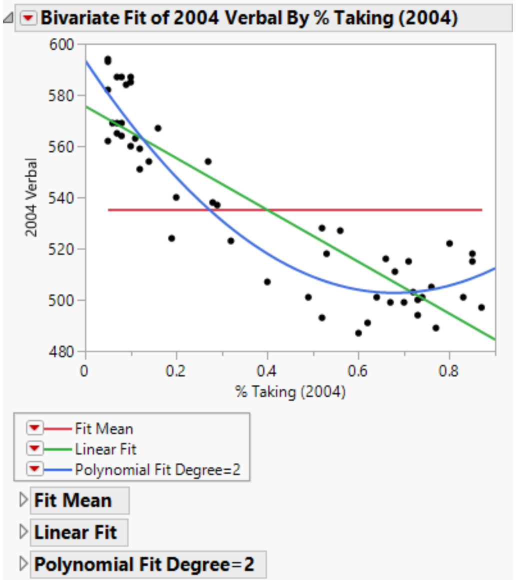
SSE Error sum of squares for the simple mean model. Appears as the sum of squares for Error in the analysis of variance tables for each model fit.

For more information about the options in the Fit Mean menu, see “Bivariate Fit Options”.

Fit Line, Fit Polynomial, and Fit Special Reports

In the Bivariate platform, use the Fit Line, Fit Polynomial, or Fit Special options to fit regression models. You can fit multiple models and then compare the fits on the scatterplot.

Figure 5.8 Example of Fit Line and Fit Polynomial



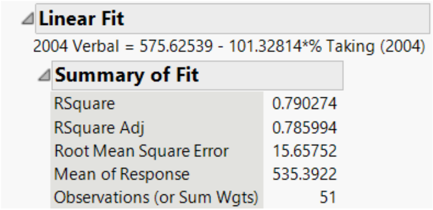
For more information about the options in the Linear Fit and Polynomial Fit Degree menus, see “Bivariate Fit Options”. For statistical details, see “Statistical Details for the Fit Line Option”.

In the Bivariate platform, there is a report for each fit that you select. The Linear, Polynomial, and Transformed Fit reports each contain a text box with the equation of the fit. Each fit report contains tables for a summary of fit, an analysis of variance (ANOVA), and parameter estimates. A fourth table, for Lack of Fit, appears if there are replicates in your data. Fits for a transformed Y variable include a summary of fit measures on the original scale table.

Summary of Fit

In the Bivariate platform fit reports, the Summary of Fit table contains numerical summaries of the model fit. The equation for the fit is shown above the Summary of Fit table.

Figure 5.9 Summary of Fit Table



Linear Fit

2004 Verbal = 575.62539 - 101.32814*% Taking (2004)

Summary of Fit

RSquare	0.790274
RSquare Adj	0.785994
Root Mean Square Error	15.65752
Mean of Response	535.3922
Observations (or Sum Wgts)	51

The Summary of Fit table contains the following statistics:

RSquare The proportion of the variation explained by the model. The remaining variation is attributed to random error. The RSquare is 1 if the model fits perfectly. See "Statistical Details for the Summary of Fit Report"

Note: A low RSquare value suggests that there might be variables not in the model that account for the unexplained variation. However, if your data are subject to a large amount of inherent variation, even a useful regression model can have a low RSquare value. Read the literature in your research area to learn about typical RSquare values.

RSquare Adj The RSquare statistic adjusted for the number of parameters in the model. RSquare Adj facilitates comparisons among models that contain different numbers of parameters. See "Statistical Details for the Summary of Fit Report".

Root Mean Square Error The estimate of the standard deviation of the random error. This quantity is the square root of the mean square for Error in the Analysis of Variance report (Figure 5.11).

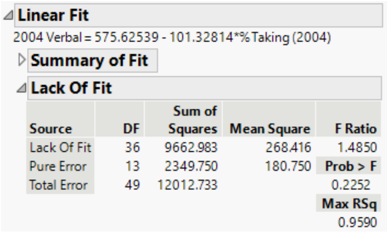
Mean of Response The sample mean (arithmetic average) of the response variable. This is the predicted response when no model effects are specified.

Observations (or Sum Wgts) The number of observations used to estimate the fit. If there is a weight variable, this is the sum of the weights.

Lack of Fit

In the Bivariate fit reports, the Lack of Fit table contains the results of a lack of fit test. The lack of fit test is available only when there are replicated X values and the model is not saturated. A sum of squares calculated from the replicates is called *pure error*. This is the portion of the overall error that cannot be explained or predicted no matter what form of model is used.

Figure 5.10 Lack of Fit Table for a Linear Fit



Linear Fit

2004 Verbal = 575.62539 - 101.32814*%Taking (2004)

Summary of Fit

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	36	9662.983	268.416	1.4850
Pure Error	13	2349.750	180.750	**Prob > F**
Total Error	49	12012.733		0.2252
				Max RSq
				0.9590

The difference between the residual error from the model and the pure error is called the *lack of fit error*. The lack of fit error can be significantly greater than the pure error if you have a misspecified model. A misspecified model is one that does a poor job of describing the data. The null hypothesis in the lack of fit test is that the lack of fit error is zero. Therefore, a small *p*-value indicates a significant lack of fit.

The Lack of Fit table contains the following columns:

Source The three sources of variation: Lack of Fit, Pure Error, and Total Error.

DF The *degrees of freedom* (DF) for each source of error.

- The Total Error DF is the degrees of freedom found on the Error line of the corresponding Analysis of Variance (ANOVA) table. See "Analysis of Variance". The Total Error DF value is the difference between the Total DF and the Model DF values found in the ANOVA table. The Error DF is partitioned into degrees of freedom for lack of fit and for pure error.
- The Pure Error DF is pooled from each replicated group of observations. See "Statistical Details for the Lack of Fit Report".
- The Lack of Fit DF is the difference between the Total Error and Pure Error DF.

Sum of Squares The sum of squares (SS) for each source of error.

- The Total Error SS is the sum of squares found on the Error line of the corresponding Analysis of Variance table. See "Analysis of Variance".
- The Pure Error SS is pooled from each replicated group of observations. The Pure Error SS divided by its DF estimates the variance of the response at a given predictor setting. This estimate is unaffected by the model. See "Statistical Details for the Lack of Fit Report".
- The Lack of Fit SS is the difference between the Total Error and Pure Error sum of squares. If the lack of fit SS is large, the model might not be appropriate for the data.

Mean Square The mean square for the Source, which is the Sum of Squares divided by the DF. A Lack of Fit mean square that is large compared to the Pure Error mean square suggests that the model is not fitting well. The F ratio can be used to conduct a formal hypothesis test.

F Ratio The ratio of the Lack of Fit mean square to the Pure Error mean square. The larger the F Ratio value, the less likely that the lack of fit error is zero.

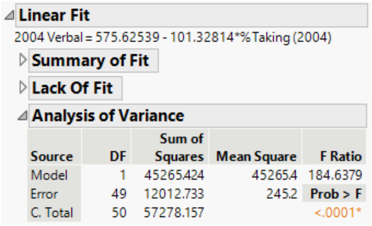
Prob > F The *p*-value for the lack of fit test. The null hypothesis is that the lack of fit error is zero. A small *p*-value indicates a significant lack of fit.

Max RSq The maximum R^2 value that can be achieved by a model using only the variables in the model. See "Statistical Details for the Lack of Fit Report".

Analysis of Variance

In the Bivariate fit reports, the Analysis of Variance table contains the calculations for comparing the fitted model to a model where all predicted values equal the response mean. The values in the analysis of variance (ANOVA) table are used to compute an *F*-ratio to evaluate the effectiveness of the model. If the *p*-value associated with the *F*-ratio is small, then the model is considered a better fit for the data than the response mean alone.

Figure 5.11 Analysis of Variance Table for a Linear Fit



2004 Verbal = 575.62539 - 101.32814*% Taking (2004)

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	45265.424	45265.4	184.6379
Error	49	12012.733	245.2	**Prob > F**
C. Total	50	57278.157		<.0001*

The Analysis of Variance table contains the following columns:

Source The three sources of variation: Model, Error, and C. Total (Corrected Total).

DF The associated degrees of freedom (DF) for each source of variation. The C. Total DF is always one less than the number of observations, and it is partitioned into degrees of freedom for the Model and Error as follows:

- The Model DF is the number of parameters (other than the intercept) used to fit the model.
- The Error DF is the difference between the C. Total DF and the Model DF.

Sum of Squares The associated Sum of Squares (SS) for each source of variation:

- The total (C. Total) SS is the sum of the squared differences between the response values and the sample mean. It represents the total variation in the response values.
- The Error SS is the sum of the squared differences between the fitted values and the actual values. It represents the variability that remains unexplained by the fitted model.

– The Model SS is the difference between C. Total SS and Error SS. It represents the variability explained by the model.

Mean Square The mean square statistics for the Model and Error sources of variation. Each Mean Square value is the sum of squares divided by its corresponding DF.

Note: The square root of the Mean Square for Error is the same as RMSE in the Summary of Fit table.

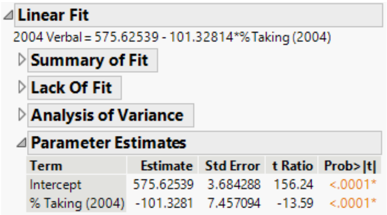
F Ratio The model mean square divided by the error mean square. The F Ratio is the test statistic for a test of whether the model differs significantly from a model where all predicted values are the response mean. The underlying hypothesis of the fit is that all the regression parameters (except the intercept) are zero. If this hypothesis is true, then both the mean square for error and the mean square for model estimate the error variance, and their ratio has an *F*-distribution.

Prob > F The observed significance probability (*p*-value) for the test. Small *p*-values are considered evidence of a regression effect.

Parameter Estimates

In the Bivariate fit reports, the Parameter Estimates table contains model parameter estimates.

Figure 5.12 Parameter Estimates Table for a Linear Fit



2004 Verbal = 575.62539 - 101.32814*%Taking (2004)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	575.62539	3.684288	156.24	<.0001*
% Taking (2004)	-101.3281	7.457094	-13.59	<.0001*

The Parameter Estimates table contains the following columns:

Term The model term that corresponds to the estimated parameter. The first term is the intercept.

Estimate The parameter estimates for each term. These are the estimates of the model coefficients.

Std Error The estimates of the standard errors of the parameter estimates.

t Ratio The test statistics for the hypothesis that each parameter is zero. It is the ratio of the parameter estimate to its standard error. Given the usual assumptions about the model, the *t* Ratio has a Student's *t*-distribution.

Prob>|t| The *p*-value for the test that the true parameter value is zero, against the two-sided alternative that it is not.

To show additional statistics, right-click in the report and select the **Columns** menu. The following statistics are not shown by default:

Lower 95% The lower 95% confidence limit for the parameter estimate.

Upper 95% The upper 95% confidence limit for the parameter estimate.

Std Beta The parameter estimates for a regression model where all of the terms have been standardized to a mean of 0 and a variance of 1. See "Statistical Details for the Parameter Estimates Report".

VIF The variance inflation factor (VIF) for each term in the model. High VIF values indicate a collinearity issue among the terms in the model.

Design Std Error The square roots of the relative variances of the parameter estimates. See "Statistical Details for the Parameter Estimates Report".

Fit Measured on Original Scale

In the Bivariate fit reports, the Fit Measured on Original Scale table contains numerical summaries of the model fit measured on the untransformed scale. This table is available only when the Y variable is transformed.

Fit Special Window

In the Bivariate platform, use the Fit Special option to fit regression models with transformed variables. You can also constrain the slope and intercept, fit a polynomial model of specific degree, or center the polynomial.

The Fit Special option opens the Specify Transformation or Constraint window that contains the following options:

Y or X Transformation Enables you to specify a transformation for the Y or X variable. The following transformations are available: natural logarithm, square root, square, reciprocal, and exponential.

Degree Enables you to fit a polynomial of a specified degree.

Centered Polynomial Enables you to center the polynomial. Deselect or select to disable or enable polynomial centering. Centering polynomials stabilizes the regression coefficients and reduces multicollinearity.

Note: Polynomial centering is not supported for transformations of the X variable.

Constrain Intercept to Enables you to constrain the model intercept to a specified value.

Constrain Slope to Enables you to constrain the model slope to a specified value.

Tip: To add the Y = X line to your plot select Fit Special and then select **Constrain Intercept to: 0** and **Constrain Slope to: 1**.

For more information about the options in the Transformed Fit menu, see "Bivariate Fit Options". For an example, see "Example of the Fit Special Option".

Fit Spline Report

In the Bivariate platform, use the **Flexible > Fit Spline** option to fit a smoothing spline that varies in smoothness (or flexibility) according to a specified lambda (λ) value. The lambda value is a tuning parameter in the spline formula. Small values of λ place high weight on the model error term for a flexible fit. Large values of λ place low weight on the error term for a stiff fit, approaching a straight line.

Select λ from the Fit Spline menu or select **Fit Spline > Other** to specify a value for λ. Alternatively, adjust λ interactively in the Smoothing Spline Fit report.

Note: To standardize your X variable, select **Fit Spline > Other** and select the **Standardize X** option.

For more information about the options in the Smoothing Spline Fit menu, see "Bivariate Fit Options". For statistical details about this fit, see "Statistical Details for the Fit Spline Option".

The Smoothing Spline Fit table contains the following columns:

R-Square The proportion of variation accounted for by the smoothing spline model. See "Statistical Details for the Smoothing Fit Report".

Sum of Squares Error Sum of squared distances from each point to the fitted spline. It is the unexplained error (*residual*) after fitting the spline model.

Change Lambda Enables you to change the λ value, either by entering a value, or by moving the slider.

Kernel Smoother Report

In the Bivariate platform, use the **Flexible > Kernel Smoother** option to fit a locally weighted least squares model to the data. The smoother is formed by repeatedly finding a locally weighted model (mean, linear or quadratic) at sampled points in the domain. The many local fits (512 in total) are combined to produce the smooth curve over the entire domain. This method is also called *LOWESS* (locally weighted scatterplot smoothing). The Kernel Smoother option implements the approach of Cleveland (1979) with a small adjustment for cases of near-perfect fits; the 6 * q50 argument in Cleveland's biweight function is replaced by max(6 * q50, 2 * q90), where q50 and q90 are the 50th and 90th percentiles, respectively. For more information about the options in the Local Smoother menu, see "Bivariate Fit Options".

The Local Smoother report contains the following columns:

R-Square Measures the proportion of variation accounted for by the smoother model. See "Statistical Details for the Smoothing Fit Report".

Sum of Squares Error Sum of squared distances from each point to the fitted smoother. It is the unexplained error (*residual*) after fitting the smoother model.

Local Fit (lambda) Enables you to specify the polynomial degree, or lambda, for each local fit.

Weight Function Enables you to specify the weight function. The LOWESS model uses tri-cube weighting. The weight function determines the influence that each x_i and y_i has on the fitting of the model.

Smoothness (alpha) Enables you to specify how many points are considered for each local fit either by entering a value, or by moving the slider. Alpha is a smoothing parameter between 0 and 1. As alpha increases, the curve becomes smoother.

Sampling Delta Enables you to specify the sampling rate that is used in the fitting process. By default, the sampling delta is zero, which means that none of the points are skipped. As the sampling delta increases, points within delta of the last sample point are skipped in the fitting process. You can use this option to reduce the number of points used when the data are dense.

Robustness Enables you to specify the robustness of the fitting routine. Each iteration re-weights the points to de-emphasize points that are farther from the fitted curve.

Fit Each Value Report

In the Bivariate platform, use the Fit Each Value option to fit a line that connects the mean response for each unique X value.

For more information about the options in the Fit Each Value menu, see "Bivariate Fit Options".

The Fit Each Value table contains summary statistics about the model fit.

Number of Observations The total number of observations.

Number of Unique Values The number of unique X values.

Degrees of Freedom The pure error degrees of freedom.

Sum of Squares The pure error sum of squares.

Mean Square The pure error mean square.

Fit Orthogonal Report

In the Bivariate platform, use the Fit Orthogonal option to fit orthogonal regression models, including Deming regression models. An orthogonal model accounts for variability in X as well as Y.

The Orthogonal Fit Ratio table contains summary statistics for the orthogonal regression model.

Variable The names of the variables.

Mean The mean of each variable.

Std Dev The standard deviation of each variable.

Variance Ratio The variance ratio used to fit the line.

Correlation The correlation between the two variables.

Intercept The intercept of the fitted line.

Slope The slope of the fitted line.

LowerCL The lower confidence limit for the slope.

UpperCL The upper confidence limit for the slope.

Alpha The alpha level used in computing the confidence interval. Enter a value in the text box to change the alpha level.

For more information about the options in the Orthogonal Fit Ratio menu, see "Bivariate Fit Options". For statistical details about this fit, see "Statistical Details for the Fit Orthogonal Option". For an example of this option, see "Example of the Fit Orthogonal Option".

Robust Fit Report

In the Bivariate platform, use the Robust options to fit robust regression models. Robust methods are used to reduce the influence of outliers on the model.

The Robust and Cauchy fit reports include two tables with the following summary statistics.

Sigma The sigma value, which is equivalent to the root mean square error (RMSE).

ChiSquare The test statistic for the hypothesis that the model fits better than the mean of the response.

PValue The p-value for the chi-square test that the slope is equal to zero.

Logworth The logworth is a transformation of the p-value defined as $-\log_{10}(p\text{-value})$.

Parameter The names of the model parameters.

Robust Estimate The parameter estimates for each term. These are the estimates of the model coefficients.

Std Error The estimates of the standard errors of the parameter estimates.

For more information about the options in the Robust Fit and Cauchy Fit menus, see "Bivariate Fit Options". For an example of this option, see "Example of the Fit Robust Option".

Passing-Bablok Fit Report

In the Bivariate platform, use the Fit Passing Bablok option to fit a regression model using the Passing-Bablok method. Passing-Bablok regression was developed for comparing measurements from two different analytical methods. The method assumes a linear relationship between the two variables with strong correlation. A line of fit and a dotted line for Y = X are added to the scatterplot. Use the Matched Pairs option in the Passing-Bablok Fit red triangle menu to perform a paired *t*-test and Bland-Altman analysis.

The Passing-Bablok Fit report contains three tables.

Nonparametric: Kendalls' τ

Use Kendall's τ to assess the correlation between the X and Y variables.

X The X variable.

Y The Y variable.

Kendall’s t A nonparametric measure of correlation between the X and Y variables. Values are between -1 and 1 with a value near 0 indicating independence between the X and Y variables.

Prob>|τ| The p-value associated with the hypothesis test of independence between the X and Y variables. A small *p*-value supports that the variables are dependent and that the Passing-Bablok method is appropriate.

CUSUM Test of Linearity

The CUSUM Test of Linearity table contains the results of a test for linearity. A small *p*-value would lead one to reject the null hypothesis of linearity, indicating that Passing-Bablok regression might not be appropriate.

Max CUSUM The maximum of the absolute values of the cumulative sums of the values $\sqrt{L/I}$ and $-\sqrt{I/L}$ that are assigned to each row based on the sign of the residual and sorted by the perpendicular distance of each point to the Passing-Bablok line. I is the number of observations with a positive residual, and L is the number of observations with a negative residual. When there is strong correlation between methods, I is equal to L. Therefore, the cumulative sums are often sums of +1 and -1. A small CUSUM value indicates a random distribution of points on either side of the Passing-Bablok Line; this result supports the hypothesis of linearity.

H The test statistic for the CUSUM test. This test statistic is defined as the Max CUSUM divided by the square root of the number of negative residuals plus 1. This test statistic has a Kolmogorov-Smirnov distribution.

Prob > H The *p*-value for the CUSUM test. Small *p*-values indicate that the Passing-Bablok procedure might not be appropriate.

Parameter Estimates

The Parameter Estimates table contains the Passing-Bablok fit estimates of the intercept and slope with corresponding 95% confidence intervals.

Matched Pairs Report

Use the Bland Altman Analysis option in the Passing-Bablok Fit red triangle menu to perform a paired *t*-test and Bland-Altman analysis. For more information about matched pairs see *Predictive and Specialized Modeling*.

Bland-Altman Analysis

The Bland-Altman Analysis table contains the value, the standard deviation, and the confidence limits for the following parameters:

Bias The mean difference between the X and Y variables.

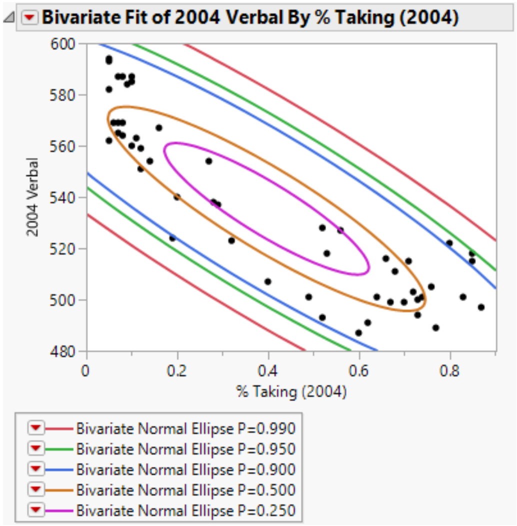
Limits of Agreement The upper and lower limits of agreement, which are set at bias ± $z_{1-\alpha/2}$*(standard deviation of the bias).

Density Ellipse Report

In the Bivariate platform, use the Density Ellipse option to draw an ellipse (or ellipses) that contains the specified mass of points and to estimate and test the correlation between factors. The number of points contained within the ellipse is determined by the probability value that you select from the Density Ellipse menu.

For more information about the options in the Bivariate Normal Ellipse menu, see “Bivariate Fit Options”. For an example of this option, see “Example of Group By Using Density Ellipses”.

Figure 5.13 Example of Density Ellipses



The density ellipsoid is computed from the bivariate normal distribution fit to the X and Y variables. The bivariate normal density is a function of the means and standard deviations of the X and Y variables and the correlation between them.

The ellipses show where a given percentage of the data is expected to lie, assuming the bivariate normal distribution. The shape of the density ellipsoid is a graphical indicator of the correlation between two variables. A narrow ellipsoid indicates strong correlation while a more circular ellipsoid indicates less correlation between the two variables.

Tip: To see a matrix of ellipses and correlations for many pairs of variables, use the **Multivariate** platform in the **Analyze > Multivariate Methods** menu. See *Multivariate Methods*.

The Bivariate Normal Ellipse report contains a table of summary statistics for the density ellipse.

Variable The names of the variables used in creating the ellipse.

Mean The mean of each variable.

Std Dev The standard deviation of each variable.

Correlation The Pearson correlation coefficient. See "Statistical Details for the Bivariate Normal Ellipse Report".

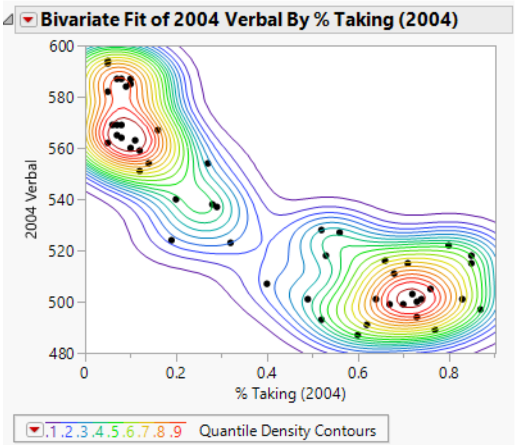
Signif. Prob The probability of obtaining a correlation with greater absolute value than the observed value if no linear relationship exists between the X and Y variables.

Number The number of observations used in the calculations.

Nonpar Density Report

In the Bivariate platform, use the Nonpar Density option to fit a smooth nonparametric bivariate surface to your data. The nonparametric density fit is visualized using a set of contours. The contours are in 5% intervals. This means that about 5% of the points generated from the estimated nonparametric distribution are within the innermost contour, 10% are within the next contour, and so on. The outermost contour contains about 95% of the points.

Figure 5.14 Example of Nonpar Density



To change the size of a nonparametric density contour grid, press Shift and select **Nonpar Density** from the Bivariate red triangle menu. Enter a larger value than the default 102 points.

For more information about the options in the Quantile Density Contours menu, see "Bivariate Fit Options".

Tip: When you have a large number of data points, use the Nonpar Density (nonparametric density) option to view patterns in the scatterplot.

The Quantile Density Contours report contains a table of the standard deviations used in creating the nonparametric density.

Variable The names of the variables used in creating the contours.

Kernel Std The kernel band width. The initial value is based on the standard deviation of the variable. Adjust the slider to increase or decrease the density of the contours.

Additional Examples of the Bivariate Platform

This section contains examples using the Bivariate platform.

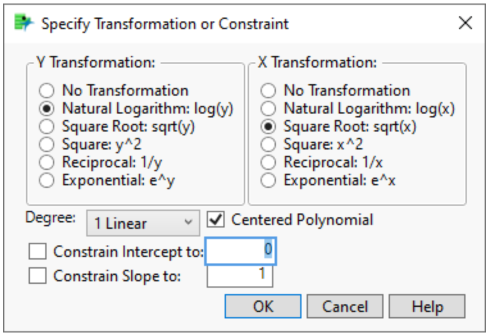
- "Example of the Fit Special Option"
- "Example of the Fit Orthogonal Option"
- "Example of the Fit Robust Option"
- "Example of Group By Using Density Ellipses"
- "Example of Group By Using Regression Lines"
- "Example of Grouping Using a By Variable"

Example of the Fit Special Option

This example shows you how to fit regression models to transformed variables using the Bivariate platform.

1. Select **Help > Sample Data Folder** and open SAT.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select 2004 Verbal and click **Y, Response**.
4. Select % Taking (2004) and click **X, Factor**.
5. Click **OK**.
6. Click the Bivariate Fit red triangle and select **Fit Special**. The Specify Transformation or Constraint window appears. For a description of this window, see "Fit Special Window".
7. For the Y Transformation, select **Natural Logarithm: log(y)**.
8. For the X Transformation, select **Square Root: sqrt(x)**.

Figure 5.15 The Specify Transformation or Constraint Window



Specify Transformation or Constraint

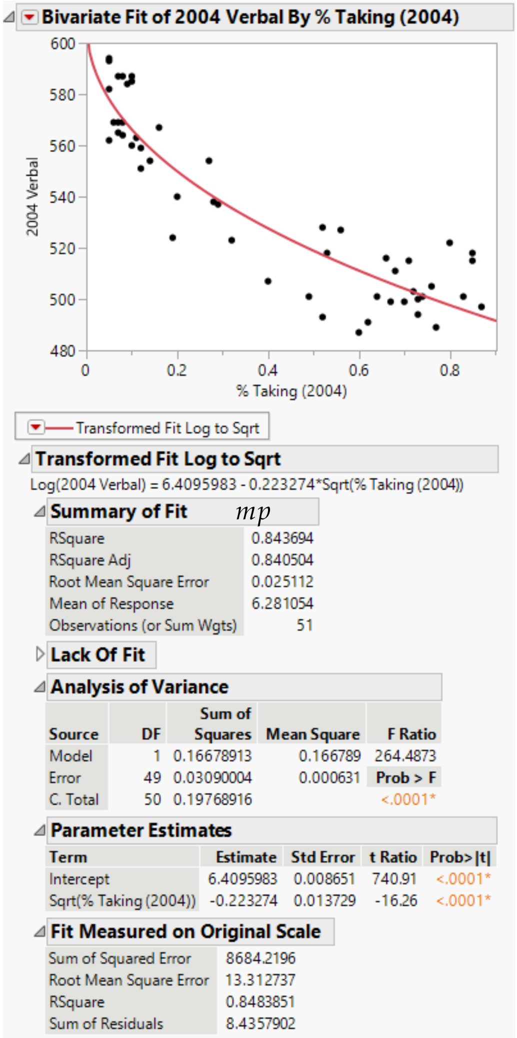
Y Transformation:
○ No Transformation
● Natural Logarithm: log(y)
○ Square Root: sqrt(y)
○ Square: y^2
○ Reciprocal: 1/y
○ Exponential: e^y

X Transformation:
○ No Transformation
○ Natural Logarithm: log(x)
● Square Root: sqrt(x)
○ Square: x^2
○ Reciprocal: 1/x
○ Exponential: e^x

Degree: 1 Linear
☑ Centered Polynomial
☐ Constrain Intercept to: 0
☐ Constrain Slope to: 1

9. Click **OK**.

Figure 5.16 Example of Fit Special Report



Transformed Fit Log to Sqrt

Log(2004 Verbal) = 6.4095983 - 0.223274*Sqrt(% Taking (2004))

Summary of Fit mp

RSquare	0.843694
RSquare Adj	0.840504
Root Mean Square Error	0.025112
Mean of Response	6.281054
Observations (or Sum Wgts)	51

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.16678913	0.166789	264.4873
Error	49	0.03090004	0.000631	**Prob > F**
C. Total	50	0.19768916		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.4095983	0.008651	740.91	<.0001*
Sqrt(% Taking (2004))	-0.223274	0.013729	-16.26	<.0001*

Fit Measured on Original Scale

Sum of Squared Error	8684.2196
Root Mean Square Error	13.312737
RSquare	0.8483851
Sum of Residuals	8.4357902

The model appears to fit the data well. The plotted fit, on the original scale, goes through the cloud of points, and the model RSquare = 0.84.

Example of the Fit Orthogonal Option

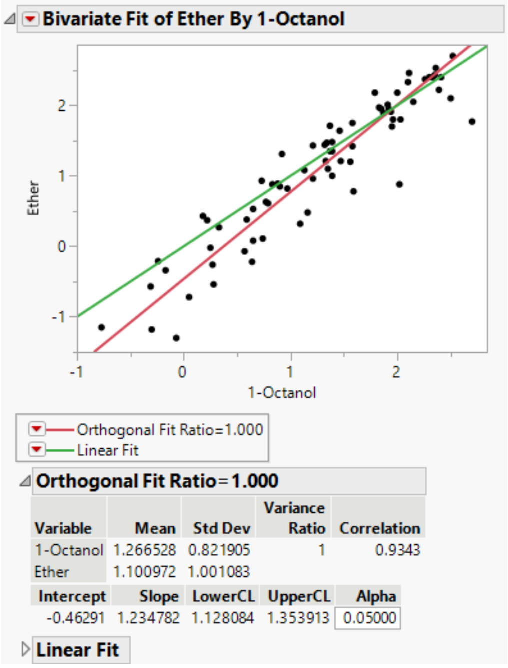
This example shows a method comparison using an orthogonal model in the Bivariate platform. One might also consider Passing-Bablok regression for method comparison data. See “Passing-Bablok Fit Report”. Chemical compounds were measured for solubility in different solvents.

1. Select **Help > Sample Data Folder** and open Solubility.jmp.

2. Select **Analyze > Fit Y by X**.
3. Select Ether and click **Y, Response**.
4. Select 1-Octanol and click **X, Factor**.
5. Click **OK**.
6. Click the Bivariate Fit red triangle and select **Fit Orthogonal > Equal Variances**.
7. For reference, click the Bivariate Fit red triangle and select **Fit Special**.
8. Select **Constrain the Intercept** and set to 0.
9. Select **Constrain the Slope** and set to 1.

   This fits a $Y = X$ line, where the intercept is 0 and the slope is 1.
10. Click **OK**.

Figure 5.17 Example of an Orthogonal Fit



Orthogonal Fit Ratio=1.000

Variable	Mean	Std Dev	Variance Ratio	Correlation
1-Octanol	1.266528	0.821905	1	0.9343
Ether	1.100972	1.001083		

Intercept	Slope	LowerCL	UpperCL	Alpha
-0.46291	1.234782	1.128084	1.353913	0.05000

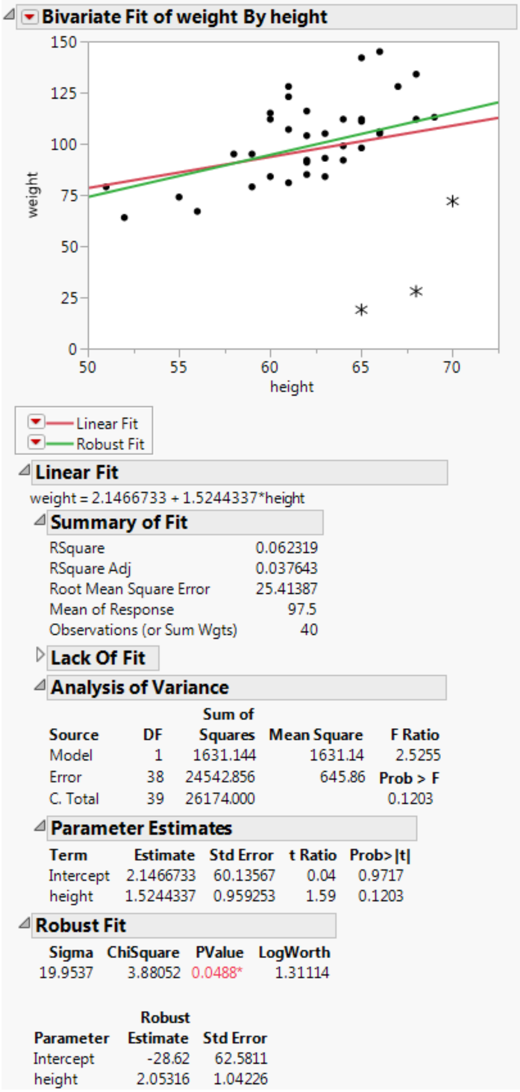
Note that red orthogonal fit line falls below the green $Y = X$ line for most of the data points and that the 95% confidence interval (Lower CL and UpperCL) for the slope for the orthogonal fit does not contain 1. The chemical compounds measured for solubility in Ether tend to measure lower than those measured in 1-Octanol.

Example of the Fit Robust Option

This example shows you how use the Bivariate platform to fit a robust model using the Huber M-estimation method.

1. Select **Help > Sample Data Folder** and open Weight Measurements.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response**.
4. Select height and click **X, Factor**.
5. Click **OK**.
6. Click the Bivariate Fit red triangle and select **Fit Line**.
7. Click the Bivariate Fit red triangle and select **Robust > Fit Robust**.

Figure 5.18 Example of Robust Fit



Linear Fit

weight = 2.1466733 + 1.5244337*height

Summary of Fit

RSquare	0.062319
RSquare Adj	0.037643
Root Mean Square Error	25.41387
Mean of Response	97.5
Observations (or Sum Wgts)	40

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1631.144	1631.14	2.5255
Error	38	24542.856	645.86	**Prob > F**
C. Total	39	26174.000		0.1203

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.1466733	60.13567	0.04	0.9717
height	1.5244337	0.959253	1.59	0.1203

Robust Fit

Sigma	ChiSquare	PValue	LogWorth
19.9537	3.88052	0.0488*	1.31114

Parameter	Robust Estimate	Std Error
Intercept	-28.62	62.5811
height	2.05316	1.04226

Three measurements are marked by stars in the scatterplot. These three observations appear to have weights that are lower than expected. When fitting a regression line (denoted by the red line), these points have leverage on the fit. The *p*-value of 0.1203 in the Analysis of Variance report suggests a weak linear relationship between height and weight. However, when you consider the Robust Fit (denoted by the green line), the linear relationship between height and weight is stronger than for the standard fit. The Robust Fit *p*-value of 0.0488 supports the hypothesis of a linear relationship between height and weight. The measurements identified with unusually low weights were influencing the analysis.

Example of Group By Using Density Ellipses

This example shows you how to use a grouping (By) variable in the Bivariate platform and add density ellipses to your data. The data table in this example identifies three different types of hot dogs: beef, meat, or poultry. You want to group the three types of hot dogs according to their cost variables.

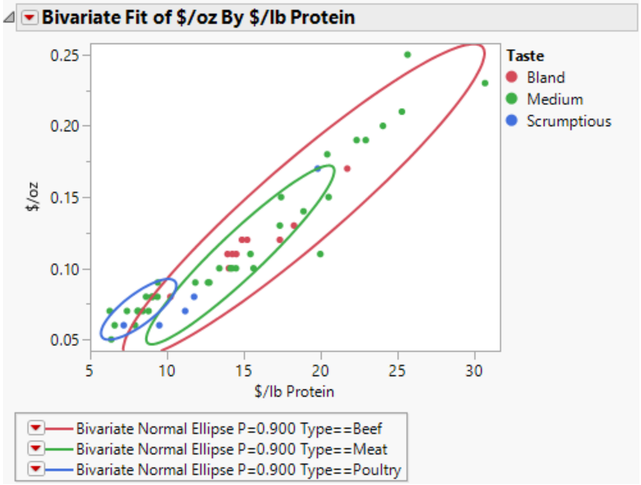
1. Select **Help > Sample Data Folder** and open Hot Dogs.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select $/oz and click **Y, Response**.
4. Select $/lb Protein and click **X, Factor**.
5. Click **OK**.
6. Click the Bivariate Fit red triangle and select **Group By**.
7. From the list, select Type.
8. Click **OK**.

   If you look at the **Group By** option again, you see it has a check mark next to it.
9. Click the Bivariate Fit red triangle and select **Density Ellipse > 0.90**.

Color the points according to Type:

10. Right-click the scatterplot and select **Row Legend**.
11. Select Type in the column list and click **OK**.

Figure 5.19 Example of Group By



The ellipses in Figure 5.19 show clearly how the different types of hot dogs cluster with respect to the cost variables.

Example of Group By Using Regression Lines

This example shows you how to use a grouping variable in the Bivariate platform to overlay regression lines so that you can compare the slopes of different groups.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response**.
4. Select height and click **X, Factor**.
5. Click **OK**.

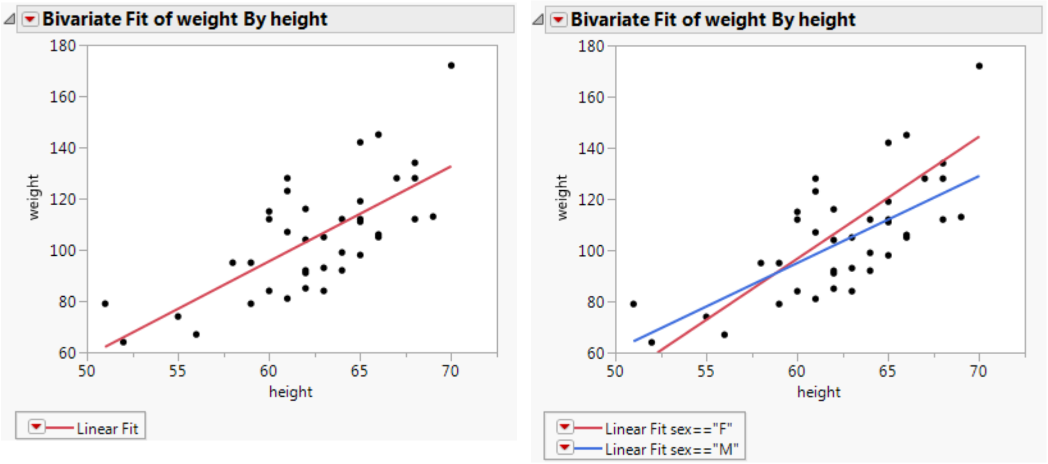
To create the example on the left in Figure 5.20:

6. Click the Bivariate Fit red triangle and select **Fit Line**.

To create the example on the right in Figure 5.20:

7. From the Linear Fit menu, select **Remove Fit**.
8. Click the Bivariate Fit red triangle and select **Group By**.
9. From the list, select sex.
10. Click **OK**.
11. Click the Bivariate Fit red triangle and select **Fit Line**.

Figure 5.20 Example of Regression Analysis for Whole Sample and Grouped Sample



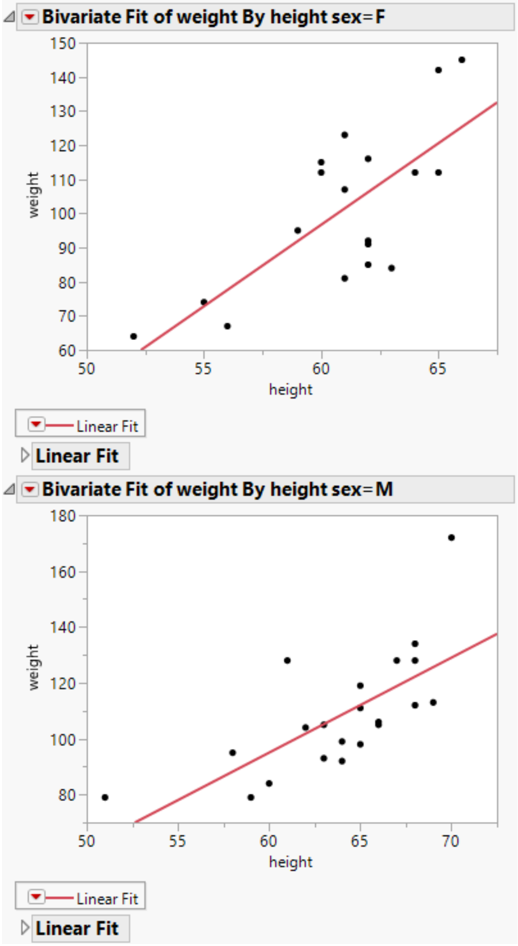
The scatterplot on the left in Figure 5.20 has a single regression line that relates weight to height. The scatterplot on the right shows separate regression lines for males and females.

Example of Grouping Using a By Variable

This example shows grouping by specifying a By variable in the Bivariate launch window. This results in separate reports and graphics for each level of the By variable (or combinations of By variables).

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response**.
4. Select height and click **X, Factor**.
5. Select sex and click **By**.
6. Click **OK**.
7. Press Ctrl, click the Bivariate Fit red triangle and select **Fit Line**.

Figure 5.21 Example of By Variable Plots



There are separate analyses for each level of the By variable (sex). You see a scatterplot for the females and a scatterplot for the males.

Statistical Details for the Bivariate Platform

This section contains statistical details for the Bivariate platform.

- "Statistical Details for the Fit Line Option"
- "Statistical Details for the Fit Polynomial Option"
- "Statistical Details for the Fit Spline Option"
- "Statistical Details for the Fit Orthogonal Option"
- "Statistical Details for the Robust Options"
- "Statistical Details for the Fit Passing Bablok Option"
- "Statistical Details for the Summary of Fit Report"
- "Statistical Details for the Lack of Fit Report"
- "Statistical Details for the Parameter Estimates Report"
- "Statistical Details for the Smoothing Fit Report"
- "Statistical Details for the Bivariate Normal Ellipse Report"

Statistical Details for the Fit Line Option

In the Bivariate platform, the Fit Line option finds the parameters β_0 and β_1 for the straight line that fits the points to minimize the residual sum of squares. The model for the i^{th} row is written $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

The mean confidence limits for a linear fit are defined as follows:

$$\hat{y}_i \pm t_{\alpha/2,\, n-2}\left[\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}\right]\hat{\sigma}$$

The individual confidence limits for a linear fit are defined as follows:

$$\hat{y}_i \pm t_{\alpha/2,\, n-2}\left[\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}\right]\hat{\sigma}$$

where

$\hat{y}_i$ is the predicted value at x_i, the i^{th} observation

$\bar{x}$ is the mean of the x variables

$$\hat{\sigma} = \sqrt{\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}}$$

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

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

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$t_{\alpha/2,n}$ is the $(\alpha/2)^{th}$ quantile of the central *t*-distribution with n–2 degrees of freedom

n is the sample size

Note: Use the Mean Confidence Limit Formula and the Individual Confidence Limit Formula options to save the limit formulas for any model fit to a new column in the data table. The formulas are saved in matrix formats.

Statistical Details for the Fit Polynomial Option

A polynomial of degree 2 is a parabola; a polynomial of degree 3 is a cubic curve. For degree k, the model for the i^{th} row is defined as follows:

$$y_i = \sum_{j=0}^{k} \beta_j x_i^j + \varepsilon_i$$

where

y_i is the response value at x_i, the i^{th} observation

β_j is the j^{th} parameter

k is the degree of the polynomial

ε_i is the error term for the model

Statistical Details for the Fit Spline Option

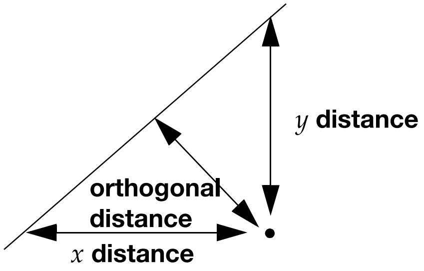
In the Bivariate platform, the cubic spline method uses a set of third-degree polynomials spliced together such that the resulting curve is continuous and smooth at the splices (knot points). The estimation is done by minimizing an objective function that is a combination of the sum of squared errors and a penalty for curvature integrated over the curve extent. See the paper by Reinsch (1967) or the text by Eubank (1999) for a description of this method.

Statistical Details for the Fit Orthogonal Option

In the Bivariate platform, you can use the Fit Orthogonal option to fit a line that minimizes the sum of the squared perpendicular differences. This can be preferable to standard least squares when there is random variation in the measurement of X. Standard least square fitting assumes that the X variable is fixed and the Y variable is a function of X plus error.

Note: The perpendicular distance depends on how X and Y are scaled, and the scaling for the perpendicular is reserved as a statistical issue, not a graphical one.

Figure 5.22 Line Perpendicular to the Line of Fit



The fit requires that you specify the ratio of the variance of the error in Y to the error in X. This is the variance of the error, not the variance of the sample points, so you must choose carefully. In standard least squares, the ratio $(\sigma_y^2)/(\sigma_x^2)$ is infinite because σ_x^2 is zero. An orthogonal fit with a large error ratio approaches the standard least squares line of fit. If you specify a ratio of zero, the fit is equivalent to the regression of X on Y, instead of Y on X.

The most common use of this technique is in comparing two measurement systems that both have errors in measuring the same value. Thus, the Y response error and the X measurement error are both the same type of measurement error. The error ratio is an assumed value, such as 1 (equal variances), or can be based on knowledge of the measurement systems.

Confidence limits are calculated as described in Tan and Iglewicz (1999).

Statistical Details for the Robust Options

This section contains details for the Fit Robust and Fit Cauchy options in the Bivariate platform.

Fit Robust

In the Bivariate platform, the Fit Robust option reduces the influence of response variable outliers on the model fit. The Huber M-estimation method is used. Huber M-estimation finds parameter estimates that minimize the Huber loss function:

$$l(\underset{\sim}{e}) = \sum_i \rho(e_i)$$

where

$$\rho(e) = \begin{cases} \frac{1}{2}e^2 & \text{if } |e| < k \\ k|e| - \frac{1}{2}k^2 & \text{if } |e| \geq k \end{cases}$$

e_i refers to the residuals

The Huber loss function penalizes outliers and increases as a quadratic for small errors and linearly for large errors. In the JMP implementation, $k = 2$. For more information about robust fitting, see Huber (1973) and Huber and Ronchetti (2009).

Fit Cauchy

In the Bivariate platform, the Fit Cauchy option estimates parameters using maximum likelihood and a Cauchy link function. This method assumes that the errors have a Cauchy distribution. A Cauchy distribution has fatter tails than the normal distribution, resulting in a reduced emphasis on outliers.

Statistical Details for the Fit Passing Bablok Option

In the Bivariate platform, the Fit Passing Bablok option uses the Passing-Bablok procedure to fit the linear equation $y = \beta_0 + \beta_1 x$ where both x and y are measured with error. The method is robust to outliers and is appropriate if there is a linear relationship between x and y. The estimate of the slope (β_1) is calculated as the median of all slopes that can be formed from all possible pairs of data points, except those rare pairs that result in an undefined slope of 0/0 or a slope of -1. To correct for estimation bias caused by the lack of independence of these slopes, the median is shifted by a factor based on *K*, the number of slopes that are less than -1. The result is an approximately unbiased estimator for β_1. The intercept (β_0) is estimated by the median of $\{y_i - \beta_1 x_i\}$. See Passing and Bablok (1983), Passing and Bablok (1984), and Bablok et. al. (1988).

Passing-Bablok is a commonly used for method comparison studies. The intercept is interpreted as the systematic bias (difference) between the two methods. The slope measures the amount of proportional bias (difference) between the two methods.

Statistical Details for the Summary of Fit Report

This section contains details for the Summary of Fit report in the Bivariate platform.

RSquare

Using quantities from the corresponding analysis of variance table, the RSquare for any continuous response fit is calculated as follows:

$$\frac{\text{Sum of Squares for Model}}{\text{Sum of Squares for C. Total}}$$

RSquare Adj

The RSquare Adj is a ratio of mean squares instead of sums of squares and is calculated as follows:

$$1-\frac{\text{Mean Square for Error}}{\text{Mean Square for C. Total}}$$

The mean square for Error is in the Analysis of Variance report (Figure 5.11). You can compute the mean square for C. Total as the Sum of Squares for C. Total divided by its respective degrees of freedom.

Note: RSquare and RSquare Adj are not reported when a constraint is placed on either the intercept or the slope.

Statistical Details for the Lack of Fit Report

This section contains details for the Lack of Fit report in the Bivariate platform.

Pure Error DF

For the Pure Error DF, consider the cases where more than one observation has the same value for height. In general, if there are g groups having multiple rows with identical values for each effect, the pooled DF, denoted DF_p, is defined as follows:

$$\mathrm{DF}_p = \sum_{i=1}^{g} (n_i - 1)$$

where n_i is the number of observations in the i^{th} group.

Pure Error SS

For the Pure Error SS, in general, if there are g groups having multiple rows with the same x value, the pooled SS, denoted SS_p, is defined as follows:

$$\mathrm{SS}_p = \sum_{i=1}^{g} \mathrm{SS}_i$$

where SS_i is the sum of squares for the i^{th} group corrected for its mean.

Max RSq

Because Pure Error is invariant to the form of the model and is the minimum possible variance, Max RSq is calculated as follows:

$$1 - \frac{\mathrm{SS(Pure\ error)}}{\mathrm{SS(Total\ for\ whole\ model)}}$$

Statistical Details for the Parameter Estimates Report

This section contains details for the Parameter Estimates report in the Bivariate platform.

Std Beta

Std Beta is calculated as follows:

$$\hat{\beta}(s_x / s_y)$$

where $\hat{\beta}$ is the estimated parameter, s_x and s_y are the standard deviations of the X and Y variables.

Design Std Error

Design Std Error is calculated as the standard error of the parameter estimate divided by the RMSE.

Statistical Details for the Smoothing Fit Report

This section contains details for the Smoothing Fit report in the Bivariate platform.

R-Square is equal to 1-(SSE/C. Total SS), where C. Total SS is available in the Fit Line ANOVA report.

Statistical Details for the Bivariate Normal Ellipse Report

This section contains details for the Bivariate Normal Ellipse report in the Bivariate platform.

The Pearson correlation coefficient is denoted *r*, and is computed as follows:

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} \text{ where } s_{xy} = \frac{\sum w_i(x_i - \overline{x_i})(y_i - \overline{y_i})}{df}$$

where w_i is either the weight of the i^{th} observation if a weight column is specified, or 1 if no weight column is assigned.

Chapter 6

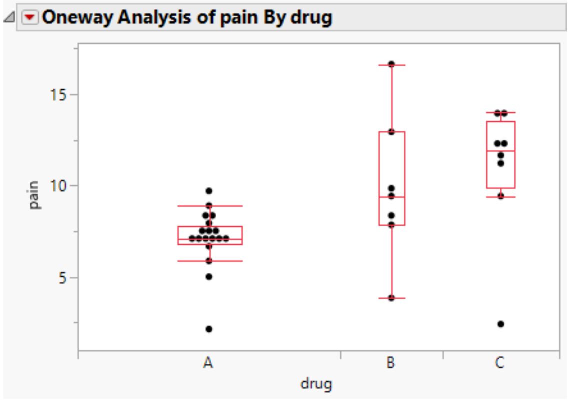
Oneway Analysis

Examine Relationships between a Continuous Y and a Categorical X Variable

Use the Oneway platform to explore how the distribution of a continuous Y variable differs across groups defined by a categorical X variable. You can use analysis of variance (ANOVA), analysis of means (ANOM), *t* tests, equivalence tests, nonparametric tests, exact tests, or multiple comparisons to evaluate differences in means or variances across groups. Density and cumulative distribution function (CDF) plots enable you to visualize the distribution of your response by the X categories. For example, you might want to find out how different categories of the same type of drug (X) affect patient pain levels on a numerical scale (Y).

The Oneway platform is the *continuous by nominal or ordinal* personality of the Fit Y by X platform.

Figure 6.1 Oneway Analysis



Contents

Overview of Oneway Platform

The Oneway platform enables you to quickly visualize the distribution of a continuous Y variable across two or more levels of a categorical X variable. You can run analysis of variance (ANOVA) models, *t* tests, analysis of means (ANOM), equivalence tests, nonparametric tests, exact tests, or multiple comparisons to evaluate differences in means or variances across the levels of the X variable. Many tests add visualization tools to the Oneway plot in the report. In addition, there are options to use density and cumulative distribution function (CDF) plots to visualize the distribution of your response across the levels of the X variable.

Tests are grouped together in the Oneway red triangle menu according to the type of test. JMP enables you to explore many types of statistical tests. The best choice of test depends on the study design, data collection, data distributions, and the hypothesis of interest.

See the following sections for details about test reports:

- “Quantiles Report”
- “Means/Anova/Pooled t Report”
- “Means and Std Deviations Report”
- “t Test Report”
- “Analysis of Means Reports”
- “Means Comparisons Reports”
- “Nonparametric Test Reports”
- “Nonparametric Multiple Comparisons Reports”
- “Unequal Variances Reports”
- “Equivalence Test Reports”
- “Robust Fit Reports”
- “Power Reports”
- “Matching Column Report”

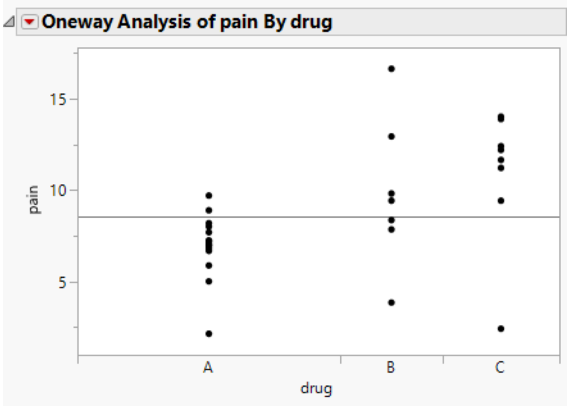
Example of Oneway Analysis

Use the Oneway platform to perform a one-way analysis of variance. This analysis enables you to test for statistically significant differences in group means. In this example, thirty-three participants were administered three different types of analgesics (A, B, and C). The participants were asked to rate their pain levels on a sliding scale. You want to find out if the mean pain levels for treatments A, B, and C are statistically significantly different.

1. Select **Help > Sample Data Folder** and open Analgesics.jmp.

2. Select **Analyze > Fit Y by X**.
3. Select pain and click **Y, Response**.
4. Select drug and click **X, Factor**.
5. Click **OK**.

Figure 6.2 Example of Oneway Analysis



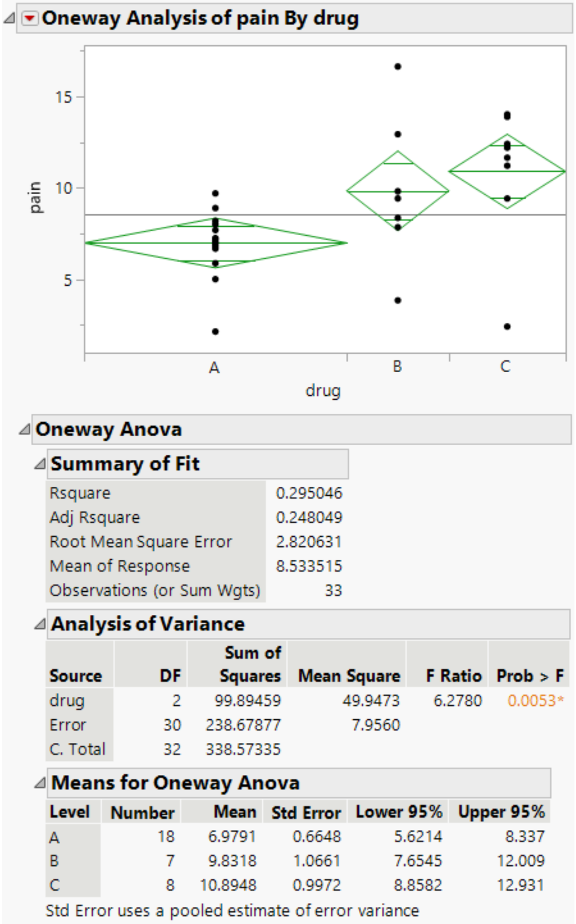
You notice that one drug (A) has consistently lower pain scores than the other drugs. You also notice that the *x*-axis ticks are unequally spaced. The length between the ticks is proportional to the number of scores (observations) for each drug.

Perform an analysis of variance on the data.

6. Click the Oneway Analysis red triangle and select **Means/Anova**.

> **Note:** If the X variable has only two levels, the **Means/Anova** option appears as **Means/Anova/Pooled t,** and adds a Pooled *t* test report to the report window.

Figure 6.3 Example of the Means/Anova Option



Oneway Analysis of pain By drug

Oneway Anova

Summary of Fit

Rsquare	0.295046
Adj Rsquare	0.248049
Root Mean Square Error	2.820631
Mean of Response	8.533515
Observations (or Sum Wgts)	33

Analysis of Variance

		Sum of			
Source	**DF**	**Squares**	**Mean Square**	**F Ratio**	**Prob > F**
drug	2	99.89459	49.9473	6.2780	0.0053*
Error	30	238.67877	7.9560		
C. Total	32	338.57335			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
A	18	6.9791	0.6648	5.6214	8.337
B	7	9.8318	1.0661	7.6545	12.009
C	8	10.8948	0.9972	8.8582	12.931

Std Error uses a pooled estimate of error variance

Note the following observations:

- Mean diamonds representing confidence intervals are added to the plot.
  - The line at the center of each diamond represents the group mean. At a glance, you can see that the mean for drug A is lower than the means for drugs B and C.
  - The vertical span of each diamond represents the 95% confidence interval for the mean of each group.

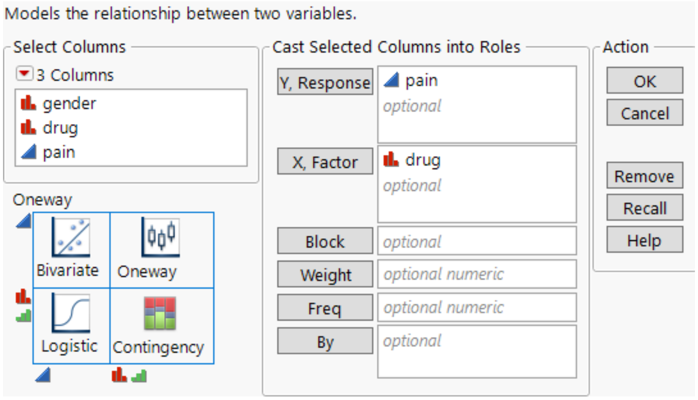
  See "Mean Diamonds and X-Axis Proportional".
- The Summary of Fit table provides overall summary information about the analysis.
- The Analysis of Variance report shows the standard ANOVA information. You notice that the Prob > F (the *p*-value) is 0.0053, which supports your visual conclusion that there are significant differences between the drugs.

- The Means for Oneway Anova report shows the mean, sample size, and standard error for each level of the categorical factor.

Launch the Oneway Platform

Launch the Oneway platform by selecting **Analyze > Fit Y by X**. The Fit Y by X launch window is used for four different types of analyses. When you enter a continuous Y variable and a categorical nominal or ordinal X variable, the Oneway platform is launched. A report for each combination of X and Y variables is launched.

Figure 6.4 The Oneway Launch Window



Models the relationship between two variables.

Y, Response: pain

X, Factor: drug

For more information about the options in the Select Columns red triangle menu, see *Using JMP*. The Oneway launch window contains the following options:

Y, Response The response variable or variables that you want to analyze. These variables must have a numeric data type.

X, Factor The predictor variable or variables that you want to analyze. These variables must have a nominal or ordinal data type.

Block A column that specifies a Blocking variable. When used, the values of the Y variable are centered by the Block variable. If there are equal counts in each block by group cell, the available Oneway platform options are for a two-way analysis. If there are unequal counts in the block by group cells, a one-way ANOVA with blocking model is fit and no Oneway platform options are available.

Weight A column containing a weight for each observation in the data table. A row is included in the analyses only when its value is greater than zero.

Freq Assigns a frequency to each row in the analysis. Assigning a frequency is useful when your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Data Format

In the Oneway platform, your data can consist of columns of unsummarized or summarized data:

Unsummarized data There is one row for each observation and columns for X and Y values.

Summarized data Each row represents a set of observations with common X and Y values. The data table must contain a frequency column of the counts of observations for each row. Enter this column as Freq in the launch window.

When one-way data are in a format other than a JMP data table, sometimes they are arranged so that a row contains information for multiple observations. To analyze the data in JMP, you must import the data and restructure it so that each row of the JMP data table contains information for a single observation. See “Example of Stacking Data for a Oneway Analysis”.

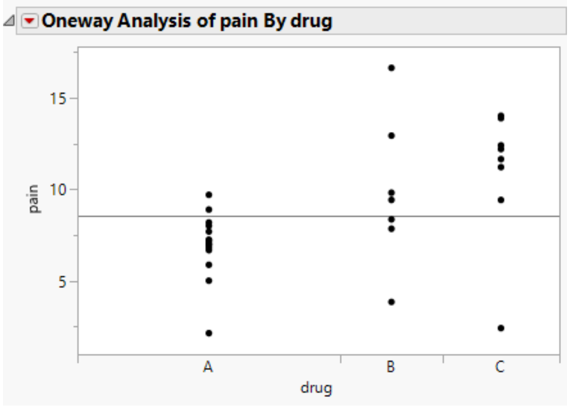
Note: The Fit Y by X launch window accommodates columns with continuous, ordinal, and nominal data types. The Oneway platform is launched for all pairs of continuous Y, Response columns and nominal or ordinal X, Factor columns. The Fit Y by X launch window launches the Bivariate, Contingency, or Logistic platforms for other column type combinations.

Oneway Report

The Oneway report initially contains a scatterplot for each pair of nominal or ordinal X variables and continuous Y variables. You can fit models, perform hypothesis tests, perform equivalence tests, compare distributions, and view statistical reports using the red triangle options. See “The Oneway Platform Options”.

Note: If a Block variable is specified in the launch window, the values of the Y variable in the Oneway plot are centered by the Block variable. The options that are available differ for analyses with blocks.

Figure 6.5 Oneway Plot



Replace Variables Interactively

You can also replace a variable by selecting a variable in the Columns panel of the associated data table and dragging it onto an axis.

The Oneway Platform Options

The Oneway Analysis red triangle menu contains testing, fitting, and display options. Options include analysis of variance (ANOVA), nonparametric tests, equivalence tests, and multiple comparisons methods to evaluate differences across groups. Density and CDF plots enable you to visualize the distribution of your response by the X categories. Options add visualizations to the scatterplot, provide statistical reports, or launches an additional analysis.

When a blocking variable is used, a one-way ANOVA model with blocking is fit to the data. Additional analysis options are limited.

Note: The Fit Group menu appears only if you have specified multiple Y or X variables. Use the Fit Group menu options to arrange reports or order them by RSquare. See *Fitting Linear Models*.

The Oneway Analysis red triangle menu contains the following options:

Quantiles Shows or hides box plots on the Oneway plot and shows or hides a quantile report. See "Oneway Analysis Reports".

Means/Anova (Available only when the X variable has more than two levels.) Shows or hides mean diamonds on the Oneway plot and shows or hides an ANOVA report. See "Means/Anova/Pooled t Report".

Means/Anova/Pooled t (Available only when the X variable has exactly two levels.) Shows or hides mean diamonds on the Oneway plot and shows or hides an ANOVA report. The ANOVA report includes a pooled *t* test report that assumes that the two groups have equal variances.

Means and Std Dev Shows or hides mean lines, error bars, and standard deviation lines on the Oneway plot and shows or hides a summary statistics table. The standard errors for the means use individual group standard deviations. For more information about these graph elements, see "Mean Lines, Error Bars, and Standard Deviation Lines".

t test (Available only when the X variable has exactly two levels.) Shows or hides a *t* test report assuming that the variances are not equal. See "t Test Report".

Analysis of Means Methods Contains options for comparing multiple groups using Analysis of Means (ANOM) methods. See "Analysis of Means Reports". For more information about ANOM methods, see Nelson et al. (2005).

Note: If you specify a Block variable in the launch window and there are equal counts for each combination of Block and X variable level, the ANOM chart is the only Analysis of Means chart available. If you specify a Block variable in the launch window and there are unequal counts, none of the Analysis of Means options are available.

ANOM Shows or hides an ANOM decision chart to compare group means to the overall mean. This method assumes that your data are approximately normally distributed. See "Example of Analysis of Means".

ANOM with Transformed Ranks Shows or hides an ANOM with transformed ranks decision chart for a nonparametric comparison of group means. ANOM with Transformed Ranks compares each group's mean transformed rank to the overall mean transformed rank. The ANOM test applies the usual ANOM procedure and critical values to the transformed observations.

Tip: Use this method if your data are clearly non-normal and cannot be transformed to normality.

ANOM for Variances (Available only when the X variable has more than two levels and each group has at least four observations.) Shows or hides an ANOMV decision chart to compare group standard deviations (or variances) to the root mean square error (or mean square error). This method assumes that your data is approximately normally distributed. For more information about the ANOM for Variances test, see Wludyka and Nelson (1997) and Nelson et al. (2005). For an example, see "Example of Analysis of Means for Variances".

ANOM for Variances with Levene (ADM) Shows or hides an ANOMV-LEV decision chart that is robust to non-normality to compare group variances. This method compares the group means of the *absolute deviations from the median* (ADM) to the overall mean ADM. For more information about the ANOM for Variances with Levene (ADM) test, see Levene (1960) or Brown and Forsythe (1974).

Tip: Use ANOM for Variances with Levene (ADM) if you suspect that your data are non-normal and cannot be transformed to normality. ANOM for Variances with Levene (ADM) is robust to outliers and non-normality.

ANOM for Ranges (Available only for balanced designs and specific group sizes.) Shows or hides an ANOMV-TR decision chart to compare group ranges to the mean of the group ranges. This is a test for scale differences based on the range as the measure of spread. See Wheeler (2003). See "Restrictions for ANOM for Ranges Test".

Tip: To select between the robust ANOMV methods, note that ANOMV-TR is more robust than ANOM-LEV. However, ANOM-LEV is more powerful than ANOM-TR.

Compare Means Contains options for multiple comparisons between specified sets of group means. See "Means Comparisons Reports".

Each Pair, Student's t Shows or hides a Fisher's LSD multiple comparison of all pairs of means report and shows or hides comparison circles on the Oneway plot.

All Pairs, Tukey HSD Shows or hides a Tukey-Kramer multiple comparison of all pairs of means report and shows or hides comparison circles on the Oneway plot.

With Best, Hsu MCHB Shows or hides a Hsu's multiple comparison with the best report and shows or hides comparison circles on the Oneway plot.

With Control, Dunnett's Shows or hides a Dunnett's multiple comparison with a control report and shows or hides comparison circles on the Oneway plot.

Tip: You can add a Control Level column property to the factor column to avoid specifying the control group each time you select With Control, Dunnett's. See *Using JMP*.

Each Pair Stepwise, Newman-Keuls Shows or hides a Newman-Keuls stepwise paired comparison report. The Newman-Keuls test does not control the family-wise error rate.

Nonparametric Contains options for nonparametric comparisons of group locations. See "Nonparametric Test Reports".

Wilcoxon / Kruskal-Wallis Test Shows or hides one or more tests that are based on Wilcoxon rank scores. The Wilcoxon rank scores are the simple ranks of the data. The Wilcoxon test is the most powerful rank test for errors with logistic distributions.

If the X variable has exactly two levels, the Wilcoxon test is equivalent to the Mann-Whitney test. In this situation, the report contains a table for the Wilcoxon two-sample normal approximation test that uses a 0.5 continuity correction and a table for the Kruskal-Wallis chi-square approximation test that does not use a continuity correction.

If the X variable has more than two levels, the test based on simple ranks is called the Kruskal-Wallis test.

For information about the report, see “Wilcoxon Kruskal - Wallis, Median, Friedman Rank, and Van der Waerden Test Reports”. For an example, see “Example of the Nonparametric Wilcoxon Test”.

Median Test Shows or hides a test based on median rank scores. The median rank scores are either 1 or 0, depending on whether a rank is above or below the median rank. The median test is the most powerful rank test for errors with double-exponential distributions. For information about the report, see “Wilcoxon Kruskal - Wallis, Median, Friedman Rank, and Van der Waerden Test Reports”.

van der Waerden Test Shows or hides a test based on Van der Waerden rank scores. The Van der Waerden rank scores are the ranks of the data divided by one plus a score value. The score value is the number of observations transformed to a normal score by applying the inverse of the normal distribution function. The Van der Waerden test is the most powerful rank test for errors with normal distributions. For information about the report, see “Wilcoxon Kruskal - Wallis, Median, Friedman Rank, and Van der Waerden Test Reports”.

Kolmogorov Smirnov Test (Available only when the X variable has exactly two levels.) Shows or hides a test based on the empirical distribution function, which tests whether the distribution of the response is the same across the groups. For information about the report, see “Kolmogorov-Smirnov Two-Sample Test Report”.

Friedman Rank Test (Available only when a Block variable with an equal number of observations within each block is specified in the launch window.) Shows or hides a test based on Friedman Rank scores. The Friedman Rank scores are the ranks of the data within each level of the blocking variable. The parametric version of this test is a repeated measures ANOVA. For information about the report, see “Wilcoxon Kruskal - Wallis, Median, Friedman Rank, and Van der Waerden Test Reports”.

Notes:

– For the Wilcoxon, Median, Van der Waerden, and Friedman Rank tests, if the X variable has more than two levels, a chi-square approximation to the one-way test is performed

  – If you specify a Block variable in the launch window and there are equal counts for each combination of Block and X variable level, the Friedman Rank Test is the only Nonparametric option available. If you specify a Block variable in the launch window and there are unequal counts, none of the Nonparametric options are available.

Exact Test (Available only when the X variable has exactly two levels.) Contains options for performing exact tests. See "Exact Test Reports".

Wilcoxon Exact Test (Available only when the X variable has exactly two levels.) Shows or hides a test of Wilcoxon scores using exact methods. See "Example of the Nonparametric Wilcoxon Test".

Median Exact Test (Available only when the X variable has exactly two levels.) Shows or hides a test of median scores using exact methods.

Van Der Waerden Exact Test (Available only when the X variable has exactly two levels.) Shows or hides a test of Van der Waerden normal scores using exact methods.

Kolmogorov Smirnov Exact Test (Available only when the X variable has exactly two levels.) Shows or hides a test based on the empirical distribution function, which tests whether the distribution of the response is the same across the groups.

Caution: Exact tests might take a long time to compute for large sample sizes.

Nonparametric Multiple Comparisons Contains options for nonparametric multiple comparisons of group locations.

Wilcoxon Each Pair Shows or hides the Wilcoxon test on each pair. This procedure does not control for the overall alpha level. This is the nonparametric version of the Each Pair, Student's t option found on the Compare Means menu. See "Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control Reports".

Steel-Dwass All Pairs Shows or hides the Steel-Dwass test on each pair. This is the nonparametric version of the All Pairs, Tukey HSD option found on the Compare Means menu. See "Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control Reports".

Steel With Control Shows or hides the Steel test to compare each level of the grouping variable to a control level. This is the nonparametric version of the With Control, Dunnett's option found on the Compare Means menu. See "Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control Reports".

Dunn All Pairs for Joint Ranks Shows or hides a comparison of each pair using the Dunn method. The Dunn method computes ranks for all the data, not just the pair being compared. The reported *p*-value reflects a Bonferroni adjustment. It is the unadjusted *p*-value multiplied by the number of comparisons. If the adjusted *p*-value exceeds 1, it is reported as 1. See "Dunn All Pairs for Joint Ranks and Dunn with Control for Joint Ranks Report".

Dunn With Control for Joint Ranks Shows or hides a comparison of each level of the grouping variable to a control level using the Dunn method. The Dunn method computes ranks for all the data, not just the pair being compared. The reported *p*-value reflects a Bonferroni adjustment. It is the unadjusted *p*-value multiplied by the number of comparisons. If the adjusted *p*-value exceeds 1, it is reported as 1. You can add a Control Level column property to the factor column to avoid specifying the control group each time you select Steel With Control or Dunn With Control for Joint Ranks. See *Using JMP*. See "Dunn All Pairs for Joint Ranks and Dunn with Control for Joint Ranks Report".

Note: The choice of a nonparametric test is dependent on your data. For guidance on test selection for a hypothesis about all pairs see Boos and Duan (2021)

Unequal Variances Shows or hides four tests for equality of group variances and Welch's ANOVA test for equal means with unequal standard deviations. The tests for equality of group variances are O-Brien, Brown-Forsythe, Levene, and Bartletts. See "Unequal Variances Reports".

Equivalence Tests Contains methods to test for equivalence, superiority, or noninferiority. See "Equivalence Test Reports". Such tests are useful when you want to detect similarities that are of practical or clinically significant interest.

Means Launches a window with options for equivalence, superiority, or noninferiority tests for means. Specify variance assumptions and the critical difference.

Standard Deviations Launches a window with options for equivalence, superiority, or noninferiority tests for standard deviations. Specify the critical ratio.

Robust Contains options for a robust mean estimates. Robust methods reduce the influence of outliers on your analysis. Robust estimates are more efficient than least squares estimates for distributions with heavy tails. See "Robust Fit Reports".

Robust Fit Shows or hides robust mean lines on the Oneway plot and shows or hides a Robust Fit report that includes Huber estimates of the group means.

Cauchy Fit Shows or hides robust mean lines on the Oneway plot and shows or hides a Cauchy Fit report that includes estimates of the group means assuming a Cauchy error distribution. This robust method is appropriate for extreme outliers.

Power Enables you to run power calculations for the ANOVA analysis. See "Power Reports". For more information about power calculations as well as examples, see *Fitting Linear Models*.

Set α Level Enables you to specify an α level. You can select from a list of common alpha levels or specify a level by selecting Other.

> **Note:** The alpha level is applied across analyses and reports. This includes confidence limits, mean diamonds, comparison circles, and multiple comparison analyses. The alpha levels for equivalence tests are set separately.

Normal Quantile Plot Contains options for plotting the quantiles of the data in each group. See "Example of a Normal Quantile Plot".

Plot Actual by Quantile Shows or hides a quantile plot to the right of the Oneway Analysis plot. The quantile plot shares the Oneway Analysis vertical axis for the Y variable. The cumulative probabilities for each group are on the horizontal axis. The plot shows quantiles computed within each level of the X variable.

Plot Quantile by Actual Shows or hides a quantile plot with the Y variable on the horizontal axis and cumulative probabilities on the vertical axis. The plot shows quantiles computed within each level of the categorical X variable.

Line of Fit (Available only when a quantile plot is open.) Shows or hides a reference line fit to the data for each level of the X variable on each open quantile plot.

Normal Quantile Label (Available only when a quantile plot is open.) Shows or hides the normal quantile scale on each open quantile plot.

CDF Plot Shows or hides the cumulative distribution function for all of the groups in the Oneway report. See "Example of a CDF Plot".

Densities Contains options for visualizing densities across groups. See "Example of the Densities Options".

Compare Densities Shows or hides a plot of overlaid probability density functions for each group.

Composition of Densities Shows or hides a plot of the summed densities, weighted by the count of each group. Across the range of the X variable, the Composition of Densities plot shows how each group contributes to the total density.

Proportion of Densities Shows or hides a plot of the contribution to the density made by each level of the X variable. The contribution is shown as a proportion of the total density across the range of the X variable.

Matching Column Shows or hides a matched fit line and a corresponding fit line on the Oneway plot based on a specified matching variable. Use this option when the data in your analysis come from matched (paired) data, such as when observations in different groups come from the same participant. See “Matching Column Report”.

Save Saves the following quantities as new columns in the current data table:

Save Residuals Saves values computed as the Y variable minus the mean of the Y variable within each level of the X variable.

Save Standardized Saves standardized values of the Y variable for each level of the X variable. The standardized value is the centered response divided by the standard deviation within each level.

Save Normal Quantiles Saves normal quantile values computed within each level of the X variable.

Save Predicted Saves the predicted mean of the Y variable for each level of the X variable.

Display Options Adds or removes elements from the plot. Some options might not appear when they are not relevant.

All Graphs Shows or hides the Oneway plot.

Points Shows or hides data points on the Oneway plot.

Box Plots Shows or hides outlier box plots for each group. See “Outlier Box Plot”. For an example, see “Conduct the Oneway Analysis”.

Mean Diamonds Shows or hides mean diamonds on the Oneway plot. Each mean diamond spans a 95% confidence interval for the corresponding group mean, with a horizontal line at the mean. The 95% confidence intervals are calculated using the pooled standard deviation. See “Mean Diamonds and X-Axis Proportional”.

Mean Lines Shows or hides a line at the mean of each group. See “Mean Lines, Error Bars, and Standard Deviation Lines”.

Mean CI Lines Shows or hides lines at the upper and lower 95% confidence levels for each group. The 95% confidence levels are calculated using the pooled standard deviation.

Mean Error Bars Shows or hides the mean of each group with error bars that are one standard error above and below the mean. See “Mean Lines, Error Bars, and Standard Deviation Lines”.

Grand Mean Shows or hides the overall mean of the Y variable.

Std Dev Lines Shows or hides lines that are one standard deviation above and below the mean of each group. See "Mean Lines, Error Bars, and Standard Deviation Lines".

Comparison Circles (Available only when a multiple comparison report is open.) Shows or hides comparison circles. See "Statistical Details for Comparison Circles". For an example, see "Conduct the Oneway Analysis".

Connect Means Shows or hides straight lines that connect the group means.

Mean of Means Shows or hides the mean of the group means.

X-Axis Proportional (Not available when the Matching Column option is selected.) Specifies the spacing on the horizontal axis. When selected, the spacing is proportional to the number of observations at each level. See "Mean Diamonds and X-Axis Proportional".

Points Spread Specifies the spread of the data points. When selected, the data points are spread across the width of the interval.

Points Jittered Specifies the spread of the data points. When selected, the data points are jittered to avoid overlapping markers. Specifically, this option is equivalent to the Centered Grid option for Jitter in Graph Builder. See *Essential Graphing*.

Matching Lines (Available only when the Matching Column option is selected.) Shows or hides lines that connect the means of each level of the matching variable.

Matching Dotted Lines (Available only when the Matching Column option is selected and values of the matching variable are all missing for a level of the X variable.) Shows or hides dotted lines that connect the means through missing levels of the matching variable. The values used in place of the missing cell means are obtained using a two-way ANOVA model.

Histograms Shows or hides side-by-side histograms to the right of the original plot.

Robust Mean Lines (Available only when a Robust option is selected.) Shows or hides a line at the robust mean of each group.

Legend Shows or hides a legend for the normal quantile, cumulative distribution function (CDF), and density plots.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Oneway Analysis Reports

The Oneway platform enables you to create visualizations, run hypothesis tests, and fit models to explore the relationship between a categorical X variable and a continuous Y variable. This section contains information about the reports that are generated for specific analysis options.

- "Quantiles Report"
- "Means/Anova/Pooled t Report"
- "Means and Std Deviations Report"
- "t Test Report"
- "Analysis of Means Reports"
- "Means Comparisons Reports"
- "Nonparametric Test Reports"
- "Nonparametric Multiple Comparisons Reports"
- "Unequal Variances Reports"
- "Equivalence Test Reports"
- "Robust Fit Reports"
- "Power Reports"
- "Matching Column Report"

Quantiles Report

In the Oneway platform, the Quantiles report lists the minimum and maximum values for each group. In addition, the 10th, 25th, 50th (median), 75th, and 90th percentiles are listed. The median is the 50th percentile, and the 25th and 75th percentiles are also called *quartiles*.

Means/Anova/Pooled t Report

In the Oneway platform, use the Means/Anova option to perform an analysis of variance. If the X variable has only two levels, this option appears as **Means/Anova/Pooled t**. The report contains tables for the summary of fit, an analysis of variance (ANOVA), and summary statistics for each group. The report includes a Pooled *t* test table if the X variable has only two levels. The report includes a Block Means table if you have specified a Block variable in the launch window and there are equal counts in each combination of block and level of the X variable. Other blocking configurations result in a Oneway Anova with Blocking report.

Summary of Fit

In the Oneway platform, the Summary of Fit table contains the following statistics:

Rsquare The proportion of the variation that is explained by the model. The remaining variation is attributed to random error. The RSquare is 1 if the model fits perfectly. See "Statistical Details for the Summary of Fit Report". R^2 is also called the *coefficient of determination*.

Note: A low RSquare value suggests that there might be variables not in the model that account for the unexplained variation. However, if your data are subject to a large amount of inherent variation, even a useful ANOVA model can have a low RSquare value. Read the literature in your research area to learn about typical RSquare values.

Adj Rsquare The RSquare statistic adjusted for the number of parameters in the model. RSquare Adj facilitates comparisons among models that contain different numbers of parameters. See "Statistical Details for the Summary of Fit Report".

Root Mean Square Error The estimate of the standard deviation of the random error. This quantity is the square root of the mean square for Error in the Analysis of Variance report.

Mean of Response The overall mean (arithmetic average) of the Y variable.

Observations (or Sum Wgts) Number of observations used in estimating the fit. If weights are used, this is the sum of the weights. See "Statistical Details for the Summary of Fit Report".

Pooled t Test

In the Oneway platform, the Pooled t Test table summarizes the results of a *t* test to compare two group means assuming equal variances across groups. This table is available only when the X variable has exactly two levels. See "t Test Report".

Analysis of Variance

In the Oneway platform, the Analysis of Variance table summarizes the results of the ANOVA analysis. An ANOVA partitions the total variation into components.

Note: If you specified a **Block** column, then the Analysis of Variance report includes the **Block** variable.

Source The sources of variation. These sources are the model source, **Error**, and **C. Total** (corrected total).

DF The degrees of freedom (DF for short) for each source of variation:

- The degrees of freedom for **C. Total** are N - 1, where N is the total number of observations used in the analysis.
- The degrees of freedom for the model is k - 1, where k is the number of levels of the X variable.

The **Error** degrees of freedom is the difference between the **C. Total** degrees of freedom and the **Model** degrees of freedom (in other words, N - k).

Sum of Squares The sum of squares (SS for short) for each source of variation:

- The total (**C. Total**) sum of squares of each response from the overall response mean. The **C. Total** sum of squares is the base model used for comparison with all other models.
- The sum of squared distances from each point to its respective group mean. This is the remaining unexplained **Error** (residual) SS after fitting the analysis of variance model.

The total SS minus the error SS gives the sum of squares attributed to the model. This tells you how much of the total variation is explained by the model.

Mean Square The sum of squares divided by its associated degrees of freedom:

- The **Model** mean square estimates the variance of the error, but only under the hypothesis that the group means are equal.
- The **Error** mean square estimates the variance of the error term independently of the model mean square and is unconditioned by any model hypothesis.

F Ratio The model mean square divided by the error mean square. If the hypothesis that the group means are equal (there is no real difference between them) is true, then both the mean square for error and the mean square for model estimate the error variance. Their ratio has an F distribution. If the analysis of variance model results in a significant reduction of variation from the total, the F ratio is higher than expected.

Prob>F Probability of obtaining an *F* value greater than the one calculated if there is no difference in the population group means. Observed significance probabilities of 0.05 or less are often considered evidence that there are differences in the group means.

Means for Oneway Anova

In the Oneway platform, the Means for Oneway Anova table summarizes response information for each level of the nominal or ordinal factor.

Level The levels of the X variable.

Number The number of observations in each group.

Mean The mean of each group.

Std Error The estimates of the standard deviations for the group means. This standard error is estimated assuming that the variance of the response is the same in each level. It is the root mean square error found in the Summary of Fit report divided by the square root of the number of values used to compute the group mean.

Lower 95% and Upper 95% The lower and upper 95% confidence interval for the group means based on a pooled estimate of the error variance and pooled degrees of freedom.

Block Means

In the Oneway platform, the Block Means report appears only if you have specified a Block variable in the launch window and there are equal counts in each combination of block and level of the X variable.

Means and Std Deviations Report

In the Oneway platform, the Means and Std Deviations report contains summary statistics for each level of the X variable. The levels of the X variable define groups, or treatments.

Level The levels of the X variable.

Number The number of observations in each group.

Mean The mean of each group.

Std Dev The standard deviation of each group.

Std Err Mean The estimates of the standard error for the group means. This value is the standard deviation divided by the square root of the number of values used to compute the group mean.

Lower 95% and Upper 95% The lower and upper 95% confidence interval for the group means based on the individual group means, standard deviations, and sample sizes.

Note: These confidence intervals are generally narrower than the ANOVA confidence intervals.

t Test Report

In the Oneway platform, the t Test report summarizes the results of a *t* test to compare two group means assuming either equal or unequal variances across groups. This report is available only when the X variable has exactly two levels.

The *t* Test or Pooled t Test table lists the difference used and the variance assumption. The report also contains the following statistics:

Difference The estimated difference between the two X levels.

Tip: Use the Value Order column property to change the order of the differences.

Std Err Dif The standard error of the difference.

Upper CL Dif The upper confidence limit for the difference.

Lower CL Dif The lower confidence limit for the difference.

Confidence The level of confidence (1-alpha). To change the level of confidence, select a new alpha level from the **Set** α **Level** option from the platform red triangle menu.

t Ratio The *t*-statistic.

DF The degrees of freedom used in the *t* test.

Prob > |t| The *p*-value associated with a two-tailed test.

Prob > t The *p*-value associated with a one-sided upper-tailed test.

Prob < t The *p*-value associated with a one-sided lower-tailed test.

t Test plot Shows the sampling distribution of the difference in the means, assuming that the null hypothesis is true. The vertical red line is the observed difference in the means. The shaded areas correspond to the *p*-values.

Analysis of Means Reports

In the Oneway platform, analysis of means (ANOM) is a multiple comparison method for means or variances. Each ANOM report contains a decision chart with the following elements:

- an upper decision limit (UDL)
- a lower decision limit (LDL)
- a horizontal (center) line; the center line position is determined by the chart type:
  - ANOM: the overall mean
  - ANOM with Transformed Ranks: the overall mean of the transformed ranks
  - ANOM for Variances: the root mean square error (or MSE when in variance scale)
  - ANOM for Variances with Levene (ADM): the overall absolute deviation from the mean
  - ANOM for Ranges: the mean of the group ranges

If a group's plotted statistic falls outside of the decision limits, then the test indicates that there is a statistical difference between that group's statistic and the overall average of the statistic for all the groups.

Analysis of Means Chart Options

In the Oneway platform, each Analysis of Means red triangle menu contains the following options:

Set Alpha Level Enables you to select a significance level from the list or specify a level with the **Other** selection. The upper and lower decision limits update when the alpha level changes.

Note: For ANOM for Ranges, only the significance levels of 0.10, 0.05, and 0.01 are available.

Show Summary Report Shows or hides a summary report that is based on the Analysis of Means method:

- For ANOM, creates a report showing group means and decision limits.
- For ANOM with Transformed Ranks, creates a report showing group mean transformed ranks and decision limits.
- For ANOM for Variances, creates a report showing group standard deviations (or variances) and decision limits.
- For ANOM for Variances with Levene (ADM), creates a report showing group mean ADMs and decision limits.

- For ANOM for Ranges, creates a report showing group ranges and decision limits.

Graph in Variance Scale (Available only for **ANOM for Variances**.) Changes the scale of the vertical axis in the decision chart between either standard deviations units or variances.

Display Options Contains the following options to customize the display:

Show Decision Limits Shows or hides decision limit lines.

Show Decision Limit Shading Shows or hides decision limit shading.

Show Center Line Shows or hides the center line statistic.

Point Options: Show Needles Shows the needles. This is the default option. **Show Connected Points** shows a line connecting the means for each group. **Show Only Points** shows only the points representing the means for each group.

Restrictions for ANOM for Ranges Test

Unlike the other ANOM decision limits, the decision limits for the ANOM for Ranges chart uses tabled critical values. For this reason, ANOM for Ranges is available only under the following conditions:

- groups of equal sizes
- groups specifically of the following sizes: 2–10, 12, 15, and 20
- number of groups between 2 and 30
- alpha levels of 0.10, 0.05, and 0.01

Means Comparisons Reports

In the Oneway platform, there are multiple options for comparing means. This section covers options for the mean comparison reports as well as details about each comparison method.

Means Comparisons Report Options

In the Oneway platform, the Means Comparisons red triangle menu contains the following options:

Difference Matrix Shows or hides a matrix of all pairwise differences between group means.

Confidence Quantile Shows or hides the critical value(s) and significance level (α) used for the means comparison procedure.

Tip: Use the Set α Level option in the Oneway Analysis red triangle menu to change the significance level.

LSD Threshold Matrix (Not available for the Each Pair Stepwise option.) Shows or hides a matrix of pairwise differences of means minus the least significant difference for those means. A positive value indicates a pair of means that are significantly different. For Hsu's MCB test, there are two LSD matrices. One for the comparison to the minimum and one for the comparison to the maximum.

Connecting Letters Report (Available only for the Each Pair, All Pairs, and Each Pair Stepwise options.) Shows or hides the traditional letter-coded report where means that do not share a letter are significantly different.

Ordered Differences Report (Available only for the Each Pair and All Pairs options.) Shows or hides all pairwise positive differences, standard error of the difference, confidence intervals, *p*-values, and a plot of the magnitude of the difference with confidence intervals. The *p*-value corresponds to the hypothesis of equal means.

Detailed Comparisons Report (Available only for the Each Pairs option.) Shows or hides a detailed report for each comparison. Each comparison shows the difference between the levels, standard error of the difference, confidence intervals, *t*-ratios, *p*-values, and degrees of freedom. A plot illustrating the comparison appears on the right of each report.

Note: The standard error of the differences is the pooled standard error based on the MSE and the sample size for each pair.

Each Pair, Student's t

In the Oneway platform, use the Each Pair, Student's t option to show Fisher's least significant difference (LSD) test based on Student's *t* test to test for differences across each pair of group levels. For an example of this test, see "Example of the Each Pair, Student's t Test".

All Pairs, Tukey HSD

In the Oneway platform, use the All Pairs, Tukey HSD option to test all pairs of differences. This method protects the overall significance level for the tests of all combinations of pairs. The HSD intervals are wider than the Student's *t* pairwise LSD intervals. Compared to the comparison circles for the All Pairs comparisons, the circles for the Tukey HSD comparisons are larger and the differences between pairs of means are less significant.

In the Confidence Quantile table, the q^* statistic is calculated as $q^* = (1/\text{sqrt}(2)) * q$ where q is the Alpha percentile of the studentized range distribution. For an example of this test, see "Example of the All Pairs, Tukey HSD Test".

The All Pairs, Tukey HSD option performs the *Tukey* or *Tukey-Kramer* HSD (honestly significant difference) test (Tukey 1953; Kramer 1956). This test is an exact alpha-level test if the sample sizes are the same, and conservative if the sample sizes are different (Hayter 1984).

With Best, Hsu MCB

In the Oneway platform, use the With Best, Hsu MCB option to test whether the mean for a given level exceeds the maximum mean of the remaining levels, or is smaller than the minimum mean of the remaining levels. See Hsu (1996). For an example of this test, see "Example of the With Best, Hsu MCB Test".

The quantiles for the Hsu MCB test vary by the level of the categorical variable. Unless the sample sizes are equal across levels, the comparison circle technique is not exact. The radius of a comparison circle is given by the standard error of the level multiplied by the largest quantile value. Use the *p*-values of the tests to obtain precise assessments of significant differences. See "Comparison with Max and Min".

The With Best, Hsu MCB option tests performs the Hsu MCB test (Hsu 1996; Hsu 1981).

Note: Means that are not regarded as the maximum or the minimum by MCB are also the means that are not contained in the selected subset of Gupta (1965) of potential maximums or minimum means.

Comparison with Max and Min

In addition to tables that are common to all of the available means comparisons methods, the Hsu's MCB report contains a Comparison with Max and Min table. This table contains the following values:

Level The level of the categorical variable.

with Max p-Value The *p*-value for the test that the mean of the level exceeds the maximum mean of the remaining levels. Use the tests in this column to screen out levels whose means are significantly smaller than or equal to the (unknown) largest true mean.

with Min p-Value The *p*-value for the test that the mean of the level is smaller than the minimum mean of the remaining levels. Use the tests in this column to screen out levels whose means are significantly greater than or equal to the (unknown) smallest true mean.

With Control, Dunnett's

In the Oneway platform, use the With Control, Dunnett's option to compare group means to a control group. The Dunnett's LSDs are between the Student's *t* and Tukey-Kramer LSDs, as they are sized on an intermediate number of comparisons. For an example of this test, see "Example of the With Control, Dunnett's Test". For more information about Dunnett's test, see Dunnett (1955).

Each Pair Stepwise, Newman-Keuls

In the Oneway platform, use the Each Pair Stepwise, Newman-Keuls option to compare the group means using an iterative, stepwise procedure. At each iteration, Tukey's HSD test is used to test the difference between two group means. For an example of this test, see "Example of the Each Pair Stepwise, Newman-Keuls Test".

The Each Pair Stepwise, Newman-Keuls option tests whether there are differences between the means using the Studentized range test in a stepwise procedure. This is the Newman-Keuls or Student-Newman-Keuls method (Keuls, 1952). This test is less conservative than a Tukey HSD test.

Caution: The Newman-Keuls test does not control the family-wise error rate. Use caution when interpreting the results of this procedure.

The following procedure is used for testing J group means:

Define the following:

J = number of groups (sorted in ascending order of group means)

N = number of observations

d = degrees of freedom, calculated as $N - J$

i = index of smallest group mean involved in a comparison

j = index of largest group mean involved in a comparison

k = minimum value of j in any comparison during the procedure

At the beginning of the procedure, set $i = 1$, $j = J$, and $k = 2$.

1. Perform Tukey's HSD test for groups i and j, where the number of groups for finding the appropriate quantile equals $j - \mathrm{i} + 1$.
   - If the test is significant, groups i and j are determined to be significantly different. Decrease j by 1.

     If this causes j to be less than k, then increase i by 1, set $k = \max(i, j) + 1$, set $j = J$, and continue to step 2.

     If this causes j to be greater than or equal to k, then continue to step 2.
   - If the test is not significant, groups i and j are not determined to be significantly different. Increase i by 1, set $k = j + 1$, set $j = J$, and continue to step 2.
2. Determine whether the procedure continues or stops based on the value of k.
   - If k is greater than i and k is less than or equal to J, repeat step 1.
   - If k is less than or equal to i or k is greater than J, stop the procedure. Any remaining untested ranges are deemed not to be significantly different.

The quantile used for Tukey's HSD is different for each test and is based on the number of group means between the sorted means being tested. In the Newman-Keuls report, the Smallest Quantile Considered (labeled Smallest q*) is the smallest studentized range quantile used in the above procedure divided by the square root of 2.

The test results are reported in the Connecting Letters Report.

For more information about the Newman-Keuls test, see Howell (2013).

Note: There are no mean circles added to the Comparison Circles graph when you use the **Each Pair Stepwise, Newman-Keuls** test. This is because the comparison circles would vary in size at each step of the procedure.

Nonparametric Test Reports

In the Oneway platform, there are nonparametric options for comparing means. This section covers reports for these methods.

The methods include testing the hypothesis of equal means or medians across groups. Nonparametric tests use functions of the response ranks, called rank scores. See Hajek (1969) and SAS Institute Inc. (2020a). Nonparametric multiple comparison procedures are also available to control the overall error rate for pairwise comparisons. See "Nonparametric Multiple Comparisons Reports".

Wilcoxon Kruskal - Wallis, Median, Friedman Rank, and Van der Waerden Test Reports

For Wilcoxon Kruskal - Wallis, Median, Friedman Rank, or Van der Waerden Tests in the Oneway platform, two or three tables provide summary statistics and test results. The summary statistics tables contain the following columns:

Level The levels of X.

Count The frequencies of each level.

Score Sum The sum of the rank score for each level.

Expected Score The expected score under the null hypothesis that there is no difference among class levels.

Score Mean The mean rank score for each level.

(Mean-Mean0)/Std0 The standardized score. Mean0 is the mean score expected under the null hypothesis. Std0 is the standard deviation of the score sum expected under the null hypothesis. The null hypothesis is that the group means or medians are equal across groups.

Wilcoxon Two-Sample Test, Normal Approximation Table

When the X variable has exactly two levels, a 2-Sample Test, Normal Approximation table contains the following columns:

S The sum of the rank scores for the level with the smaller number of observations.

Z The test statistic for the normal approximation test that uses a 0.5 continuity correction. See "Two-Sample Normal Approximations".

Prob>|Z| The *p*-value for the normal approximation test that uses a 0.5 continuity correction. This *p*-value is based on a standard normal distribution.

Kruskal-Wallis Test, ChiSquare Approximation

The Kruskal-Wallis Test table contains the results of a chi-square test for location. See Conover (1999). If the number of groups is two, the Kruskal-Wallis test is equivalent to the Wilcoxon test.

ChiSquare The values of the chi-square test statistic. See "One-Way ChiSquare Approximations".

DF The degrees of freedom for the test.

Prob>ChiSq The *p*-value for the test. The *p*-value is based on a ChiSquare distribution with degrees of freedom equal to the number of levels of X minus 1. If the number of groups is two, this *p*-value is equal to the *p*-value for the normal approximation test without a 0.5 continuity correction.

Kolmogorov-Smirnov Two-Sample Test Report

In the Oneway platform, the Kolmogorov-Smirnov test report contains two tables: a summary table and an asymptotic test table. The summary table contains the following columns:

Level The two levels of the X variable.

Count The frequencies of each level.

EDF at Maximum The value of the empirical cumulative distribution function (EDF) for that level of the X variable for which the difference between the two EDFs is a maximum. For the Total, the value is the pooled EDF (the EDF for the entire data set) at the value of the X variable for which the difference between the two EDFs is a maximum.

Deviation from Mean at Maximum For each level, the value obtained by the following steps:

- Compute the difference between the EDF at Maximum for the given level and the EDF at maximum for the pooled data set (Total).
- Multiply this difference by the square root of the count of observations for that level.

The asymptotic test table contains the following columns:

KS The Kolmogorov-Smirnov statistic computed as follows:

$$\mathrm{KS} = \underset{j}{max} \sqrt{\frac{1}{n}\sum_{i} n_i(F_i(x_j) - F(x_j))^2}$$

The formula uses the following notation:

– x_j, $j = 1,\dots, n$ are the observations

– n_i is the number of observations in the ith level of X

– F is the pooled cumulative empirical distribution function

– $F_{\boldsymbol{i}}$ is the cumulative empirical distribution function for the i^{th} level of X

Note: Although this version of the Kolmogorov-Smirnov statistic applies even when the X variable has more than two levels, in JMP the Kolmogorov-Smirnov option is available only when the X variable has exactly two levels.

KSa The asymptotic Kolmogorov-Smirnov statistic computed as $\mathrm{KS}\sqrt{n}$, where n is the total number of observations.

D=max|F1-F2| The maximum absolute deviation between the EDFs for the two levels. This is the version of the Kolmogorov-Smirnov statistic typically used to compare two samples.

Prob > D The *p*-value for the test. This is the probability that D exceeds the computed value under the null hypothesis of no difference between the levels.

D+ = max(F1-F2) The one-sided test statistic for the alternative hypothesis that the level of the first group exceeds the level of the second group.

Prob > D+ The *p*-value for the test of D+.

D- = max(F2-F1) The one-sided test statistic for the alternative hypothesis that the level of the second group exceeds the level of the first group

Prob > D- The *p*-value for the test of D-.

Exact Test Reports

When the X variable has exactly two levels, the Oneway platform can perform an exact test for each nonparametric test type. For the Wilcoxon Kruskal -Wallis, Median, and Van der Waerden Exact Tests, the 2-Sample: Exact Test table contains the following columns:

S The sum of the rank scores for the observations in the smaller group. If the two levels of X have the same numbers of observations, then the value of S corresponds to the last level of X in the value ordering.

Prob ≤ S The one-sided *p*-value for the test.

Prob ≥ |S-Mean| The two-sided *p*-value for the test.

Note: For the Kolmogorov-Smirnov exact test, the table provides the same statistics as the asymptotic test. However, the *p*-values are computed to be exact. See “Kolmogorov-Smirnov Two-Sample Test Report”.

Nonparametric Multiple Comparisons Reports

In the Oneway platform, the Nonparametric Multiple Comparisons option provides several methods for performing nonparametric multiple comparisons. These tests are based on ranks and, except for the Wilcoxon Each Pair test, control for the overall experimentwise error rate. For more information about these tests, see See Dunn (1964) and Hsu (1996). This section covers reports for these methods.

Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control Reports

The reports for these multiple comparison procedures include test results and confidence intervals. For these tests, observations are ranked within the sample obtained by considering only the two levels used in a given comparison.

q* The quantile used in computing the confidence intervals.

Alpha The alpha level used in computing the confidence interval. You can change the confidence level by selecting the Set α Level option from the Oneway menu.

Level The first level of the X variable used in the pairwise comparison.

- Level The second level of the X variable used in the pairwise comparison.

Score Mean Difference The mean of the rank score of the observations in the first level (Level) minus the mean of the rank scores of the observations in the second level (-Level), where a continuity correction is applied.

Denote the number of observations in the first level by n_1 and the number in the second level by n_2. The observations are ranked within the sample consisting of these two levels. Tied ranks are averaged. Denote the sum of the ranks for the first level by ScoreSum_1 and for the second level by ScoreSum_2.

If the difference in mean scores is positive, then the Score Mean Difference is defined as follows:

$$\text{Score Mean Difference} = (\text{ScoreSum}_1 - 0.5)/n_1 - (\text{ScoreSum}_2 + 0.5)/n_2$$

If the difference in mean scores is negative, then the Score Mean Difference is defined as follows:

$$\text{Score Mean Difference} = (\text{ScoreSum}_1 + 0.5)/n_1 - (\text{ScoreSum}_2 - 0.5)/n_2$$

Std Error Dif The standard error of the Score Mean Difference.

Z The standardized test statistic, which has an asymptotic standard normal distribution under the null hypothesis of no difference in means.

p-Value The *p*-value for the asymptotic test based on Z.

Hodges-Lehmann The Hodges-Lehmann estimator of the location shift. All paired differences consisting of observations in the first level minus observations in the second level are constructed. The Hodges-Lehmann estimator is the median of these differences. The Difference Plot bar chart shows the size of the Hodges-Lehmann estimate.

Lower CL The lower confidence limit for the Hodges-Lehmann statistic.

Note: Not computed if group sample sizes are large enough to cause memory issues.

Upper CL The upper confidence limit for the Hodges-Lehmann statistic.

Note: Not computed if group sample sizes are large enough to cause memory issues.

Difference Plot A bar chart of the score mean differences.

Dunn All Pairs for Joint Ranks and Dunn with Control for Joint Ranks Report

The Dunn comparison procedures are based on the rank of an observation in the entire data set. For the Dunn with Control for Joint Ranks tests, you must select a control level.

Level The first level of the X variable used in the pairwise comparison.

- Level The second level of the X variable used in the pairwise comparison.

Score Mean Difference The mean of the rank score of the observations in the first level (Level) minus the mean of the rank scores of the observations in the second level (-Level), where a continuity correction is applied. The ranks are obtained by ranking the observations within the entire sample. Tied ranks are averaged. The continuity correction is described in "Score Mean Difference".

Std Error Dif The standard error of the Score Mean Difference.

Z The standardized test statistic, which has an asymptotic standard normal distribution under the null hypothesis of no difference in means.

p-Value The *p*-value for the asymptotic test based on Z.

Unequal Variances Reports

The Oneway platform provides four tests for equality of group variances. The concept behind the first three tests of unequal variances is to perform an analysis of variance on a new response variable constructed to measure the spread in each group. The fourth test is Bartlett's test, which is similar to the likelihood ratio test under normal distributions. The Unequal Variances option is not available when a Block variable is specified in the launch window.

Note: Another method to test for unequal variances is ANOMV. See "Analysis of Means Reports".

The following Tests for Equal Variances are available:

O'Brien Constructs a dependent variable so that the group means of the new variable equal the group sample variances of the original response. An ANOVA on the O'Brien variable is actually an ANOVA on the group sample variances (O'Brien 1979; Olejnik and Algina 1987).

Brown-Forsythe Shows the *F* test from an ANOVA where the response is the absolute value of the difference of each observation and the group median (Brown and Forsythe 1974).

Levene Shows the *F* test from an ANOVA where the response is the absolute value of the difference of each observation and the group mean (Levene 1960). The spread is measured as $z_{ij} = |y_{ij} - \bar{y}_i|$.

Bartlett Compares the weighted arithmetic average of the sample variances to the weighted geometric average of the sample variances. The geometric average is always less than or equal to the arithmetic average with equality holding only when all sample variances are equal. The more variation there is among the group variances, the more these two averages differ. A function of these two averages is created, which approximates a χ^2-distribution (or, in fact, an *F* distribution under a certain formulation). Large values correspond to large values of the arithmetic or geometric ratio, and therefore to widely varying group variances. Dividing the Bartlett Chi-square test statistic by the degrees of freedom gives the *F* value shown in the table. Bartlett's test is not very robust to violations of the normality assumption (Bartlett and Kendall 1946).

F Test 2-sided (Available only if there are two levels of the X variable.) If there are only two groups tested, then a standard *F* test for unequal variances is also performed. The *F* test is the ratio of the larger to the smaller variance estimate. The *p*-value from the *F* distribution is doubled to make it a two-sided test.

Note: If you have specified a **Block** column, then the variance tests are performed on data after it has been adjusted for the Block means.

See "Example of the Unequal Variances Option".

Tip: If the unequal variances test reveals that the group variances are significantly different, consider Welch's test instead of the standard ANOVA test. The Welch statistic is based on the usual ANOVA F test. However, the means are weighted by the reciprocal of the group mean variances (Welch 1951; Brown and Forsythe 1974; Asiribo and Gurland 1990). If there are only two levels, the Welch ANOVA is equivalent to an unequal variance t test.

Tests That the Variances Are Equal Report

In the Oneway platform, the Tests That the Variances Are Equal report shows a plot of standard deviations and provides summary tables. The first table contains the following columns:

Level The factor levels occurring in the data.

Count The frequencies of each level.

Std Dev The standard deviations of the response for each factor level. The standard deviations are equal to the means of the O'Brien variable. If a level occurs only once in the data, no standard deviation is calculated.

MeanAbsDif to Mean The mean absolute difference of the response and group mean. The mean absolute differences are equal to the group means of the Levene variable.

MeanAbsDif to Median The absolute difference of the response and group median. The mean absolute differences are equal to the group means of the Brown-Forsythe variable.

The second table summarizes the tests for equal variances and contains the following columns:

Test The name of each test.

F Ratio The calculated F statistic. See "Statistical Details for Tests That the Variances Are Equal".

DFNum The degrees of freedom in the numerator. If a factor has k levels, the numerator has k - 1 degrees of freedom. Levels occurring only once in the data are not used in calculating test statistics for O'Brien, Brown-Forsythe, or Levene. The numerator degrees of freedom in this situation is the number of levels used in calculations minus one.

DFDen The degrees of freedom used in the denominator. For O'Brien, Brown-Forsythe, and Levene, a degree of freedom is subtracted for each factor level used in calculating the test statistic. If a factor has k levels, the denominator degrees of freedom is n - k.

p-Value The probability of obtaining an F-ratio value larger than the one calculated if the variances are equal across all levels.

Note: A warning appears if any level of the X variable contains fewer than 5 observations. For more information about the performance of the above tests with small sample sizes, see Brown and Forsythe (1974) and Miller (1972).

Welch's Test Report

F Ratio The *F* test statistic for the equal means test.

DFNum The degrees of freedom in the numerator of the test. If a factor has *k* levels, the numerator has *k* - 1 degrees of freedom. Levels occurring only once in the data are not used in calculating the Welch ANOVA. The numerator degrees of freedom in this situation is the number of levels used in calculations minus one.

DFDen The degrees of freedom in the denominator of the test. See "Statistical Details for Tests That the Variances Are Equal".

Prob>F The probability of obtaining an *F* value larger than the one calculated if the means are equal across all levels.

t Test (Available only when the X variable has exactly two levels.) Shows the relationship between the *F* ratio and the *t* Test. Calculated as the square root of the *F* ratio.

Equivalence Test Reports

In the Oneway platform, the options in the Equivalence Test submenu enable you to perform equivalence, superiority, or noninferiority tests for means or standard deviations.

For equivalence tests, the two one-sided tests (TOST) method is used to test for practical equivalence between the means or the ratio of two standard deviations. For equivalence tests about means, see Schuirmann (1987). Two one-sided *t* tests are constructed for the null hypotheses that the true difference or true ratio exceeds the threshold values. If both tests reject, the difference in the means or ratio does not statistically exceed either threshold value. Therefore, the groups are considered practically equivalent. See "Example of an Equivalence Test".

The equivalence test reports contain plots and summary tables. When you select an equivalence test option in the Equivalence Tests submenu, you must specify test characteristics in the Equivalence Test Specification window.

Equivalence Test Specification Window

Selecting an equivalence test option in the Equivalence Tests submenu launches a window that enables you to define your test.

Alternative Hypothesis Defines the structure of the equivalence test.

- **Equivalence (Two-Sided)** Specifies an equivalence test. Use this option when the goal is to show that group differences are not bigger than the equivalence margin.
- **Superiority (One-Sided)** Specifies a superiority test. Use this option when the goal is to show that a group is superior, or better than, another group.
- **Noninferiority (One-Sided)** Specifies a noninferiority test. Use this option when the goal is to show that a group is not inferior to another group.

Side of the Alternative Hypothesis (Available only for one-sided tests.) Specifies the direction of the alternative hypothesis.

Hypotheses Plot Provides a graphical depiction of the hypothesis test.

Type of Comparisons Defines the type of comparison to be performed.

- **All Pairwise** Specifies comparisons between all pairs of group levels.
- **With Control** Specifies comparisons of each level to a specified control level.

Variance Assumption Defines the variance assumption used for the test calculations.

- **Pooled Variance** Specifies that calculations be based on a pooled variance. Use this option when variances are assumed equal across groups.
- **Unequal Variances** Specifies that calculations be based on unequal group variances.

Margin and Alpha Defines the significance levels for the test.

- **Difference** (Available only for tests for means.) Specifies the equivalence, superiority, or noninferiority margin. This margin, or delta, is the difference that has practical significance. For equivalence tests, the difference must be greater than zero.
- **Ratio** (Available only for tests for standard deviations.) Specifies the equivalence, superiority, or noninferiority margin as a ratio of standard deviations. This margin defines a ratio in standard deviations that has practical significance. The range of values is defined as (Ratio, 1/Ratio). For equivalence tests, the ratio must be different from 1.
- **Alpha** Specifies the significance level for the test.

Equivalence Tests Report

The test report begins with a description of the alternative hypothesis begin tested. For each comparison, the Tests report contains the following columns:

Difference (Available only for tests for means.) The estimated difference in the means.

Std Error of Difference (Available only for tests for means.) The estimated standard error of the difference in the means.

Ratio (Available only for tests for standard deviations.) The estimated ratio of standard deviations.

Lower Bound t Ratio, Upper Bound t Ratio (Available only for tests for means. Only one bound appears for one-sided tests.) The lower or upper bound *t* ratios for the one-sided significance tests.

Lower Bound F-value, Upper Bound F-Value (Available only for tests for standard deviations. Only one bound appears for one-sided tests.) The lower or upper bound *t* ratios for the one-sided significance tests.

Lower Bound p-Value, Upper Bound p-Value (Only one *p*-value appears for one-sided tests.) The significance probabilities (*p*-values) that correspond to the lower or upper bound *t* ratios.

Max p-Value (Appears only for equivalence tests.) Maximum of the lower and upper bound *p*-values.

Lower 90%, Upper 90% Limits for a $1–2\alpha$ confidence interval for the difference in the means or the ratio of standard deviations.

Assessment An assessment of the hypothesis test for the specified alpha level.

Equivalence Tests Options

The Equivalence, Superiority, or Noninferiority Tests red triangle menu contains the following options:

Test Report Shows or hides a report that summarizes the equivalence tests. See "Equivalence Tests Report".

Scatterplot (Available only for tests for means.) Shows or hides a scatterplot. The pairwise comparison confidence intervals are plotted on a scatterplot that is colored by acceptance regions. The intervals are plotted on a mean-mean scale with shading that indicates the equivalent, superior, or noninferiority regions. This plot is sometimes called a *diffogram* or a mean-mean scatterplot.

Tip: Hover over a point to show the groups being compared and the estimated difference.

The Scatterplot has the following option:

Show Reference Lines Shows or hides reference lines for the points on the scatterplot. This is not recommended if there are many points in the scatterplot. If there are many points, it is recommended that you hover over the points to view the labels.

Forest Plot Shows or hides a forest plot. The comparison confidence intervals are plotted versus the difference in means or ratio of standard deviations. The intervals are plotted on a difference of means or ratio of standard deviations scale. Shading indicates the equivalent, superior, or noninferiority regions.

Tip: Hover over a point to show the groups being compared and the estimated difference or ratio.

Pairwise Comparisons (Available only for equivalence tests for means.) Shows or hides the Practical Equivalence reports for all pairwise comparisons. In addition to the values provided in the "Equivalence Tests Report", the report includes a visual representation of each paired test.

Remove Removes the test report from the Oneway Analysis report window.

Robust Fit Reports

In the Oneway platform, the Robust option provides two methods to reduce the influence of outliers or extreme data points in your data set: Robust Fit and Cauchy Fit. Each method adds a line to the Oneway plot at the robust mean.

- The Robust Fit option reduces the influence of outliers in the response variable. The Huber M-estimation method is used. Huber M-estimation finds parameter estimates that minimize the Huber loss function. The Huber loss function penalizes outliers and increases as a quadratic for small errors and linearly for large errors. For more information about robust fitting, see Huber (1973) and Huber and Ronchetti (2009). See "Example of the Robust Fit Option". For more information about the Huber loss function, see "Statistical Details for the Robust Fit".
- The Cauchy Fit option assumes that the errors have a Cauchy distribution. A Cauchy distribution has fatter tails than the normal distribution, resulting in a reduced emphasis on outliers. This option can be useful if you have a large proportion of outliers in your data. However, if your data are close to normal with only a few outliers, this option can lead to incorrect inferences. The Cauchy option estimates parameters using maximum likelihood and a Cauchy link function.

Note: The Robust options are not available when a Block variable is specified in the launch window.

There are two tables in the reports that are created for the Robust options. The tables contain the following columns:

Sigma The sigma value, which is equivalent to the root mean square error (RMSE).

ChiSquare The test statistic for the hypothesis that the model fits better than the mean of the response.

PValue The *p*-value for the chi-square test that the slope is equal to zero.

Logworth he logworth is a transformation of the *p*-value defined as $-\log_{10}(p\text{-value})$.

Level The levels of the X variable.

Robust Mean The robust mean estimated based on either the Huber or the Cauchy method.

Std Error The estimates of the standard errors of the parameter estimates.

Power Reports

In the Oneway platform, the Power option calculates statistical power and other details about a given hypothesis test. See "Example of the Power Option". For statistical details, see "Statistical Details for Power".

- *LSV* (the Least Significant Value) is the value of some parameter or function of parameters that would produce a certain *p*-value alpha. Said another way, you want to know how small an effect would be declared significant at some *p*-value alpha. The LSV provides a measuring stick for significance on the scale of the parameter, rather than on a probability scale. It shows how sensitive the design and data are.
- *LSN* (the Least Significant Number) is the total number of observations that would produce a specified *p*-value alpha given that the data has the same form. The LSN is defined as the number of observations needed to reduce the variance of the estimates enough to achieve a significant result with the given values of alpha, sigma, and delta (the significance level, the standard deviation of the error, and the effect size). If you need more data to achieve significance, the LSN helps tell you how many more. The LSN is the total number of observations that yields approximately 50% power.
- *Power* is the probability of getting significance (p-value < alpha) when a real difference exists between groups. It is a function of the sample size, the effect size, the standard deviation of the error, and the significance level. The power tells you how likely your experiment is to detect a difference (effect size), at a given alpha level.

Note: When there are only two groups in a one-way layout, the LSV computed by the power facility is the same as the least significant difference (LSD) shown in the multiple-comparison tables.

Power Details

The Power Details window and tables in the Oneway platform are the same as those in the Fit Model platform. For more information about power calculation, see *Fitting Linear Models*.

For each of four columns Alpha, Sigma, Delta, and Number, fill in a single value, two values, or the start, stop, and increment for a sequence of values (Figure 6.29). Power calculations are performed on all possible combinations of the values that you specify.

Alpha (α) Significance level, between 0 and 1 (usually 0.05, 0.01, or 0.10). Initially, a value of 0.05 shows.

Sigma (σ) Standard error of the residual error in the model. Initially, RMSE, the estimate from the square root of the mean square error is supplied here.

Delta (δ) Raw effect size. For more information about effect size computations, see *Fitting Linear Models*. The first position is initially set to the square root of the sums of squares for the hypothesis divided by n (that is, $\delta = \sqrt{SS/n}$).

Number (n) Total sample size across all groups. Initially, the actual sample size is put in the first position.

Solve for Power Solves for the power (the probability of a significant result) as a function of all four values: α, σ, δ, and n.

Solve for Least Significant Number Solves for the number of observations needed to achieve approximately 50% power given α, σ, and δ.

Solve for Least Significant Value Solves for the value of the parameter or linear test that produces a p-value of α. This is a function of α, σ, n, and the standard error of the estimate. This feature is available only when the X variable has exactly two levels and is usually used for individual parameters.

Adjusted Power and Confidence Interval When you look at power retrospectively, you use estimates of the standard error and the test parameters.

- Adjusted power is the power calculated from a more unbiased estimate of the non-centrality parameter.
- The confidence interval for the adjusted power is based on the confidence interval for the non-centrality estimate.

Adjusted power and confidence limits are computed only for the original Delta, because that is where the random variation is.

Matching Column Report

In the Oneway platform, use the Matching Column option to specify a matching (ID) variable for a matching model analysis. The Matching Column option addresses the case when the data in a one-way analysis come from matched (paired) data. Matched data can occur when observations in different groups come from the same participant. The Matching Column option is not available when a Block variable is specified in the launch window. See “Example of the Matching Column Option”.

Note: A special case of matching leads to the paired *t* test. The **Matched Pairs** platform handles this type of data, but the data must be organized with the pairs in different columns, not in different rows.

The **Matching Column** option performs two primary actions:

- It fits an additive model (using an iterative proportional fitting algorithm) that includes both the grouping variable (the X variable in the Fit Y by X analysis) and the matching variable that you select. The iterative proportional fitting algorithm makes a difference if there are hundreds of participants, because the equivalent linear model would be very slow and would require huge memory resources.
- It adds lines between the points that match across the groups to the Oneway plot. If there are multiple observations with the same matching ID value, lines connect the group means. To remove the lines from the Oneway plot, select **Display Options > Matching Lines**.

The Matching Fit report shows the effects with *F* tests. These are equivalent to the tests that you get with the Fit Model platform if you run two models, one with the interaction term and one without. If there are only two levels, then the *F* test is equivalent to the paired *t* test.

Note: For more information about the Fit Model platform, see *Fitting Linear Models*.

Oneway Plot Elements

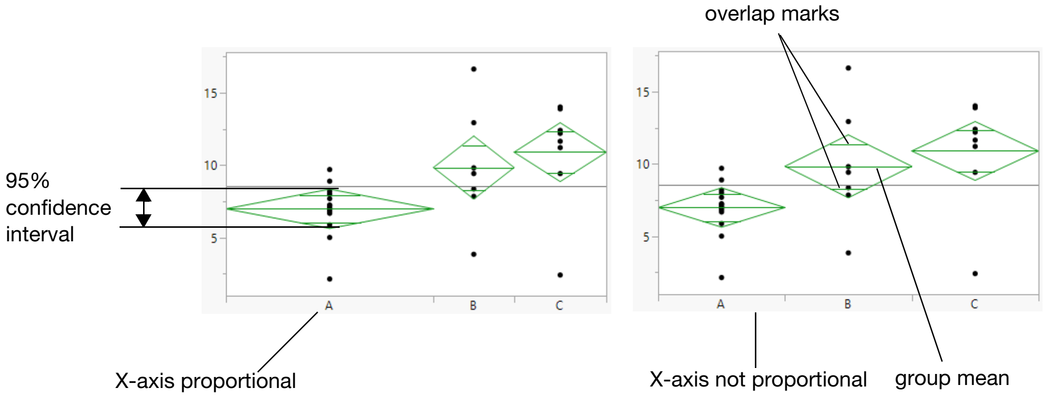
The Oneway platform includes a plot of a continuous Y variable versus a categorical X variable. This plot enables you to visualize the distribution of the Y variable for each level of the X variable. Many of the Oneway platform options add visualization elements to the plot. The Oneway red triangle menu contains a Display Options submenu that enables you to add or remove the display elements. This section contains details for the following display elements:

- "Mean Diamonds and X-Axis Proportional"
- "Mean Lines, Error Bars, and Standard Deviation Lines"
- "Comparison Circles"

Mean Diamonds and X-Axis Proportional

In the Oneway platform, the horizontal axis defaults to a proportional axis. That is, the spacing of the group levels is proportional to the number of data points in each group. The mean diamond display options, or the ANOVA test, adds mean diamonds to the Oneway plot. A mean diamond shows the sample mean and confidence interval for a level of the X variable.

Figure 6.6 Mean Diamonds and X-Axis Proportional Options



Mean Diamond Details

- The top and bottom of each diamond span the (1-alpha)x100 confidence interval for the mean of each group. The confidence interval computation assumes that the variances are equal across observations. Therefore, the height of the diamond is proportional to the reciprocal of the square root of the number of observations in the group.
- If the **X-Axis proportional** option is selected, the horizontal extent of each group along the horizontal axis (the horizontal size of the diamond) is proportional to the sample size for

each level of the X variable. Therefore, the narrower diamonds are usually taller, because fewer data points results in a wider confidence interval estimates.

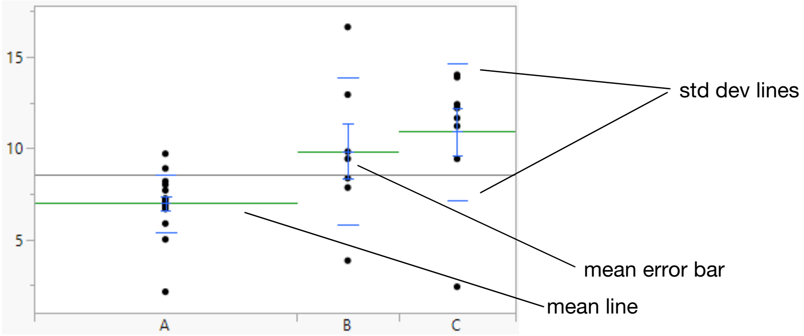
- The mean line across the middle of each diamond is drawn at the group mean.
- The lines near the top and bottom of the diamonds are overlap marks. For groups with equal sample sizes, diamonds that overlap within the overlap marks indicate that the two group means are not significantly different at the given confidence level. Overlap marks are computed as group mean ± $(\sqrt{2})/2 \times \mathrm{CI}/2$.
- The mean diamonds are added to the Oneway plot when you select the **Means/Anova/Pooled t** or **Means/Anova** option from the platform menu. However, you can show or hide them at any time by selecting **Display Options > Mean Diamonds** from the red triangle menu.

Mean Lines, Error Bars, and Standard Deviation Lines

In the Oneway platform, select the **Display Options > Mean Lines** option to show mean lines on the Oneway plot. Mean lines are drawn at the mean of the response for each level of the X variable.

Select the **Means and Std Dev** option from the red triangle menu to add mean error bars and standard deviation lines to the plot. To turn each option on or off singly, select **Display Options > Mean Error Bars** or **Std Dev Lines**.

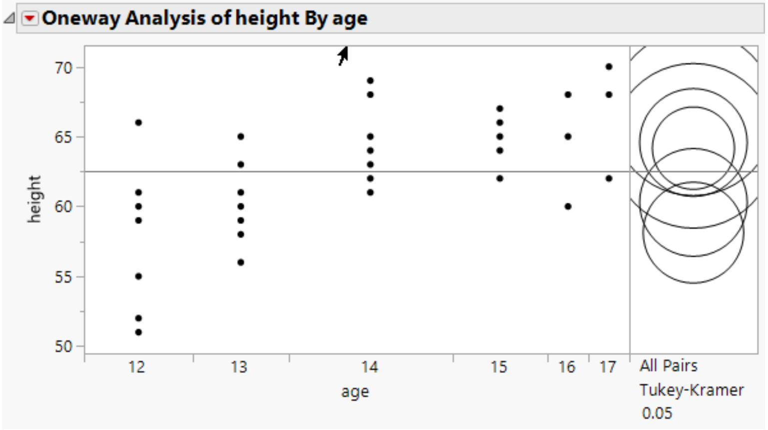
Figure 6.7 Mean Lines, Mean Error Bars, and Std Dev Lines



Comparison Circles

In the Oneway platform, each multiple comparison test, except for the Each Pair Stepwise, Newman-Keuls method, adds *comparison circles* to the Oneway plot. Comparison circles provide an interactive visual representation of group mean comparisons. Figure 6.8 shows the comparison circles for the All Pairs, Tukey HSD method. The radii of the circles are based on the test and significance level. See “Statistical Details for Comparison Circles”.

Figure 6.8 Visual Comparison of Group Means



Compare each pair of group means visually by examining the intersection of the comparison circles. The outside angle of intersection indicates whether the group means are significantly different (Figure 6.9).

- Circles for means that are significantly different either do not intersect, or intersect slightly, so that the outside angle of intersection is less than 90 degrees.
- If the circles intersect by an angle of more than 90 degrees, or if they are nested, the corresponding group means are not significantly different.
- Click a circle to highlight the corresponding group. Circles that are gray correspond to groups with means that are significantly different from the highlighted group (Figure 6.10). To deselect circles, click in the white space outside the circles.

Figure 6.9 Angles of Intersection and Significance

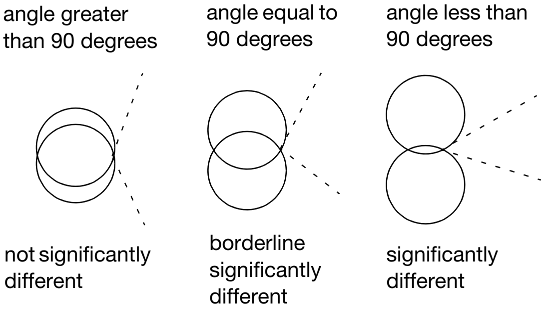
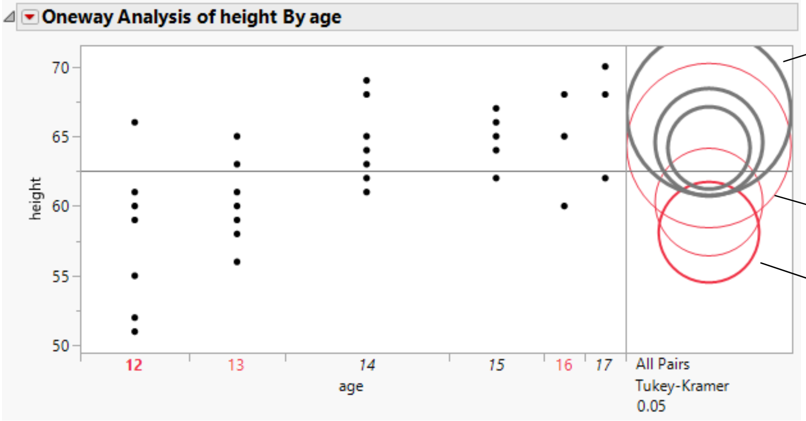


Figure 6.10 Highlighting Comparison Circles



Groups that are different from the selected group appear as thick gray circles.

Groups that are not different from the selected group appear as thin red circles.

The selected group appears as a thick red circle.

Additional Examples of the Oneway Platform

This section contains examples using the Oneway platform.

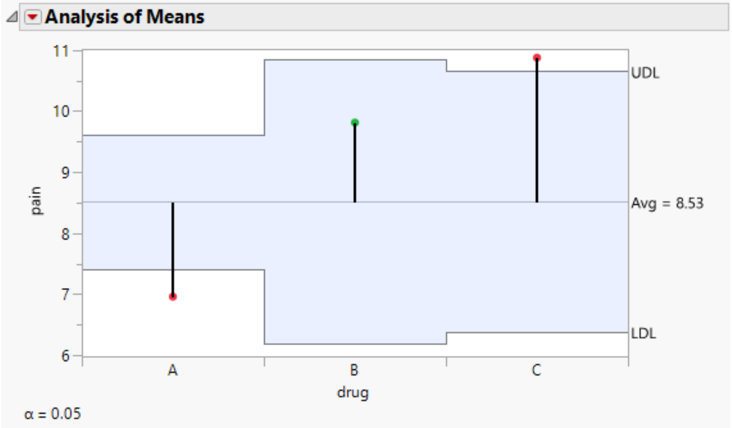
- “Example of Analysis of Means”
- “Example of Analysis of Means for Variances”
- “Example of the Each Pair, Student’s t Test”
- “Example of the All Pairs, Tukey HSD Test”
- “Example of the With Best, Hsu MCB Test”
- “Example of the With Control, Dunnett’s Test”
- “Example of the Each Pair Stepwise, Newman-Keuls Test”
- “Example Contrasting Four Compare Means Tests”
- “Example of the Nonparametric Wilcoxon Test”
- “Example of the Unequal Variances Option”
- “Example of an Equivalence Test”
- “Example of the Robust Fit Option”
- “Example of the Power Option”
- “Example of a Normal Quantile Plot”
- “Example of a CDF Plot”
- “Example of the Densities Options”
- “Example of the Matching Column Option”
- “Example of Stacking Data for a Oneway Analysis”

Example of Analysis of Means

Use the Oneway platform to test for differences in means using the multiple comparison analysis of means (ANOM) method.

1. Select **Help > Sample Data Folder** and open Analgesics.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select pain and click **Y, Response**.
4. Select drug and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle and select **Analysis of Means Methods > ANOM**.

Figure 6.11 Analysis of Means Decision Chart



From the ANOM decision chart, notice that the means for drug A and C are statistically different from the overall mean. The drug A mean is lower and the drug C mean is higher. Note the decision limits for the drug types are not the same, due to different sample sizes.

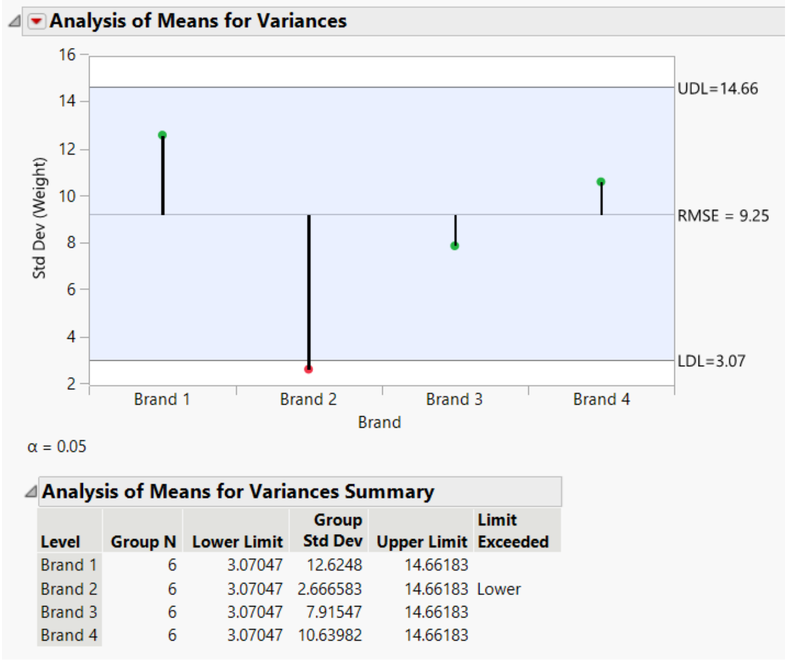
Example of Analysis of Means for Variances

Use the Oneway platform to test for differences in variances using analysis of means for variances (ANOMV). Four different brands of springs were tested to see what weight is required to extend a spring 0.10 inches. Six springs of each brand were tested. The data were checked for normality, since the ANOMV test is not robust to non-normality. Examine the brands to determine whether the variability is significantly different between brands.

1. Select **Help > Sample Data Folder** and open Spring Data.jmp.
2. Select **Analyze > Fit Y by X**.

3. Select Weight and click **Y, Response**.
4. Select Brand and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Analysis of Means Methods > ANOM for Variances**.
7. Click the Analysis of Means for Variances red triangle menu and select **Show Summary Report**.

Figure 6.12 Analysis of Means for Variances Chart



Analysis of Means for Variances

α = 0.05

Analysis of Means for Variances Summary

Level	Group N	Lower Limit	Group Std Dev	Upper Limit	Limit Exceeded
Brand 1	6	3.07047	12.6248	14.66183	
Brand 2	6	3.07047	2.666583	14.66183	Lower
Brand 3	6	3.07047	7.91547	14.66183	
Brand 4	6	3.07047	10.63982	14.66183	

Note that the standard deviation for Brand 2 exceeds the lower decision limit. Therefore, Brand 2 has significantly lower variance than the other brands.

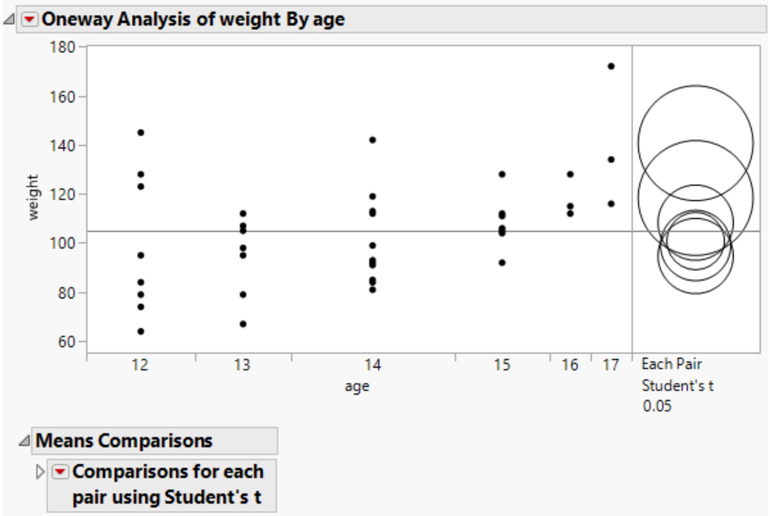
Example of the Each Pair, Student's t Test

This example uses the Oneway platform to illustrate the use of all possible *t* tests.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.

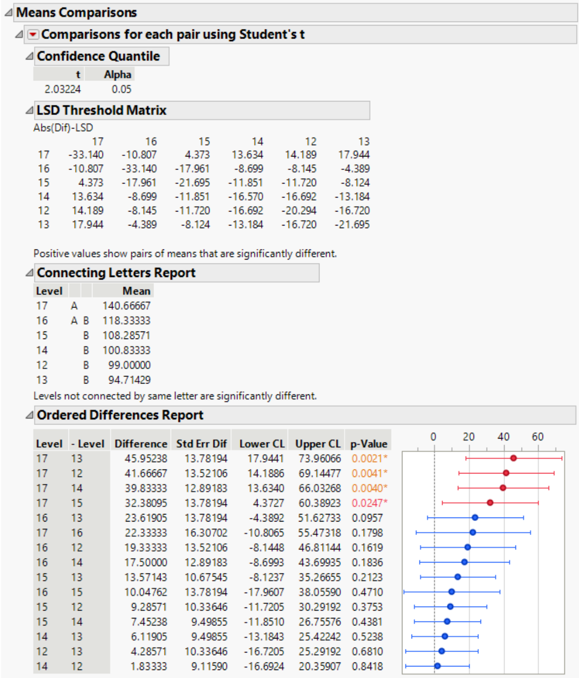
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Compare Means > Each Pair, Student’s t**.

Figure 6.13 Example of Each Pair, Student’s t Comparison Circles



The means comparison method uses Fisher’s least significant difference (LSD) method to evaluate differences between individual pairs of means. The method does not maintain the error rate for simultaneous comparisons. For a large number of groups consider other comparison methods.

Figure 6.14 Example of Means Comparisons Report for Each Pair, Student's *t*



Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
2.03224	0.05

LSD Threshold Matrix

Abs(Dif)-LSD

	17	16	15	14	12	13
17	-33.140	-10.807	4.373	13.634	14.189	17.944
16	-10.807	-33.140	-17.961	-8.699	-8.145	-4.389
15	4.373	-17.961	-21.695	-11.851	-11.720	-8.124
14	13.634	-8.699	-11.851	-16.570	-16.692	-13.184
12	14.189	-8.145	-11.720	-16.692	-20.294	-16.720
13	17.944	-4.389	-8.124	-13.184	-16.720	-21.695

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level			Mean
17	A		140.66667
16	A	B	118.33333
15		B	108.28571
14		B	100.83333
12		B	99.00000
13		B	94.71429

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
17	13	45.95238	13.78194	17.9441	73.96066	0.0021*
17	12	41.66667	13.52106	14.1886	69.14477	0.0041*
17	14	39.83333	12.89183	13.6340	66.03268	0.0040*
17	15	32.38095	13.78194	4.3727	60.38923	0.0247*
16	13	23.61905	13.78194	-4.3892	51.62733	0.0957
17	16	22.33333	16.30702	-10.8065	55.47318	0.1798
16	12	19.33333	13.52106	-8.1448	46.81144	0.1619
16	14	17.50000	12.89183	-8.6993	43.69935	0.1836
15	13	13.57143	10.67545	-8.1237	35.26655	0.2123
16	15	10.04762	13.78194	-17.9607	38.05590	0.4710
15	12	9.28571	10.33646	-11.7205	30.29192	0.3753
15	14	7.45238	9.49855	-11.8510	26.75576	0.4381
14	13	6.11905	9.49855	-13.1843	25.42242	0.5238
12	13	4.28571	10.33646	-16.7205	25.29192	0.6810
14	12	1.83333	9.11590	-16.6924	20.35907	0.8418

In Figure 6.14, the LSD threshold table shows the difference between the absolute difference in the means and the LSD (least significant difference). If the values are positive, the difference in the two means is larger than the LSD, and the two groups are statistically significantly different.

Example of the All Pairs, Tukey HSD Test

This example uses the Oneway platform to illustrate the Tukey HSD test.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Compare Means > All Pairs, Tukey HSD**.

Figure 6.15 All Pairs, Tukey HSD Comparison Circles

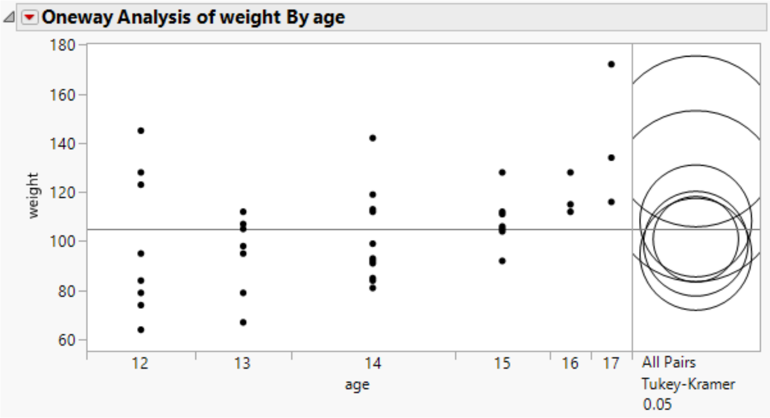
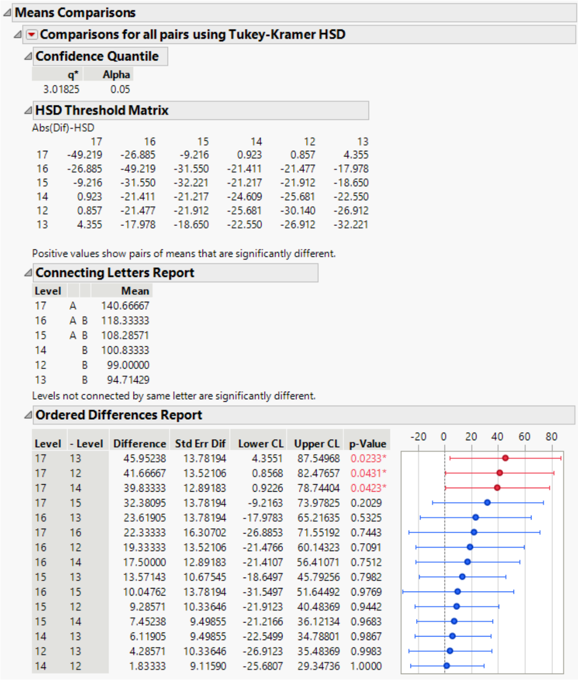


Figure 6.16 Means Comparisons Report for All Pairs, Tukey HSD



Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
3.01825	0.05

HSD Threshold Matrix

Abs(Dif)-HSD

	17	16	15	14	12	13
17	-49.219	-26.885	-9.216	0.923	0.857	4.355
16	-26.885	-49.219	-31.550	-21.411	-21.477	-17.978
15	-9.216	-31.550	-32.221	-21.217	-21.912	-18.650
14	0.923	-21.411	-21.217	-24.609	-25.681	-22.550
12	0.857	-21.477	-21.912	-25.681	-30.140	-26.912
13	4.355	-17.978	-18.650	-22.550	-26.912	-32.221

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level			Mean
17	A		140.66667
16	A	B	118.33333
15	A	B	108.28571
14		B	100.83333
12		B	99.00000
13		B	94.71429

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
17	13	45.95238	13.78194	4.3551	87.54968	0.0233*
17	12	41.66667	13.52106	0.8568	82.47657	0.0431*
17	14	39.83333	12.89183	0.9226	78.74404	0.0423*
17	15	32.38095	13.78194	-9.2163	73.97825	0.2029
16	13	23.61905	13.78194	-17.9783	65.21635	0.5325
17	16	22.33333	16.30702	-26.8853	71.55192	0.7443
16	12	19.33333	13.52106	-21.4766	60.14323	0.7091
16	14	17.50000	12.89183	-21.4107	56.41071	0.7512
15	13	13.57143	10.67545	-18.6497	45.79256	0.7982
16	15	10.04762	13.78194	-31.5497	51.64492	0.9769
15	12	9.28571	10.33646	-21.9123	40.48369	0.9442
15	14	7.45238	9.49855	-21.2166	36.12134	0.9683
14	13	6.11905	9.49855	-22.5499	34.78801	0.9867
12	13	4.28571	10.33646	-26.9123	35.48369	0.9983
14	12	1.83333	9.11590	-25.6807	29.34736	1.0000

In Figure 6.16, the Tukey-Kramer HSD Threshold matrix shows the actual absolute difference in the means minus the HSD. Pairs with a positive value are statistically significantly different at the specified alpha significance level. The q^* value in the confidence quantile table represents the quantile that is used to scale the HSDs. It has a computational role comparable to the quantile in a Student's t test.

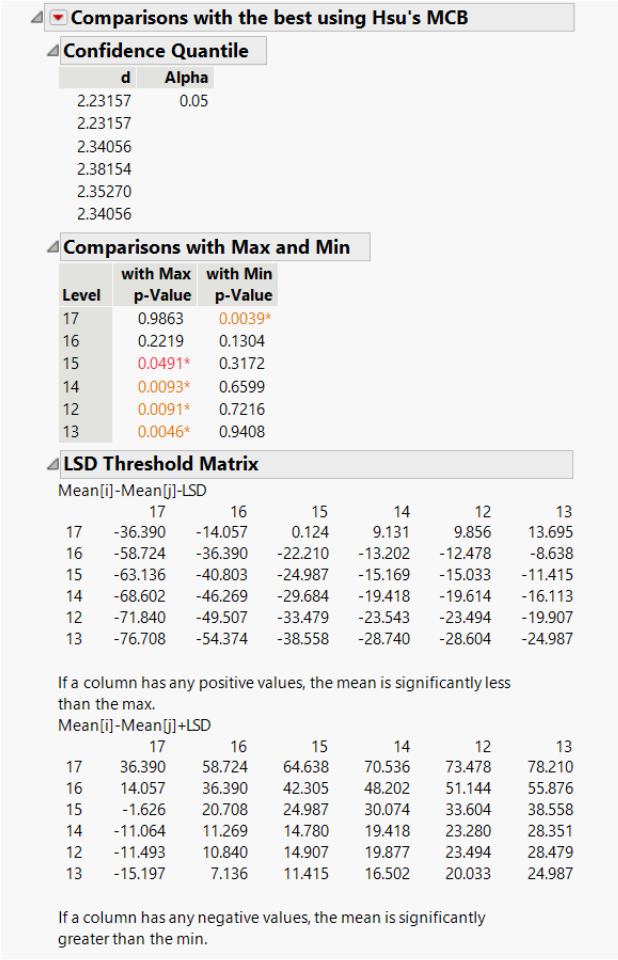
Example of the With Best, Hsu MCB Test

This example uses the Oneway platform to illustrate the Hsu multiple comparison with the best (MCB) test.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Compare Means > With Best, Hsu MCB**.

Figure 6.17 Examples of With Best, Hsu MCB Comparison Circles

Figure 6.18 Example of Means Comparisons Report for With Best, Hsu MCB



Comparisons with the best using Hsu's MCB

Confidence Quantile

d	Alpha
2.23157	0.05
2.23157	
2.34056	
2.38154	
2.35270	
2.34056	

Comparisons with Max and Min

Level	with Max p-Value	with Min p-Value
17	0.9863	0.0039*
16	0.2219	0.1304
15	0.0491*	0.3172
14	0.0093*	0.6599
12	0.0091*	0.7216
13	0.0046*	0.9408

LSD Threshold Matrix

Mean[i]-Mean[j]-LSD

	17	16	15	14	12	13
17	-36.390	-14.057	0.124	9.131	9.856	13.695
16	-58.724	-36.390	-22.210	-13.202	-12.478	-8.638
15	-63.136	-40.803	-24.987	-15.169	-15.033	-11.415
14	-68.602	-46.269	-29.684	-19.418	-19.614	-16.113
12	-71.840	-49.507	-33.479	-23.543	-23.494	-19.907
13	-76.708	-54.374	-38.558	-28.740	-28.604	-24.987

If a column has any positive values, the mean is significantly less than the max.

Mean[i]-Mean[j]+LSD

	17	16	15	14	12	13
17	36.390	58.724	64.638	70.536	73.478	78.210
16	14.057	36.390	42.305	48.202	51.144	55.876
15	-1.626	20.708	24.987	30.074	33.604	38.558
14	-11.064	11.269	14.780	19.418	23.280	28.351
12	-11.493	10.840	14.907	19.877	23.494	28.479
13	-15.197	7.136	11.415	16.502	20.033	24.987

If a column has any negative values, the mean is significantly greater than the min.

The Comparison with Max and Min report compares the mean of each level to the maximum and the minimum of the means of the remaining levels. For example, the mean for age 15 differs significantly from the maximum of the means of the remaining levels. The mean for age 17 differs significantly from the minimum of the means of the remaining levels. The maximum mean could occur for age 16 or age 17, because neither mean differs significantly from the maximum mean. By the same reasoning, the minimum mean could correspond to any of the ages other than age 17.

Example of the With Control, Dunnett's Test

This example uses the Oneway platform to illustrate Dunnett's test with a control group.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.

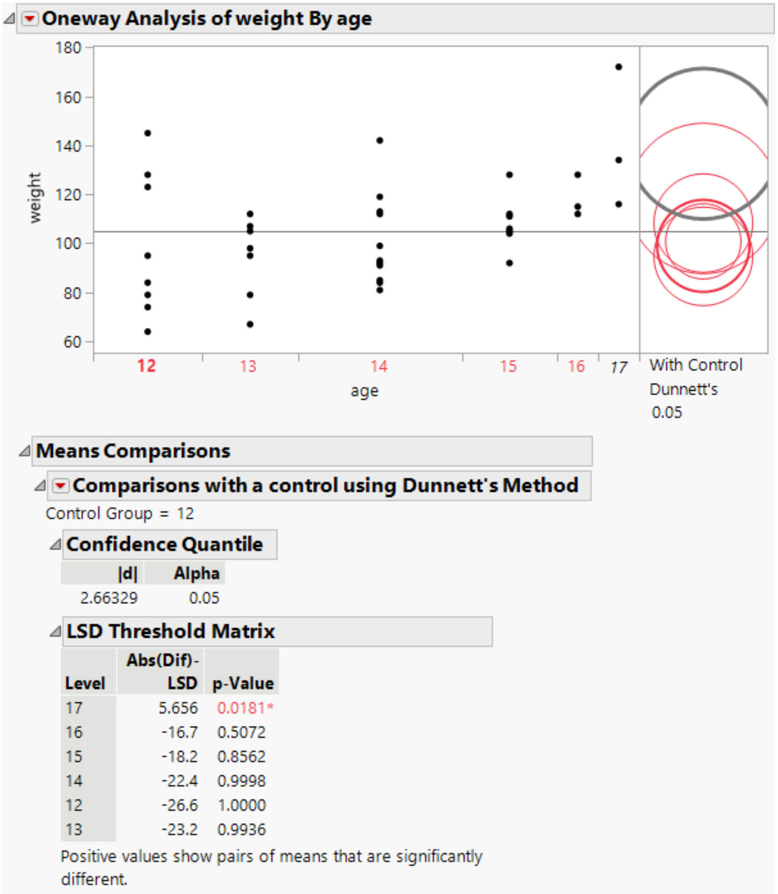
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Compare Means > With Control, Dunnett's**.
7. Select 12 as the age to use as the control group.

Tip: Click a point in the control group to highlight it in the scatterplot before selecting the **Compare Means > With Control, Dunnett's** option. The test uses the selected group as the control group.

---

8. Click **OK**.

Figure 6.19 Example of With Control, Dunnett's Comparison Circles



Oneway Analysis of weight By age

Means Comparisons

Comparisons with a control using Dunnett's Method

Control Group = 12

Confidence Quantile

|d|	Alpha
2.66329	0.05

LSD Threshold Matrix

Level	Abs(Dif)-LSD	p-Value
17	5.656	0.0181*
16	-16.7	0.5072
15	-18.2	0.8562
14	-22.4	0.9998
12	-26.6	1.0000
13	-23.2	0.9936

Positive values show pairs of means that are significantly different.

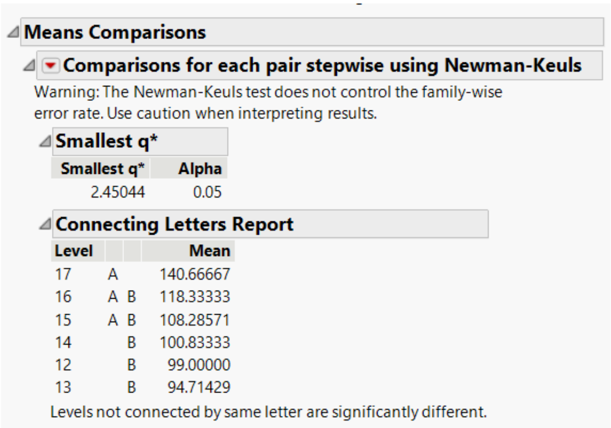
Using the comparison circles or the results in the LSD Threshold Matrix, you can conclude that level 17 is the only level that is significantly different from the control level of 12.

Example of the Each Pair Stepwise, Newman-Keuls Test

This example uses the Oneway platform to illustrate the Newman-Keuls test.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response**.
4. Select age and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Compare Means > Each Pair Stepwise, Newman-Keuls**.

Figure 6.20 Example of Means Comparisons Report for Each Pair Stepwise, Newman-Keuls



Means Comparisons

Comparisons for each pair stepwise using Newman-Keuls

Warning: The Newman-Keuls test does not control the family-wise error rate. Use caution when interpreting results.

Smallest q*

Smallest q*	Alpha
2.45044	0.05

Connecting Letters Report

Level			Mean
17	A		140.66667
16	A	B	118.33333
15	A	B	108.28571
14		B	100.83333
12		B	99.00000
13		B	94.71429

Levels not connected by same letter are significantly different.

The Connecting Letters Report shows that Level 17 is significantly different from all other levels except 16 and 15.

Example Contrasting Four Compare Means Tests

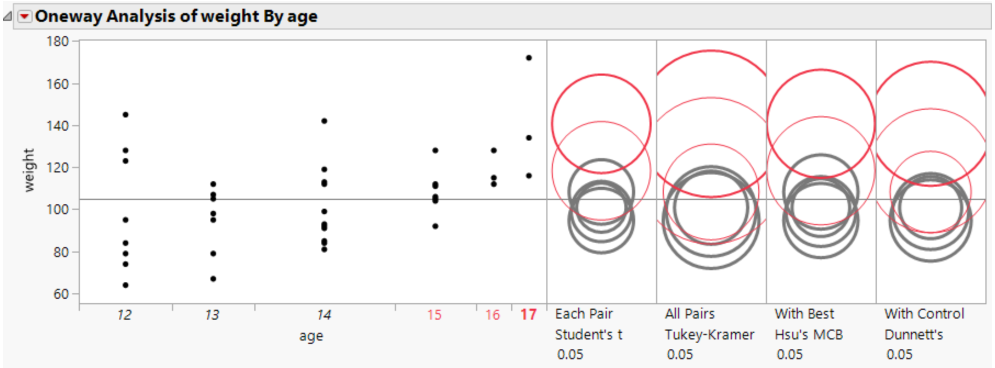
This example uses the Oneway test to compare four different means comparison tests.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.

5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select each one of the **Compare Means** options. For the With Control, Dunnett's option, select age 17 as the control group.

The four methods all test differences between group means. Each test is used for a specific hypothesis and different findings can occur.

Figure 6.21 Comparison Circles for Four Multiple Comparison Tests



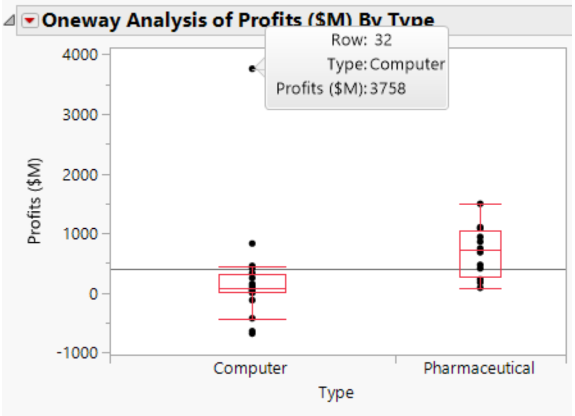
In Figure 6.21, age group 17 is highlighted. The other control circles are colored in relation to age group 17. Notice that for the Student's t and Hsu methods, age group 15 (the third circle from the top) is gray. This indicates that it is significantly different from age group 17. However, for the Tukey and Dunnett methods, age group 15 is red, which indicates that it is not significantly different from age group 17.

Example of the Nonparametric Wilcoxon Test

In the Oneway platform, use a Wilcoxon test to determine whether the mean profit earned by companies differs by type of company. The data consist of various metrics on two types of companies, Pharmaceutical (12 companies) and Computer (20 companies).

1. Select **Help > Sample Data Folder** and open Companies.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Profits ($M) and click **Y, Response.**
4. Select Type and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle and select **Display Options > Box Plots**.

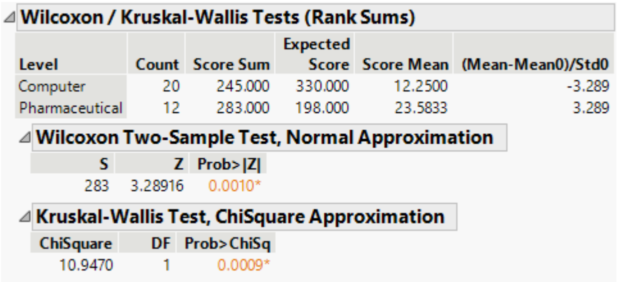
Figure 6.22 Computer Company Profit Distribution



The box plots suggest that the distributions of profits within groups are not normal or symmetric. There is a very large value for the company in row 32 that might affect parametric tests.

7. Click the Oneway Analysis red triangle menu and select **Nonparametric > Wilcoxon / Kruskal-Wallis Test**.

Figure 6.23 Wilcoxon Test Results



Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Computer	20	245.000	330.000	12.2500	-3.289
Pharmaceutical	12	283.000	198.000	23.5833	3.289

Wilcoxon Two-Sample Test, Normal Approximation

S	Z	Prob>|Z|
283	3.28916	0.0010*

Kruskal-Wallis Test, ChiSquare Approximation

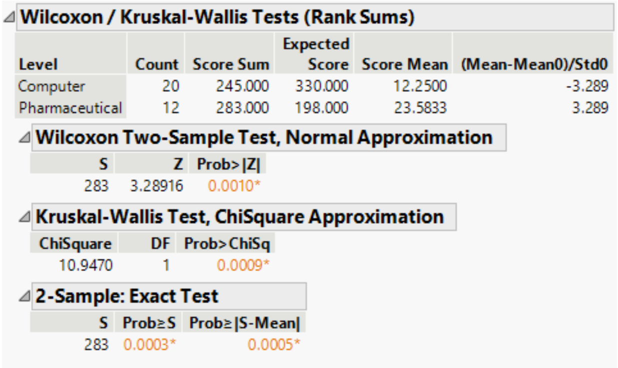
ChiSquare	DF	Prob>ChiSq
10.9470	1	0.0009*

Both the normal approximation that uses a 0.5 continuity correction and the chi-square approximation for the Wilcoxon test statistic indicate significance at a *p*-value of 0.0010. You conclude that there is a significant difference in the location of the distributions, and conclude that mean profit differs based on company type.

The normal and chi-square tests are based on the asymptotic distributions of the test statistics. You can also conduct an exact test.

8. Click the Oneway Analysis red triangle menu and select **Nonparametric > Exact Test > Wilcoxon Exact Test**.

Figure 6.24 Wilcoxon Exact Test Results



Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Computer	20	245.000	330.000	12.2500	-3.289
Pharmaceutical	12	283.000	198.000	23.5833	3.289

Wilcoxon Two-Sample Test, Normal Approximation

S	Z	Prob>|Z|
283	3.28916	0.0010*

Kruskal-Wallis Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
10.9470	1	0.0009*

2-Sample: Exact Test

S	Prob≥S	Prob≥|S-Mean|
283	0.0003*	0.0005*

The observed value of the test statistic is S = 283. This is the sum of the ranks for the level of Type with the smaller sample size (pharmaceuticals). The probability of observing an absolute difference from the mean midrank that exceeds the absolute value of S minus the mean of the midranks is 0.0005. This is a two-sided test for a difference in location and supports rejecting the hypothesis that profits do not differ by type of company.

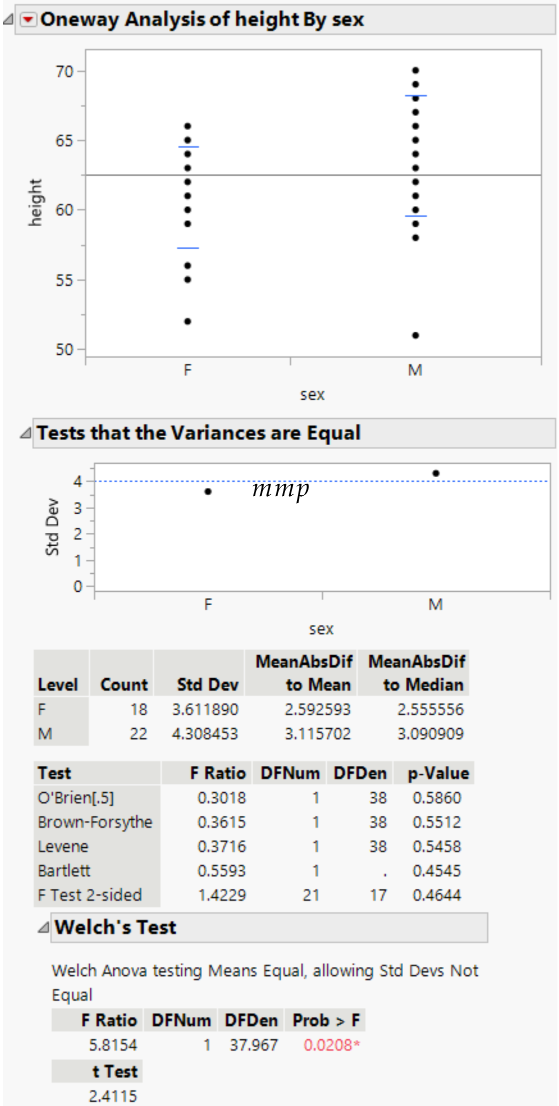
In this example, the nonparametric tests are more appropriate than the normality-based ANOVA test or the unequal variances *t* test. The nonparametric tests are resistant to the large value in row 32 and do not require the assumption of normality.

Example of the Unequal Variances Option

Use the Oneway platform to test whether two groups have equal variance.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select height and click **Y, Response.**
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Unequal Variances**.

Figure 6.25 Example of the Unequal Variances Report



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
F	18	3.611890	2.592593	2.555556
M	22	4.308453	3.115702	3.090909

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.3018	1	38	0.5860
Brown-Forsythe	0.3615	1	38	0.5512
Levene	0.3716	1	38	0.5458
Bartlett	0.5593	1	.	0.4545
F Test 2-sided	1.4229	21	17	0.4644

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
5.8154	1	37.967	0.0208*

t Test
2.4115

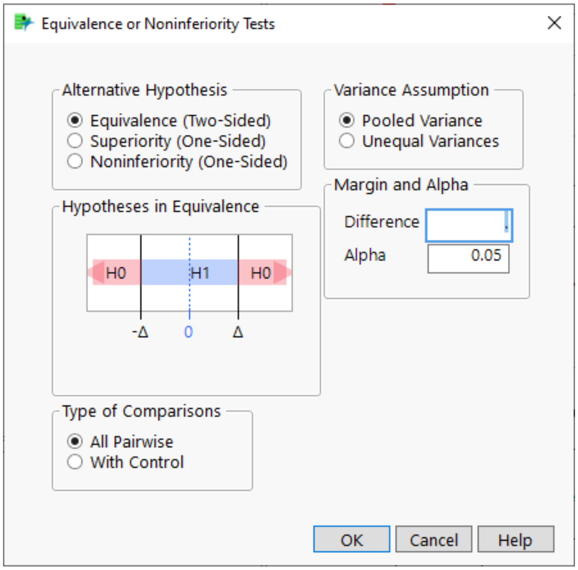
Since the *p*-value from the 2-sided *F* Test is large, you cannot conclude that the variances are unequal.

Example of an Equivalence Test

Use the Oneway platform to investigate whether an alternative drug is not inferior to morphine for a procedure in dogs. We compare the difference in the first two time points.

1. Select **Help > Sample Data Folder** and open Dogs.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select diff and click **Y, Response.**
4. Select drug and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Equivalence Tests > Means**.

Figure 6.26 Equivalence or Noninferiority Tests Dialog


Equivalence or Noninferiority Tests

Alternative Hypothesis
● Equivalence (Two-Sided)
○ Superiority (One-Sided)
○ Noninferiority (One-Sided)

Variance Assumption
● Pooled Variance
○ Unequal Variances

Margin and Alpha
Difference
Alpha 0.05

Hypotheses in Equivalence

Type of Comparisons
● All Pairwise
○ With Control

7. Select **Noninferiority (One-Sided)**.
8. Leave the variance assumption set to **Pooled Variance**.
9. Set the **Difference** to 1. This is the noninferiority margin.
10. Click **OK**.

Figure 6.27 Example of a Test for Noninferiority

Oneway Analysis of diff By drug

Noninferiority Tests with Pooled Variance

Test	Alternative Hypothesis
Noninferiority	μT - μC > -1

Noninferiority Tests

drug	-drug	Difference	Std Error of Difference	Lower Bound t Ratio	Lower Bound p-Value	Lower 95%	Upper 95%	Assessment (α=0.05)
morphine	trimeth	-0.639474	0.7351705	0.4903973	0.3157	-1.93434	0.6553888	Inferior

Forest Plot

The observed difference in the response for the two drugs was -0.639 units. The lower bound on the confidence interval of -1.93 falls below the lower noninferiority limit of -1.0. We do not conclude that the alternative drug is not inferior to the standard drug.

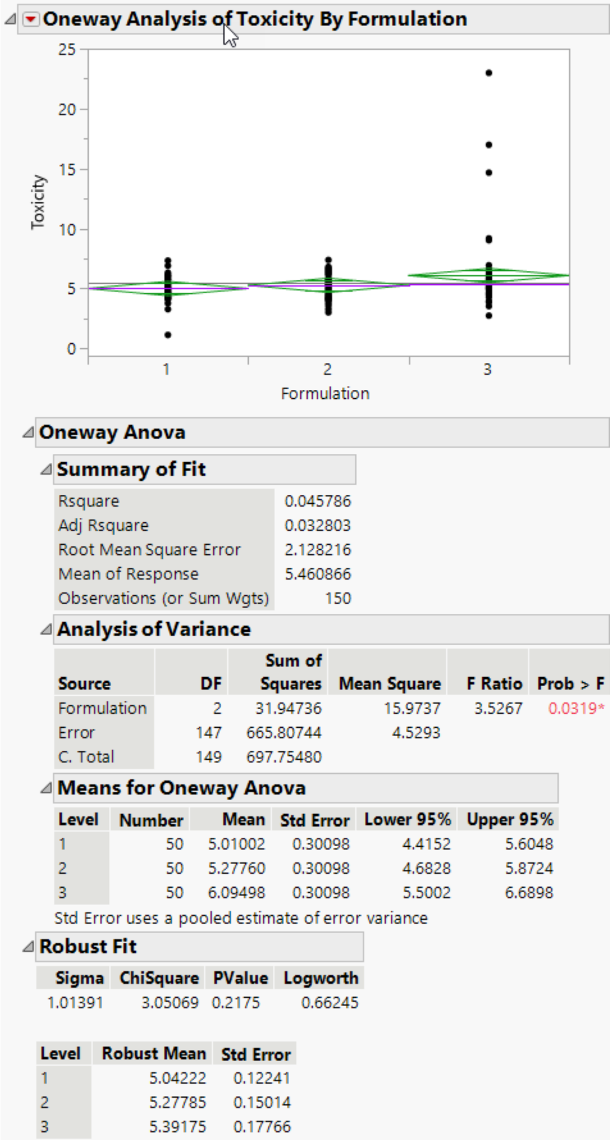
Example of the Robust Fit Option

Use the Oneway platform with a robust fit in this example where one of three groups contains outliers. The data contain the toxicity levels for three different formulations of a drug.

1. Select **Help > Sample Data Folder** and open Drug Toxicity.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Toxicity and click **Y, Response**.
4. Select Formulation and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Means/Anova**.

7. Click the Oneway Analysis red triangle menu and select **Robust > Robust Fit**.

Figure 6.28 Example of Robust Fit



Oneway Analysis of Toxicity By Formulation

Oneway Anova

Summary of Fit

Rsquare	0.045786
Adj Rsquare	0.032803
Root Mean Square Error	2.128216
Mean of Response	5.460866
Observations (or Sum Wgts)	150

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Formulation	2	31.94736	15.9737	3.5267	0.0319*
Error	147	665.80744	4.5293		
C. Total	149	697.75480			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	50	5.01002	0.30098	4.4152	5.6048
2	50	5.27760	0.30098	4.6828	5.8724
3	50	6.09498	0.30098	5.5002	6.6898

Std Error uses a pooled estimate of error variance

Robust Fit

Sigma	ChiSquare	PValue	Logworth
1.01391	3.05069	0.2175	0.66245

Level	Robust Mean	Std Error
1	5.04222	0.12241
2	5.27785	0.15014
3	5.39175	0.17766

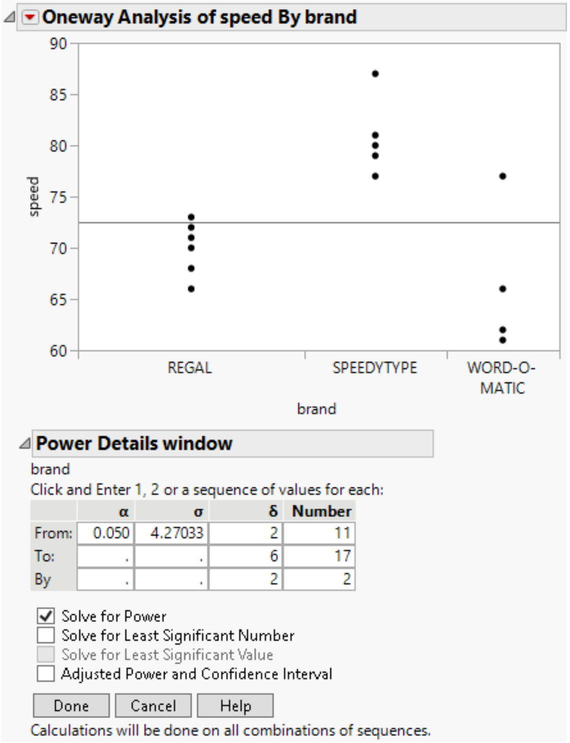
If you look at the standard Analysis of Variance report, you might conclude that there is a difference between the three formulations, since the *p*-value is 0.0319. However, when you look at the Robust Fit report, you would not conclude that the three formulations are significantly different, because the *p*-value there is 0.2175. It appears that the toxicity for a few of the observations is unusually high, creating the undue influence on the data.

Example of the Power Option

This example illustrates the Power option in the Oneway platform.

1. Select **Help > Sample Data Folder** and open Typing Data.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select speed and click **Y, Response.**
4. Select brand and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Power**.
7. Within the From row, type 2 for Delta (the third box) and type 11 for Number.
8. Within the To row, type 6 for Delta, and type 17 in the Number box.
9. Within the By row, type 2 for both Delta and Number.
10. Select the **Solve for Power** check box.

Figure 6.29 Example of the Power Details Window



brand

Click and Enter 1, 2 or a sequence of values for each:

	α	σ	δ	Number
From:	0.050	4.27033	2	11
To:	.	.	6	17
By	.	.	2	2

☑ Solve for Power
☐ Solve for Least Significant Number
☐ Solve for Least Significant Value
☐ Adjusted Power and Confidence Interval

Calculations will be done on all combinations of sequences.

11. Click **Done**.

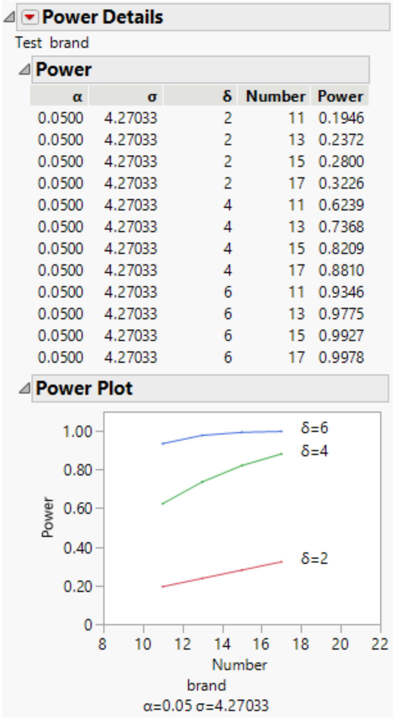
Note: The **Done** button remains dimmed until all of the necessary options are applied.

Power is computed for each combination of Delta and Number, and appears in the Power report.

To plot the Power values:

12. Click the red triangle below the table and select **Power Plot**.

Figure 6.30 Example of the Power Report



Power Details

Test brand

Power

α	σ	δ	Number	Power
0.0500	4.27033	2	11	0.1946
0.0500	4.27033	2	13	0.2372
0.0500	4.27033	2	15	0.2800
0.0500	4.27033	2	17	0.3226
0.0500	4.27033	4	11	0.6239
0.0500	4.27033	4	13	0.7368
0.0500	4.27033	4	15	0.8209
0.0500	4.27033	4	17	0.8810
0.0500	4.27033	6	11	0.9346
0.0500	4.27033	6	13	0.9775
0.0500	4.27033	6	15	0.9927
0.0500	4.27033	6	17	0.9978

Power Plot

13. You might need to click and drag vertically on the Power axis to see all of the data in the plot.

Power is plotted for each combination of Delta and Number. As you might expect, the power rises for larger Number (sample sizes) values and for larger Delta values (difference in means).

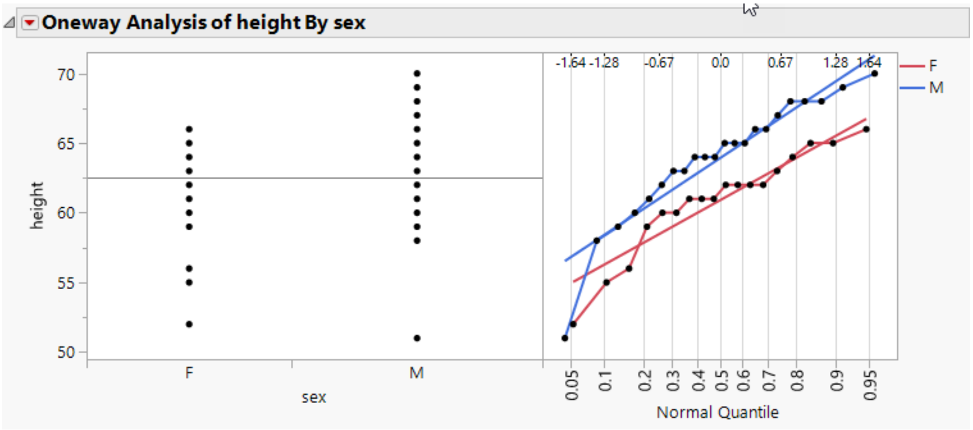
Example of a Normal Quantile Plot

Use the Oneway platform to produce a normal quantile plot.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.

2. Select **Analyze > Fit Y by X**.
3. Select height and click **Y, Response.**
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Normal Quantile Plot > Plot Actual by Quantile**.

Figure 6.31 Example of a Normal Quantile Plot



Note the following:

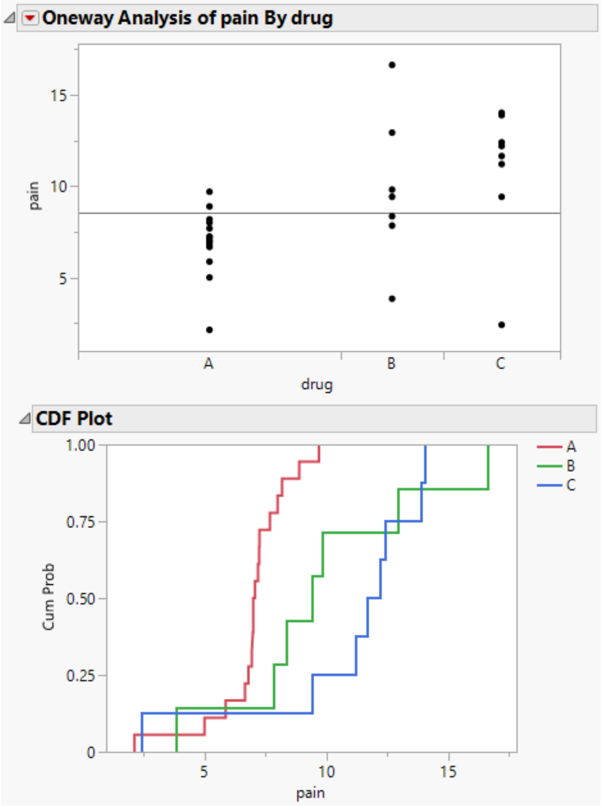
- The Line of Fit appears by default.
- The data points track very closely to the line of fit, indicating a normal distribution.

Example of a CDF Plot

Use the Oneway platform to produce a cumulative distribution function (CDF) plot.

1. Select **Help > Sample Data Folder** and open Analgesics.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select pain and click **Y, Response**.
4. Select drug and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **CDF Plot**.

Figure 6.32 Example of a CDF Plot



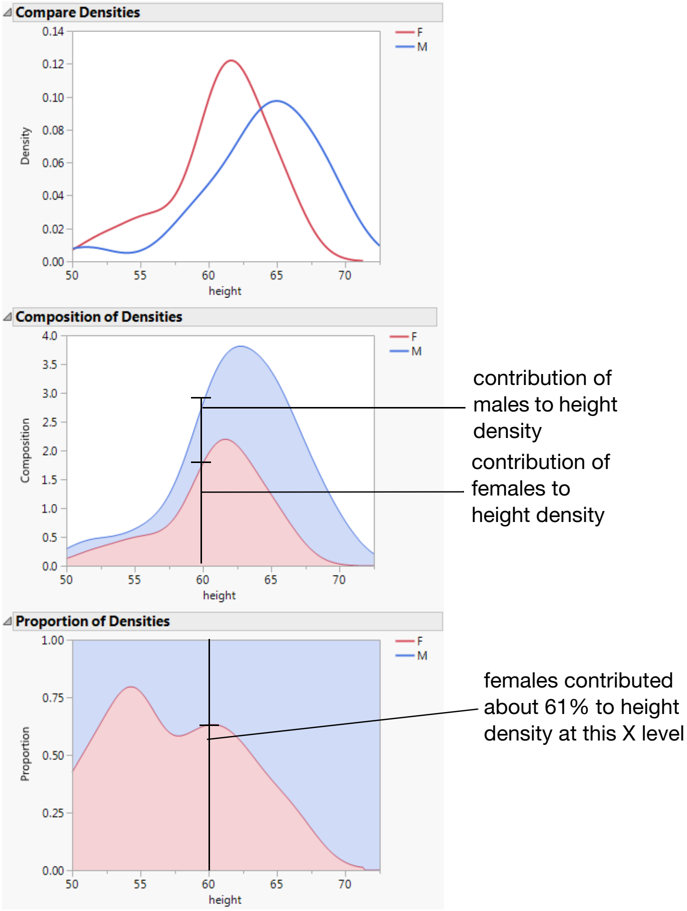
The CDF plot has a curve for each level of the X variable.

Example of the Densities Options

This example illustrates the Densities options in the Oneway platform.

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select height and click **Y, Response.**
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select all three options: **Densities > Compare Densities**, **Densities > Composition of Densities**, and **Densities > Proportion of Densities**.

Figure 6.33 Example of the Densities Options



contribution of males to height density

contribution of females to height density

females contributed about 61% to height density at this X level

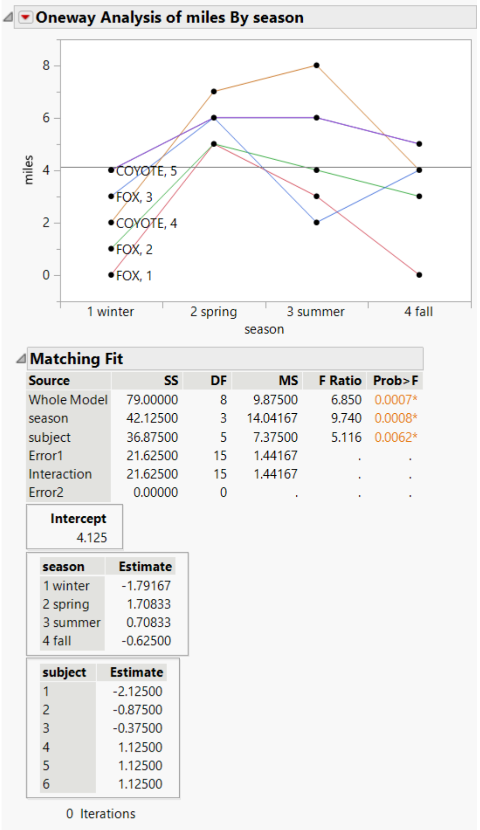
Example of the Matching Column Option

Use the Oneway platform to analyze data on six animals and the miles that they travel during different seasons.

1. Select **Help > Sample Data Folder** and open Matching.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select miles and click **Y, Response**.
4. Select season and click **X, Factor**.
5. Click **OK**.
6. Click the Oneway Analysis red triangle menu and select **Matching Column**.
7. Select subject as the matching column.

8. Click **OK**.

Figure 6.34 Example of the Matching Column Report



Source	SS	DF	MS	F Ratio	Prob>F
Whole Model	79.00000	8	9.87500	6.850	0.0007*
season	42.12500	3	14.04167	9.740	0.0008*
subject	36.87500	5	7.37500	5.116	0.0062*
Error1	21.62500	15	1.44167	.	.
Interaction	21.62500	15	1.44167	.	.
Error2	0.00000	0	.	.	.

Intercept
4.125

season	Estimate
1 winter	-1.79167
2 spring	1.70833
3 summer	0.70833
4 fall	-0.62500

subject	Estimate
1	-2.12500
2	-0.87500
3	-0.37500
4	1.12500
5	1.12500
6	1.12500

0 Iterations

The plot graphs the miles traveled by season, with subject as the matching variable. The labels next to the first measurement for each subject on the graph are determined by the species and subject variables.

The Matching Fit report shows the season and subject effects with *F* tests. These are equivalent to the tests that you get with the Fit Model platform if you run two models, one with the interaction term and one without. If there are only two levels, then the *F* test is equivalent to the paired *t* test.

Note: For more information about the Fit Model platform, see *Fitting Linear Models*.

Example of Stacking Data for a Oneway Analysis

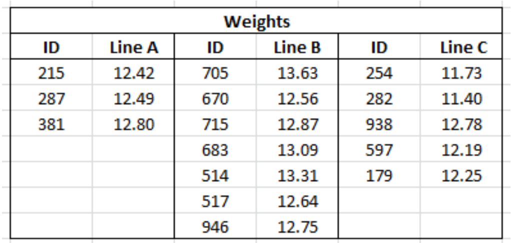
When your data are in a format other than a JMP data table, sometimes they are arranged so that a row contains information for multiple observations. To analyze the data in JMP, you must import the data and restructure it so that each row of the JMP data table contains information for a single observation. For example, suppose that your data are in a spreadsheet. The data for parts produced on three production lines are arranged in three sets of columns. In your JMP data table, you need to stack the data from the three production lines into a single set of columns so that each row represents the data for a single part.

Description and Goals

This example uses the file Fill Weights.xlsx, which contains the weights of cereal boxes randomly sampled from three different production lines. Figure 6.35 shows the format of the data.

- The ID columns contain an identifier for each cereal box that was measured.
- The Line columns contain the weights (in ounces) for boxes sampled from the corresponding production line.

Figure 6.35 Data Format



Weights					
ID	Line A	ID	Line B	ID	Line C
215	12.42	705	13.63	254	11.73
287	12.49	670	12.56	282	11.40
381	12.80	715	12.87	938	12.78
		683	13.09	597	12.19
		514	13.31	179	12.25
		517	12.64		
		946	12.75		

The target fill weight for the boxes is 12.5 ounces. Although you are interested in whether the three production lines are meeting the target, initially you want to see whether the three lines are achieving the same mean fill rate. You can use Oneway to test for differences among the mean fill weights.

To use the Oneway platform, you need to do the following:

1. Import the data into JMP. See "Import the Data".
2. Reshape the data so that each row in the JMP data table reflects only a single observation. Reshaping the data requires that you stack the cereal box IDs, the line identifiers, and the weights into columns. See "Stack the Data".

Import the Data

This example illustrates two ways to import data from Microsoft Excel into JMP. Select one method or explore both:

- Use the **File > Open** option to import data from a Microsoft Excel file using the Excel Import Wizard. See “Import the Data Using the Excel Import Wizard”. This method is convenient for any Excel file.
- Copy and paste data from Microsoft Excel into a new JMP data table. See “Copy and Paste the Data from Excel”. You can use this method with small data files.

For more information about how to import data from Microsoft Excel, see *Using JMP*.

Import the Data Using the Excel Import Wizard

1. Select **Help > Sample Data Folder** and open Fill Weights.xlsx located in the Samples/Import Data folder.

   The file opens in the Excel Import Wizard.

2. Type 3 next to **Column headers start on row**.

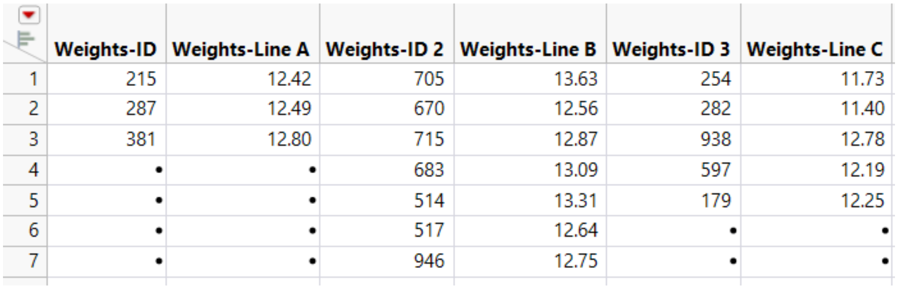
   In the Excel file, row 1 contains information about the table and row 2 is blank. The column heading information starts on row 3.

3. Type 2 for **Number of rows with column headers**.

   In the Excel file, rows 3 and 4 both contain column heading information.

4. Click **Import**.

Figure 6.36 JMP Table Created Using Excel Import Wizard



	Weights-ID	Weights-Line A	Weights-ID 2	Weights-Line B	Weights-ID 3	Weights-Line C
1	215	12.42	705	13.63	254	11.73
2	287	12.49	670	12.56	282	11.40
3	381	12.80	715	12.87	938	12.78
4	•	•	683	13.09	597	12.19
5	•	•	514	13.31	179	12.25
6	•	•	517	12.64	•	•
7	•	•	946	12.75	•	•

   The data are placed in seven rows and multiple IDs appear in each row. For each of the three lines, there are an ID and Weight column, giving a total of six columns.

   Notice that the “Weights” part of the ID column name is unnecessary and misleading. You could rename the columns now, but it will be more efficient to rename the columns after you stack the data.

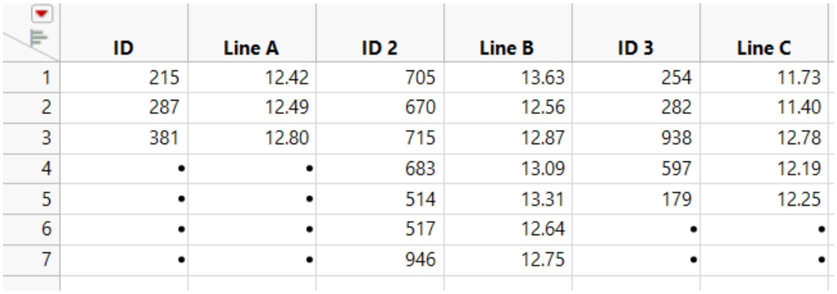
5. Proceed to “Stack the Data”.

Copy and Paste the Data from Excel

1. Open Fill Weights.xlsx in Microsoft Excel.
2. Select the data inside the table but exclude the unnecessary “Weights” heading.
3. Right-click and select **Copy**.
4. In JMP, select **File > New > Data Table**.
5. Select **Edit > Paste with Column Names**.

   The **Edit > Paste with Column Names** option is used when your column names are included in the selection on the clipboard.

Figure 6.37 JMP Table Created Using Paste with Column Names



	ID	Line A	ID 2	Line B	ID 3	Line C
1	215	12.42	705	13.63	254	11.73
2	287	12.49	670	12.56	282	11.40
3	381	12.80	715	12.87	938	12.78
4	•	•	683	13.09	597	12.19
5	•	•	514	13.31	179	12.25
6	•	•	517	12.64	•	•
7	•	•	946	12.75	•	•

6. Proceed to “Stack the Data”.

Stack the Data

Use the Stack option to place one observation in each row of a new data table. For more information about the Stack option, see *Using JMP*.

1. In the JMP data table, select **Tables > Stack**.
2. Select all six columns and click **Stack Columns**.
3. Select **Multiple Series Stack**.

   You are stacking two series, ID and Line, so you do not change the Number of Series, which is set to 2 by default. The columns that contain the series are not contiguous. They alternate (ID, Line A, ID, Line B, ID, Line C). For this reason, you do not check Contiguous.
4. Deselect **Stack By Row**.
5. Select **Eliminate Missing Rows**.
6. Enter Stacked next to **Output table name**.
7. Click **OK**.

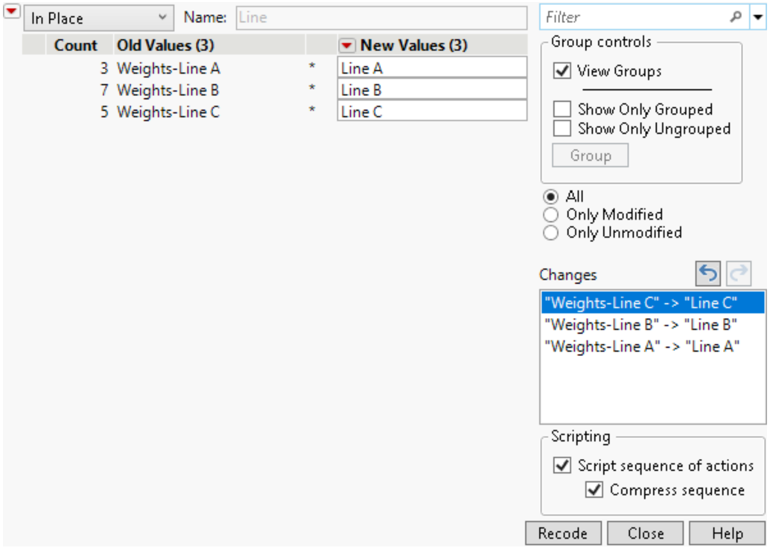
   In the new data table, Data and Data 2 are columns containing the ID and Weight data.
8. Right-click the Label column heading and select **Delete Columns**.

The entries in the Label column were the column headings for the box IDs in the imported data table. These entries are not needed.

9. Rename each column by double-clicking on the column heading:
   - Data to ID
   - Label 2 to Line
   - Data 2 to Weight
10. In the Columns panel, click the icon to the left of ID and select **Nominal**.

    Although ID is given as a number, it is an identifier and should be treated as nominal when modeling. This is not an issue in this example, but it is good practice to assign the appropriate modeling type to a column.
11. (Applies only if you imported the data from Excel using **File > Open.)** Do the following:
    1. Click the Line column heading to select the column and select **Cols > Recode**.
    2. Click **New Column** and select **In Place**.
    3. Change the values in the **New Values** column to match those in Figure 6.38 below.

Figure 6.38 Recode Column Values



In Place Name: Line

Count	Old Values (3)	New Values (3)
3	Weights-Line A	Line A
7	Weights-Line B	Line B
5	Weights-Line C	Line C

Group controls
☑ View Groups
☐ Show Only Grouped
☐ Show Only Ungrouped

● All
○ Only Modified
○ Only Unmodified

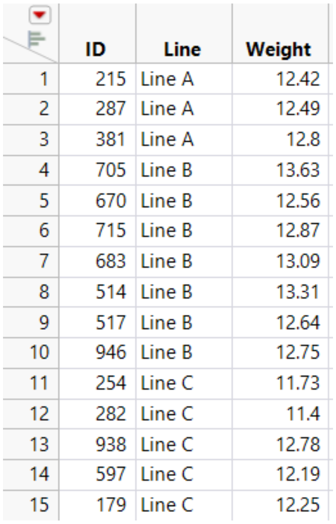
Changes
"Weights-Line C" -> "Line C"
"Weights-Line B" -> "Line B"
"Weights-Line A" -> "Line A"

Scripting
☑ Script sequence of actions
☑ Compress sequence

    4. Click **Recode**.

Your new data table is now properly structured for JMP analysis. Each row contains data for a single cereal box. The first column gives the box ID, the second gives the production line, and the third gives the weight of the box (Figure 6.39).

Figure 6.39 Recoded Data Table



	ID	Line	Weight
1	215	Line A	12.42
2	287	Line A	12.49
3	381	Line A	12.8
4	705	Line B	13.63
5	670	Line B	12.56
6	715	Line B	12.87
7	683	Line B	13.09
8	514	Line B	13.31
9	517	Line B	12.64
10	946	Line B	12.75
11	254	Line C	11.73
12	282	Line C	11.4
13	938	Line C	12.78
14	597	Line C	12.19
15	179	Line C	12.25

Conduct the Oneway Analysis

This part of the example contains the following tasks:

- Conduct a Oneway Analysis of Variance to test for differences in the mean fill weights among the three production lines.
- Obtain Comparison Circles to explore which lines might differ.
- Label points by ID in case you want to reweigh or further examine their boxes.

Before beginning, verify that you are using the Stacked data table.

1. Select **Analyze > Fit Y by X**.
2. Select Weight and click **Y, Response**.
3. Select Line and click **X, Factor**.
4. Click **OK**.
5. Click the Oneway Analysis red triangle menu and select **Means/Anova**.

   The mean diamonds in the plot show 95% confidence intervals for the production line means. The points that fall outside the mean diamonds might seem like outliers. However, they are not. To see this, add box plots to the plot.
6. Click the Oneway Analysis red triangle menu and select **Display Options > Box Plots**.

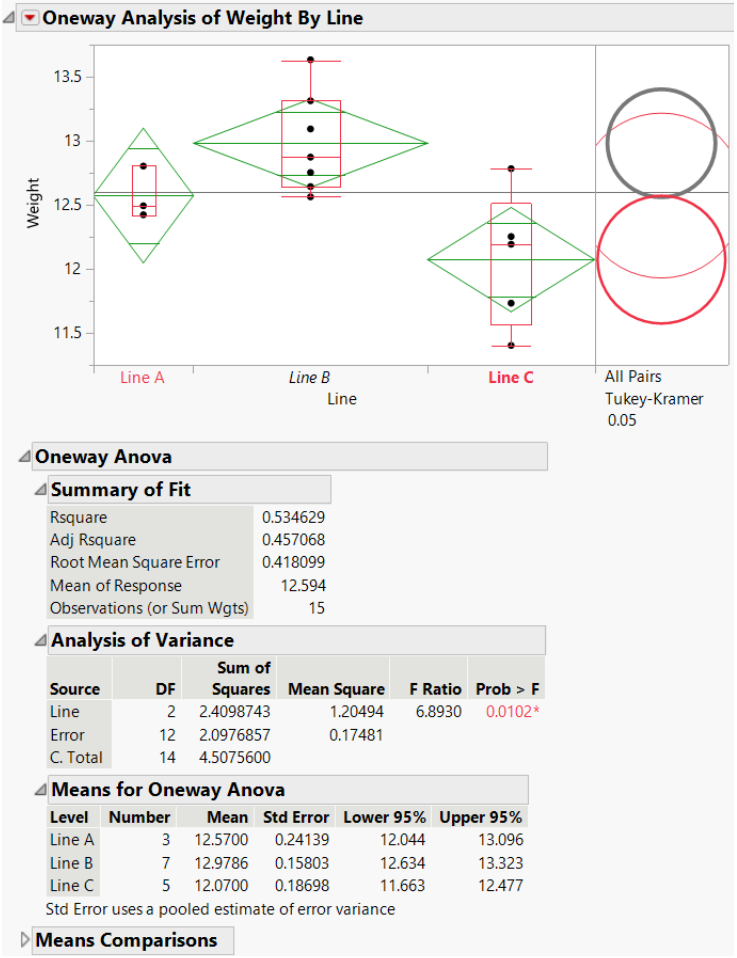
   All points fall within the box plots boundaries. Therefore, they are not outliers.
7. From the data table, in the Columns panel, right-click ID and select **Label/Unlabel**.
8. In the plot, hover over the points to see their ID values, as well as their Line and Weight data (Figure 6.40).

9. Click the Oneway Analysis red triangle menu and select **Compare Means > All Pairs, Tukey HSD**.

   Comparison circles appear in a panel to the right of the plot.

10. Click the bottom comparison circle.

Figure 6.40 Oneway Analysis of Weight by Line



Oneway Analysis of Weight By Line

Oneway Anova

Summary of Fit

Rsquare	0.534629
Adj Rsquare	0.457068
Root Mean Square Error	0.418099
Mean of Response	12.594
Observations (or Sum Wgts)	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Line	2	2.4098743	1.20494	6.8930	0.0102*
Error	12	2.0976857	0.17481		
C. Total	14	4.5075600			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Line A	3	12.5700	0.24139	12.044	13.096
Line B	7	12.9786	0.15803	12.634	13.323
Line C	5	12.0700	0.18698	11.663	12.477

Std Error uses a pooled estimate of error variance

Means Comparisons

In the Analysis of Variance report, the *p*-value of 0.0102 provides evidence that the means are not all equal. In the plot, the comparison circle for Line C is selected and appears red. Since the circle for Line B appears as thick gray, the mean for Line C differs from the mean for Line B at the 0.05 significance level. The means for Lines A and B do not show a statistically significant difference.

The mean diamonds shown in the plot span 95% confidence intervals for the means. The numeric bounds for the 95% confidence intervals are given in the Means for Oneway ANOVA report. Both of these indicate that the confidence intervals for Lines B and C do not contain the target fill weight of 12.5: Line B seems to overfill and Line C seems to underfill. For these two production lines, the underlying causes that result in off-target fill weights must be addressed.

Statistical Details for the Oneway Platform

This section contains statistical details for the Oneway platform.

- "Statistical Details for Comparison Circles"
- "Statistical Details for Power"
- "Statistical Details for ANOM"
- "Statistical Details for the Summary of Fit Report"
- "Statistical Details for Tests That the Variances Are Equal"
- "Statistical Details for Nonparametric Test Statistics"
- "Statistical Details for the Robust Fit"

Statistical Details for Comparison Circles

In the Oneway platform, comparison circles are a graphical representation of the *least significant difference* (LSD) in a multiple comparison test. This least significant difference is a probability quantile multiplied by the standard error of the difference of the two means. For the Each Pairs option, the Fisher LSD is used and the probability quantile is the Student's *t*-statistic. The comparison circles calculations are demonstrated for this case. The LSD is defined as follows:

$$\mathrm{LSD} = t_{\alpha/2}\mathrm{std}(\hat{\mu}_1 - \hat{\mu}_2)$$

The standard error of the difference of two independent means is calculated from the following relationship:

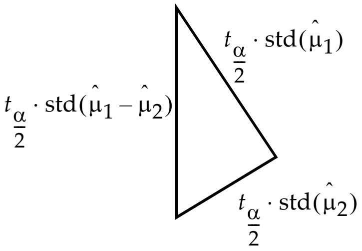
$$[\mathrm{std}(\hat{\mu}_1 - \hat{\mu}_2)]^2 = [\mathrm{std}(\hat{\mu}_1)]^2 + [\mathrm{std}(\hat{\mu}_2)]^2$$

When the means are uncorrelated, these quantities have the following relationship:

$$\mathrm{LSD}^2 = [t_{\alpha/2}\mathrm{std}((\hat{\mu}_1 - \hat{\mu}_2))]^2 = [t_{\alpha/2}\mathrm{std}(\hat{\mu}_1)]^2 + [t_{\alpha/2}\mathrm{std}\hat{\mu}_2]^2$$

These squared values form a Pythagorean relationship, illustrated graphically by the right triangle shown in Figure 6.41.

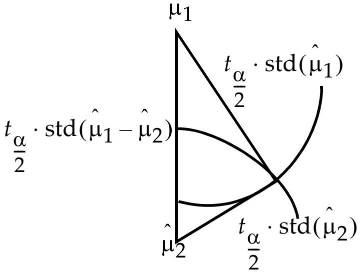
Figure 6.41 Relationship of the Difference between Two Means



The hypotenuse of this triangle is a measuring stick for comparing means. The means are significantly different if and only if the actual difference is greater than the hypotenuse (LSD).

Suppose that you have two means that are exactly on the borderline, where the actual difference is the same as the least significant difference. Draw the triangle with vertices at the means measured on a vertical scale. Also, draw circles around each mean so that the diameter of each is equal to the confidence interval for that mean.

Figure 6.42 Geometric Relationship of *t* Test Statistics



The radius of each circle is the length of the corresponding leg of the triangle, which is $t_{\alpha/2}\text{std}(\hat{\mu}_i)$.

The circles must intersect at the same right angle as the triangle legs, giving the following relationship:

- If the means differ exactly by their least significant difference, then the confidence interval circles around each mean intersect at a right angle. That is, the angle of the tangents is a right angle.

Now, consider how these circles must intersect if the means are different by greater than or less than the least significant difference:

- If the circles intersect so that the outside angle is greater than a right angle, then the means *are not* significantly different. If the circles intersect so that the outside angle is less than a right angle, then the means *are* significantly different. An outside angle of less than 90 degrees indicates that the means are farther apart than the least significant difference.

- If the circles do not intersect, then they are significantly different. If they nest, they are not significantly different (Figure 6.9).

The same graphical technique works for many multiple-comparison tests, substituting a different probability quantile value for the Student's *t*.

Statistical Details for Power

To compute power, the Oneway platform uses the noncentral *F* distribution. The formula (O'Brien and Lohr 1984) is defined as follows:

$$\text{Power} = \text{Prob}(F > F_{crit}, \nu_1, \nu_2, nc)$$

where:

F is distributed as the noncentral $F(nc, \nu_1, \nu_2)$ and $F_{crit} = F_{(1-\alpha, \nu 1, \nu 2)}$ is the 1 - α quantile of the F distribution with ν_1 and ν_2 degrees of freedom.

$\nu_1 = r - 1$ is the numerator df.

$\nu_2 = r(n - 1)$ is the denominator df.

n is the number per group.

r is the number of groups.

$nc = n(CSS)/\sigma^2$ is the non-centrality parameter.

$CSS = \sum_{g=1}^{r} (\mu_g - \mu)^2$ is the corrected sum of squares.

μ_g is the mean of the g^{th} group.

μ is the overall mean.

σ^2 is estimated by the mean squared error (MSE).

Statistical Details for ANOM

Transformed Ranks

Suppose that there are n observations. The transformed observations are computed as follows:

- Rank all observations from smallest to largest, accounting for ties. For tied observations, assign each one the average of the block of ranks that they share.
- Denote the ranks by $R_1, R_2, \ldots, R_n$.
- The transformed rank corresponding to the i^{th} observation is:

$$\text{Transformed } R_i = \textit{Normal Quantile}\left[\left(\frac{R_i}{2n+1}\right) + 0.5\right]$$

The ANOM procedure is applied to the values Transformed R_i. Since the ranks have a uniform distribution, the transformed ranks have a folded normal distribution. See Nelson et al. (2005).

Statistical Details for the Summary of Fit Report

This section contains details for the Summary of Fit report in the Oneway platform.

Rsquare

Using quantities from the Analysis of Variance report for the model, the R^2 for any continuous response fit is calculated as follows:

$$\frac{\text{Sum of Squares (Model)}}{\text{Sum of Squares (C Total)}}$$

Adj Rsquare

Adj Rsquare is a ratio of mean squares instead of sums of squares and is calculated as follows:

$$1 - \frac{\text{Mean Square (Error)}}{\text{Mean Square (C Total)}}$$

The mean square for Error is found in the Analysis of Variance report and the mean square for **C. Total** can be computed as the **C. Total** Sum of Squares divided by its respective degrees of freedom. See "Analysis of Variance".

Statistical Details for Tests That the Variances Are Equal

This section contains details for the Tests That the Variances Are Equal report in the Oneway platform.

F Ratio

O'Brien's test constructs a dependent variable so that the group means of the new variable equal the group sample variances of the original response. The O'Brien variable is computed as follows:

$$r_{ijk} = \frac{(n_{ij} - 1.5)n_{ij}(y_{ijk} - \overline{y_{ij}})^2 - 0.5s_{ij}^2(n_{ij} - 1)}{(n_{ij} - 1)(n_{ij} - 2)}$$

where n represents the number of y_{ijk} observations.

Brown-Forsythe is the model F statistic from an ANOVA on $z_{ij} = |y_{ij} - \tilde{y}_i|$ where $\tilde{y}_i$ is the median response for the i^{th} level.

The Levene F is the model F statistic from an ANOVA on $z_{ij} = |y_{ij} - \overline{y_{i.}}|$ where $\overline{y_{i.}}$ is the mean response for the i^{th} level.

Bartlett's test is calculated as follows:

$$T = \frac{v\log\left(\sum_i \frac{v_i}{v} s_i^2\right) - \sum_i v_i \log(s_i^2)}{1 + \left(\frac{\sum_i \frac{1}{v_i} - \frac{1}{v}}{3(k-1)}\right)} \quad \text{where } v_i = n_i - 1 \text{ and } v = \sum_i v_i$$

and n_i is the count on the i^{th} level and s_i^2 is the response sample variance on the i^{th} level. The Bartlett statistic has a χ^2-distribution. Dividing the Chi-square test statistic by the degrees of freedom results in the reported F value.

Welch's Test F Ratio

The Welch's Test F Ratio is computed as follows:

$$F = \frac{\left[\frac{\sum_i w_i(\bar{y}_i - \tilde{y}_{..})^2}{k-1}\right]}{\left\{1 + \frac{2(k-2)}{k^2-1}\left[\sum_i \frac{\left(1 - \frac{w_i}{u}\right)^2}{n_i - 1}\right]\right\}} \quad \text{where } w_i = \frac{n_i}{s_i^2},\ u = \sum_i w_i,\ \tilde{y}_{..} = \sum_i \frac{w_i \bar{y}_{i.}}{u},$$

and n_i is the count on the i^{th} level, $\bar{y}_{i.}$ is the mean response for the i^{th} level, and s_i^2 is the response sample variance for the i^{th} level.

Welch's Test DF Den

The Welch approximation for the denominator degrees of freedom is defined as follows:

$$df = \frac{1}{\left(\frac{3}{k^2-1}\right)\left[\sum_i \frac{\left(1 - \frac{w_i}{u}\right)^2}{n_i - 1}\right]}$$

where w_i, n_i, and u are defined as in the F ratio formula.

Statistical Details for Nonparametric Test Statistics

This section provides formulas for the test statistics used in the Wilcoxon, Median, van der Waerden, and Friedman Rank tests in the Oneway platform.

Notation

The tests are based on scores and use the following notation.

$j = 1,\ldots, n$ The observations in the entire sample.

$i = 1,\ldots, k$ The levels of X, where k is the total number of levels.

$n_1, n_2,\ldots, n_k$ The number of observations in each of the k levels of X.

R_j The midrank of the j^{th} observation. The *midrank* is the observation's rank if it is not tied and its average rank if it is tied.

α A function of the midranks used to define scores for the various tests.

The following notation is used when a Block variable is specified in the launch window.

$b = 1,\ldots, B$ The levels of the blocking variable, where B is the total number of blocks.

R_{bi} The midrank of the i^{th} level of X within block b.

The function α defines scores as follows:

Wilcoxon Scores

$$\alpha(R_j) = R_j$$

Median Scores

$$\alpha(R_j) = \begin{cases} 1 \text{ if } R_j > \text{median} \\ 0 \text{ if } R_j < \text{median} \\ t \text{ if } R_j = \text{median} \end{cases}$$

Let n_t denote the number of observations tied at the median and let n_u denote the number of observations greater than the median. Then t is given by the following:

$$t = \frac{\text{floor}(n/2) - n_u}{n_t}$$

van der Waerden Scores

$$\alpha(R_j) = \text{Standard Normal Quantile}(R_j/(n+1))$$

Friedman Rank Scores

$$\alpha(R_{bi}) = R_{bi}$$

Two-Sample Normal Approximations

Tests based on the normal approximation are given only when X has exactly two levels. The notation used in this section is defined in "Notation". The statistics that appear in the Two-Sample Normal Approximation report are defined below.

S The statistic *S* is the sum of the values $\alpha(R_j)$ for the observations in the smaller group. If the two levels of X have the same numbers of observations, then the value of S corresponds to the last level of X in the value ordering.

Z The value of Z is defined as follows:

$$Z = (S - E(S))/\sqrt{Var(S)}$$

Note: The Wilcoxon test adds a continuity correction. If $(S - E(S))$ is greater than zero, then 0.5 is subtracted from the numerator. If $(S - E(S))$ is less than zero, then 0.5 is added to the numerator.

***E*(*S*)** The expected value of S under the null hypothesis. Denote the number of observations in the smaller level, or in the last level in the value ordering if the two groups have the same number of observations, by n_l:

$$E(S) = \frac{n_l}{n}\sum_{j=1}^{n}\alpha(R_j)$$

***Var*(*S*)** Define *ave* to be the average score across all observations. Then the variance of S is defined as follows:

$$Var(S) = \frac{n_1 n_2}{n(n-1)}\sum_{j=1}^{n}(\alpha(R_j) - ave)^2$$

Two-Sample Normal Approximations for Friedman Rank Test

When you use the Friedman Rank test, the calculations for the two-sample normal approximation is the same as above, except that the variance of S is different. The variance of S is computed as follows:

$$Var(S) = \frac{B}{(n-1)} \sum_{j=1}^{n} (\alpha(R_j) - ave)^2$$

One-Way ChiSquare Approximations

Note: The ChiSquare test based on the Wilcoxon scores is known as the Kruskal-Wallis test.

The notation used in this section is defined in "Notation". The following quantities are used in calculating the ChiSquare statistic:

T_i The total of the scores for the i^{th} level of X.

$E(T_i)$ The expected value of the total score for level i under the null hypothesis of no difference in levels, defined as follows:

$$E(T_i) = \frac{n_i}{n} \sum_{j=1}^{n} \alpha(R_j)$$

$Var(T)$ Define *ave* to be the average score across all observations. Then the variance of T is defined as follows:

$$Var(T) = \frac{1}{(n-1)} \sum_{j=1}^{n} (\alpha(R_j) - ave)^2$$

The value of the test statistic is given below. This statistic has an asymptotic chi-square distribution with $k - 1$ degrees of freedom.

$$C = \left(\sum_{i=1}^{k} (T_i - E(T_i))^2 / n_i \right) / Var(T)$$

One-Way ChiSquare Approximations for Friedman Rank Test

The ChiSquare test statistic for the Friedman Rank Test is calculated as follows:

$$C = \frac{\sum_{i=1}^{\kappa} (T_i - E(T_i))^2 / n_i}{\frac{1}{(k-1)} \sum_{j=1}^{n} (\alpha(R_j) - ave)^2 / n_i}$$

Statistical Details for the Robust Fit

In the Oneway platform, the Robust Fit option reduces the influence of outliers in the response variable. The Huber M-estimation method is used. Huber M-estimation finds parameter estimates that minimize the Huber loss function:

$$l(\underset{\sim}{e}) = \sum_{i} \rho(e_i)$$

where

$$\rho(e) = \begin{cases} \frac{1}{2}e^2 & \text{if } |e| < k \\ k|e| - \frac{1}{2}k^2 & \text{if } |e| \geq k \end{cases}$$

e_i refers to the residuals

The Huber loss function penalizes outliers and increases as a quadratic for small errors and linearly for large errors. For more information about robust fitting, see Huber (1973) and Huber and Ronchetti (2009). See “Example of the Robust Fit Option”.

Chapter 7

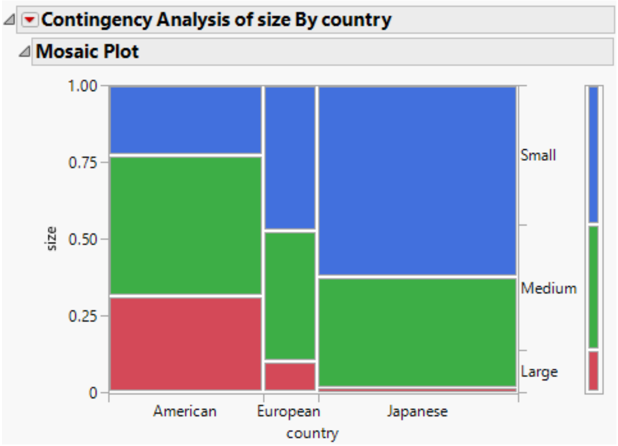
Contingency Analysis

Examine Relationships between Two Categorical Variables

Use the Contingency platform to investigate the relationship between two categorical variables. The categorical variables can be ordinal or nominal. The analysis results include a mosaic plot, a contingency table of frequency counts and proportions, and chi-square tests of significance. You can interactively perform additional analyses and tests on your data, such as analysis of means for proportions, correspondence analysis, and measures of association.

The Contingency platform is the *categorical by categorical* personality of the Fit Y by X platform.

Figure 7.1 Example of Contingency Analysis



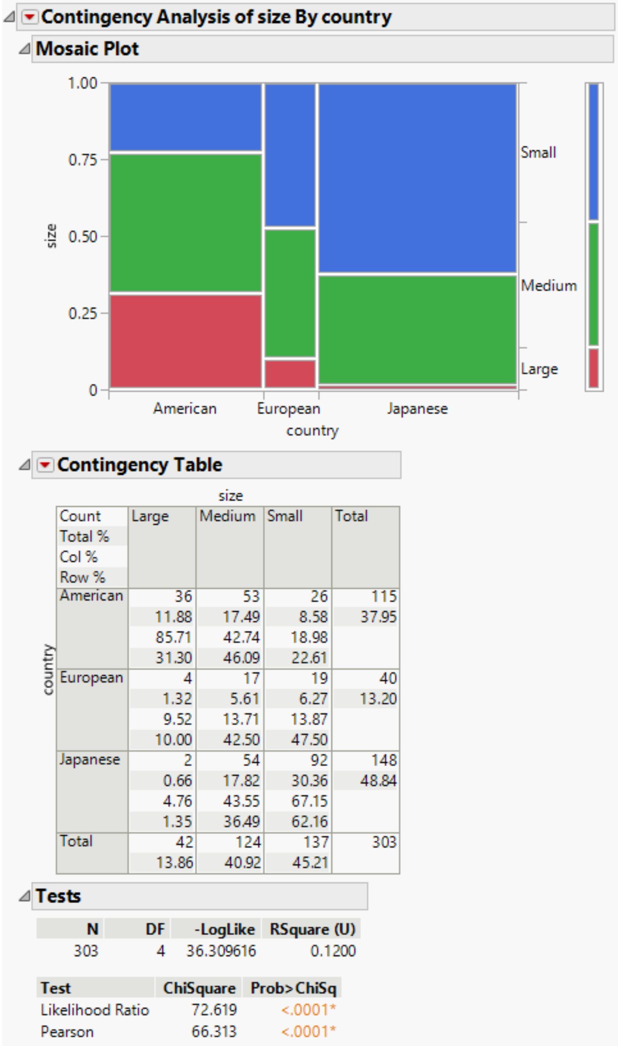
Contents

Example of Contingency Analysis

Use the Contingency platform to examine the relationship between two categorical variables. This example uses data collected from a survey about car ownership. The data include respondent attributes: sex, marital status, and age. The data also include attributes of the respondent’s car: country of origin, the size, and the type. Examine the relationship between car sizes (small, medium, and large) and the cars’ country of origin.

1. Select **Help > Sample Data Folder** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select size and click **Y, Response**.
4. Select country and click **X, Factor**.
5. Click **OK**.

Figure 7.2 Example of Contingency Analysis



Contingency Analysis of size By country

Mosaic Plot

Contingency Table

country	Count Total % Col % Row %	Large	Medium	Small	Total
American	Count	36	53	26	115
	Total %	11.88	17.49	8.58	37.95
	Col %	85.71	42.74	18.98	
	Row %	31.30	46.09	22.61	
European	Count	4	17	19	40
	Total %	1.32	5.61	6.27	13.20
	Col %	9.52	13.71	13.87	
	Row %	10.00	42.50	47.50	
Japanese	Count	2	54	92	148
	Total %	0.66	17.82	30.36	48.84
	Col %	4.76	43.55	67.15	
	Row %	1.35	36.49	62.16	
Total	Count	42	124	137	303
	Total %	13.86	40.92	45.21	

Tests

N	DF	-LogLike	RSquare (U)
303	4	36.309616	0.1200

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	72.619	<.0001*
Pearson	66.313	<.0001*

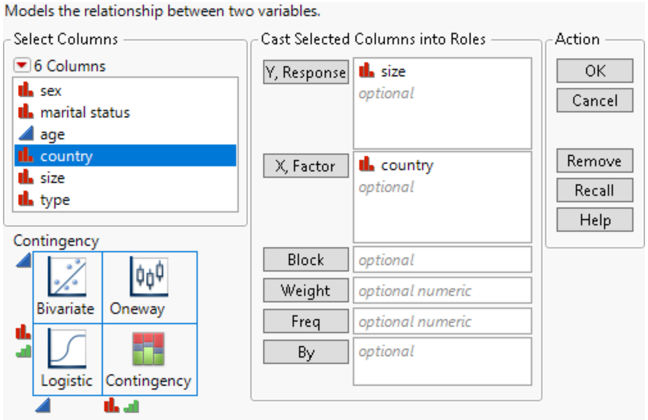
From the mosaic plot and legend, notice the following:

- Very few Japanese cars fall into the Large size category.
- The majority of the European cars fall into the Small and Medium size categories.
- The majority of the American cars fall into the Large and Medium size categories.

Launch the Contingency Platform

Launch the Contingency platform by selecting **Analyze > Fit Y by X**. The Fit Y by X launch window is used for four different types of analyses. When you enter an ordinal or nominal Y variable and an ordinal or nominal X variable, the Contingency platform is launched.

Figure 7.3 The Contingency Launch Window



Models the relationship between two variables.

For more information about the options in the Select Columns red triangle menu, see *Using JMP*. The Contingency launch window contains the following options:

Y, Response The response variable or variables that you want to analyze. These variables must have an ordinal or nominal modeling type.

X, Factor The predictor variable or variables that you want to analyze. These variables must have an ordinal or nominal modeling type.

Block A column that specifies a Blocking variable. Specifying a Block variable identifies a second factor and performs a Cochran-Mantel-Haenszel test. Blocking across an additional classification variable is also called stratification.

Weight A column containing a weight for each observation in the data table. A row is included in the analyses only when its value is greater than zero.

Freq Assigns a frequency to each row in the analysis. Assigning a frequency is useful when your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Data Format

In the Contingency platform, your data can consist of unsummarized or summarized data:

Unsummarized data There is one row for each observation and columns for X and Y values.

Summarized data Each row represents a set of observations with common X and Y values. The data table must contain a frequency column that contains the counts of observations for each row. Enter this column as Freq in the launch window.

For an example of summarized categorical data, see "Example of Analysis of Means for Proportions".

Note: The Fit Y by X launch window accommodates columns with continuous, ordinal, and nominal modeling types. The Contingency platform is launched for all pairs of Y, Response and X, Factor columns that have the ordinal or nominal modeling types. The Fit Y by X launch window launches the Bivariate, Oneway, or Logistic platforms for other column type combinations.

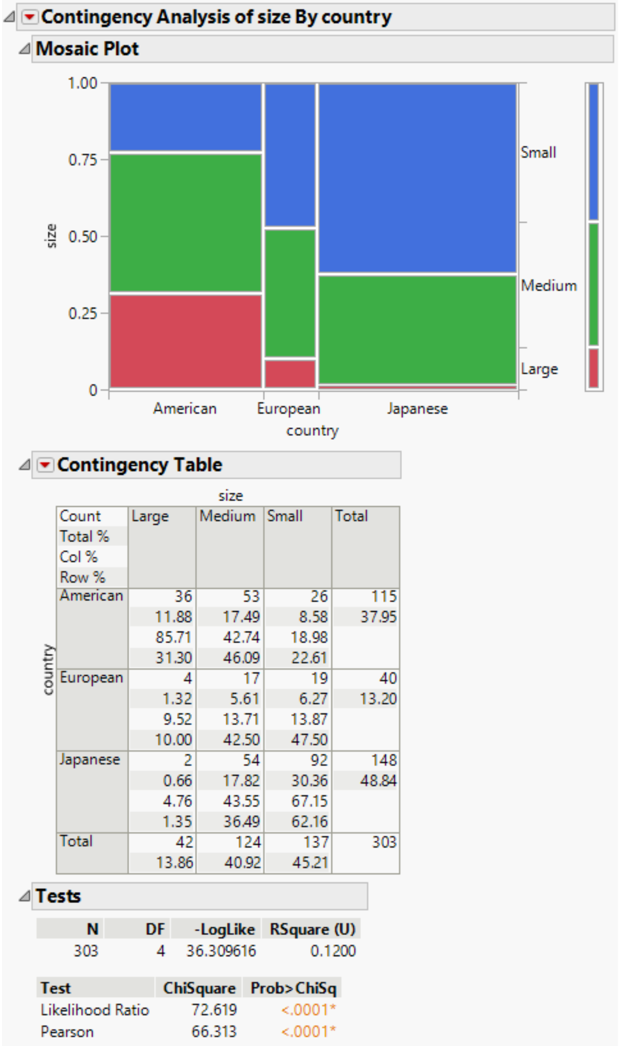
The Contingency Report

The Contingency report initially contains a mosaic plot, a contingency table, and tests to determine whether the levels of the Y response variable differ across the levels of the X factor variables. You can run additional analyses and tests using the red triangle menu options. See "Contingency Platform Options".

This section contains information about the following sections of the Contingency report:

- "Mosaic Plot"
- "Contingency Table"
- "Tests Report"

Figure 7.4 Example of a Contingency Report



Contingency Analysis of size By country

Mosaic Plot

Contingency Table

		size			
country	Count Total % Col % Row %	Large	Medium	Small	Total
	American	36	53	26	115
		11.88	17.49	8.58	37.95
		85.71	42.74	18.98	
		31.30	46.09	22.61	
	European	4	17	19	40
		1.32	5.61	6.27	13.20
		9.52	13.71	13.87	
		10.00	42.50	47.50	
	Japanese	2	54	92	148
		0.66	17.82	30.36	48.84
		4.76	43.55	67.15	
		1.35	36.49	62.16	
	Total	42	124	137	303
		13.86	40.92	45.21	

Tests

N	DF	-LogLike	RSquare (U)
303	4	36.309616	0.1200

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	72.619	<.0001*
Pearson	66.313	<.0001*

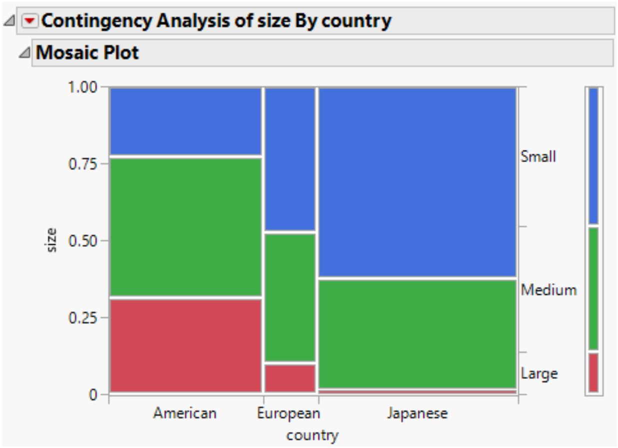
Replace Variables Interactively

You can interactively replace variables in the plot by dragging and dropping a variable from one axis to the other axis. You can also replace a variable by selecting a variable in the Columns panel of the associated data table and dragging it onto an axis.

Mosaic Plot

In the Contingency platform, the mosaic plot is a graphical representation of the two-way frequency table, which is also called a contingency table. A mosaic plot is divided into rectangles of varying dimensions; the vertical length of each rectangle is proportional to the proportions of the Y variable within each level of the X variable. The mosaic plot was introduced by Hartigan and Kleiner (1981) and refined by Friendly (1994).

Figure 7.5 Example of a Mosaic Plot



Note the following about the mosaic plot:

- The width of the partitions on the horizontal axis represent the number of observations for each level of the X variable.
- The proportions on the vertical axis on the right side of the plot represent the overall proportions of the levels of the Y variable for the combined levels of the X variable. These proportions represent the null hypothesis of no association between the X and Y variables.
- The scale of the vertical axis on the left side of the plot shows the response probability. The whole axis is equivalent to a probability of one, which represents the total sample.

Tip: You can click on a rectangle in the mosaic plot to highlight the selection and highlight the corresponding rows in the associated data table. You can right-click on a rectangle to add labels to the plot.

Mosaic Plot Pop-Up Menu

In the Contingency report, right-click the mosaic plot to change colors and label the cells. Each portion of the plot has an independent menu.

Set Colors Shows the current assignment of colors to levels and enables you to update the assignments. See "Set Colors".

Cell Labeling Specifies the label to be drawn in each cell of the mosaic plot.

- **No Labels** Shows no labels, and removes any labels that were previously added.
- **Label by Count** Shows the number of observations in each cell.
- **Label by Percent** Shows the percent of observations in each cell.
- **Label by Value** Shows the levels of the Y variable corresponding to each cell.
- **Label by Row** Shows the row labels for all of the rows represented by the cell.

Line Color Specifies a line color for the lines around each cell.

Line Style Specifies a line style for the lines around each cell.

Line Width Specifies the line width for the lines around each cell.

Transparency Specifies the transparency for the cell colors.

Note: For descriptions of the remainder of the options in the pop-up menu, see *Using JMP*.

Set Colors

When you select the Set Colors option in the mosaic plot pop-up menu, the Select Colors for Values window appears. When this window initially opens, it shows the current assignment of colors to levels.

The default mosaic colors depend on whether the Y response column is ordinal or nominal, and whether it has an existing Value Colors column property. To change the color for any level, click the oval in the second column of colors and select a new color.

The Select Colors for Values window contains the following options:

Macros Updates the colors using one of the following methods:

- **Gradient Between Ends** Applies a color gradient to all levels of the variable.
- **Gradient Between Selected Points** Applies a color gradient only to the levels that you have selected. You can select a range of levels by dragging the cursor over the levels that you want to select, or by pressing the Shift key and clicking the first and last level.

Reverse Colors Reverses the order of the colors.

Revert to Old Colors Reverts any changes that have been made since the Select Colors for Values window was opened.

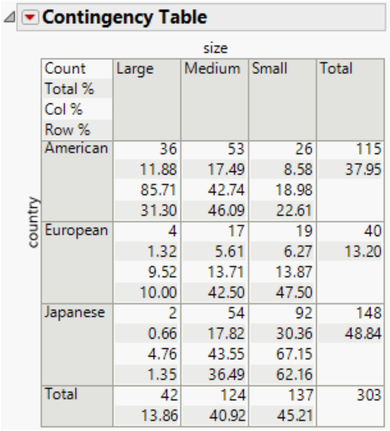
Color Theme Updates the colors for each value based on a color theme.

Save Colors to Column Enables you to save the updated colors to the data table. If you change the color theme and select this check box, a Value Colors column property is added to the column in the associated data table. To edit this property from the data table, select **Cols > Column Info**.

Contingency Table

The Contingency Table is a two-way frequency table. There is a row for each level of the X variable and a column for each level of the Y variable.

Figure 7.6 Example of a Contingency Table



Contingency Table

		size			
	Count Total % Col % Row %	Large	Medium	Small	Total
country	American	36	53	26	115
		11.88	17.49	8.58	37.95
		85.71	42.74	18.98	
		31.30	46.09	22.61	
	European	4	17	19	40
		1.32	5.61	6.27	13.20
		9.52	13.71	13.87	
		10.00	42.50	47.50	
	Japanese	2	54	92	148
		0.66	17.82	30.36	48.84
		4.76	43.55	67.15	
		1.35	36.49	62.16	
	Total	42	124	137	303
		13.86	40.92	45.21	

Note the following about contingency tables:

- The Count, Total%, Col%, and Row% correspond to the data within each cell that has row and column headings.
- The last column contains the total counts and percentages for each row.
- The bottom row contains total counts and percentages for each column.

In Figure 7.6, focus on the cars that are large and come from America. The following table explains the conclusions that you can make about these cars using the contingency table.

Table 7.1 Conclusions Based on Example of a Contingency Table

Number	Description	Label in Table
36	Number of cars that are both large and come from America.	Count
11.88%	Percentage of all 303 cars that are both large and come from America (36/303).	Total %
85.71%	Percentage of the 42 large cars that come from America (36/42).	Col %
31.30%	Percentage of the 115 American cars that are large (36/115).	Row %
37.95%	Percentage of all 303 cars that come from America (115/303).	Total %
13.86%	Percentage of all 303 cars that are large (42/303).	Total %

Tips:

- To show or hide statistics in the Contingency Table, from the Contingency Table red triangle menu, select the statistic that you want to show or hide.
- To change the format of statistics in the Contingency Table, right-click the value in the table and select Format.

Description of Statistics in the Contingency Table

In the Contingency platform, the Contingency Table contains the following statistics:

Count The cell frequency, margin total frequencies, and grand total. The grand total is also the total sample size.

Total% The percent of cell counts and margin totals relative to the grand total.

Row% The percent of each cell count relative to its row total.

Col% The percent of each cell count relative to its column total.

Expected The expected frequency (E) of each cell under the assumption of independence. Computed as the product of the corresponding row total and column total divided by the grand total.

Deviation The observed cell frequency (O) minus the expected cell frequency (E).

Cell Chi Square The chi-square values computed for each cell as $(O - E)^2 / E$.

Col Cum The cumulative column total, calculated from the top of the table to the bottom.

Col Cum% The cumulative column percentage, calculated from the top of the table to the bottom.

Row Cum The cumulative row total, calculated from the left of the table to the right.

Row Cum% The cumulative row percentage, calculated from the left of the table to the right.

Make Into Data Table Creates a new data table that contains the statistics that currently appear in the contingency table.

Format Opens a window that enables you to change the formatting of the statistics in the contingency table.

Tests Report

In the Contingency platform, the Tests report shows the results for two tests that can be used to evaluate whether the response level rates are independent of the levels of the X variable.

The Tests report contains the following statistics:

N The total number of observations.

DF The degrees of freedom associated with the tests. The degrees of freedom are equal to $(c - 1)(r - 1)$, where c is the number of columns and r is the number of rows.

-LogLike The negative log-likelihood, which measures fit and uncertainty (much like sums of squares in continuous response situations). See *Fitting Linear Models*.

Rsquare (U) The portion of the total uncertainty attributed to the model fit.

- An R^2 of 1 means that the levels of the X variable completely predict the levels of the Y variable.
- An R^2 of 0 means that the model predicts no better than using fixed background response rates.

Test The names of the chi-square tests of the hypothesis that the response rates are the same in each sample category.

ChiSquare The test statistic for the chi-square test.

Prob>ChiSq The probability of obtaining a chi-square value greater than the one computed if no relationship exists between the response and factor. If both variables have only two levels, Fisher’s exact probabilities for the one-tailed tests and the two-tailed test also appear.

For more information about the statistics in this report, see "Statistical Details for the Tests Report".

Fisher's Exact Test

This report gives the results of Fisher's exact test for a 2 x 2 table. The results appear automatically for 2 x 2 tables. For more information about Fisher's exact test and the test for *r* x *c* tables, see "Fisher's Exact Test Report".

Contingency Platform Options

The Contingency Analysis red triangle menu contains options to perform additional analyses.

Note: The Fit Group menu appears only if you have specified multiple Y or multiple X variables. Use the Fit Group menu options to arrange reports or order them by RSquare. See *Fitting Linear Models*.

The Contingency Analysis red triangle menu contains the following options:

Mosaic Plot Shows or hides a graphical representation of the contingency table. See "Mosaic Plot".

Contingency Table Shows or hides a two-way frequency table. The table contains a row for each level of the X variable and a column for each level of the Y variable. See "Contingency Table".

Tests Shows or hides tests that measure if the response level rates are the same across the levels of the X variable. These tests are analogous to the Analysis of Variance table for continuous data. See "Tests Report".

Set α Level Changes the alpha level used in confidence intervals. Select one of the common values (0.10, 0.05, 0.01) or select a specific value using the **Other** option.

Analysis of Means for Proportions (Available only if the response has exactly two levels.) Shows or hides an analysis of means for proportions (ANOMP) decision chart to compare group proportions. ANOMP is a multiple comparison procedure, which compares the response proportions for the levels of the X variable to the overall response proportion. See "Analysis of Means for Proportions Report".

Correspondence Analysis Shows or hides a correspondence analysis, which identifies rows or columns of a frequency table that have similar patterns of counts. In the correspondence analysis plot, there is a point for each row and for each column of the contingency table. See "Correspondence Analysis Report".

Cochran Mantel Haenszel Shows or hides a test that determines whether there is a relationship between two categorical variables after blocking across a third classification variable. When you select this option, a window appears that enables you to specify the grouping column for the test. See "Cochran-Mantel-Haenszel Test Report".

Agreement Statistic (Available only when the X and Y variables have the same levels.) Shows or hides a report that contains statistics that measure the agreement between levels. The report includes the Kappa statistic (Agresti 1990), as well as the standard error, confidence interval, and hypothesis test for the statistic. The report also includes Bowker's test of symmetry, which is also known as McNemar's test. See "Agreement Statistic Report".

Relative Risk (Available only when the X and Y variables each have exactly two levels.) Shows or hides the relative risk between the levels of the response. See "Relative Risk Report".

The Relative Risk report also gives a confidence interval for this ratio. You can change the alpha level using the Set α Level option.

Risk Difference (Available only when the X and Y variables each have exactly two levels.) Shows or hides the difference in risks between the levels of the response.

The Risk Difference report also gives a confidence interval for this ratio. You can change the alpha level using the Set α Level option.

Odds Ratio (Available only when the X and Y variables each have exactly two levels.) Shows or hides a report of the odds ratio. See "Statistical Details for the Odds Ratio".

The Odds Ratio report also gives a confidence interval for this ratio. You can change the alpha level using the Set α Level option.

Two Sample Test for Proportions (Available only when the X and Y variables each have exactly two levels.) Shows or hides a two-sample test for proportions. This test compares the proportions of the Y variable between the two levels of the X variable. See "Two Sample Test for Proportions Report".

Measures of Association Shows or hides a report that contains measures of the association between the variables in the contingency table. See "Measures of Association Report".

Cochran Armitage Trend Test (Available only when one variable has exactly two levels and the other variable is ordinal.) Shows or hides a test for trends in binomial proportions across levels of a single variable. See "Cochran Armitage Trend Test Report".

Exact Test Contains options for exact versions of the following tests:

Fisher's Test Shows or hides Fisher's exact test for testing the association between two categorical variables. This test does not depend on any large-sample distributional assumptions. See "Fisher's Exact Test Report".

Cochran Armitage Trend Test (Available only when one of the variables has exactly two levels.) Shows or hides the exact version of the Cochran-Armitage trend test. See "Cochran Armitage Trend Test Report".

Agreement Statistic (Available only when one of the variables has exactly two levels.) Shows or hides the exact version of the agreement statistic Kappa. See "Agreement Statistic Report".

Note: Exact tests are not available when there are non-integer values in the Freq variable. Also, if the overall sample size is larger than 32767 and the contingency table is larger than 2 x 2, the exact test options are not available.

Equivalence Tests (Available only when the X and Y variables each have exactly two levels.) Contains the following options for testing equivalence, superiority, or non-inferiority between response categories. See "Equivalence Test Reports".

Risk Difference Launches a window with options for equivalence, superiority, or non-inferiority tests for the difference in risks between the levels of the response.

Relative Risk Launches a window with options for equivalence, superiority, or non-inferiority tests for the relative risk between the levels of the response.

Display Options Contains the following option to modify the mosaic plot:

Horizontal Mosaic Rotates the mosaic plot horizontally or vertically.

Make Into Data Table Creates a JMP data table from the report table.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Contingency Analysis Reports

The Contingency platform enables you to create visualizations and investigate the relationship between two categorical variables. This section contains information about the reports that are generated for specific analysis options.

- "Analysis of Means for Proportions Report"
- "Correspondence Analysis Report"
- "Cochran-Mantel-Haenszel Test Report"
- "Agreement Statistic Report"
- "Relative Risk Report"
- "Two Sample Test for Proportions Report"
- "Measures of Association Report"
- "Cochran Armitage Trend Test Report"
- "Fisher's Exact Test Report"
- "Equivalence Test Reports"

Analysis of Means for Proportions Report

In the Contingency platform, the analysis of means for proportions (ANOMP) is a multiple comparison method for testing if individual group proportions differ from the overall proportion. For more information about analysis of means methods, see Nelson et al. (2005). See also "Example of Analysis of Means for Proportions".

If the Y variable has two levels, you can use this option to compare response proportions for the levels of the X variable to the overall response proportion. This method uses a normal approximation to the binomial distribution. Therefore, if the sample sizes are too small, a warning appears in the results.

The following options appear in the Analysis of Means for Proportions red triangle menu:

Set Alpha Level Specifies the alpha level used to determine the decision limits in the Analysis of Means for Proportions chart.

Show Summary Report Shows or hides a report that contains the response proportions and decision limits for each level of the X variable. The report also indicates whether a limit has been exceeded.

Switch Response Level for Proportion Changes the response category that is used in the analysis.

Display Options Contains the following options to modify the analysis of means for proportions chart:

Show Decision Limits Shows or hides the decision limit lines in the Analysis of Means for Proportions chart.

Show Decision Limit Shading Shows or hides the decision limit shading in the Analysis of Means for Proportions chart.

Show Center Line Shows or hides the center line for the Analysis of Means in Proportions chart.

Point Options Specifies the drawing style of the points in the Analysis of Means for Proportions chart. You can choose between vertical needles, connected points, and points only. By default, the chart is drawn with needles that connect the points to the horizontal line that is drawn at the average.

Correspondence Analysis Report

In the Contingency platform, the Correspondence Analysis option provides a graphical technique to visualize which rows or columns of a frequency table have similar patterns of counts. In the correspondence analysis plot, there is a point for each row and for each column. Use Correspondence Analysis when you have many levels, which might make it difficult to derive useful information from the mosaic plot. See "Example of Correspondence Analysis".

Correspondence Analysis plots contain row and column profiles. The *row profile* can be defined as the set of rowwise rates, or in other words, the counts in a row divided by the total count for that row. If two rows have very similar row profiles, their points in the correspondence analysis plot are close together. Squared distances between row points are approximately proportional to chi-square test statistics that test the homogeneity between the pair of rows.

Column and row profiles are alike because the problem is defined symmetrically. The distance between a row point and a column point has no meaning. However, the directions of columns and rows from the origin are meaningful, and the relationships between the points can help you interpret the plot.

Use the options in the Correspondence Analysis red triangle menu to add a three-dimensional scatterplot to the Contingency report and to add column properties to the data table.

3D Correspondence Analysis Shows or hides a three-dimensional scatterplot.

Save Value Order Saves a Value Ordering column property to both the X and Y variable columns in the data table. The column property specifies the order of the levels sorted by the first correspondence score coefficient.

Details Report

In the Contingency platform, the Details report contains statistical information about the correspondence analysis and shows the values used in the plot.

Singular Value The singular value decomposition of the contingency table. See “Statistical Details for Correspondence Analysis”.

Inertia The square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

Portion The portion of inertia with respect to the total inertia.

Cumulative The cumulative portion of inertia.

Tip: If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

X Variable c1, c2, c3 The values plotted on the Correspondence Analysis plot.

Y Variable c1, c2, c3 The values plotted on the Correspondence Analysis plot.

Cochran-Mantel-Haenszel Test Report

In the Contingency platform, the Cochran-Mantel-Haenszel test evaluates the relationship between two categorical variables after blocking across a third classification. Blocking across a third classification variable is also called stratification.

Note: See “Example of a Cochran-Mantel-Haenszel Test”.

The Cochran-Mantel-Haenszel Tests report contains a table with the results of various tests. Each test has a chi-square statistic (ChiSquare), associated degrees of freedom (DF), and significance probability (Prob>Chisq). The following tests are reported:

Correlation of Scores (Applicable when both the Y and X variables are ordinal or interval.) Tests the alternative hypothesis that there is a linear association between the Y and X variables for at least one level of the blocking variable.

Row Score by Col Categories (Applicable when the Y variable is ordinal or interval.) Tests the alternative hypothesis that, for at least one level of the blocking variable, the mean scores of the rows are unequal.

Col Score by Row Categories (Applicable when X is ordinal or interval.) Tests the alternative hypothesis that, for at least one level of the blocking variable, the mean scores of the columns are unequal.

General Assoc. of Categories Tests that for at least one level of the blocking variable, there is some type of association between the Y and X variables.

Tip: You can perform a row means score test by specifying a column that contains a constant value as the blocking variable in the Cochran-Mantel-Haenszel test.

Agreement Statistic Report

When two categorical variables have the same levels, use the Agreement Statistic option in the Contingency platform to compute the Kappa statistic (Agresti 1990). This option also computes the Kappa statistic standard error, confidence interval, hypothesis test, and Bowker's test of symmetry. See "Example of the Agreement Statistic Option".

The Kappa statistic and associated *p*-value in this report are approximate. See "Statistical Details for the Agreement Statistic". An exact version of the agreement statistic is available. See "Fisher's Exact Test Report".

The Kappa Coefficient report contains the following statistics:

Kappa The Kappa statistic.

Std Err The standard error of the Kappa statistic.

Lower 95% The lower endpoint of the confidence interval for Kappa.

Upper 95% The upper endpoint of the confidence interval for Kappa.

Prob>Z The significance probability (*p*-value) for a one-sided asymptotic test for Kappa. The null hypothesis tests if Kappa equals zero.

Prob>|Z| The significance probability (*p*-value) for a two-sided asymptotic test for Kappa.

The Bowker's Test report contains the following statistics:

ChiSquare The test statistic for Bowker's test. For Bowker's test of symmetry, the null hypothesis is that the probabilities in the square table satisfy symmetry, or that $p_{ij}=p_{ji}$ for all pairs of table cells. When both the Y and X variables have exactly two levels, Bowker's test is equivalent to McNemar's test.

Prob>ChiSq The significance probability (*p*-value) for the Bowker's test statistic. The null hypothesis of Bowker's test is that the probabilities in the square table satisfy symmetry.

Relative Risk Report

In the Contingency platform, use the Relative Risk option to calculate risk ratios for 2 x 2 contingency tables. Confidence intervals also appear in the report. For more information about this method, see Agresti (1990, sect. 3.4.2). See "Example of the Relative Risk Option".

When you select the Relative Risk option, the Choose Relative Risk Categories window appears. You can select a single response and factor combination, or you can calculate the risk ratios for all combinations of response and factor levels.

Two Sample Test for Proportions Report

In the Contingency platform, you can construct a confidence interval and perform a hypothesis test for the difference between two proportions. This analysis is available when both the X and Y variables have exactly two levels. See "Example of a Two Sample Test for Proportions".

Description The description of the test being performed.

Proportion Difference The difference in the proportions between the levels of the X variable.

Lower 95% The lower endpoint of the confidence interval for the difference. This value is based on the adjusted Wald confidence interval.

Upper 95% The upper endpoint of the confidence interval for the difference. This value is based on the adjusted Wald confidence interval.

Adjusted Wald Test (Null Hypothesis) The description of the null hypotheses for the one-tailed and two-tailed tests.

Prob The significance probabilities (*p*-values) for the tests.

Tip: Use the radio buttons below the table to change the response level of interest for the test.

Measures of Association Report

In the Contingency platform, the Measures of Association option provides association statistics. See "Example of the Measures of Association Option".

The Measures of Association report contains the value, standard error, and confidence intervals for the following statistics:

Gamma A measure of ordinal association. Defined as the difference in probability of concordant and discordant pairs disregarding tied pairs. Takes values in the range -1 to 1.

Kendall's Tau-b Similar to Gamma with a correction for the number of ties. Takes values in the range -1 to 1.

Stuart's Tau-c Similar to Gamma with an adjustment for table size and a correction for ties. Takes values in the range -1 to 1.

Somers' D An asymmetric modification of Tau-b. Somers' D uses a correction for ties only when the pair is tied on the independent variable. Takes values in the range -1 to 1.

- The C|R denotes that the row variable X is regarded as an independent variable and the column variable Y is regarded as a dependent variable.
- Similarly, the R|C denotes that the column variable Y is regarded as an independent variable and the row variable X is a dependent variable.

Note: Gamma, Kendall's Tau-b, Stuart's Tau-c, and Somers' D are measures of ordinal association that consider whether the variable Y tends to increase as X increases. They classify pairs of observations as concordant or discordant. A pair is concordant if an observation with a larger value of X also has a larger value of Y. A pair is discordant if an observation with a larger value of X has a smaller value of Y. These measures are appropriate only when both variables are ordinal.

Lambda Asymmetric Differs for C|R and R|C. Takes values in the range 0 to 1.

- For C|R, this measure is interpreted as the probable improvement in predicting the column variable Y given knowledge of the row variable X.
- For R|C, this measure is interpreted as the probable improvement in predicting the row variable X given knowledge about the column variable Y.

Lambda Symmetric Loosely interpreted as the average of the two Lambda Asymmetric measures. Takes values in the range 0 to 1.

Uncertainty Coef Differs for C|R and R|C. Takes values in the range 0 to 1.

- For C|R, this measure is the proportion of uncertainty in the column variable Y that is explained by the row variable X.
- For R|C, this measure is interpreted as the proportion of uncertainty in the row variable X that is explained by the column variable Y.

Uncertainty Coef Symmetric Symmetric version of the two Uncertainty Coef measures. Takes values in the range 0 to 1.

Note: The Lambda and Uncertainty measures are appropriate for both ordinal and nominal variables.

For computational details about the measures of association statistics, see the FREQ Procedure chapter in SAS Institute Inc. (2020b). The following references also contain additional information:

- Brown and Benedetti (1977)
- Goodman and Kruskal (1979)
- Kendall and Stuart (1979)
- Snedecor and Cochran (1980)
- Somers (1962)

Cochran Armitage Trend Test Report

In the Contingency platform, the Cochran Armitage Trend Test option tests for trends in the binomial proportions across the levels of a single variable. This test is appropriate only when one variable has two levels and the other variable is ordinal. The two-level variable represents the response, and the other represents an explanatory variable with ordered levels. The null hypothesis is the hypothesis of no trend, which means that the binomial proportion is the same for all levels of the explanatory variable. See "Example of the Cochran Armitage Trend Test".

Note: The test statistic and significance probabilities (*p*-values) given in this test are approximate. An exact version of the trend test is also available.

Fisher's Exact Test Report

Fisher's exact test for an *r* x *c* table tests for association between two variables. An *r* x *c* table is one where the variable assigned to the rows has *r* levels and the variable assigned to the columns has *c* levels. Fisher's exact test assumes that the row and column totals are fixed, and uses the hypergeometric distribution to compute probabilities.

This test does not depend on any large-sample distribution assumptions. This means that it is appropriate for situations where the likelihood ratio and Pearson tests become less reliable, such as small sample sizes or sparse tables.

The Fisher's Exact Test report includes the following information:

Table Probability (P) The probability for the observed table. This is not the *p*-value for the test.

Two-sided Prob ≤ P The significance probability (*p*-value) for the two-sided test.

For 2 x 2 tables, the Fisher's Exact test is automatically performed, unless one row or column contains all zeros (in this case, the test cannot be calculated). See "Tests Report".

Equivalence Test Reports

In the Contingency platform, the options in the Equivalence Tests submenu enable you to perform equivalence, superiority, or noninferiority tests for risk differences or relative risks.

The equivalence test reports contain plots and summary tables. When you select an option in the Equivalence Tests submenu, you must specify test characteristics in the Equivalence Test Specification window.

Equivalence Test Specification Window

Selecting an equivalence test option in the Equivalence Tests submenu launches a window that enables you to define your test.

Alternative Hypothesis Defines the structure of the equivalence test.

- **Equivalence (Two-Sided)** Specifies an equivalence test. Use this option when the goal is to show that group differences are not bigger than the equivalence margin.
- **Superiority (One-Sided)** Specifies a superiority test. Use this option when the goal is to show that a group is superior, or better than, another group.
- **Noninferiority (One-Sided)** Specifies a non-inferiority test. Use this option when the goal is to show that a group is not inferior to another group.

Side of the Alternative Hypothesis (Available only for one-sided tests.) Specifies the direction of the alternative hypothesis.

Hypotheses Plot Provides a graphical depiction of the hypothesis test.

Y Category of Interest Defines the level of the Y variable that is the numerator in the ratio.

X Category of Interest Defines the level of the X variable that is the denominator in the ratio of the control.

Margin and Alpha Defines the significance levels for the test.

- **Difference** (Available only for tests for risk difference.) Specifies the equivalence, superiority, or non-inferiority margin. This margin, or delta, is the difference that has practical significance. For equivalence tests, the difference must be greater than zero.
- **Ratio** (Available only for tests for relative risk.) Specifies the equivalence, superiority, or non-inferiority margin as a risk ratio. This margin, or delta, defines a risk ratio that has practical significance. The range of values is defined as (Ratio, 1/Ratio). For equivalence tests, the ratio must be different from 1.
- **Alpha** Specifies the significance level for the test.

Equivalence Tests Report

The test report begins with a description of the alternative hypothesis begin tested. For each comparison, the Tests report contains the following columns:

Difference (Available only for tests for risk difference.) The estimated risk difference.

Ratio (Available only for tests for relative risk.) The estimated risk ratio.

Std Error of Difference (Available only for tests for risk difference.) The estimated standard error of the risk difference.

Lower Bound Std Error, Upper Bound Std Error (Only one bound appears for one-sided tests.) The standard errors estimated under the lower or upper hypothetical value.

Lower Bound z Ratio, Upper Bound z Ratio (Only one bound appears for one-sided tests.) The lower or upper bound *z* ratios for the one-sided significance tests.

Lower Bound p-Value, Upper Bound p-Value (Only one *p*-value appears for one-sided tests.) The significance probabilities (*p*-values) that correspond to the lower or upper bound *z* ratios.

Max p-Value (Appears only for two-sided tests.) Maximum of the lower and upper bound *p*-values.

Lower 90%, Upper 90% Limits for a 1–2α confidence interval for the risk difference or risk ratio.

Assessment An assessment of the hypothesis test for the specified alpha level.

Equivalence Tests Options

The Equivalence, Superiority, or Non-inferiority Tests red triangle menu contains the following options:

Test Report Shows or hides a report that summarizes the equivalence tests, superiority tests, or non-inferiority tests for risk differences or risk ratios. See "Equivalence Test Reports".

Forest Plot Shows or hides a forest plot. The comparison confidence intervals are plotted versus the risk difference or relative risk. The intervals are plotted on a risk difference or relative risk scale. Shading indicates the equivalent, superior, or non-inferiority regions.

Tip: Hover over a point to show the groups being compared and the estimated risk difference or relative risk.

Remove Removes the test report from the Contingency Analysis report window.

Additional Examples of the Contingency Platform

This section contains examples using the Contingency platform.

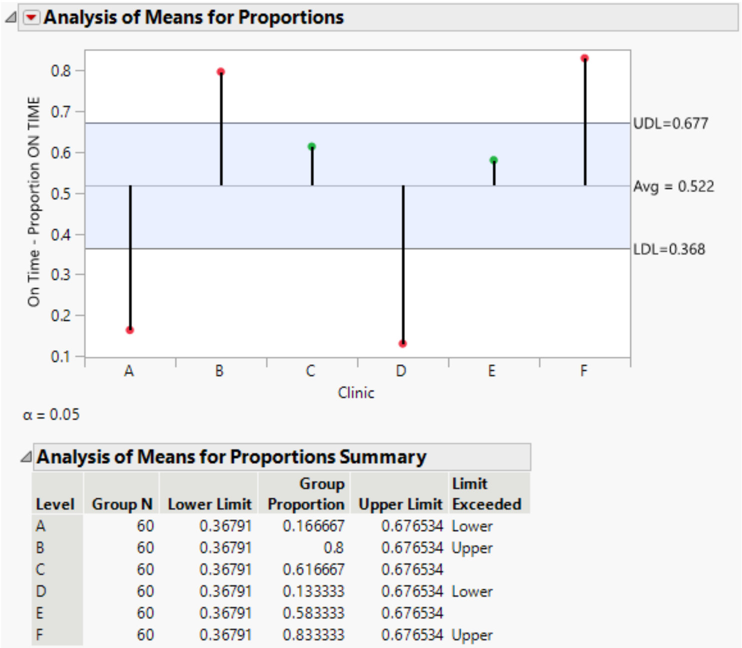
- "Example of Analysis of Means for Proportions"
- "Example of Correspondence Analysis"
- "Example of a Cochran-Mantel-Haenszel Test"
- "Example of the Agreement Statistic Option"
- "Example of the Relative Risk Option"
- "Example of a Two Sample Test for Proportions"
- "Example of the Measures of Association Option"
- "Example of the Cochran Armitage Trend Test"

Example of Analysis of Means for Proportions

Use the Contingency platform to examine the proportion of patients that arrived on-time to their appointments at six clinics in a geographic region. 60 random appointments were selected from 1 week of records for each of the six clinics. To be considered on-time, the patient must be taken to an exam room within five minutes of their scheduled appointment time.

1. Select **Help > Sample Data Folder** and open Office Visits.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select On Time and click **Y, Response**.
4. Select Clinic and click **X, Factor**.
5. Select Frequency and click **Freq**.
6. Click **OK**.
7. Click the Contingency Analysis red triangle menu and select **Analysis of Means for Proportions**.
8. Click the Analysis of Means for Proportions red triangle menu and select **Show Summary Report** and **Switch Response Level for Proportion**.

Figure 7.7 Example of Analysis of Means for Proportions



Analysis of Means for Proportions Summary

Level	Group N	Lower Limit	Group Proportion	Upper Limit	Limit Exceeded
A	60	0.36791	0.166667	0.676534	Lower
B	60	0.36791	0.8	0.676534	Upper
C	60	0.36791	0.616667	0.676534	
D	60	0.36791	0.133333	0.676534	Lower
E	60	0.36791	0.583333	0.676534	
F	60	0.36791	0.833333	0.676534	Upper

The analysis of means plot shows the proportion of patients who were on-time from each clinic. Notice the following:

- The proportion of on-time arrivals is the highest for clinic F, followed by clinic B.
- Clinic D has the lowest proportion of on-time arrivals, followed by clinic A.
- Clinic E and clinic C are close to the average, and do not exceed the decision limits.

Example of Correspondence Analysis

Use the Contingency platform to perform a correspondence analysis for a cheese tasting experiment. The experiment recorded the counts of nine different response levels across four different cheese additives. Correspondence analysis provides a visualization of the associations across levels when an association across variables is established.

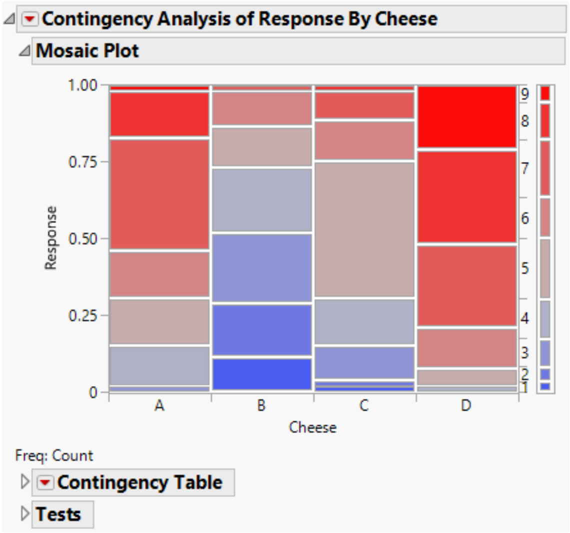
1. Select **Help > Sample Data Folder** and open Cheese.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.

   The Response values range from one to nine, where one is the least liked, and nine is the best liked.
4. Select Cheese and click **X, Factor**.

A, B, C, and D represent four different cheese additives.

5. Select Count and click **Freq**.
6. Click **OK**.

Figure 7.8 Mosaic Plot for the Cheese Data



Freq: Count

From the mosaic plot, notice that the rankings differ across the cheese types. In particular, Cheese B is ranked consistently lower than the other cheeses. A correspondence analysis might be the next step in the analysis.

7. To see the correspondence analysis plot, click the Contingency Analysis red triangle menu and select **Correspondence Analysis**.

Figure 7.9 Example of a Correspondence Analysis Plot

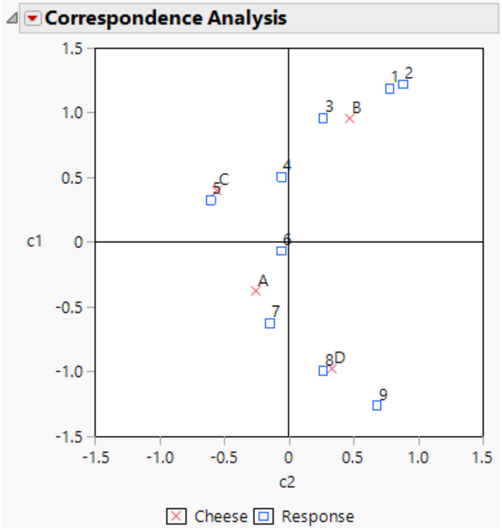


Figure 7.9 shows the correspondence analysis graphically, where the plot axes are labeled c1 and c2. Notice the following:

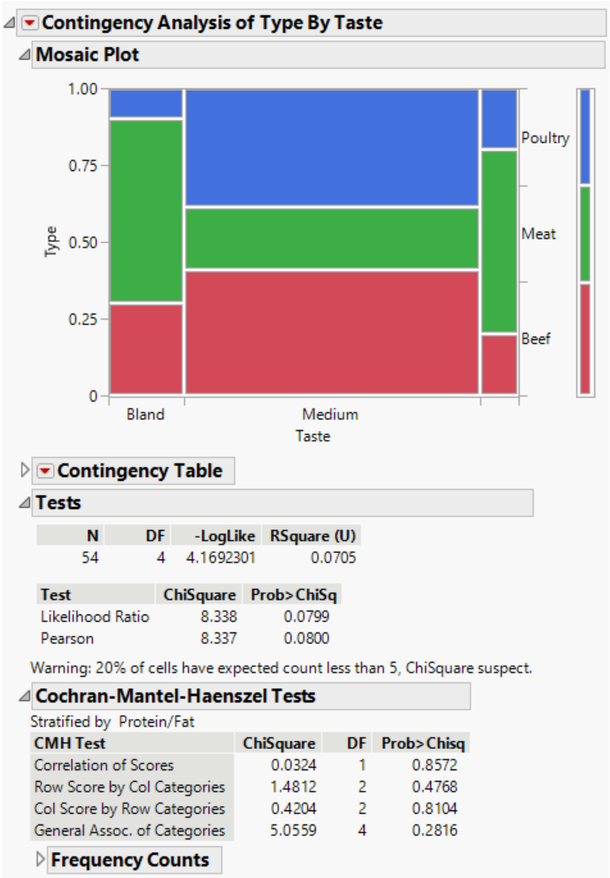
- The c1 axis appears to explain the general satisfaction level. The cheeses on the c1 axis go from least liked at the top to most liked at the bottom.
- Cheese D is the most liked cheese, with responses of 8 and 9.
- Cheese B is the least liked cheese, with responses of 1, 2, and 3.
- Cheeses C and A are in the middle, with responses of 4, 5, 6, and 7.

Example of a Cochran-Mantel-Haenszel Test

Use the Contingency platform to examine the relationship between hot dog type and taste.

1. Select **Help > Sample Data Folder** and open Hot Dogs.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Type and click **Y, Response**.
4. Select Taste and click **X, Factor**.
5. Click **OK**.
6. Click the Contingency Analysis red triangle menu and select **Cochran Mantel Haenszel**.
7. Select Protein/Fat as the grouping variable and click **OK**.

Figure 7.10 Example of a Cochran-Mantel-Haenszel Test



N	DF	-LogLike	RSquare (U)
54	4	4.1692301	0.0705

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	8.338	0.0799
Pearson	8.337	0.0800

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Stratified by Protein/Fat

CMH Test	ChiSquare	DF	Prob>Chisq
Correlation of Scores	0.0324	1	0.8572
Row Score by Col Categories	1.4812	2	0.4768
Col Score by Row Categories	0.4204	2	0.8104
General Assoc. of Categories	5.0559	4	0.2816

Notice the following:

- The Tests report shows a marginally significant chi-square probability of about 0.0799, indicating some significance in the relationship between hot dog taste and type.
- The Cochran-Mantel-Haenszel report shows that the *p*-value for the general association of categories is 0.2816. There does not appear to be a relationship between hot dog taste and type when the comparison is controlled by the protein/fat content.

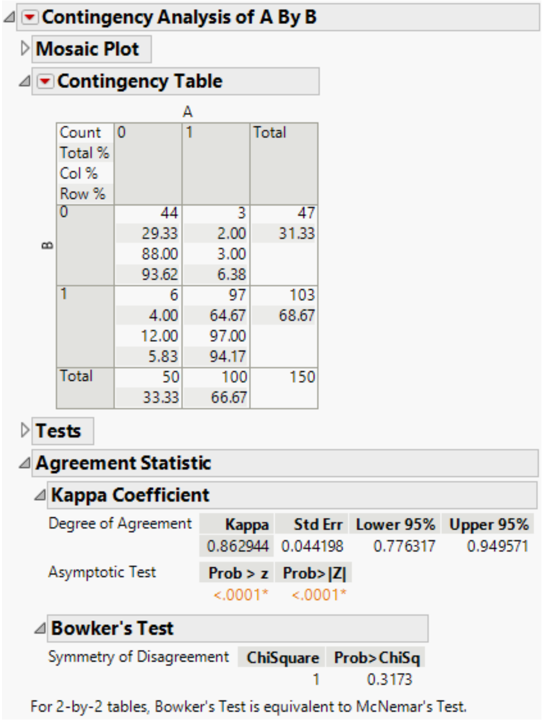
Example of the Agreement Statistic Option

Use the Contingency platform to examine the relationship between two raters. The data table contains results from three people (raters) who rated fifty parts three times each. Examine the relationship between raters A and B.

1. Select **Help > Sample Data Folder** and open Attribute Gauge.jmp.

2. Select **Analyze > Fit Y by X**.
3. Select A and click **Y, Response**.
4. Select B and click **X, Factor**.
5. Click **OK**.
6. Click the Contingency Analysis red triangle menu and select **Agreement Statistic**.

Figure 7.11 Example of the Agreement Statistic Report



Contingency Analysis of A By B

Mosaic Plot

Contingency Table

		A		
	Count Total % Col % Row %	0	1	Total
B	0	44 29.33 88.00 93.62	3 2.00 3.00 6.38	47 31.33
	1	6 4.00 12.00 5.83	97 64.67 97.00 94.17	103 68.67
	Total	50 33.33	100 66.67	150

Tests

Agreement Statistic

Kappa Coefficient

Degree of Agreement	Kappa	Std Err	Lower 95%	Upper 95%
	0.862944	0.044198	0.776317	0.949571

Asymptotic Test	Prob > z	Prob>|Z|
	<.0001*	<.0001*

Bowker's Test

Symmetry of Disagreement	ChiSquare	Prob>ChiSq
	1	0.3173

For 2-by-2 tables, Bowker's Test is equivalent to McNemar's Test.

You notice that the agreement statistic of 0.86 is high (close to 1) and the *p*-value of <.0001 is small. This reinforces the high agreement seen by looking at the diagonal of the contingency table. Agreement between the raters occurs when both raters give a rating of 0 or both give a rating of 1.

Example of the Relative Risk Option

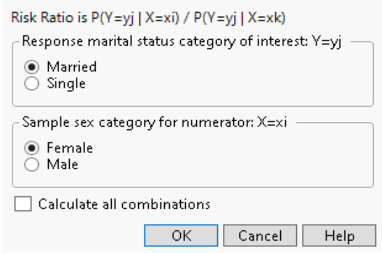
Use the Contingency platform to examine the relative probabilities of being married and single for the participants in a survey.

1. Select **Help > Sample Data Folder** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.

3. Select marital status and click **Y, Response**.
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. Click the Contingency Analysis red triangle menu and select **Relative Risk**.

   The Choose Relative Risk Categories window appears.

Figure 7.12 The Choose Relative Risk Categories Window



Risk Ratio is P(Y=yj | X=xi) / P(Y=yj | X=xk)

Response marital status category of interest: Y=yj
● Married
○ Single

Sample sex category for numerator: X=xi
● Female
○ Male

☐ Calculate all combinations

Note the following about the Choose Relative Risk Categories window:

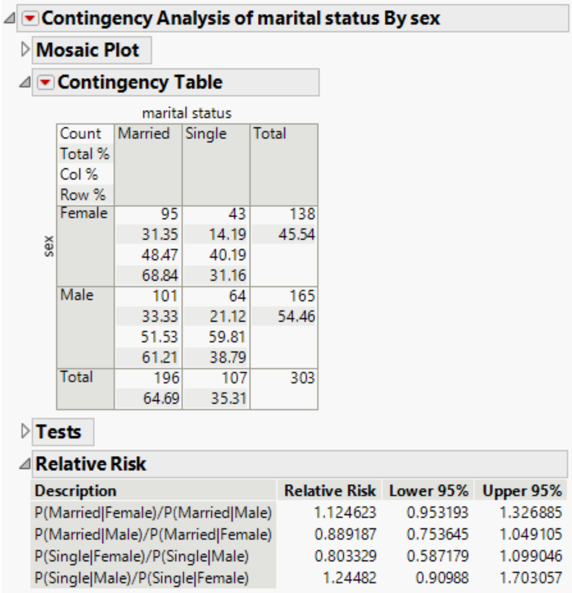
- If you are interested in only a single response and factor combination, you can select that here. For example, if you clicked **OK** in the window in Figure 7.12, this is the calculation:

$$\frac{P(Y = \text{Married} | X = \text{Female})}{P(Y = \text{Married} | X = \text{Male})}$$

- If you would like to calculate the risk ratios for all (2 × 2 = 4) combinations of response and factor levels, select the **Calculate All Combinations** check box (Figure 7.13).

7. Ask for all combinations by selecting the **Calculate All Combinations** check box. Leave all other default selections as is.

Figure 7.13 Example of the Risk Ratio Report



		marital status		
	Count Total % Col % Row %	Married	Single	Total
sex	Female	95 31.35 48.47 68.84	43 14.19 40.19 31.16	138 45.54
	Male	101 33.33 51.53 61.21	64 21.12 59.81 38.79	165 54.46
	Total	196 64.69	107 35.31	303

Description	Relative Risk	Lower 95%	Upper 95%
P(Married|Female)/P(Married|Male)	1.124623	0.953193	1.326885
P(Married|Male)/P(Married|Female)	0.889187	0.753645	1.049105
P(Single|Female)/P(Single|Male)	0.803329	0.587179	1.099046
P(Single|Male)/P(Single|Female)	1.24482	0.90988	1.703057

The relative risk is calculated as follows:

1. Examine the first entry in the Relative Risk report, which is P(Married | Female)/P(Married | Male).
2. You can find these probabilities in the Contingency Table. Since the probabilities are computed based on two levels of sex, which differs across the rows of the table, use the Row% to read the probabilities:

   $$P(\text{Married} \mid \text{Female}) = 0.6884$$

   $$P(\text{Married} \mid \text{Male}) = 0.6121$$

   Therefore, the relative risk is calculated as follows:

   $$P(\text{Married} \mid \text{Female})/P(\text{Married} \mid \text{Male}) = \frac{0.6884}{0.6121} = 1.1247$$

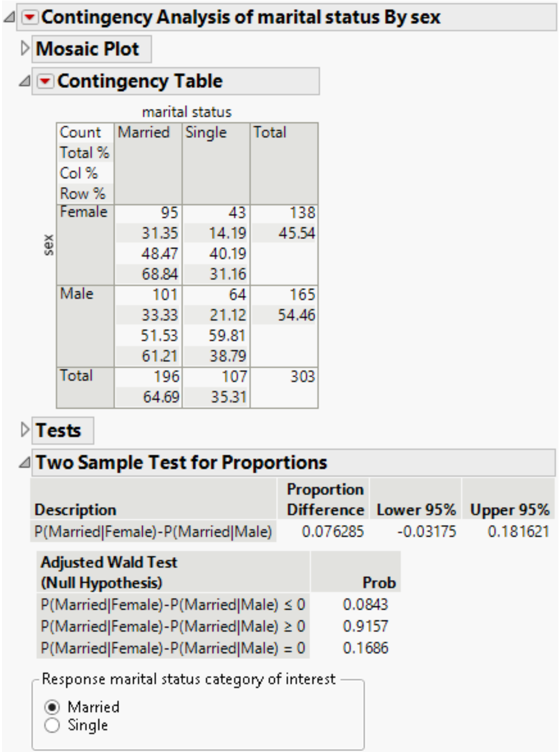
Example of a Two Sample Test for Proportions

Use the Contingency platform to examine the probability of being married for males and females.

1. Select **Help > Sample Data Folder** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select marital status and click **Y, Response**.
4. Select sex and click **X, Factor**.

5. Click **OK**.
6. Click the Contingency Analysis red triangle menu and select **Two Sample Test for Proportions**.

Figure 7.14 Example of the Two Sample Test for Proportions Report



Contingency Analysis of marital status By sex

Mosaic Plot

Contingency Table

marital status

sex	Count Total % Col % Row %	Married	Single	Total
	Female	95 31.35 48.47 68.84	43 14.19 40.19 31.16	138 45.54
	Male	101 33.33 51.53 61.21	64 21.12 59.81 38.79	165 54.46
	Total	196 64.69	107 35.31	303

Tests

Two Sample Test for Proportions

Description	Proportion Difference	Lower 95%	Upper 95%
P(Married|Female)-P(Married|Male)	0.076285	-0.03175	0.181621

Adjusted Wald Test (Null Hypothesis)	Prob
P(Married|Female)-P(Married|Male) ≤ 0	0.0843
P(Married|Female)-P(Married|Male) ≥ 0	0.9157
P(Married|Female)-P(Married|Male) = 0	0.1686

Response marital status category of interest

● Married
○ Single

In this example, you are comparing the probability of being married between females and males. See the Row% in the Contingency Table to obtain the following:

$$P(\text{Married} \mid \text{Female}) = 0.6884$$

$$P(\text{Married} \mid \text{Male}) = 0.6121$$

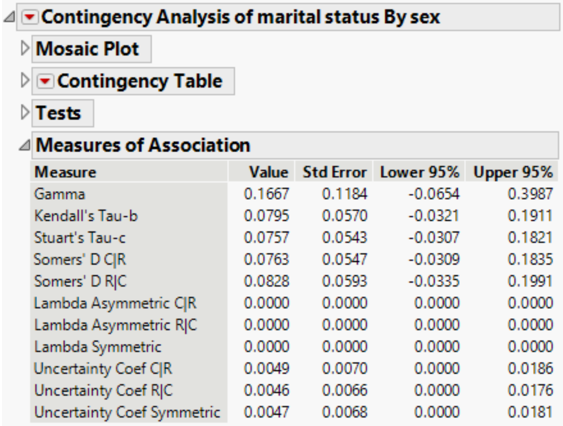
The difference between these two numbers, 0.0763, is the Proportion Difference shown in the report. The two-sided confidence interval is [-0.03175, 0.181621]. The *p*-value by the adjusted Wald method corresponding to the confidence interval is 0.1686, which is close to the *p*-value (0.1665) by Pearson's chi-square test. The Pearson's chi-square test is more common than the modified Wald's test for testing the difference of two proportions.

Example of the Measures of Association Option

Use the Contingency platform to examine the association of being married with gender.

1. Select **Help > Sample Data Folder** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select marital status and click **Y, Response**.
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. Click the Contingency Analysis red triangle menu and select **Measures of Association**.

Figure 7.15 Example of the Measures of Association Report



Measure	Value	Std Error	Lower 95%	Upper 95%
Gamma	0.1667	0.1184	-0.0654	0.3987
Kendall's Tau-b	0.0795	0.0570	-0.0321	0.1911
Stuart's Tau-c	0.0757	0.0543	-0.0307	0.1821
Somers' D C|R	0.0763	0.0547	-0.0309	0.1835
Somers' D R|C	0.0828	0.0593	-0.0335	0.1991
Lambda Asymmetric C|R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000	0.0000	0.0000
Uncertainty Coef C|R	0.0049	0.0070	0.0000	0.0186
Uncertainty Coef R|C	0.0046	0.0066	0.0000	0.0176
Uncertainty Coef Symmetric	0.0047	0.0068	0.0000	0.0181

Since the variables that you want to examine (sex and marital status) are nominal, use the Lambda and Uncertainty measures. All of them are small, so it seems that there is a weak association between sex and marital status.

Example of the Cochran Armitage Trend Test

Use the Contingency platform to investigate whether there is a relationship between the proportion of males and females that buy different sizes of cars.

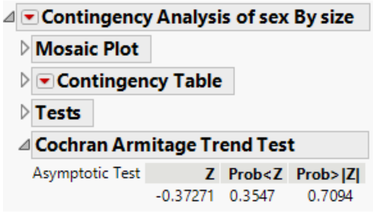
1. Select **Help > Sample Data Folder** and open Car Poll.jmp.

For the purposes of this test, change size to an ordinal variable:

2. In the Columns panel, right-click the icon next to size and select **Ordinal**.
3. Select **Analyze > Fit Y by X**.
4. Select sex and click **Y, Response**.

5. Select size and click **X, Factor**.
6. Click **OK**.
7. Click the Contingency Analysis red triangle menu and select **Cochran Armitage Trend Test**.

Figure 7.16 Example of the Cochran Armitage Trend Test Report



Asymptotic Test	Z	Prob<Z	Prob>|Z|
	-0.37271	0.3547	0.7094

The two-sided *p*-value (0.7094) is large. From this, you cannot conclude that there is a relationship in the proportion of male and females that purchase different sizes of cars.

Statistical Details for the Contingency Platform

This section contains statistical details for the Contingency platform.

- "Statistical Details for the Agreement Statistic"
- "Statistical Details for the Odds Ratio"
- "Statistical Details for the Tests Report"
- "Statistical Details for Correspondence Analysis"

Statistical Details for the Agreement Statistic

This section contains details for the agreement statistic in the Contingency platform. Viewing the two response variables as two independent ratings of the *n* subjects, the Kappa coefficient equals +1 when there is complete agreement of the raters. When the observed agreement exceeds the amount of agreement expected just by chance, the Kappa coefficient is positive and its magnitude reflects the strength of agreement. Although unusual in practice, Kappa is negative when the observed agreement is less than the amount of agreement expected just by chance. The minimum value of Kappa depends on the marginal proportions, but it is always between -1 and 0.

The Kappa coefficient is computed as follows:

$$\hat{\kappa} = \frac{P_0 - P_c}{1 - P_c} \text{ where } P_0 = \sum_i p_{ii} \text{ and } P_c = \sum_i p_{i+} p_{+i}$$

Note that p_{ij} is the proportion of subjects in the $(i, j)^{\text{th}}$ cell, such that $\sum_i \sum_j p_{ij} = 1$.

The asymptotic variance of the simple kappa coefficient is estimated by the following:

$$\text{var} = \frac{A + B - C}{(1 - P_c)^2 n} \text{ where } A = \sum_i p_{ii}[1 - (p_{i+} + p_{+i})(1 - \hat{\kappa})]^2, B = (1 - \hat{\kappa})^2 \sum_{i \neq j} \sum p_{ij}(p_{+i} + p_{j+})^2 \text{ and}$$

$$C = [\hat{\kappa} - P_c(1 - \hat{\kappa})]^2$$

See Cohen (1960) and Fleiss et al. (1969).

For Bowker's test of symmetry, the null hypothesis is that the probabilities in the two-by-two table satisfy symmetry ($p_{ij}=p_{ji}$).

Statistical Details for the Odds Ratio

In the Contingency platform, the Odds Ratio is calculated as follows:

$$\frac{p_{11} \times p_{22}}{p_{12} \times p_{21}}$$

where p_{ij} is the count in the i^{th} row and j^{th} column of the 2 x 2 table.

Statistical Details for the Tests Report

This section contains details for the Tests report in the Contingency platform.

Rsquare (U)

Rsquare (U) is computed as follows:

$$\frac{\text{negative log-likelihood for Model}}{\text{negative log-likelihood for Corrected Total}}$$

The total negative log-likelihood is found by fitting fixed response rates across the total sample.

Test

The computations for the two chi-square tests are described in this section.

The likelihood ratio chi-square test statistic is computed as twice the negative log-likelihood for Model in the Tests table. Some books use the notation G^2 for this statistic. The difference of two negative log-likelihoods, one with *whole-population* response probabilities and one with *each-population* response rates, is defined as follows:

$$G^2 = 2\left[\sum_{ij}(-n_{ij})\ln(p_j) - \sum_{ij}-n_{ij}\ln(p_{ij})\right] \text{ where } p_{ij} = \frac{n_{ij}}{N} \text{ and } p_j = \frac{N_j}{N}$$

This formula can be more compactly written as follows:

$$G^2 = 2\sum_i\sum_j n_{ij}\ln\left(\frac{n_{ij}}{e_{ij}}\right)$$

The Pearson chi-square test statistic is calculated by summing the squares of the differences between the observed and expected cell counts. The Pearson chi-square test exploits the property that frequency counts tend to a normal distribution in very large samples. The Pearson chi-square statistic is defined as follows:

$$\chi^2 = \sum\frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$$

where O is the observed cell counts and E is the expected cell counts. The summation is over all cells. There is no continuity correction done here, as is sometimes done in 2 x 2 tables.

Statistical Details for Correspondence Analysis

This section contains details for correspondence analysis in the Contingency platform.

Lists values from the following singular value decomposition (SVD):

$$\mathbf{D}_r^{-0.5}(\mathbf{P} - rc')\mathbf{D}_c^{-0.5} = \mathbf{U}Diag(\Lambda)\mathbf{V}'$$

where:

P is the matrix of counts divided by the total frequency

r and *c* are row and column sums of **P**

$\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices of the values of *r* and *c*, respectively

Λ is the column vector of the singular values reported in the details report

For more information about singular value decomposition, see *Multivariate Methods*.

The row coordinates (*rc*) and column coordinates (*cc*) in the Details report are computed as follows:

$$rc = \mathbf{D}_r^{-0.5}\mathbf{U}Diag(\Lambda)$$

$$cc = \mathbf{D}_c^{-0.5}\mathbf{V}Diag(\Lambda)$$

Chapter 8

Logistic Analysis

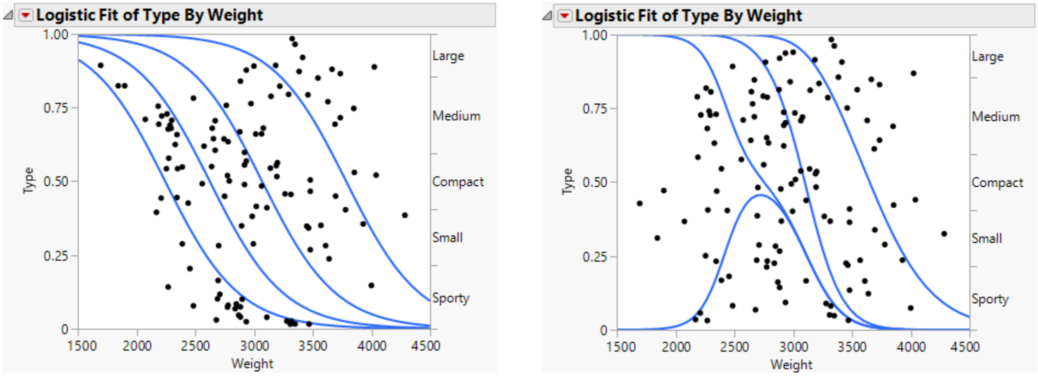
Examine Relationships between a Categorical Y and a Continuous X Variable

Use the Logistic platform to fit a logistic regression model to a categorical Y variable with a continuous X variable. You can view ROC curves, lift curves, and odds ratio estimates. The fitted model provides estimated probabilities for each value of the X variable. You can also perform inverse prediction, which enables you to predict the X value for a specific probability value of the Y variable.

The Logistic platform is the *nominal or ordinal by continuous* personality of the Fit Y by X platform. There is a distinction between nominal and ordinal responses in this platform:

- Nominal logistic regression models estimate a set of curves that partition the probability among the levels of a nominal response variable. An example of a nominal logistic regression model is shown on the right side of Figure 8.1.
- Ordinal logistic regression models estimate the probability of being less than or equal to a target level of an ordinal response variable. The model estimates a single logistic curve that gets shifted horizontally to produce probabilities for the ordered categories. This model is less complex and is recommended for ordered responses. An example of an ordinal logistic regression model is shown on the left side of Figure 8.1.

Figure 8.1 Examples of Ordinal and Nominal Logistic Regression



Contents

Overview of the Logistic Platform

Logistic regression enables you to model the probabilities of the levels of a categorical Y variable based on the values of a continuous X variable. Logistic regression has long been used in a variety of applications such as modeling dose-response data and purchase-choice data. Many texts in categorical statistics cover logistic regression (Agresti 1990), in addition to texts specifically focused on logistic regression (Hosmer and Lemeshow 1989).

In many logistic regression settings, you can also use discriminant analysis, especially if you prefer to think of the continuous variables as X variables and the categories as Y variables and work backward. However, discriminant analysis assumes that the continuous data are normally distributed random responses, rather than fixed regressors. See *Multivariate Methods*.

Simple logistic regression in the Fit Y by X platform is a more graphical and simplified version of the general models for categorical responses in the Fit Model platform. For examples of more complex logistic regression models, see *Fitting Linear Models*. For logistic regression using the normal distribution function, also called *probit analysis*, see *Predictive and Specialized Modeling*.

Nominal Logistic Regression

Nominal logistic regression models estimate the probability of choosing one of the levels of the Y variable as a smooth function of the continuous X variable. The fitted probabilities must be between 0 and 1, and they must sum to 1 across the levels of the Y variable for a given value of the X variable.

In a logistic probability plot, the vertical axis represents probability. For k levels of the Y variable, k - 1 smooth curves partition the total probability (which equals 1) among the levels of the Y variable. The fitting principle for a logistic regression minimizes the sum of the negative natural logarithms of the probabilities fitted to the response events that occur (that is, maximum likelihood).

Ordinal Logistic Regression

When the Y variable is ordinal, a modified version of logistic regression is used for fitting. The cumulative probability of being at or below each level of the Y variable is modeled by a curve. The curves for each level have the same shape, but they are shifted to the right or left.

The ordinal logistic model fits a different intercept, but the same slope, for each of r - 1 cumulative logistic comparisons, where r is the number of levels of the Y variable. The ordinal model is preferred to the nominal model when it is appropriate because it has fewer parameters to estimate.

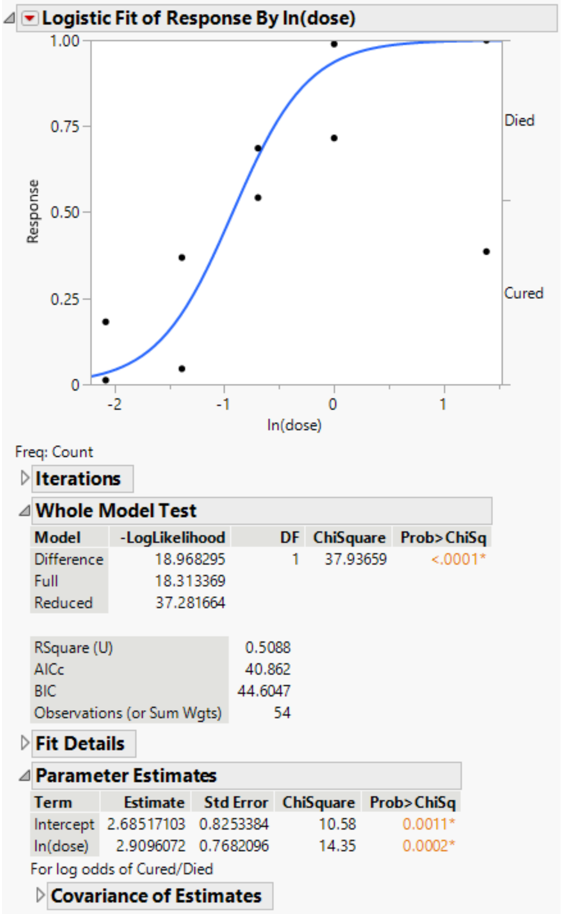
Example of Nominal Logistic Regression

Use the Logistic platform to examine the relationship between a nominal Y response variable and a continuous X factor variable. The data in this example come from an experiment where 5 groups, each containing 12 rabbits, were injected with streptococci bacteria. Once the rabbits were confirmed to have the bacteria in their system, they were given different doses of penicillin. You want to find out whether the natural log of dosage amounts has any effect on whether the rabbits are cured.

1. Select **Help > Sample Data Folder** and open Penicillin.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.
4. Select ln(dose) and click **X, Factor**.

   Notice that JMP automatically fills in Count for **Freq**, because the Count column was previously assigned the role of Freq.
5. From the Target Level list, select Cured.
6. Click **OK**.

Figure 8.2 Example of Nominal Logistic Report



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	18.968295	1	37.93659	<.0001*
Full	18.313369			
Reduced	37.281664			

RSquare (U)	0.5088
AICc	40.862
BIC	44.6047
Observations (or Sum Wgts)	54

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.68517103	0.8253384	10.58	0.0011*
ln(dose)	2.9096072	0.7682096	14.35	0.0002*

For log odds of Cured/Died

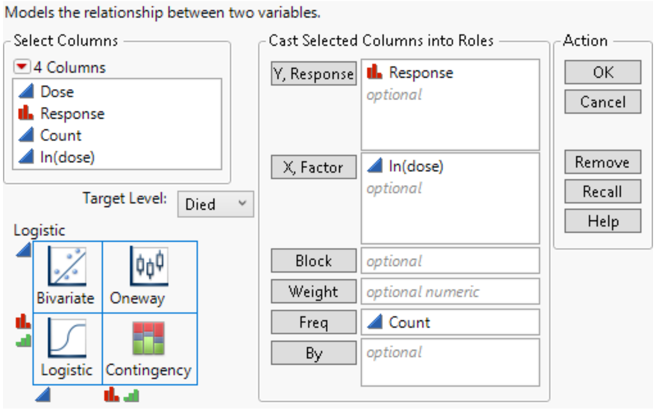
The plot shows the fitted model as a function of ln(dose). The fitted model is the predicted probability of being cured. The *p*-value is significant, which indicates that the dosage amounts have a significant effect on whether the rabbits are cured. The marginal distribution that is shown on the right vertical axis represents the probabilities for the levels of the Y variable if there is no association between the X and Y variables.

Tip: To change the response level that is analyzed, use the Target Level option in the launch window or use the Value Order column property.

Launch the Logistic Platform

Launch the Logistic platform by selecting **Analyze > Fit Y by X**. The Fit Y by X launch window is used for four different types of analyses. When you enter an ordinal or nominal Y variable and a continuous X variable, the Logistic platform is launched.

Figure 8.3 The Logistic Launch Window



Models the relationship between two variables.

Select Columns
4 Columns
Dose
Response
Count
ln(dose)

Target Level: Died

Cast Selected Columns into Roles
Y, Response: Response
X, Factor: ln(dose)
Block: optional
Weight: optional numeric
Freq: Count
By: optional

For more information about the options in the Select Columns red triangle menu, see *Using JMP*. The Logistic launch window contains the following options:

Y, Response The response variable or variables that you want to analyze. The response variable is often called the dependent variable. These variables must have an ordinal or nominal modeling type.

X, Factor The predictor variable or variables that you want to analyze. The factor variables are often called independent variables. These variables must have a continuous modeling type.

Block (Not applicable for a logistic analysis.) A column that specifies a Blocking variable.

Weight A column containing a weight for each observation in the data table. A row is included in the analyses only when its value is greater than zero.

Freq Assigns a frequency to each row in the analysis. Assigning a frequency is useful when your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Target Level (Available only when there is one binary response that has a nominal modeling type.) Enables you to specify the level of the response whose probability you want to model.

Tip: The default level that is used in the logistic model as the target value is based on the order of the levels. Use the Value Order column property to change the default target value.

Data Format

In the Logistic platform, your data can consist of unsummarized or summarized data:

Unsummarized data There is one row for each observation and columns for X and Y values.

Summarized data Each row represents a set of observations with common X and Y values. The data table must contain a frequency column that contains the counts of observations for each row. Enter this column as Freq in the launch window.

For an example of summarized data in the Logistic platform, see "Example of Nominal Logistic Regression".

Note: The Fit Y by X launch window accommodates columns with continuous, ordinal, and nominal modeling types. The Logistic platform is launched for all pairs of ordinal or nominal Y, Response and continuous X, Factor columns. The Fit Y by X launch window launches the Bivariate, Oneway, or Contingency platforms for other column type combinations.

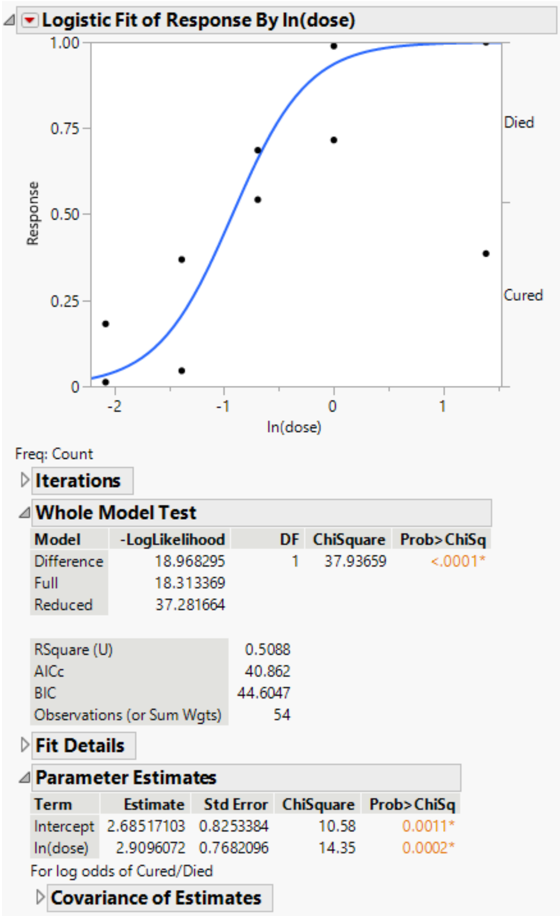
The Logistic Report

The Logistic report initially contains a logistic plot and the Iterations, Whole Model Test, Fit Details, and Parameter Estimates reports. You can change the appearance of the logistic plot and request other plots using the red triangle menu options. See "Logistic Platform Options".

This section contains information about the following sections of the Logistic report:

- "Logistic Plot"
- "Iterations Report"
- "Whole Model Test Report"
- "Fit Details Report"
- "Parameter Estimates Report"

Figure 8.4 Example of a Logistic Report



Logistic Fit of Response By ln(dose)

Freq: Count

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	18.968295	1	37.93659	<.0001*
Full	18.313369			
Reduced	37.281664			

RSquare (U)	0.5088
AICc	40.862
BIC	44.6047
Observations (or Sum Wgts)	54

Fit Details

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.68517103	0.8253384	10.58	0.0011*
ln(dose)	2.9096072	0.7682096	14.35	0.0002*

For log odds of Cured/Died

Covariance of Estimates

Replace Variables Interactively

You can also replace a variable by selecting a variable in the Columns panel of the associated data table and dragging it onto an axis.

Logistic Plot

In the Logistic platform, the logistic probability plot illustrates the model fit. The vertical axis on the right side of the plot is partitioned proportional to the count of observations for each level of the Y variable. There is a curve in the plot for all but one of the levels of the Y variable. The curve is the predicted cumulative probability up to and including the corresponding level of the Y variable. The vertical axis on the left side of the plot is on a probability scale. The predicted probability for a specific level is measured as the vertical distances between the curve for that level and the curve below the level.

The points in the logistic plot represent the observations from the data table. The horizontal position of each point is determined by the value of the continuous X variable for that point. The vertical position of each point is determined by the value of the categorical Y variable for that point. Each point falls at a random vertical position between the curve for its level and the next lower curve. This vertical jittering of the points makes it easier to see where the points are most dense, but the vertical position does not directly correspond to the values on the vertical axis. Because a fixed random seed is used, the vertical positions do not differ across multiple fits of the same model.

Iterations Report

In the Logistic platform, the Iterations report shows each iteration and the evaluated criteria that determine whether the logistic model has converged. Iterations appear only for nominal logistic regression models.

Whole Model Test Report

In the Logistic platform, the Whole Model Test report shows whether the model fits better than simply using a constant for all of the response probabilities. This report is analogous to the Analysis of Variance report for a continuous response model. The test shown in this report is a likelihood ratio chi-square test that evaluates how well the logistic regression model fits the data.

The negative sum of natural logs of the observed probabilities is called the negative log-likelihood (–LogLikelihood). The negative log-likelihood for categorical data is similar to the sums of squares for continuous data. Twice the difference in the negative log-likelihood between the model fitted to the data and the model with equal probabilities for each level is a chi-square statistic. This test statistic is for the hypothesis that the value of the X variable has no association with the levels of the Y variable.

Values of RSquare (U) (sometimes denoted as R^2) range from 0 to 1. Higher R^2 values are indicative of a better model fit. Note that high values of R^2 are rare in categorical models.

The Whole Model Test report contains the following columns:

Model The label of the sources of variation.

- **Difference** The difference between the Full model and the Reduced model. This model is used to measure the significance that the X variable contributes to the fit.
- **Full** The complete model that includes the intercepts and the X variable.
- **Reduced** The model that includes only the intercept parameters.

–LogLikelihood The negative log-likelihood, which measures variation, for the respective models. See *Fitting Linear Models*.

DF The degrees of freedom (DF) for the Difference between the Full and Reduced model.

Chi-Square The likelihood ratio chi-square test statistic for the hypothesis that the model fits no better than fixed response rates across the whole sample. The test statistic is computed by taking twice the difference in negative log-likelihoods between the fitted model and the reduced model that has only intercepts. See "Statistical Details for the Logistic Platform".

Prob>ChiSq The probability of obtaining a greater chi-square value if the specified model fits no better than the model that includes only intercepts. Models are often judged significant if this probability is below 0.05.

RSquare (U) The proportion of the total uncertainty that is attributed to the model fit, defined as the Difference negative log-likelihood value divided by the Reduced negative log-likelihood value. An RSquare (U) value of 1 indicates that the predicted probabilities for events that occur are equal to one: There is no uncertainty in predicted probabilities. Because certainty in the predicted probabilities is rare for logistic models, RSquare (U) tends to be small. See "Statistical Details for the Logistic Platform".

RSquare (U) is sometimes referred to as *U*, the uncertainty coefficient, or as *McFadden's pseudo* R^2.

AICc The corrected Akaike Information Criterion. See *Fitting Linear Models*.

BIC The Bayesian Information Criterion. See *Fitting Linear Models*.

Observations (or Sum Weights) Total number of observations in the sample. If a Freq or Weight column is specified in the Fit Model window, this value is the sum of the values of a column assigned to the Freq or Weight role.

Fit Details Report

The Fit Details report in the Logistic platform contains the values of the following statistics, including an algebraic definition of each measure of fit:

Entropy RSquare Compares the log-likelihoods from the fitted model and the constant probability model. This is equivalent to Rsquare (U). See "Statistical Details for the Logistic Platform".

Generalized RSquare A measure that can be applied to general regression models. It is based on the likelihood function *L* and is scaled to have a maximum value of 1. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the

Nagelkerke or Craig and Uhler R^2, which is a normalized version of Cox and Snell's pseudo R^2. See Nagelkerke (1991).

Mean -Log p The average of -log(p), where p is the fitted probability associated with the event that occurred.

RASE The root average square error, where the differences are between the response and p (the fitted probability for the event that actually occurred).

Mean Abs Dev The average of the absolute values of the differences between the response and p (the fitted probability for the event that actually occurred).

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category.

N The number of observations.

For Entropy RSquare and Generalized RSquare, values closer to 1 indicate a better fit. For Mean -Log p, RASE, Mean Abs Dev, and Misclassification Rate, smaller values indicate a better fit.

Parameter Estimates Report

The nominal logistic model fits a parameter for the intercept and slope for each of $k - 1$ logistic comparisons, where k is the number of response levels. The ordinal logistic model fits a single slope for each comparison. The Parameter Estimates report in the Logistic platform lists these estimates. Each parameter estimate can be examined and tested individually, although this is seldom of much interest.

Term The parameters in the logistic model. There is an intercept and a slope term for the factor at each level of the response variable, except the last level. Ordinal logistic models contain only one slope term.

Estimate The parameter estimates given by the logistic model.

Std Error The standard error of each parameter estimate. The standard errors are used to compute the statistical tests that compare each term to zero.

Chi-Square The Wald tests for the hypotheses that each of the parameters is zero. The Wald chi-square is computed as (Estimate / Std Error)^2.

Prob>ChiSq The observed significance probabilities (p-values) for the chi-square tests.

Covariance of Estimates

Reports the estimated variances of the parameter estimates, and the estimated covariances between the parameter estimates. The square root of the variance estimates is the same as those that appear in the Std Error column of the Parameter Estimates table.

Note: The Covariance of Estimates report appears only for nominal response variables, and does not appear for ordinal response variables.

Logistic Platform Options

The Logistic Fit red triangle menu contains options for the Logistic plot and the Logistic Fit report.

Note: The Fit Group menu appears only if you have specified multiple Y or multiple X variables. Use the Fit Group menu options to arrange reports or order them by RSquare. See *Fitting Linear Models*.

The Logistic Fit red triangle menu contains the following options:

Odds Ratios (Available only for a response with two levels.) Adds or removes columns that contain odds ratios to the Parameter Estimates report. See *Fitting Linear Models*.

Inverse Prediction (Available only for two-level nominal responses.) Enables you to predict values of the predictor variable for one or more values of the response variable. See "Inverse Prediction".

Logistic Plot Shows or hides the logistic plot. See "Logistic Plot".

Plot Options Contains the following options that affect the logistic plot:

Show Points Shows or hides the points in the logistic plot.

Show Rate Curve Shows or hides the rate curve in the logistic plot. The rate curve is useful only if you have several points for each value of the X variable. In these cases, you get reasonable estimates of the rate at each value, and you can compare this rate with the fitted logistic curve. To prevent too many degenerate points, usually at zero or one, JMP shows only the rate value if there are at least three points at the x-value.

Line Color Enables you to select the color of the plot curves.

ROC Curve Shows or hides the Receiver Operating Characteristic (ROC) curve for the model. The ROC curve is a plot of sensitivity versus (1 - specificity) for each value of the X variable. See "ROC Curves".

Lift Curve Shows or hides a lift curve for the model. A lift curve shows the predictive ability of the model. The lift curve plots the lift versus the portion of the observations. The lift curve contains a point for each unique predicted probability value. Each predicted probability of a response level defines a portion of the observations that have a predicted probability greater than or equal to the unique predicted probability value. For a particular level of the response, the *lift* value is the ratio of the proportion of observed responses in that portion to the overall proportion of observed responses. See *Predictive and Specialized Modeling* for more information about lift curves.

Save Probability Formula Saves new columns to the data table. The new columns contain the formula for the probability that is predicted by the model. The new columns contain the following:

- formulas for linear combinations (typically called logits) of the factor variable
- prediction formulas for the response level probabilities
- a prediction formula for the most likely response

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Logistic Analysis Reports

The Logistic platform enables you to create visualizations and fit models to explore the relationship between a continuous X variable and a categorical Y variable. This section contains information about the reports that are generated for specific analysis options.

- "ROC Curves"
- "Inverse Prediction"

ROC Curves

The Logistic platform contains an option to fit a receiver operating characteristic (ROC) curve for the logistic regression model. The ROC Curve option in the Logistic platform uses the Target Level from the platform launch window as the *positive* response level in the ROC curve. See "Example of ROC Curves".

Suppose you have a value of the X variable that is a diagnostic measurement and you want to determine a threshold value of the X variable that indicates the following:

- A condition exists if the value of the X variable is greater than the threshold.
- A condition does not exist if the value of the X variable is less than the threshold.

For example, you could measure a blood component level as a diagnostic test to predict a type of cancer. Consider the diagnostic test as you vary the threshold and thus cause more or fewer false positives and false negatives. You then plot those rates. The ideal is to have a very narrow range of values of the X variable that best divides true negatives and true positives. The Receiver Operating Characteristic (ROC) curve shows how rapidly this transition happens. The goal of the ROC curve is to have diagnostics that maximize the area under the curve.

Two standard definitions are used in medicine:

- *Sensitivity* is the probability that a given value of the X variable correctly predicts an existing condition. For a given x, the probability of incorrectly predicting the existence of a condition is 1 – sensitivity.
- *Specificity* is the probability that a test correctly predicts that a condition does not exist.

A ROC curve is a plot of sensitivity by (1 – specificity) for each value of the X variable. The area under the ROC curve is a common index used to summarize the information contained in the curve.

If a test predicted perfectly, it would have a value above which the entire abnormal population would fall and below which all normal values would fall. It would be perfectly sensitive and then pass through the point (0, 1) on the grid. The closer the ROC curve comes to this ideal point, the better its discriminating ability. A test with no predictive ability produces a curve that follows the diagonal of the grid (DeLong et al. 1988).

The ROC curve is a graphical representation of the relationship between the false-positive and true-positive rates. A standard way to evaluate the relationship is using the area under the curve, which appears below the plot in the report. In the plot, a yellow 45-degree angle line is drawn at a tangent to the ROC Curve. This marks the cutoff point that maximizes the sum of sensitivity and specificity.

Inverse Prediction

You can use the Logistic platform for inverse prediction, which enables you to make predictions in the opposite direction than usual. Instead of predicting the value of a Y response variable from the value of an X factor variable, inverse prediction is the prediction of the values of an X factor variable from the value of a Y response variable. In logistic regression, instead of a predicting a value of the Y response variable, you predict the probability attributed to each of the levels of the Y response variable. This feature is available only for two-level nominal responses.

The Fit Model platform also has an option that provides an inverse prediction with confidence limits. See *Fitting Linear Models* for more information about inverse prediction.

Note: See also "Example of Inverse Prediction".

Additional Examples of Logistic Regression

This section contains examples using the Logistic platform.

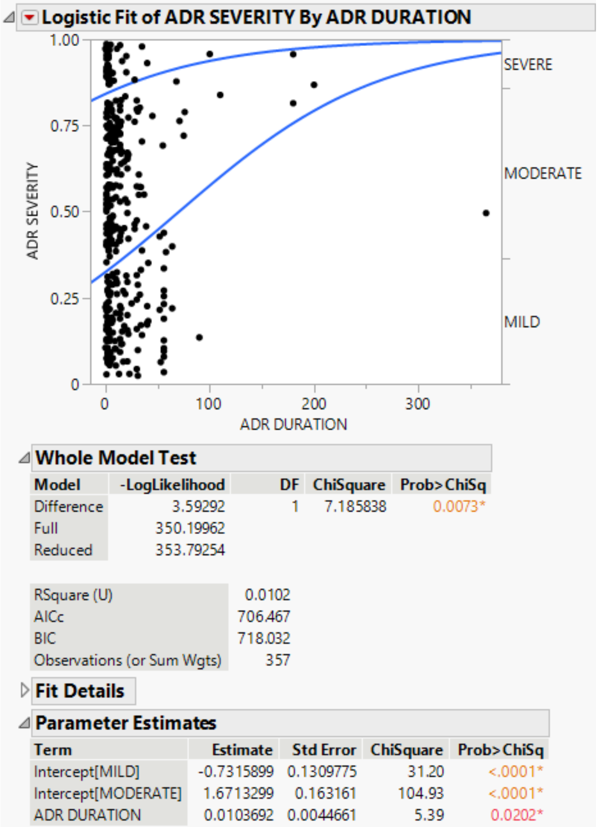
- "Example of Ordinal Logistic Regression"
- "Example of a Logistic Plot"
- "Example of ROC Curves"
- "Example of Inverse Prediction"

Example of Ordinal Logistic Regression

This example shows you how to examine the relationship between an ordinal response and a continuous factor. In this example, suppose you want to model the severity of an adverse event as a function of duration of adverse events.

1. Select **Help > Sample Data Folder** and open AdverseR.jmp.
2. Right-click the icon to the left of ADR SEVERITY and change the modeling type to ordinal.
3. Select **Analyze > Fit Y by X**.
4. Select ADR SEVERITY and click **Y, Response**.
5. Select ADR DURATION and click **X, Factor**.
6. Click **OK**.

Figure 8.5 Example of Ordinal Logistic Report



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	3.59292	1	7.185838	0.0073*
Full	350.19962			
Reduced	353.79254			

RSquare (U)	0.0102
AICc	706.467
BIC	718.032
Observations (or Sum Wgts)	357

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[MILD]	-0.7315899	0.1309775	31.20	<.0001*
Intercept[MODERATE]	1.6713299	0.163161	104.93	<.0001*
ADR DURATION	0.0103692	0.0044661	5.39	0.0202*

In the plot, markers for the data are drawn at their X coordinate value along the horizontal axis and at a random vertical position below the curve for their Y value.

For more information about the Whole Model Test report and the Parameter Estimates report, see "The Logistic Report". In the Parameter Estimates report, an intercept parameter is estimated for every response level except the last, but there is only one slope parameter. The intercept parameters show the spacing of the response levels. They always increase monotonically.

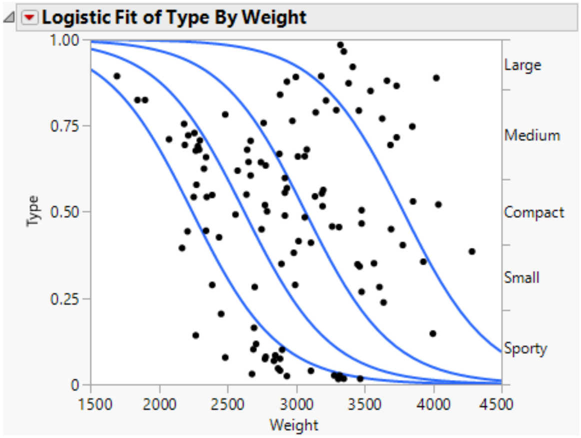
Example of a Logistic Plot

In this example, suppose you want to use car weight to predict car type for 116 cars. Car type can be one of the following, from smallest to largest: Sporty, Small, Compact, Medium, or Large.

1. Select **Help > Sample Data Folder** and open Car Physical Data.jmp.
2. In the Columns panel, right-click the icon to the left of Type, and select **Ordinal**.
3. Right-click Type and select **Column Info**.
4. From the Column Properties menu, select **Value Order**.
5. Verify that the data are in the following top-down order: Sporty, Small, Compact, Medium, Large.
6. Click **OK**.
7. Select **Analyze > Fit Y by X**.
8. Select Type and click **Y, Response**.
9. Select Weight and click **X, Factor**.
10. Click **OK**.

    The report window appears.

Figure 8.6 Example of Type by Weight Logistic Plot

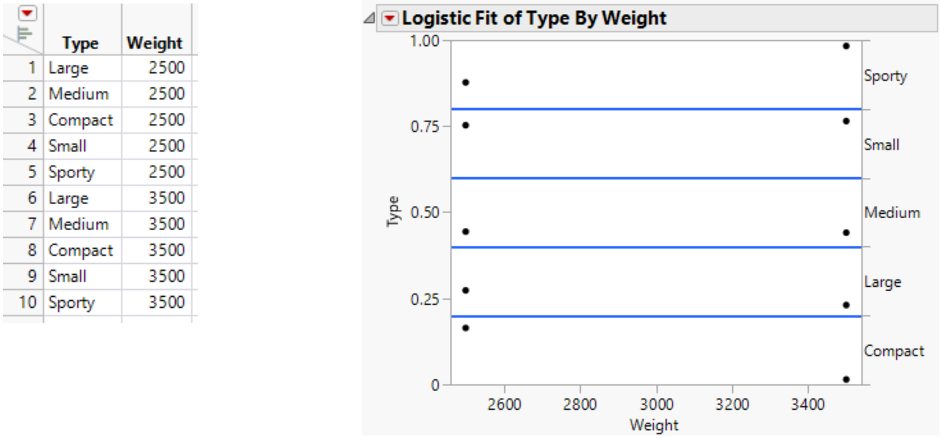


Note the following observations:

- The first (bottom) curve represents the probability that a car at a given weight is Sporty.
- The second curve represents the probability that a car is Small or Sporty. Looking only at the distance between the first and second curves corresponds to the probability of being Small.
- As you might expect, heavier cars are more likely to be Large.
- Markers for the data are drawn at their *x*-coordinate. The *y* position is jittered randomly within the range corresponding to the response category for that row.

If the X variable has no effect on the response, then the fitted lines are horizontal and the probabilities are constant for each response across the continuous factor range. Figure 8.7 shows a logistic plot where Weight is not useful for predicting Type.

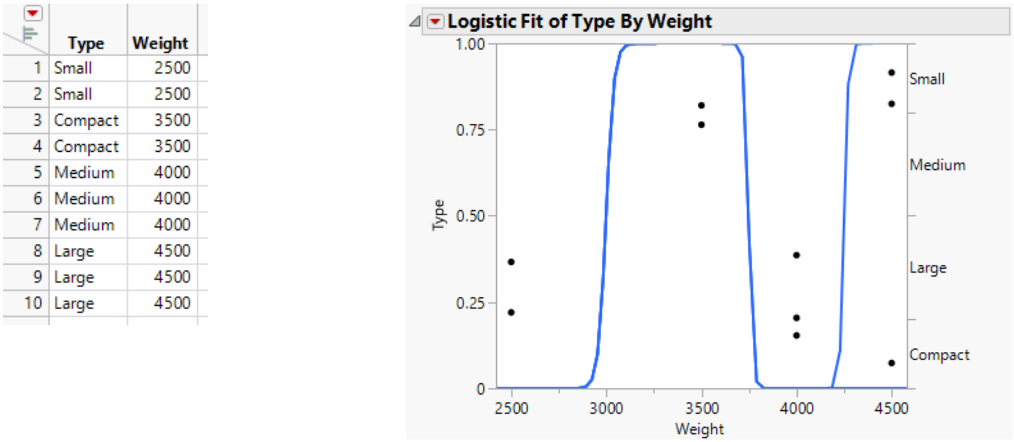
Figure 8.7 Examples of Sample Data Table and Logistic Plot Showing No *y* by *x* Relationship



	Type	Weight
1	Large	2500
2	Medium	2500
3	Compact	2500
4	Small	2500
5	Sporty	2500
6	Large	3500
7	Medium	3500
8	Compact	3500
9	Small	3500
10	Sporty	3500

If the response is completely predicted by the value of the factor, then the logistic curves are effectively vertical. The prediction of a response is near certain (the probability is almost 1) at each of the factor levels. Figure 8.8 shows a logistic plot where Weight almost perfectly predicts Type.

Figure 8.8 Examples of Sample Data Table and Logistic Plot Showing an Almost Perfect *y* by *x* Relationship



	Type	Weight
1	Small	2500
2	Small	2500
3	Compact	3500
4	Compact	3500
5	Medium	4000
6	Medium	4000
7	Medium	4000
8	Large	4500
9	Large	4500
10	Large	4500

In this case, the parameter estimates become very large and are labeled *unstable* in the regression report. In these cases, you might consider using the Generalized Linear Model personality with Firth bias-adjusted estimates. See *Fitting Linear Models*.

Note: To re-create the plots in Figure 8.7 and Figure 8.8, you must first create the data tables shown here, and then perform steps 7-10 at the beginning of this section.

Example of ROC Curves

Use the Logistic platform to fit an ROC curve.

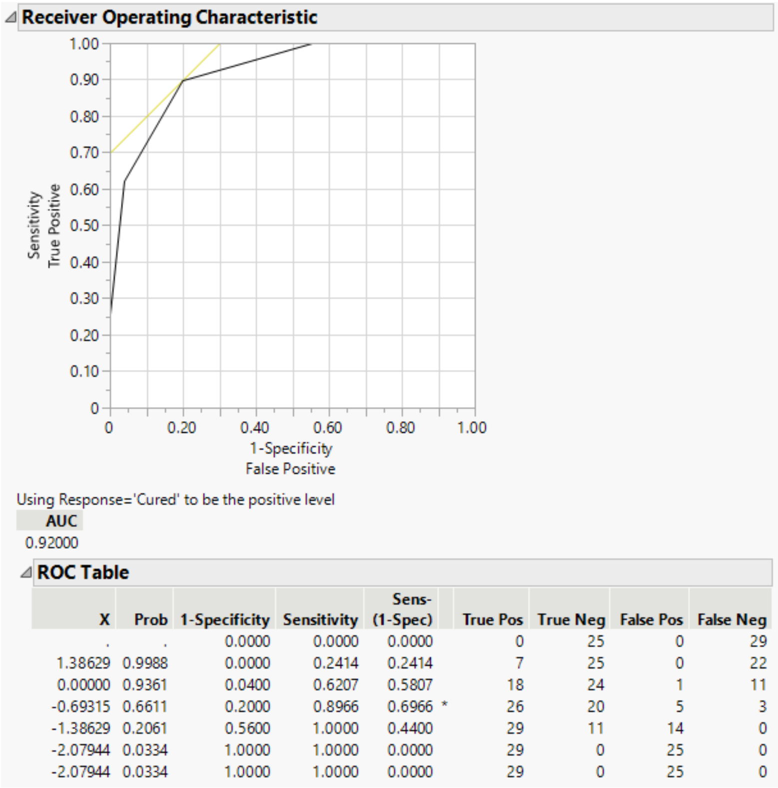
1. Select **Help > Sample Data Folder** and open Penicillin.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.
4. Select ln(dose) and click **X, Factor**.

   Notice that JMP automatically fills in Count for **Freq**. Count was previously assigned the role of Freq.
5. From the Target Level list, select Cured.
6. Click **OK**.
7. Click the Logistic Fit red triangle menu and select **ROC Curve**.

Note: This example shows a ROC Curve for a nominal response. For more information about ordinal ROC curves, see *Predictive and Specialized Modeling*.

The results for the response by ln(dose) example are shown here. The ROC curve plots the probabilities described above, for predicting response. Note that in the ROC Table, the row with the highest Sens-(1-Spec) is marked with an asterisk. The X value that corresponds to this point is the threshold that maximizes the sum of sensitivity and specificity. The yellow line on the ROC curve plot is the tangent line to this point.

Figure 8.9 Examples of ROC Curve and Table



Using Response='Cured' to be the positive level

AUC
0.92000

ROC Table

X	Prob	1-Specificity	Sensitivity	Sens-(1-Spec)		True Pos	True Neg	False Pos	False Neg
.	.	0.0000	0.0000	0.0000		0	25	0	29
1.38629	0.9988	0.0000	0.2414	0.2414		7	25	0	22
0.00000	0.9361	0.0400	0.6207	0.5807		18	24	1	11
-0.69315	0.6611	0.2000	0.8966	0.6966	*	26	20	5	3
-1.38629	0.2061	0.5600	1.0000	0.4400		29	11	14	0
-2.07944	0.0334	1.0000	1.0000	0.0000		29	0	25	0
-2.07944	0.0334	1.0000	1.0000	0.0000		29	0	25	0

Since the ROC curve falls in the upper left quadrant of the plot and the AUC is greater than 0.5, you conclude that the model has predictive ability.

Example of Inverse Prediction

In the Logistic platform, you can use the Crosshair tool to visually approximate an inverse prediction. If your response has exactly two levels, the Inverse Prediction option enables you to request an exact inverse prediction. Inverse prediction gives you an estimated value of the X variable that corresponds to a given probability of the target level of the Y variable, as well as a confidence interval for that estimate.

1. Select **Help > Sample Data Folder** and open Penicillin.jmp.

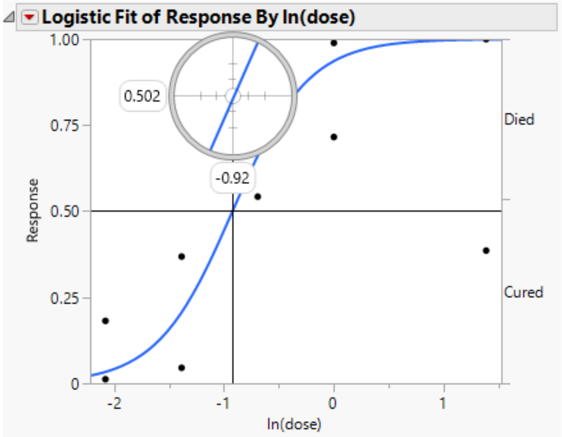
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.
4. Select ln(dose) and click **X, Factor**.

   Notice that JMP automatically fills in Count for **Freq**. Count was previously assigned the role of Freq.
5. From the Target Level list, select Cured.
6. Click **OK**.

Crosshair Tool

7. Click the Crosshair tool.
8. Place the crosshair at about 0.5 on the vertical (Response) probability axis on the left side of the plot.
9. Move the crosshair intersection to the prediction line, and read the ln(dose) value that shows below the crosshair.

Figure 8.10 Example of Crosshair Tool on Logistic Plot



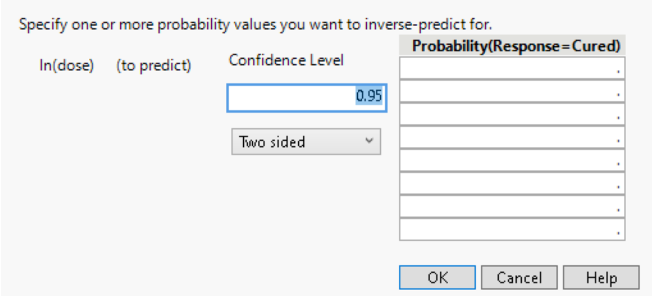
In this example, a rabbit with a ln(dose) of approximately -0.9 is equally likely to be cured as it is to die. To obtain a precise inverse prediction, use the Inverse Prediction option.

Inverse Prediction Option

10. Click the Logistic Fit red triangle menu and select **Inverse Prediction** (Figure 8.11).
11. Type 0.95 for the **Confidence Level**.
12. Select **Two sided** for the confidence interval.

13. Request the response probability of interest. Type 0.5 and 0.9 for this example, which indicates you are requesting the values for ln(dose) that correspond to a 0.5 and 0.9 probability of being cured.

Figure 8.11 Inverse Prediction Window



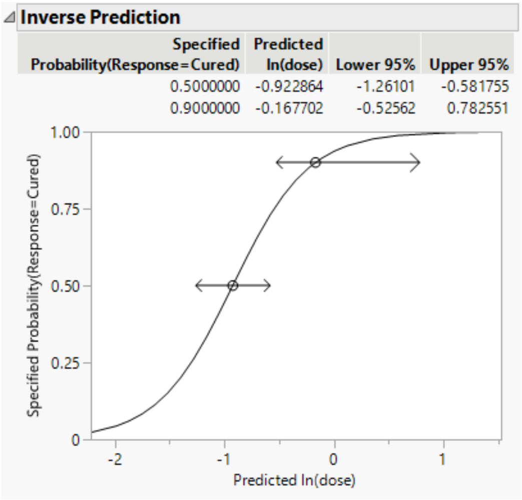
Specify one or more probability values you want to inverse-predict for.

Confidence Level 0.95

Two sided

14. Click **OK**.

Figure 8.12 Example of Inverse Prediction Plot



Inverse Prediction

Specified Probability(Response=Cured)	Predicted ln(dose)	Lower 95%	Upper 95%
0.5000000	-0.922864	-1.26101	-0.581755
0.9000000	-0.167702	-0.52562	0.782551

The estimates of the values of the X variable and the corresponding confidence intervals are shown in the Inverse Prediction report in both tabular and graphic format. For example, the value of ln(dose) that results in a 90% probability of being cured is estimated to be between -0.526 and 0.783.

Statistical Details for the Logistic Platform

This section contains statistical details for the Whole Model Test report.

Chi-Square

The Chi-Square statistic is sometimes denoted G^2 and is defined as follows:

$$G^2 = 2(\sum -\ln p(\text{background}) - \sum -\ln p(\text{model}))$$

The summations are over all observations instead of all cells.

RSquare (U)

RSquare (U) is computed as follows:

$$\frac{\text{negative log-likelihood for Difference}}{\text{negative log-likelihood for Reduced}}$$

These quantities appear in the Whole Model Test report.

Note: RSquare (U) is also known as *McFadden's pseudo R-square*.

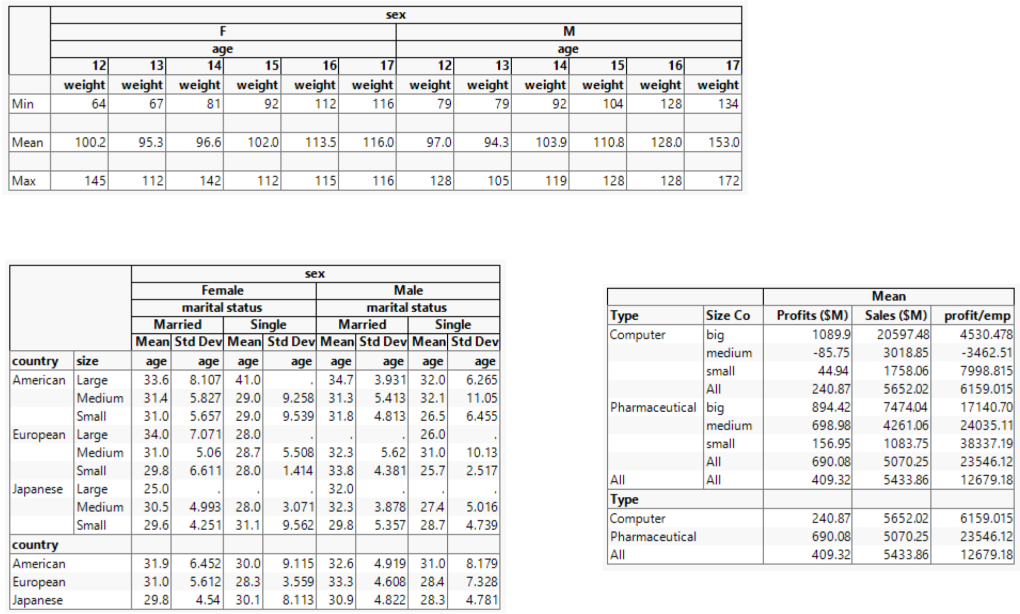
Chapter 9

Tabulate

Create Summary Tables Interactively

Use the Tabulate platform to interactively construct summary tables, or pivot tables, of descriptive statistics. The Tabulate platform is an easy and flexible way to present summary data in tabular form. Tables are built from assigning data table columns as rows or columns in the tabulation and then assigning the desired summary statistics.

Figure 9.1 Tabulate Examples



	sex											
	F						M					
	age						age					
	12	13	14	15	16	17	12	13	14	15	16	17
	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight
Min	64	67	81	92	112	116	79	79	92	104	128	134
Mean	100.2	95.3	96.6	102.0	113.5	116.0	97.0	94.3	103.9	110.8	128.0	153.0
Max	145	112	142	112	115	116	128	105	119	128	128	172

		sex							
		Female				Male			
		marital status				marital status			
		Married		Single		Married		Single	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
country	**size**	**age**	**age**	**age**	**age**	**age**	**age**	**age**	**age**
American	Large	33.6	8.107	41.0	.	34.7	3.931	32.0	6.265
	Medium	31.4	5.827	29.0	9.258	31.3	5.413	32.1	11.05
	Small	31.0	5.657	29.0	9.539	31.8	4.813	26.5	6.455
European	Large	34.0	7.071	28.0	.	.	.	26.0	.
	Medium	31.0	5.06	28.7	5.508	32.3	5.62	31.0	10.13
	Small	29.8	6.611	28.0	1.414	33.8	4.381	25.7	2.517
Japanese	Large	25.0	.	.	.	32.0	.	.	.
	Medium	30.5	4.993	28.0	3.071	32.3	3.878	27.4	5.016
	Small	29.6	4.251	31.1	9.562	29.8	5.357	28.7	4.739
country									
American		31.9	6.452	30.0	9.115	32.6	4.919	31.0	8.179
European		31.0	5.612	28.3	3.559	33.3	4.608	28.4	7.328
Japanese		29.8	4.54	30.1	8.113	30.9	4.822	28.3	4.781

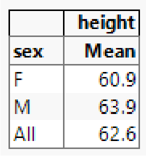
		Mean		
Type	**Size Co**	**Profits ($M)**	**Sales ($M)**	**profit/emp**
Computer	big	1089.9	20597.48	4530.478
	medium	-85.75	3018.85	-3462.51
	small	44.94	1758.06	7998.815
	All	240.87	5652.02	6159.015
Pharmaceutical	big	894.42	7474.04	17140.70
	medium	698.98	4261.06	24035.11
	small	156.95	1083.75	38337.19
	All	690.08	5070.25	23546.12
All	All	409.32	5433.86	12679.18
Type				
Computer		240.87	5652.02	6159.015
Pharmaceutical		690.08	5070.25	23546.12
All		409.32	5433.86	12679.18

Contents

Example of the Tabulate Platform

Learn how to summarize data in a tabular format. In this example, you have data containing height measurements for male and female students. You want to create a table that shows the mean height for males and females and the aggregate mean for both sexes.

Figure 9.2 Table Showing Mean Height

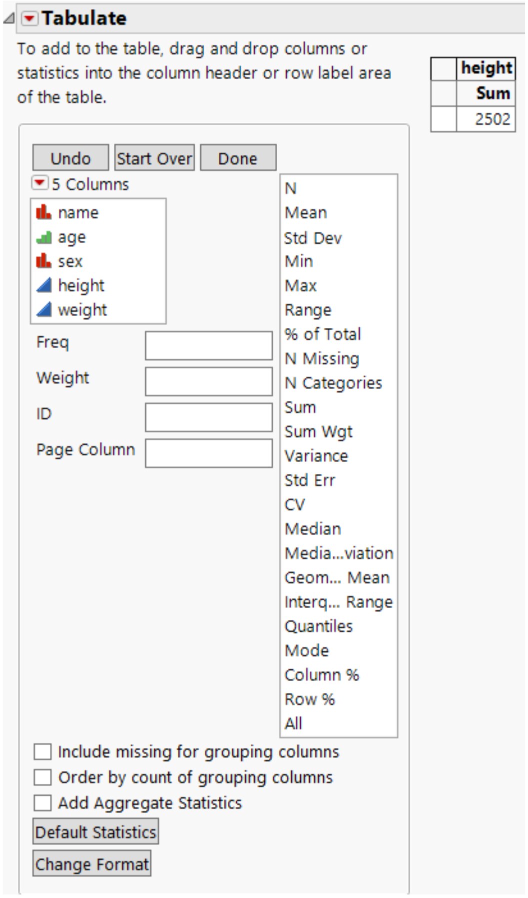


	height
sex	Mean
F	60.9
M	63.9
All	62.6

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Tabulate**.

   Since height is the variable that you are examining, you want it to appear at the top of the table.
3. Click height and drag it into the Drop zone for columns.

Figure 9.3 Height Variable Added



Tabulate

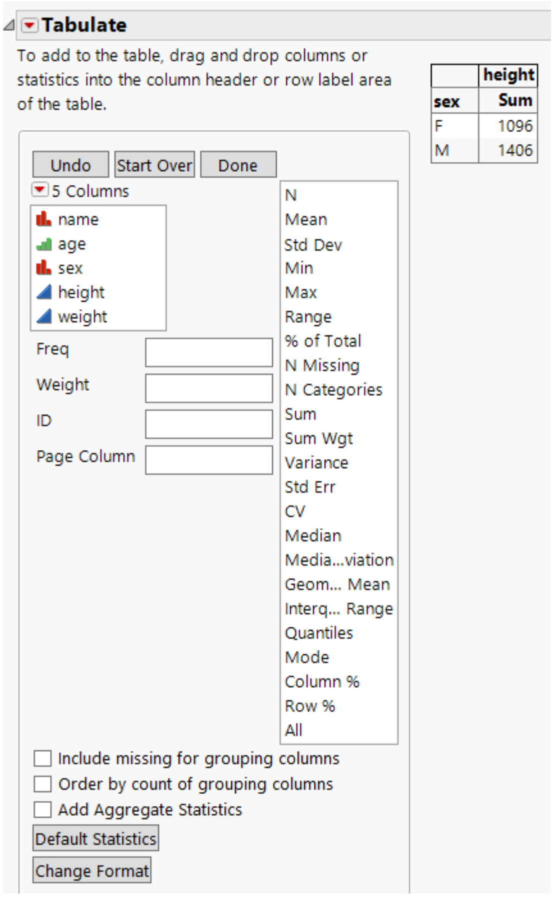
To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

	height
	Sum
	2502

You want the statistics by sex, and you want sex to appear on the side.

4. Click sex and drag it into the blank cell next to the number 2502.

Figure 9.4 Sex Variable Added



To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

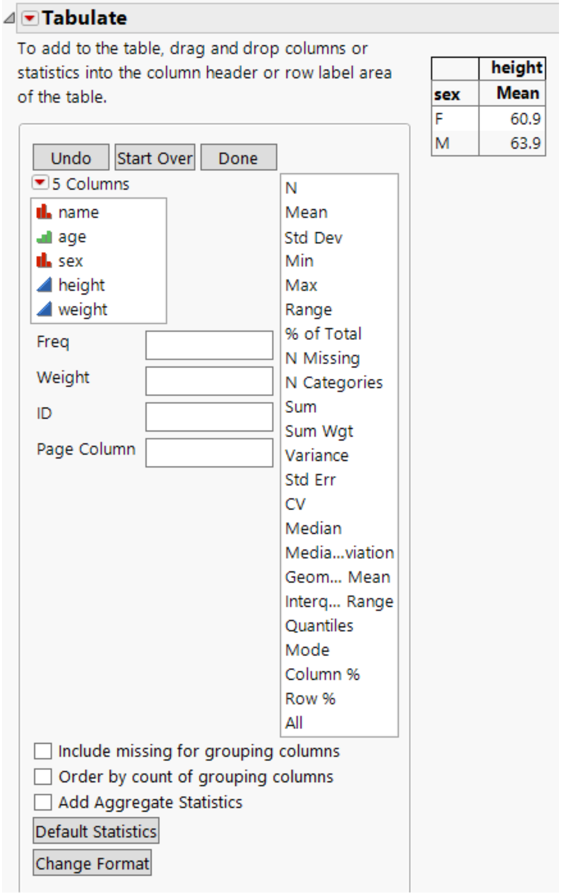
	height
sex	Sum
F	1096
M	1406

☐ Include missing for grouping columns
☐ Order by count of grouping columns
☐ Add Aggregate Statistics

Instead of the sum, you want it to show the mean.

5. Click Mean and drag it on top of Sum.

Figure 9.5 Mean Statistic Added



To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

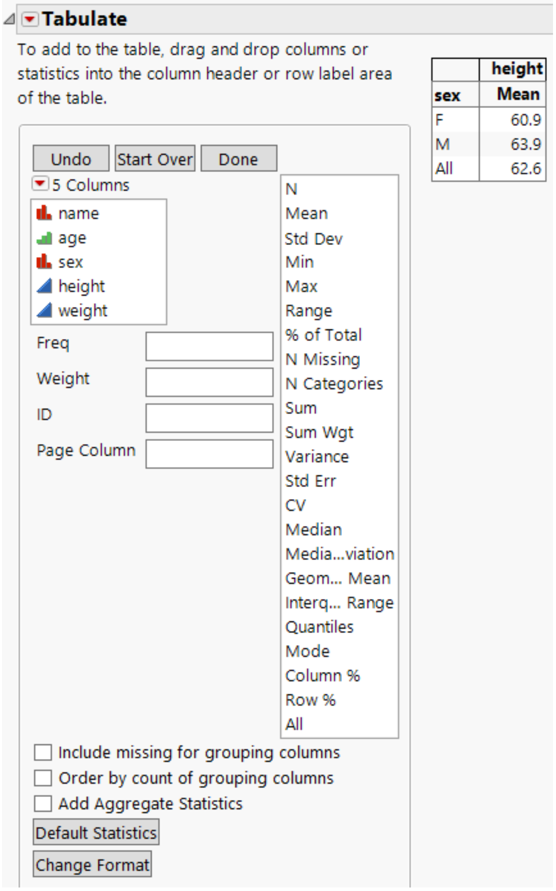
	height
sex	Mean
F	60.9
M	63.9

☐ Include missing for grouping columns
☐ Order by count of grouping columns
☐ Add Aggregate Statistics

You also want to see the combined mean for males and females.

6. Click All and drag it on top of sex. Or, you can simply select the **Add Aggregate Statistics** check box.

Figure 9.6 All Statistic Added



To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

	height
sex	Mean
F	60.9
M	63.9
All	62.6

☐ Include missing for grouping columns

☐ Order by count of grouping columns

☐ Add Aggregate Statistics

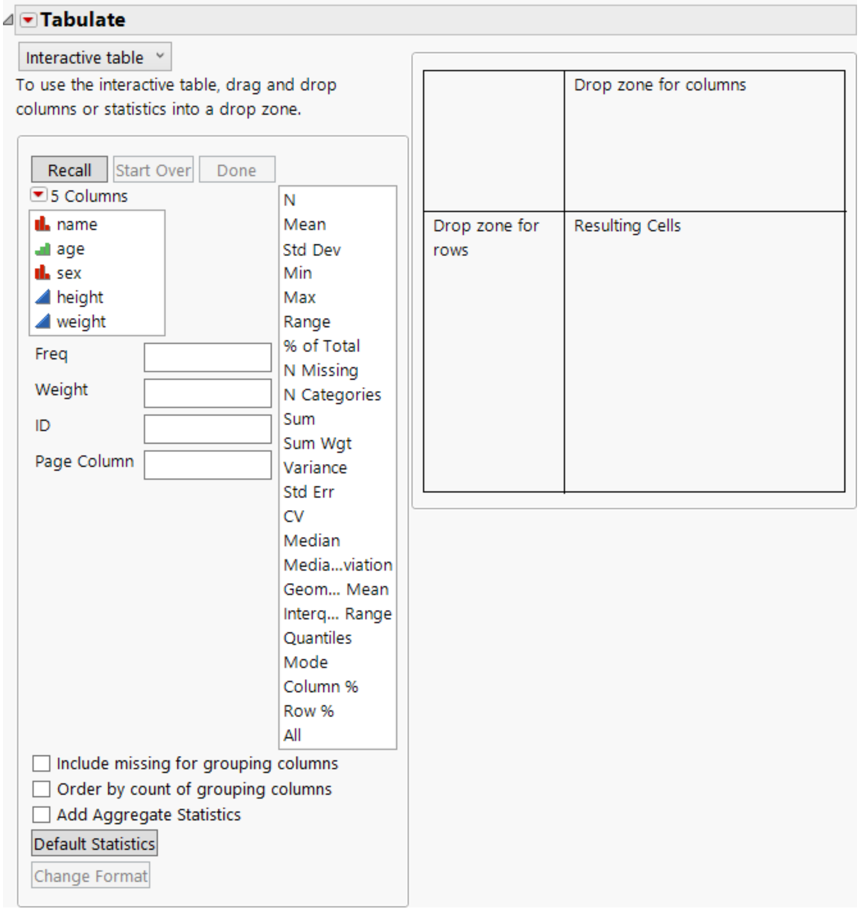
7. (Optional) Click **Done**.

The completed table shows the mean height for females, males, and the combined mean height for both.

Launch the Tabulate Platform

Launch the Tabulate platform by selecting **Analyze > Tabulate**.

Figure 9.7 The Tabulate Interactive Table



For more information about red triangle options, see “Tabulate Platform Options”. For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Tabulate window contains the following options:

Interactive table/Dialog Switch between the two modes. Use the interactive table mode to drag and drop items, creating a custom table. Use the dialog mode to create a simple table using a fixed format. See “Tabulate Platform Dialog Window”.

Statistics options Lists standard statistics. Drag any statistic from the list to the table to incorporate it. See “Add Statistics”.

Drop zone for columns Drag and drop columns or statistics here to create columns.

Note: If the data table contains columns with names equal to those in the Statistics options, be sure to drag and drop the column name from the column list. Otherwise, JMP might substitute the statistic of the same name in the table.

Drop zone for rows Drag and drop columns or statistics here to create rows.

Tip: You can also select one or more columns in the columns list, select one or more of the statistics, and then Alt-click (Option-click on macOS) on a drop zone to create rows or columns in the table.

Resulting cells Shows the resulting cells based on the columns or statistics that you drag and drop.

Freq Identifies the data table column whose values assign a frequency to each row. This option is useful when a frequency is assigned to each row in summarized data.

Weight Identifies the data table column whose values assign weight (such as importance or influence) to the data.

ID Identifies the data table column whose values identify unique occurrences.

Page Column Generates separate tables for each category of a nominal or ordinal column. See "Example Using a Page Column".

Include missing for grouping columns Creates a separate group for missing values in grouping columns. When unchecked, missing values are not included in the table. Note that any missing value codes that you have defined as column properties are taken into account.

Order by count of grouping columns Changes the order of the table to be in descending order of the values of the grouping columns.

Add Aggregate Statistics Adds aggregate statistics for all rows and columns.

Default Statistics Enables you to change the default statistics that appear when you drag and drop analysis or non-analysis (for example, grouping) columns.

Change Format Enables you to change the numeric format for displaying specific statistics. See "Change Numeric Formats".

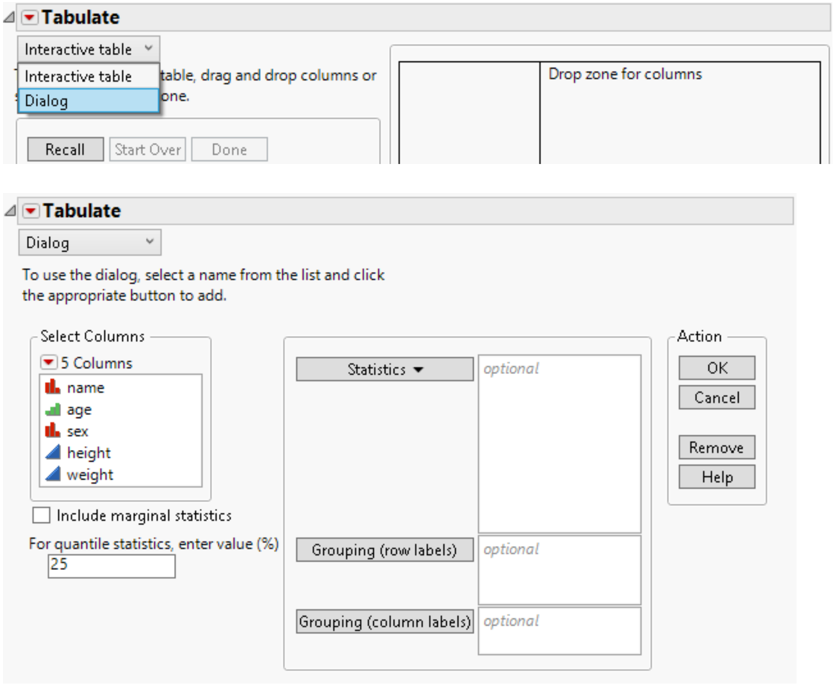
Change Plot Scale (Appears only if **Show Chart** is selected from the red triangle menu.) Enables you to specify a uniform custom scale.

Uniform plot scale (Appears only if **Show Chart** is selected from the red triangle menu.) Deselect this box for each column of bars to use the scale determined separately from the data in each displayed column.

Tabulate Platform Dialog Window

If you prefer not to use drag and drop operations to build a table interactively, you can create a simple table using the Dialog interface in the Tabulate platform. After selecting **Analyze > Tabulate**, select **Dialog** from the menu, as shown in Figure 9.8. You can make changes to the table by selecting **Show Control Panel** from the red triangle menu, and then drag and drop new items into the table.

Figure 9.8 Using the Dialog



To use the dialog, select a name from the list and click the appropriate button to add.

The dialog contains the following options:

Include marginal statistics Aggregates summary information for categories of a grouping column.

For quantile statistics, enter value (%) Enter the value at which the specific percentage of the argument is less than or equal to. For example, 75% of the data is less than the 75th quantile. This applies to all grouping columns.

Statistics Once you have selected a column, select a standard statistic to apply to that column. See "Add Statistics".

Grouping (row labels) Select the column to use as the row label.

Grouping (column labels) Select the column to use as the column label.

Add Statistics

The Tabulate platform supports a list of standard statistics that appear in the control panel. You can drag any keyword from that list to the table, just like you do with the columns. Note the following:

Tip: You can select both a column and a statistic at the same time and drag them into the table.

- The statistics associated with each cell are calculated on values of the analysis columns from all observations in that category, as defined by the grouping columns.
- All of the requested statistics have to reside in the same dimension, either in the row table or in the column table.
- If you drag a continuous column into a data area, it is treated as an analysis column.

Note: Analysis columns are numeric, continuous columns for which you want to compute statistics. See "Analysis Columns".

Tabulate uses the following keywords:

N Provides the number of nonmissing values in the column. This is the default statistic when there is no analysis column.

Mean Provides the arithmetic mean of a column's values. It is the sum of nonmissing values (and if defined, multiplied by the **weight** variable) divided by the **Sum Wgt**.

Std Dev Provides the sample standard deviation, computed for the nonmissing values. It is the square root of the sample variance.

Min Provides the smallest nonmissing value in a column.

Max Provides the largest nonmissing value in a column.

Range Provides the difference between **Max** and **Min**.

% of Total Computes the percentage of total of the whole population. The denominator used in the computation is the total of all the included observations, and the numerator is the total for the category. If there is no analysis column, the % of Total is the percentage of total of counts. If there is an analysis column, the % of Total is the percentage of the total of the sum of the analysis column. Thus, the denominator is the sum of the analysis column over all the included observations, and the numerator is the sum of the analysis column for that category. You can request different percentages by dragging the keyword into the table.

- Dropping one or more grouping columns from the table to the % of Total heading changes the denominator definition. For this, Tabulate uses the sum of these grouping columns for the denominator.
- To get the percentage of the column total, drag all the grouping columns on the row table and drop them onto the **% of Total** heading (same as Column %). Similarly, to get the percentage of the row total, drag all grouping columns on the column table and drop them onto the **% of Total** heading (same as Row %).

N Missing Provides the number of missing values.

N Categories Provides the number of distinct categories in the analysis column.

Sum Provides the sum of all values in the column. This is the default statistic for analysis columns when there are no other statistics for the table.

Sum Wgt Provides the sum of all weight values in a column. Or, if no column is assigned the weight role, **Sum Wgt** is the total number of nonmissing values.

Variance Provides the sample variance, computed for the nonmissing values. It is the sum of squared deviations from the mean, divided by the number of nonmissing values minus one.

Std Err Provides the standard error of the mean. It is the standard deviation divided by the square root of **N**. If a column is assigned the role of weight, then the denominator is the square root of the sum of the weights.

CV (Coefficient of Variation) Provides the measure of dispersion, which is the standard deviation divided by the mean multiplied by one hundred.

Median Provides the 50th percentile, which is the value where half the data are below and half are above or equal to the 50th quantile (median).

Geometric Mean The *n*th root of the product of the data. For example, geometric means are often used to calculate interest rates. The statistic is also helpful when the data contains a large value in a skewed distribution.

Note: Negative values result in missing numbers, and zero values (with no negative values) result in zero.

Interquartile Range Provides the difference between the 3rd quartile and 1st quartile.

Quantiles Provides the value at which the specific percentage of the argument is less than or equal to. For example, 75% of the data is less than the 75th quantile. You can request different quantiles by clicking and dragging the **Quantiles** keyword into the table, and then entering the quantile into the box that appears.

Column % Provides the percent of each cell count to its column total if there is no analysis column. If there is an analysis column, the Column % is the percent of the column total of the sum of the analysis column. For tables with statistics on the top, you can add Column % to tables with multiple row tables (stacked vertically).

Row % Provides the percent of each cell count to its row total if there is no analysis column. If there is an analysis column, the Row % is the percent of the row total of the sum of the analysis column. For tables with statistics on the side, you can add Row % to tables with multiple column tables (side by side tables).

All Aggregates summary information for categories of a grouping column.

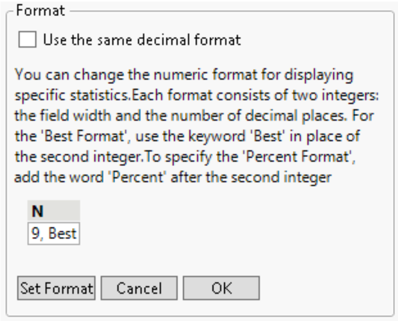
Change Numeric Formats

The formats of each cell depend on the analysis column and the statistics. For counts, the default format has no decimal digits. For each cell defined by some statistics, JMP tries to determine a reasonable format using the format of the analysis column and the statistics requested. To override the default format:

1. Click the **Change Format** button at the bottom of the Tabulate window.
2. In the panel that appears, enter the field width, a comma, and then the number of decimal places that you want displayed in the table (Figure 9.9).
3. To exhibit the cell value in Percent format, add a comma after the number of decimal places and type the word **Percent**.
4. (Optional) If you would like JMP to determine the best format for you to use, type the word **Best** in the text box.

   JMP now considers the precision of each cell value and selects the best way to show it.
5. Click **OK** to implement the changes and close the Format section, or click **Set Format** to see the changes implemented without closing the Format section.

Figure 9.9 Changing Numeric Formats



☐ Use the same decimal format

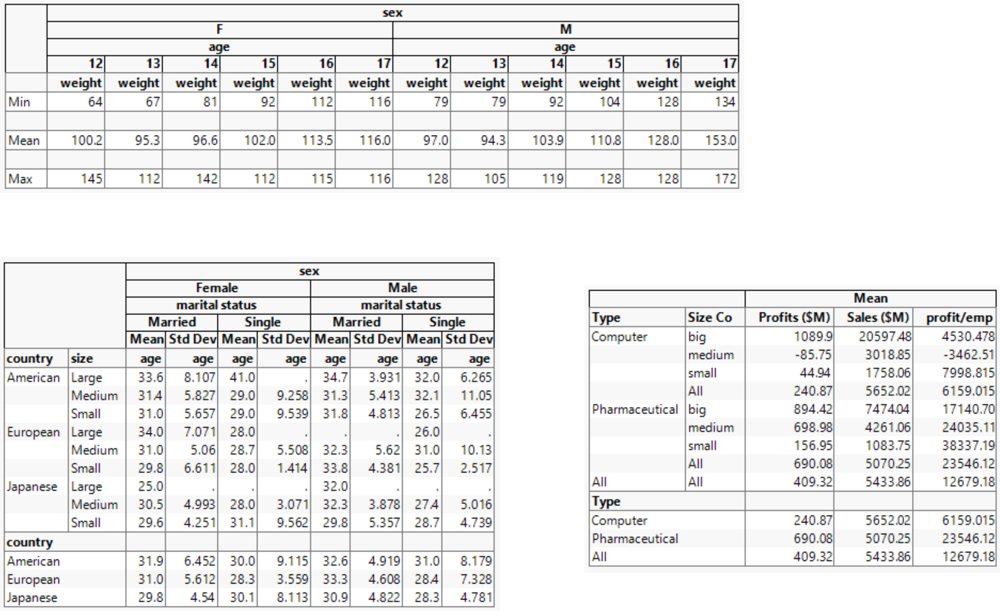
You can change the numeric format for displaying specific statistics.Each format consists of two integers: the field width and the number of decimal places. For the 'Best Format', use the keyword 'Best' in place of the second integer.To specify the 'Percent Format', add the word 'Percent' after the second integer

N
9, Best

Tabulate Platform Report

The Tabulate platform report consists of one or more column tables concatenated side by side, and one or more row tables concatenated top to bottom. The output might have only a column table or a row table.

Figure 9.10 Tabulate Output



	sex											
	F						M					
	age						age					
	12	13	14	15	16	17	12	13	14	15	16	17
	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight
Min	64	67	81	92	112	116	79	79	92	104	128	134
Mean	100.2	95.3	96.6	102.0	113.5	116.0	97.0	94.3	103.9	110.8	128.0	153.0
Max	145	112	142	112	115	116	128	105	119	128	128	172

		sex							
		Female				Male			
		marital status				marital status			
		Married		Single		Married		Single	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
country	**size**	age	age	age	age	age	age	age	age
American	Large	33.6	8.107	41.0	.	34.7	3.931	32.0	6.265
	Medium	31.4	5.827	29.0	9.258	31.3	5.413	32.1	11.05
	Small	31.0	5.657	29.0	9.539	31.8	4.813	26.5	6.455
European	Large	34.0	7.071	28.0	.	.	.	26.0	.
	Medium	31.0	5.06	28.7	5.508	32.3	5.62	31.0	10.13
	Small	29.8	6.611	28.0	1.414	33.8	4.381	25.7	2.517
Japanese	Large	25.0	.	.	.	32.0	.	.	.
	Medium	30.5	4.993	28.0	3.071	32.3	3.878	27.4	5.016
	Small	29.6	4.251	31.1	9.562	29.8	5.357	28.7	4.739
country									
American		31.9	6.452	30.0	9.115	32.6	4.919	31.0	8.179
European		31.0	5.612	28.3	3.559	33.3	4.608	28.4	7.328
Japanese		29.8	4.54	30.1	8.113	30.9	4.822	28.3	4.781

		Mean		
Type	**Size Co**	**Profits ($M)**	**Sales ($M)**	**profit/emp**
Computer	big	1089.9	20597.48	4530.478
	medium	-85.75	3018.85	-3462.51
	small	44.94	1758.06	7998.815
	All	240.87	5652.02	6159.015
Pharmaceutical	big	894.42	7474.04	17140.70
	medium	698.98	4261.06	24035.11
	small	156.95	1083.75	38337.19
	All	690.08	5070.25	23546.12
All	All	409.32	5433.86	12679.18
Type				
Computer		240.87	5652.02	6159.015
Pharmaceutical		690.08	5070.25	23546.12
All		409.32	5433.86	12679.18

Creating a table interactively is an iterative process:

- Click the items (columns or statistics) from the appropriate list, and drag them into the drop zone (for rows or columns). See “Edit Tables”, and “Column and Row Tables”.

- Add to the table by repeating the drag and drop process. The table updates to reflect the latest addition. If there are already column headings or row labels, you can decide where the addition goes relative to the existing items.

Note the following about clicking and dragging:

- JMP uses the modeling type to determine a column's role. Continuous columns are assumed to be analysis columns. See "Analysis Columns". Ordinal or nominal columns are assumed to be grouping columns. See "Grouping Columns".
- When you drag and drop multiple columns into the initial table:
  - If the columns share a set of common values, they are combined into a single table. A crosstabulation of the column names and the categories gathered from these columns is generated. Each cell is defined by one of the columns and one of the categories.
  - If the columns do not share common values, they are put into separate tables.
  - You can always change the default action by right-clicking on a column and selecting **Combine Tables** or **Separate Tables**. See "Right-Click Menu for Columns".
- To nest columns, create a table with the first column, and then drag the additional columns into the first column.
- In a properly created table, all grouping columns are together, all analysis columns are together, and all statistics are together. Therefore, JMP does not intersperse a statistics keyword within a list of analysis columns. JMP also does not insert an analysis column within a list of grouping columns.
- You can drag columns from the Table panel in the data table onto a Tabulate table instead of using the Tabulate Control Panel.

Note: The Tabulate table is updated when you add data to the open data table, delete rows, and recode the data.

Analysis Columns

In the Tabulate platform, analysis columns are any numeric (continuous) columns for which you want to compute statistics. Tabulate computes statistics on the analysis columns for each category formed from the grouping columns.

Note: All the analysis columns must reside in the same dimension, either in the row table or in the column table.

Grouping Columns

In the Tabulate platform, grouping columns are columns (nominal or ordinal) that you want to use to classify your data into categories of information. They can have character, integer, or even decimal values, but the number of unique values should be limited.

Note the following:

- If grouping columns are nested, the Tabulate platform constructs distinct categories from the hierarchical nesting of the values of the columns. For example, from the grouping columns Sex with values F and M, and the grouping column Marital Status with values Married and Single, Tabulate constructs four distinct categories: F and Married, F and Single, M and Married, M and Single.
- If grouping columns are nested, you can stack the nested values to save horizontal space in the table. To stack the grouping columns, right-click the grouping column and select the Stack Grouping Columns option. The stacked grouping columns are consolidated based on the hierarchical nesting of the values of the columns.
- You can specify grouping columns for column tables as well as row tables. Together they generate the categories that define each table cell.
- Tabulate does not include observations with a missing value for one or more grouping columns by default. You can include them by checking the **Include missing for grouping columns** option.
- To specify codes or values that should be treated as missing, use the Missing Value Codes column property. You can include these by checking the **Include missing for grouping columns** option. For more information about Missing Value Codes, see *Using JMP*.

Column and Row Tables

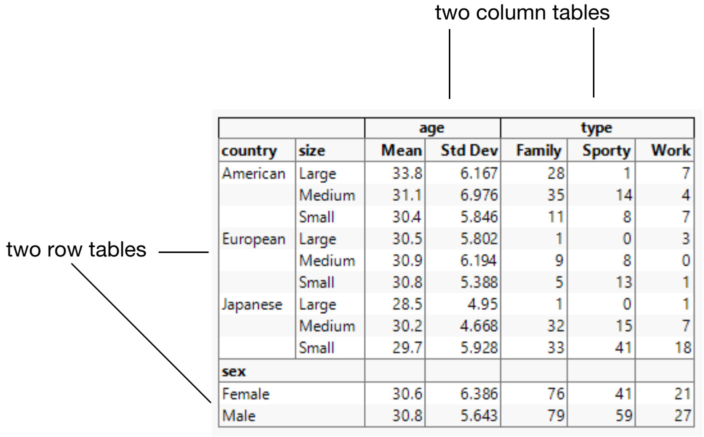
In the Tabulate platform, a table is defined by its column headings and row labels. These sub-tables are referred to as *row tables* and *column tables* (Figure 9.11).

Example of Row and Column Tables in the Tabulate Platform

1. Select **Help > Sample Data Folder** and open Car Poll.jmp.
2. Select **Analyze > Tabulate**.
3. Drag size into the Drop zone for rows.
4. Drag country to the left of the size heading.
5. Drag Mean over the N heading.
6. Drag Std Dev below the Mean heading.
7. Drag age above the Mean heading.

8. Drag type to the far right of the table.
9. Drag sex under the table.

Figure 9.11 Row and Column Tables



two column tables

two row tables

		age		type		
country	**size**	**Mean**	**Std Dev**	**Family**	**Sporty**	**Work**
American	Large	33.8	6.167	28	1	7
	Medium	31.1	6.976	35	14	4
	Small	30.4	5.846	11	8	7
European	Large	30.5	5.802	1	0	3
	Medium	30.9	6.194	9	8	0
	Small	30.8	5.388	5	13	1
Japanese	Large	28.5	4.95	1	0	1
	Medium	30.2	4.668	32	15	7
	Small	29.7	5.928	33	41	18
sex						
Female		30.6	6.386	76	41	21
Male		30.8	5.643	79	59	27

For multiple column tables, the labels on the side are shared across the column tables. In this instance, country and sex are shared across the tables. Similarly, for multiple row tables, the headings on the top are shared among the row tables. In this instance, both age and type are shared among the tables.

Edit Tables

In the Tabulate platform, there are several ways to edit items that you add to a table. You can delete items, remove column labels, or edit statistical key words or labels.

Delete Items

After you add items to the table, you can remove them in any one of the following ways:

- Drag the item away from the table.
- To remove the last item, click **Undo**.
- Right-click an item and select **Delete**.

Remove Column Labels

Grouping columns display the column name on top of the categories associated with that column. For some columns, the column name might seem redundant. Remove the column name from the column table by right-clicking on the column name and selecting **Remove Column Label**. To re-insert the column label, right-click one of its associated categories and select **Restore Column Label**.

Edit Statistical Key Words and Labels

You can edit a statistical key word or a statistical label. For example, instead of Mean, you might want to use the word Average. Right-click the word that you want to edit and select **Change Item Label**. In the box that appears, enter the new label. Alternatively, you can type directly into the edit box.

If you change one statistics keyword to another statistics keyword, JMP assumes that you actually want to change the statistics, not just the label. It would be as if you have deleted the statistics from the table and added the latter.

Tabulate Platform Options

The following options are available from the Tabulate red triangle menu:

Show Control Panel Displays the control panel for further interaction.

Show Table Displays the summarized data in tabular form.

Show Chart Displays the summarized data in bar charts that mirrors the table of summary statistics. The simple bar chart enables visual comparison of the relative magnitude of the summary statistics. By default, all columns of bars share the same scale. You can have each column of bars use the scale determined separately from the data in each displayed column, by clearing the **Uniform plot scale** check box. You can specify a uniform custom scale using the **Change Plot Scale** button. The charts are either 0-based or centered on 0. If the data are all nonnegative, or all non-positive, the charts baseline is at 0. Otherwise, the charts are centered on 0.

Show Shading Displays gray shading boxes in the table when there are multiple rows.

Show Tooltip Displays tips that appear when you hover over areas of the table.

Show Test Build Panel Displays the control area that lets you create a test build using a random sample from the original table. This is particularly useful when you have large amounts of data. See “Show Test Build Panel”.

Make Into Data Table Makes a data table from the report. There is one data table for each row table, because labels of different row tables might not be mapped to the same structure.

Full Path Column Name Uses the fully qualified column names of grouping columns for the column name in the created data table.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

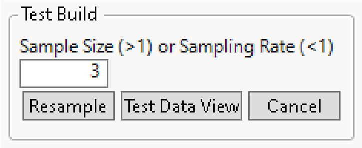
For a description of the options in the Select Columns red triangle menu, see *Using JMP*.

Show Test Build Panel

If you are summarizing a very large data table in the Tabulate platform, you might want to use a small subset of the data table to try out different table layouts to find the best one. In this case, JMP generates a random subset of the size as specified and uses that subset when it builds the table. To use the test build feature:

1. Click the Tabulate red triangle and select **Show Test Build Panel**.
2. Enter the size of the sample that you want in the box under **Sample Size (>1) or Sampling Rate (<1)**, as shown in Figure 9.12. The size of the sample can be either the proportion of the active table that you enter or the number of rows from the active table.

Figure 9.12 The Test Build Panel



Test Build

Sample Size (>1) or Sampling Rate (<1): 3

3. Click **Resample**.

4. To see the sampled data in a JMP data table, click the **Test Data View** button. When you dismiss the test build panel, Tabulate uses the full data table to regenerate the tables as designed.

Right-Click Menu for Columns

In the Tabulate report table, right-click a column to see the following options:

Delete Deletes the selected column.

Pack Columns (Available only when you right-click on a statistic column.) Contains options for consolidating statistical columns in the table for a more compact display.

Pack Consolidates the selected statistical columns into a single column.

Unpack Separates the selected statistics into separate columns.

Template Enables you to define the display format of the consolidated statistics. The leftmost selected column when you select Pack is designated as ^FIRST, and the rest of the selected columns are designated as ^OTHERS. You can also choose a delimiter to separate the list of statistics that appear inside the parentheses.

Use as Grouping column Changes the analysis column to a grouping column.

Use as Analysis column Changes the grouping column to an analysis column.

Change Item Label (Appears only for separate or nested columns.) Enter a new label.

Combine Tables (Columns by Categories) (Appears only for separate or nested columns.) Combines separate or nested columns. See "Example of Combining Columns into a Single Table".

Nest Grouping Columns Nests grouping columns vertically or horizontally.

Separate Tables (Appears only for combined tables.) Creates a separate table for each column.

Remove Column Label Removes the column name from the column table.

Restore Column Label Restores a hidden column name to the column table.

Order By Count Orders the levels of a grouping column by the count of each level, from most to least. This option can also be used to override the setting of the Order by count of grouping columns option for a particular grouping column.

Stack Grouping Columns (Available only when there are multiple grouping columns.) Specifies that the values of the nested grouping columns appear in a consolidated format to conserve horizontal space in the table. The values of the nested grouping columns are indented to show the hierarchy of the grouping columns. See “Example of Stacking Grouping Columns”.

Additional Examples of the Tabulate Platform

This section contains examples using the Tabulate platform.

- “Example of Creating Different Tables and Rearranging Contents”
- “Example of Combining Columns into a Single Table”
- “Example Using a Page Column”
- “Example of Stacking Grouping Columns”

Example of Creating Different Tables and Rearranging Contents

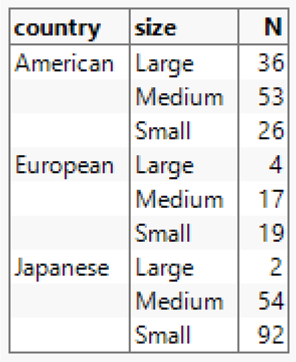
This example contains the following steps to create and modify tables using the Tabulate platform:

1. “Create a Table of Counts”
2. “Create a Table Showing Statistics”
3. “Rearrange the Table Contents”

Create a Table of Counts

Suppose that you would like to create a table that contains counts for how many people in a survey own Japanese, European, and American cars. You also want the counts broken down by the size of the car.

Figure 9.13 Table Showing Counts of Car Ownership

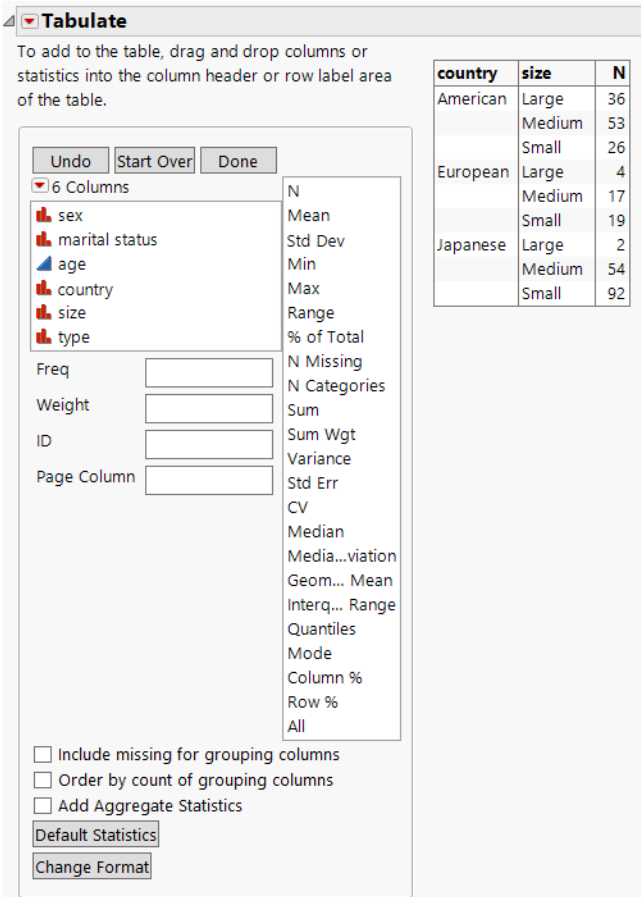


country	size	N
American	Large	36
	Medium	53
	Small	26
European	Large	4
	Medium	17
	Small	19
Japanese	Large	2
	Medium	54
	Small	92

1. Select **Help > Sample Data Folder** and open Car Poll.jmp.

2. Select **Analyze > Tabulate**.
3. Click country and drag it into the Drop zone for rows.
4. Click size and drag it to the right of the country heading.

Figure 9.14 Country and Size Added to the Table



To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

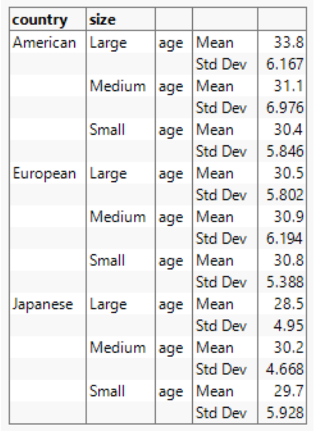
country	size	N
American	Large	36
	Medium	53
	Small	26
European	Large	4
	Medium	17
	Small	19
Japanese	Large	2
	Medium	54
	Small	92

☐ Include missing for grouping columns
☐ Order by count of grouping columns
☐ Add Aggregate Statistics

Create a Table Showing Statistics

Suppose that you would like to see the mean (average) and the standard deviation of the age of people who own each size of car.

Figure 9.15 Table Showing Mean and Standard Deviation by Age

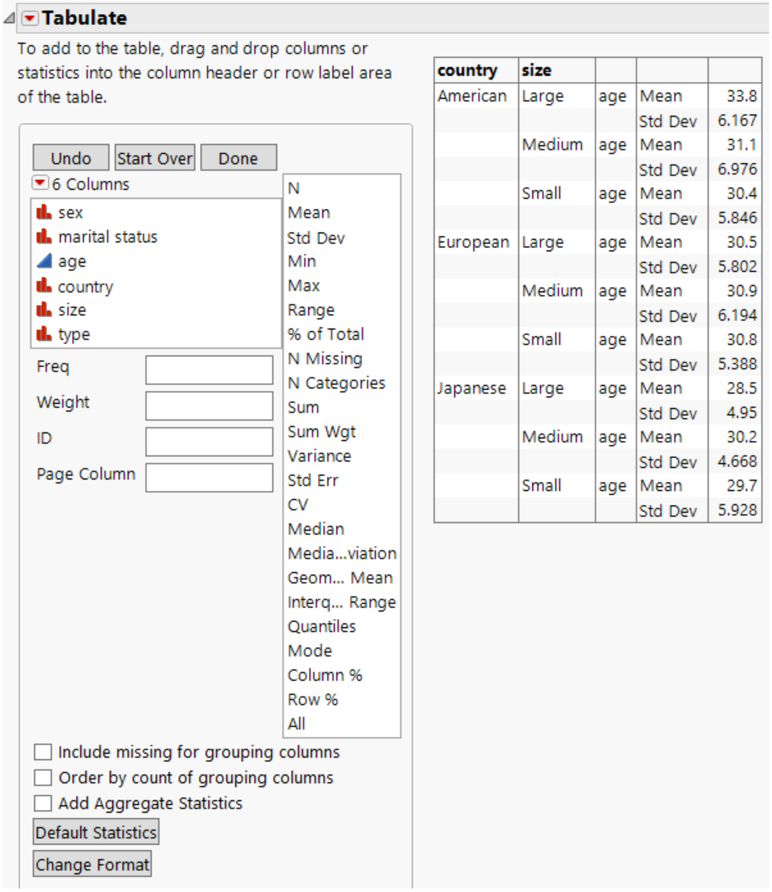


country	size			
American	Large	age	Mean	33.8
			Std Dev	6.167
	Medium	age	Mean	31.1
			Std Dev	6.976
	Small	age	Mean	30.4
			Std Dev	5.846
European	Large	age	Mean	30.5
			Std Dev	5.802
	Medium	age	Mean	30.9
			Std Dev	6.194
	Small	age	Mean	30.8
			Std Dev	5.388
Japanese	Large	age	Mean	28.5
			Std Dev	4.95
	Medium	age	Mean	30.2
			Std Dev	4.668
	Small	age	Mean	29.7
			Std Dev	5.928

1. Start from Figure 9.14. Click age and drag it to the right of the size heading.
2. Click Mean and drag it over Sum.
3. Click Std Dev and drag it below Mean.

   Std Dev is placed below Mean in the table. Dropping Std Dev above Mean places Std Dev above Mean in the table.

Figure 9.16 Age, Mean, and Std Dev Added to the Table



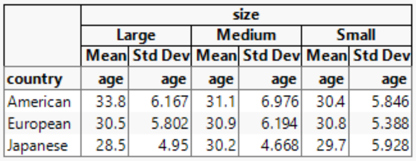
To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

country	size			
American	Large	age	Mean	33.8
			Std Dev	6.167
	Medium	age	Mean	31.1
			Std Dev	6.976
	Small	age	Mean	30.4
			Std Dev	5.846
European	Large	age	Mean	30.5
			Std Dev	5.802
	Medium	age	Mean	30.9
			Std Dev	6.194
	Small	age	Mean	30.8
			Std Dev	5.388
Japanese	Large	age	Mean	28.5
			Std Dev	4.95
	Medium	age	Mean	30.2
			Std Dev	4.668
	Small	age	Mean	29.7
			Std Dev	5.928

Rearrange the Table Contents

Suppose that you would prefer size to be on top, showing a crosstab layout.

Figure 9.17 Size on Top

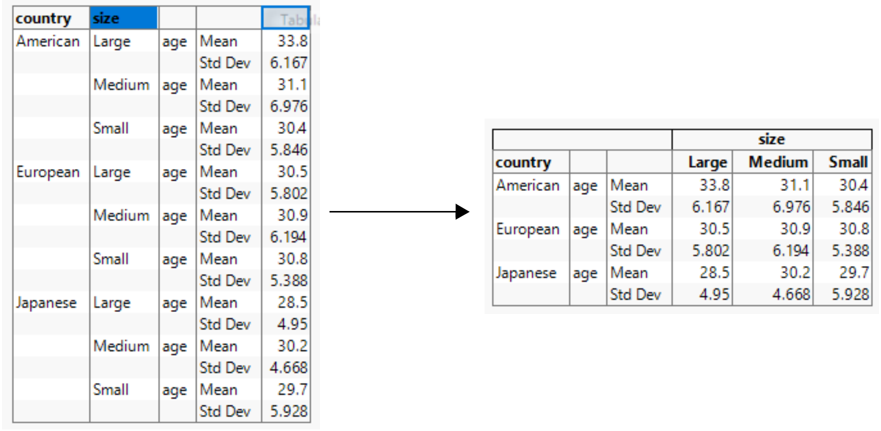


	size					
	Large		Medium		Small	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
country	age	age	age	age	age	age
American	33.8	6.167	31.1	6.976	30.4	5.846
European	30.5	5.802	30.9	6.194	30.8	5.388
Japanese	28.5	4.95	30.2	4.668	29.7	5.928

Rearrange the table contents.

1. Start from Figure 9.16. Click the size heading and drag it to the right of the table headings.

Figure 9.18 Moving size



country	size			Tab
American	Large	age	Mean	33.8
			Std Dev	6.167
	Medium	age	Mean	31.1
			Std Dev	6.976
	Small	age	Mean	30.4
			Std Dev	5.846
European	Large	age	Mean	30.5
			Std Dev	5.802
	Medium	age	Mean	30.9
			Std Dev	6.194
	Small	age	Mean	30.8
			Std Dev	5.388
Japanese	Large	age	Mean	28.5
			Std Dev	4.95
	Medium	age	Mean	30.2
			Std Dev	4.668
	Small	age	Mean	29.7
			Std Dev	5.928

			size		
country			Large	Medium	Small
American	age	Mean	33.8	31.1	30.4
		Std Dev	6.167	6.976	5.846
European	age	Mean	30.5	30.9	30.8
		Std Dev	5.802	6.194	5.388
Japanese	age	Mean	28.5	30.2	29.7
		Std Dev	4.95	4.668	5.928

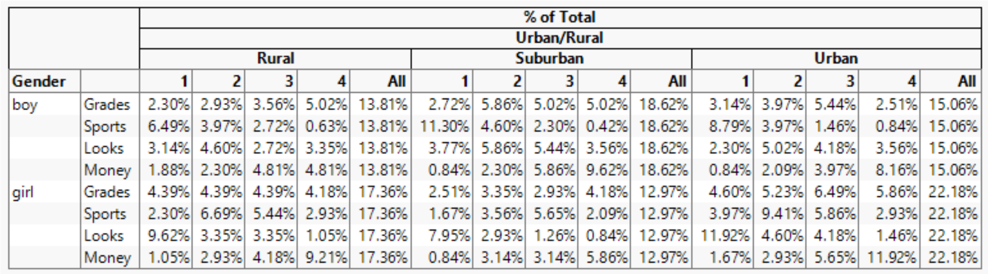
2. Click age and drag it under the Large Medium Small heading.
3. Select both Mean and Std Dev, and then drag them under the Large heading.

Now your table clearly presents the data. It is easier to see the mean and standard deviation of the car owner age broken down by car size and country.

Example of Combining Columns into a Single Table

In this example, you have data from students indicating the importance of self-reported factors in children's popularity (grades, sports, looks, money). Using the Tabulate platform, you want to see all of these factors in a single, combined table with additional statistics and factors.

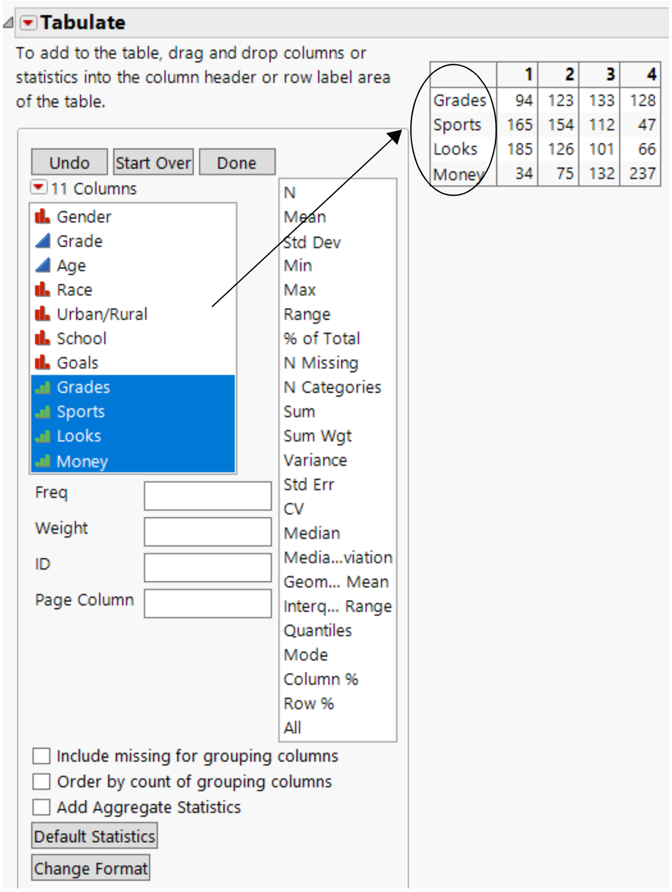
Figure 9.19 Adding Demographic Data



		% of Total														
		Urban/Rural														
		Rural					Suburban					Urban				
Gender		1	2	3	4	All	1	2	3	4	All	1	2	3	4	All
boy	Grades	2.30%	2.93%	3.56%	5.02%	13.81%	2.72%	5.86%	5.02%	5.02%	18.62%	3.14%	3.97%	5.44%	2.51%	15.06%
	Sports	6.49%	3.97%	2.72%	0.63%	13.81%	11.30%	4.60%	2.30%	0.42%	18.62%	8.79%	3.97%	1.46%	0.84%	15.06%
	Looks	3.14%	4.60%	2.72%	3.35%	13.81%	3.77%	5.86%	5.44%	3.56%	18.62%	2.30%	5.02%	4.18%	3.56%	15.06%
	Money	1.88%	2.30%	4.81%	4.81%	13.81%	0.84%	2.30%	5.86%	9.62%	18.62%	0.84%	2.09%	3.97%	8.16%	15.06%
girl	Grades	4.39%	4.39%	4.39%	4.18%	17.36%	2.51%	3.35%	2.93%	4.18%	12.97%	4.60%	5.23%	6.49%	5.86%	22.18%
	Sports	2.30%	6.69%	5.44%	2.93%	17.36%	1.67%	3.56%	5.65%	2.09%	12.97%	3.97%	9.41%	5.86%	2.93%	22.18%
	Looks	9.62%	3.35%	3.35%	1.05%	17.36%	7.95%	2.93%	1.26%	0.84%	12.97%	11.92%	4.60%	4.18%	1.46%	22.18%
	Money	1.05%	2.93%	4.18%	9.21%	17.36%	0.84%	3.14%	3.14%	5.86%	12.97%	1.67%	2.93%	5.65%	11.92%	22.18%

1. Select **Help > Sample Data Folder** and open Children's Popularity.jmp.
2. Select **Analyze > Tabulate**.
3. Select Grades, Sports, Looks, and Money and drag them into the Drop zone for rows.

Figure 9.20 Columns by Categories



To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

	1	2	3	4
Grades	94	123	133	128
Sports	165	154	112	47
Looks	185	126	101	66
Money	34	75	132	237

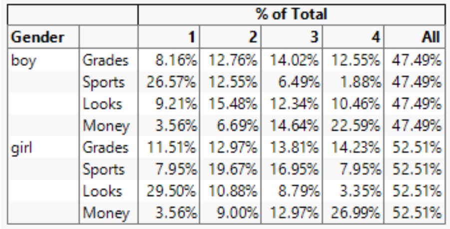
☐ Include missing for grouping columns
☐ Order by count of grouping columns
☐ Add Aggregate Statistics

Notice that a single, combined table appears.

Tabulate the percentage of the one to four ratings of each category.

4. Drag **Gender** into the empty heading at left.
5. Drag **% of Total** above the numbered headings.
6. Drag **All** beside the number 4.

Figure 9.21 Gender, % of Total, and All Added to the Table

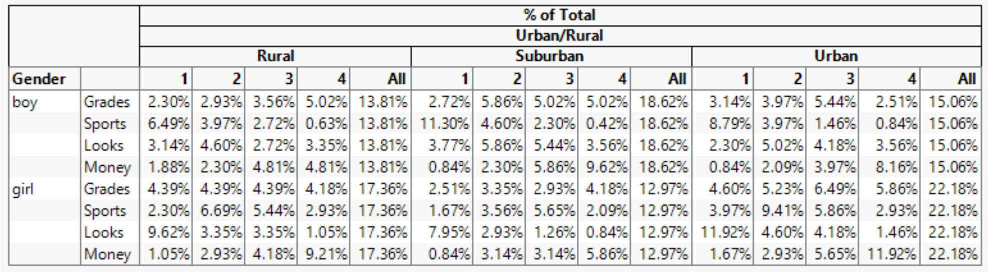


		% of Total				
Gender		**1**	**2**	**3**	**4**	**All**
boy	Grades	8.16%	12.76%	14.02%	12.55%	47.49%
	Sports	26.57%	12.55%	6.49%	1.88%	47.49%
	Looks	9.21%	15.48%	12.34%	10.46%	47.49%
	Money	3.56%	6.69%	14.64%	22.59%	47.49%
girl	Grades	11.51%	12.97%	13.81%	14.23%	52.51%
	Sports	7.95%	19.67%	16.95%	7.95%	52.51%
	Looks	29.50%	10.88%	8.79%	3.35%	52.51%
	Money	3.56%	9.00%	12.97%	26.99%	52.51%

Break down the tabulation further by adding demographic data.

7. Drag Urban/Rural below the % of Total heading.

Figure 9.22 Urban/Rural Added to the Table



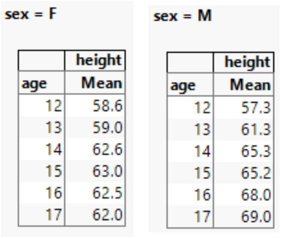
		% of Total														
		Urban/Rural														
		Rural					Suburban					Urban				
Gender		**1**	**2**	**3**	**4**	**All**	**1**	**2**	**3**	**4**	**All**	**1**	**2**	**3**	**4**	**All**
boy	Grades	2.30%	2.93%	3.56%	5.02%	13.81%	2.72%	5.86%	5.02%	5.02%	18.62%	3.14%	3.97%	5.44%	2.51%	15.06%
	Sports	6.49%	3.97%	2.72%	0.63%	13.81%	11.30%	4.60%	2.30%	0.42%	18.62%	8.79%	3.97%	1.46%	0.84%	15.06%
	Looks	3.14%	4.60%	2.72%	3.35%	13.81%	3.77%	5.86%	5.44%	3.56%	18.62%	2.30%	5.02%	4.18%	3.56%	15.06%
	Money	1.88%	2.30%	4.81%	4.81%	13.81%	0.84%	2.30%	5.86%	9.62%	18.62%	0.84%	2.09%	3.97%	8.16%	15.06%
girl	Grades	4.39%	4.39%	4.39%	4.18%	17.36%	2.51%	3.35%	2.93%	4.18%	12.97%	4.60%	5.23%	6.49%	5.86%	22.18%
	Sports	2.30%	6.69%	5.44%	2.93%	17.36%	1.67%	3.56%	5.65%	2.09%	12.97%	3.97%	9.41%	5.86%	2.93%	22.18%
	Looks	9.62%	3.35%	3.35%	1.05%	17.36%	7.95%	2.93%	1.26%	0.84%	12.97%	11.92%	4.60%	4.18%	1.46%	22.18%
	Money	1.05%	2.93%	4.18%	9.21%	17.36%	0.84%	3.14%	3.14%	5.86%	12.97%	1.67%	2.93%	5.65%	11.92%	22.18%

You can see that for boys in rural, suburban, and urban areas, sports are the most important factor for popularity. For girls in rural, suburban, and urban areas, looks are the most important factor for popularity.

Example Using a Page Column

In this example, you have data containing height measurements for male and female students. Using the Tabulate platform, you want to create a table that shows the mean of the heights by the age of the students. Then you want to stratify your data by sex in different tables. To do so, add the stratification column as a page column, which builds the pages for each group.

Figure 9.23 Mean Height of Students by Sex



sex = F

	height
age	Mean
12	58.6
13	59.0
14	62.6
15	63.0
16	62.5
17	62.0

sex = M

	height
age	Mean
12	57.3
13	61.3
14	65.3
15	65.2
16	68.0
17	69.0

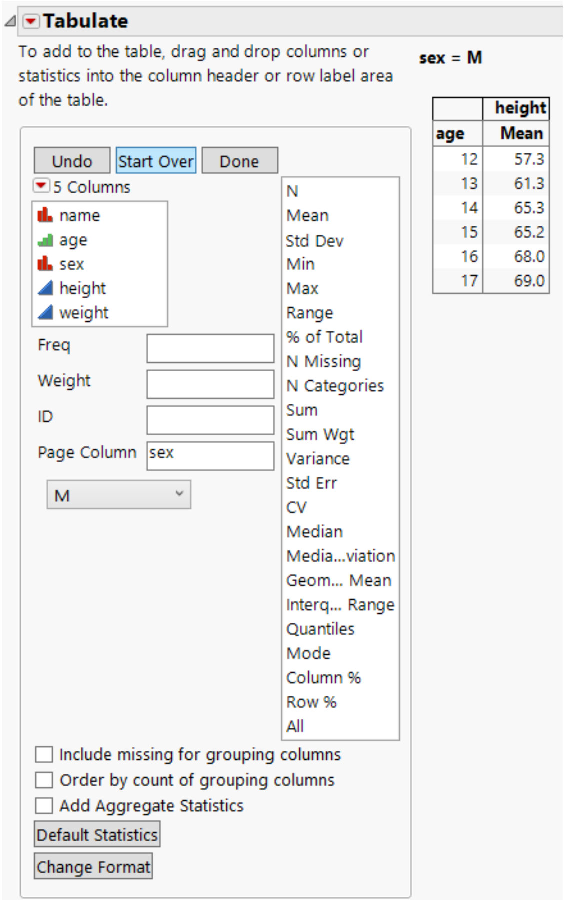
Females Males

1. Select **Help > Sample Data Folder** and open Big Class.jmp.
2. Select **Analyze > Tabulate**.

   Since height is the variable that you are examining, you want it to appear at the top of the table.
3. Click height and drag it into the Drop zone for columns.

   You want the statistics by age, and you want age to appear on the side.
4. Click age and drag it into the blank cell next to the number 2502.
5. Click Mean and drag it into the cell that says Sum.
6. Click sex and drag it into **Page Column**.
7. Select F from the Page Column list to show the mean of the heights for only females.
8. Select M from the Page Column list to show the mean of the heights for only males. You can also select **None Selected** to show all values.

Figure 9.24 Using a Page Column



To add to the table, drag and drop columns or statistics into the column header or row label area of the table.

Page Column: sex

M

☐ Include missing for grouping columns

☐ Order by count of grouping columns

☐ Add Aggregate Statistics

sex = M

	height
age	**Mean**
12	57.3
13	61.3
14	65.3
15	65.2
16	68.0
17	69.0

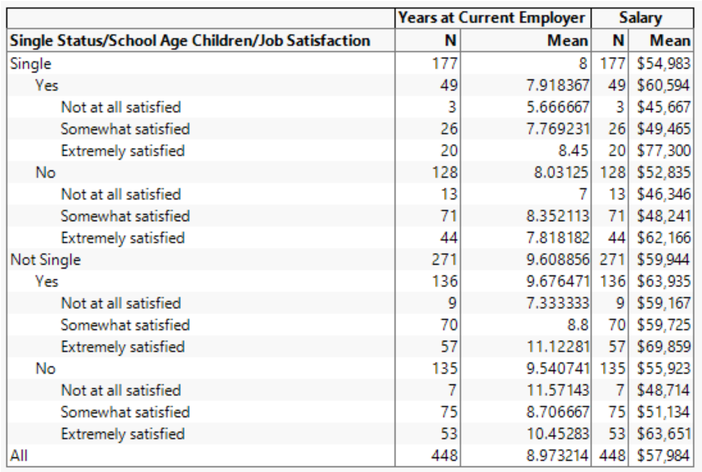
Example of Stacking Grouping Columns

In this example, you want to examine employment information for survey respondents based on a set of grouping variables. Because some of the variable names and values are long, you want to stack the grouping columns to conserve horizontal space in the report table.

1. Select **Help > Sample Data Folder** and open Consumer Preferences.jmp.
2. Select **Analyze > Tabulate**.
3. Select Years at Current Employer and Salary and drag them into the Drop zone for columns.
4. Select N and Mean and drag them into the cell that says Sum below Years at Current Employer.
5. Select Single Status, School Age Children, and Job Satisfaction and drag them to the blank cell on the left side of the table.

6. Right-click on the Single Status heading and select **Stack Grouping Columns**.
7. Select the **Add Aggregate Statistics** check box.

Figure 9.25 Stacked Grouping Columns



	Years at Current Employer		Salary	
Single Status/School Age Children/Job Satisfaction	**N**	**Mean**	**N**	**Mean**
Single	177	8	177	$54,983
Yes	49	7.918367	49	$60,594
Not at all satisfied	3	5.666667	3	$45,667
Somewhat satisfied	26	7.769231	26	$49,465
Extremely satisfied	20	8.45	20	$77,300
No	128	8.03125	128	$52,835
Not at all satisfied	13	7	13	$46,346
Somewhat satisfied	71	8.352113	71	$48,241
Extremely satisfied	44	7.818182	44	$62,166
Not Single	271	9.608856	271	$59,944
Yes	136	9.676471	136	$63,935
Not at all satisfied	9	7.333333	9	$59,167
Somewhat satisfied	70	8.8	70	$59,725
Extremely satisfied	57	11.12281	57	$69,859
No	135	9.540741	135	$55,923
Not at all satisfied	7	11.57143	7	$48,714
Somewhat satisfied	75	8.706667	75	$51,134
Extremely satisfied	53	10.45283	53	$63,651
All	448	8.973214	448	$57,984

The table shows a breakdown of mean salary, mean number of years at current employer, and number of respondents for each combination of the grouping variables. For example, there were 20 respondents who were single with school age children and who were extremely satisfied with their job. These 20 respondents have been at their current employer for an average of 8.45 years and have an average salary of $77,300.

The aggregate values for the Single Status and School Age Children variables also appear in the table. For example, not taking job satisfaction into account, there were 128 respondents who were single without school age children. These 128 respondents have been at their current employer for an average of 8.03 years and have an average salary of $52,835.

Chapter 10

Simulate

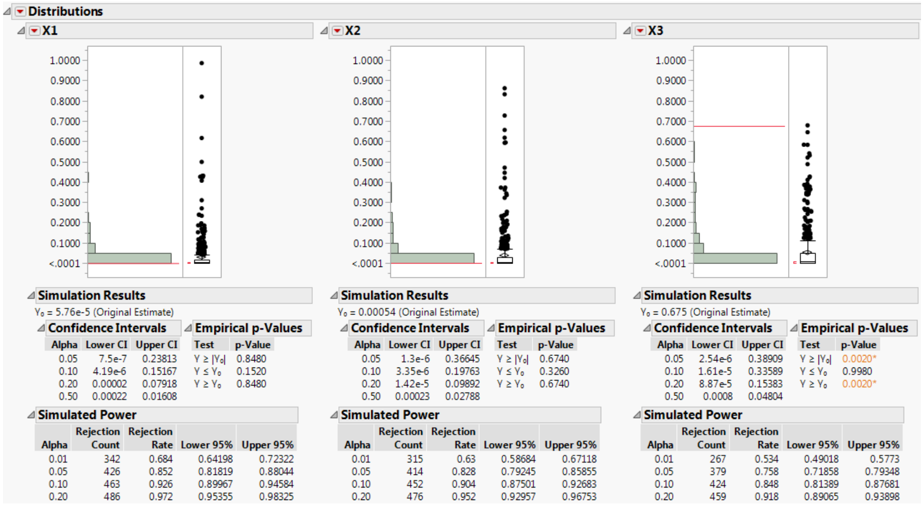
Answer Challenging Questions with Parametric Resampling

The Simulate feature provides powerful parametric and nonparametric simulation capability. Use Simulate to do the following:

- Expand on the bootstrap to provide parametric bootstrapping.
- Obtain power calculations in nonstandard situations.
- Approximate the distribution of statistics, such as predicted values, and confidence intervals, in nonstandard situations.
- Conduct permutation tests.
- Explore the effect of assumptions about predictors on models.
- Explore various "what if" scenarios relative to your models.
- Evaluate new or existing statistical methods.

The Simulate feature is available in many reports, including all of those that support Bootstrap. To access Simulate, right-click in a report.

Figure 10.1 Power Analysis Using Simulate



X1

Simulation Results

Y_0 = 5.76e-5 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	7.5e-7	0.23813
0.10	4.19e-6	0.15167
0.20	0.00002	0.07918
0.50	0.00022	0.01608

Empirical p-Values

Test	p-Value
$Y \ge \lvert Y_0 \rvert$	0.8480
$Y \le Y_0$	0.1520
$Y \ge Y_0$	0.8480

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	342	0.684	0.64198	0.72322
0.05	426	0.852	0.81819	0.88044
0.10	463	0.926	0.89967	0.94584
0.20	486	0.972	0.95355	0.98325

X2

Simulation Results

Y_0 = 0.00054 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	1.3e-6	0.36645
0.10	3.35e-6	0.19763
0.20	1.42e-5	0.09892
0.50	0.00023	0.02788

Empirical p-Values

Test	p-Value
$Y \ge \lvert Y_0 \rvert$	0.6740
$Y \le Y_0$	0.3260
$Y \ge Y_0$	0.6740

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	315	0.63	0.58684	0.67118
0.05	414	0.828	0.79245	0.85855
0.10	452	0.904	0.87501	0.92683
0.20	476	0.952	0.92957	0.96753

X3

Simulation Results

Y_0 = 0.675 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	2.54e-6	0.38909
0.10	1.61e-5	0.33589
0.20	8.87e-5	0.15383
0.50	0.0008	0.04804

Empirical p-Values

Test	p-Value
$Y \ge \lvert Y_0 \rvert$	0.0020*
$Y \le Y_0$	0.9980
$Y \ge Y_0$	0.0020*

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	267	0.534	0.49018	0.5773
0.05	379	0.758	0.71858	0.79348
0.10	424	0.848	0.81389	0.87681
0.20	459	0.918	0.89065	0.93898

Contents

Overview of the Simulate Feature

The Simulate feature provides simulated results for a column of statistics in a report. Right-click a column of statistics in a report and select Simulate. In the Simulate window, specify a column in your data table that forms the basis for your simulation. This is the column that you *switch out*. This column can have any role in the analysis. In particular, it can be a response or a predictor in a model. You then specify a column in your data table that contains a formula that you want to use for the simulation. This is the column that you *switch in*. It functions as a surrogate for the column that you switched out.

Note: Your data table must contain a column that has a random component.

The method works as follows. A column of simulated values is generated based on the formula in the formula column that you switch in. The entire analysis that generated the report containing the statistics of interest is rerun using this new column of simulated values to replace the column that you switched out. This process is repeated *N* times, where *N* is the total number of samples that you specify.

The Simulate analysis produces an output data table showing a summary of the analysis.

- Each row of the data table represents the results of the analysis for one column of simulated values.
- There is a column for each row of the report table involved in the simulation.
- There are scripts to facilitate your analysis.

Tip: The Simulate feature reruns the entire analysis that appears in the platform report from which Simulate is invoked. As a result, Simulate might run slowly for your selected column because of extraneous analyses in the report. If Simulate is taking a long time, remove extraneous options from the platform report before running Simulate.

The Simulate feature is available in all statistical platforms except the following: Association Analysis, Diagram, Multidimensional Scaling, Multiple Factor Analysis, Reliability Block Diagram, Reliability Forecast, Repairable Systems Simulation, Response Screening, and Text Explorer.

Example of the Simulate Feature

In this example, your goal is to obtain semiparametric confidence intervals for the variance components of hard-to-change variables. For this data, you are interested in the effects of temperature, time, and the amount of catalyst on a reaction. Temperature is a very-hard-to-change variable (whole plot factor), time is hard-to-change (subplot factor), and the amount of catalyst is easy-to-change. For information about whole plot and subplot factors, see the *Design of Experiments Guide*.

Previous studies have suggested that the whole-plot standard deviation is about twice the error standard deviation, and the sub-plot error is about 1.5 times the error standard deviation. The Wald intervals given in the REML report, which assume that the variance components are asymptotically normal, have poor coverage properties. You obtain confidence intervals using percentiles of the simulated distributions of the variance components.

Construct the Design

In this section, you construct a custom design for your split-split-plot experiment.

Tip: If you prefer to skip the steps in this section, select **Help > Sample Data Folder** and open Design Experiment/Catalyst Design.jmp. In the Catalyst Design.jmp data table, click the green triangle next to the **DOE Simulate** script. Then go to "Fit the Model".

1. Select **DOE > Custom Design**.
2. In the Factors outline, type 3 next to **Add N Factors**.
3. Click **Add Factor > Continuous**.
4. Double-click to rename these factors Temperature, Time, and Catalyst.

   Keep the default Values of –1 and 1 for these factors.
5. For Temperature, click **Easy** and select **Very Hard**.

   This defines Temperature to be a whole plot factor.
6. For Time, click **Easy** and select **Hard** for Time.

   This defines Time to be a subplot factor.
7. Click **Continue**.
8. In the Model outline, select **Interactions > 2nd**.

   This adds all two-way interactions to the model.
9. Click the Custom Design red triangle and select **Simulate Responses**.

   This opens the Simulate Responses window after you select Make Table to construct the design table.

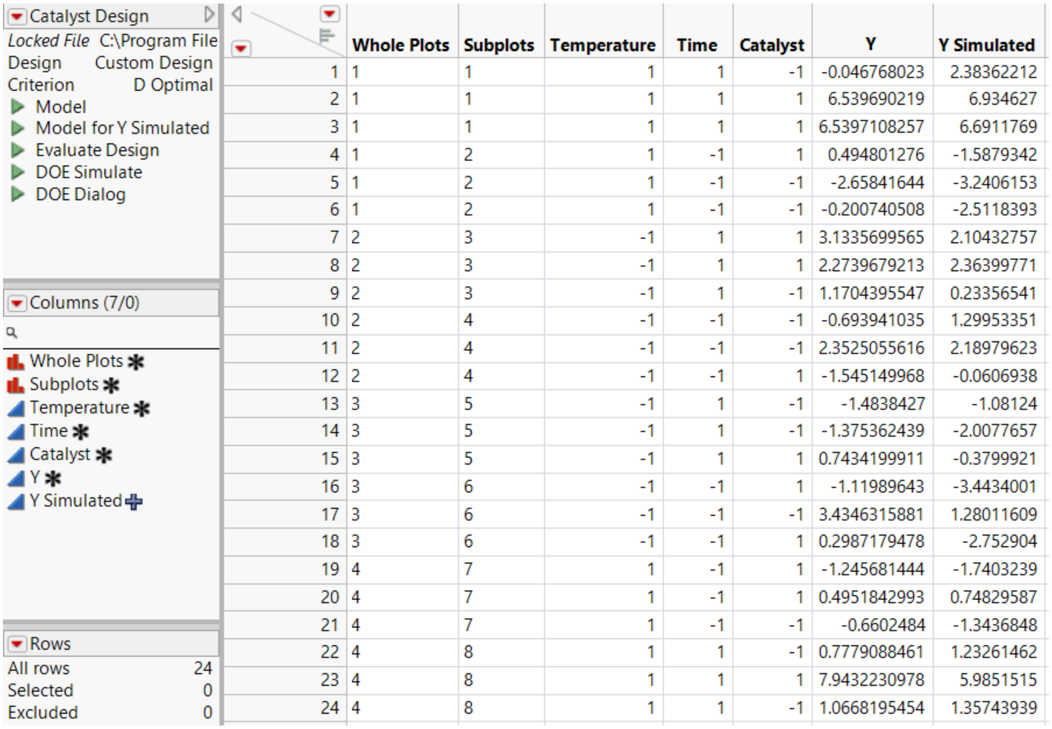
Note: Setting the Random Seed in step 10 and Number of Starts in step 11 reproduces the same design shown in this example. In constructing a design on your own, these steps are not necessary.

---

10. (Optional) Click the Custom Design red triangle and select **Set Random Seed**. Type 12345 and click **OK**.
11. (Optional) Click the Custom Design red triangle and select **Number of Starts**. Type 1000 and click **OK**.
12. Click **Make Design**.
13. Click **Make Table**.

Note: The entries in your Y and Y Simulated columns might differ from those that appear in Figure 10.2.

Figure 10.2 Design Table



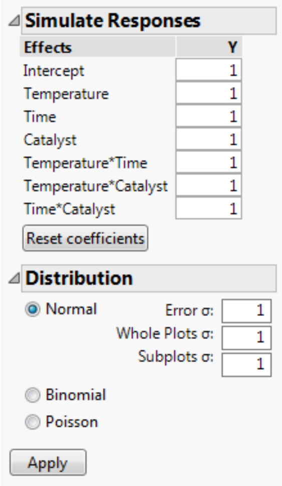
Catalyst Design
Locked File C:\Program File
Design Custom Design
Criterion D Optimal
▶ Model
▶ Model for Y Simulated
▶ Evaluate Design
▶ DOE Simulate
▶ DOE Dialog

Columns (7/0)
Whole Plots ✱
Subplots ✱
Temperature ✱
Time ✱
Catalyst ✱
Y ✱
Y Simulated ✚

Rows
All rows 24
Selected 0
Excluded 0

	Whole Plots	Subplots	Temperature	Time	Catalyst	Y	Y Simulated
1	1	1	1	1	-1	-0.046768023	2.38362212
2	1	1	1	1	1	6.539690219	6.934627
3	1	1	1	1	1	6.5397108257	6.6911769
4	1	2	1	-1	1	0.494801276	-1.5879342
5	1	2	1	-1	-1	-2.65841644	-3.2406153
6	1	2	1	-1	-1	-0.200740508	-2.5118393
7	2	3	-1	1	1	3.1335699565	2.10432757
8	2	3	-1	1	1	2.2739679213	2.36399771
9	2	3	-1	1	-1	1.1704395547	0.23356541
10	2	4	-1	-1	-1	-0.693941035	1.29953351
11	2	4	-1	-1	-1	2.3525055616	2.18979623
12	2	4	-1	-1	1	-1.545149968	-0.0606938
13	3	5	-1	1	-1	-1.4838427	-1.08124
14	3	5	-1	1	-1	-1.375362439	-2.0077657
15	3	5	-1	1	1	0.7434199911	-0.3799921
16	3	6	-1	-1	1	-1.11989643	-3.4434001
17	3	6	-1	-1	-1	3.4346315881	1.28011609
18	3	6	-1	-1	1	0.2987179478	-2.752904
19	4	7	1	-1	1	-1.245681444	-1.7403239
20	4	7	1	-1	1	0.4951842993	0.74829587
21	4	7	1	-1	-1	-0.6602484	-1.3436848
22	4	8	1	1	-1	0.7779088461	1.23261462
23	4	8	1	1	1	7.9432230978	5.9851515
24	4	8	1	1	-1	1.0668195454	1.35743939

Figure 10.3 Simulate Responses Window



Effects	Y
Intercept	1
Temperature	1
Time	1
Catalyst	1
Temperature*Time	1
Temperature*Catalyst	1
Time*Catalyst	1

● Normal Error σ: 1
Whole Plots σ: 1
Subplots σ: 1
○ Binomial
○ Poisson

The design table and a Simulate Responses window appear. Notice that the design table contains a **DOE Simulate** script. At any time, you can run this script to specify different parameter values.

Continue to the next section, where you specify standard deviations for the whole plot and subplot errors, and fit a REML model to the first set of simulated values.

Fit the Model

In this section, you fit a model using the REML method. Assume that the whole plot and subplot errors are normal. Based on your estimates of their standard deviations, if the error standard deviation is about 1 unit, the whole plot standard deviation is about 2 units and the subplot standard deviation is about 1.5 units. Since you are interested only in the whole- and sub-plot variation, you do not need to change the values assigned to Effects in the Simulate Responses outline.

1. In the Distribution panel (Figure 10.3), next to **Whole Plots** σ, type 2.

   Notice that the Normal distribution is selected by default. As a result, normal error is added to the formula.

2. Next to **Subplots** σ, type 1.5.
3. Click **Apply**.

   In the data table, the formula for Y Simulated updates to reflect your specifications. To view the formula, click the plus sign to the right of the column name in the Columns panel.

4. In the data table, click the green triangle next to the **Model** script.
5. Click the Y variable next to the **Y** button and click **Remove**.

6. Click Y Simulated and click the **Y** button.

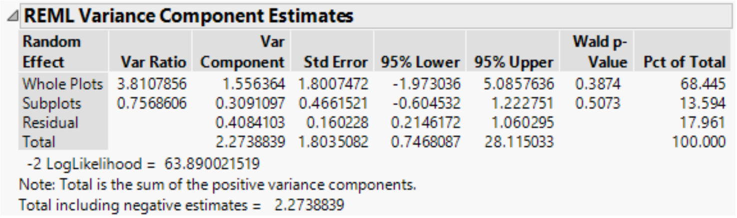
   This action replaces Y with a column that contains a simulation formula.

7. Click **Run**.

   The model that is fit is based on a single set of simulated responses.

Note: Because the values in Y Simulated are randomly generated, the entries in your report might differ from those that appear in Figure 10.4.

Figure 10.4 REML Report Showing Wald Confidence Intervals



REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
Whole Plots	3.8107856	1.556364	1.8007472	-1.973036	5.0857636	0.3874	68.445
Subplots	0.7568606	0.3091097	0.4661521	-0.604532	1.222751	0.5073	13.594
Residual		0.4084103	0.160228	0.2146172	1.060295		17.961
Total		2.2738839	1.8035082	0.7468087	28.115033		100.000

-2 LogLikelihood = 63.890021519

Note: Total is the sum of the positive variance components.

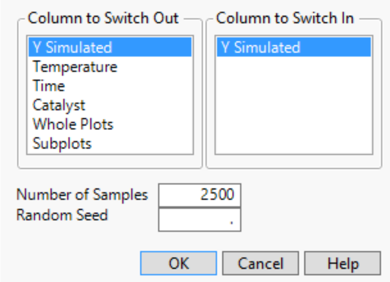
Total including negative estimates = 2.2738839

Generate Confidence Intervals

In this section, you simulate variance component estimates and use these to construct simulated percentile confidence intervals.

1. In the REML Variance Components Estimates outline, right-click in the **Var Component** column and select **Simulate**.

Figure 10.5 Simulate Window



Number of Samples 2500

Random Seed .

In your simulations, you replace the column Y Simulated, which you used to run your model, with a new instance of the column Y Simulated, which generates a new column of simulated values for each simulation. The column on which you right-clicked and that appears as selected, **Var Component**, is simulated for each effect listed in the Parameter Estimates table.

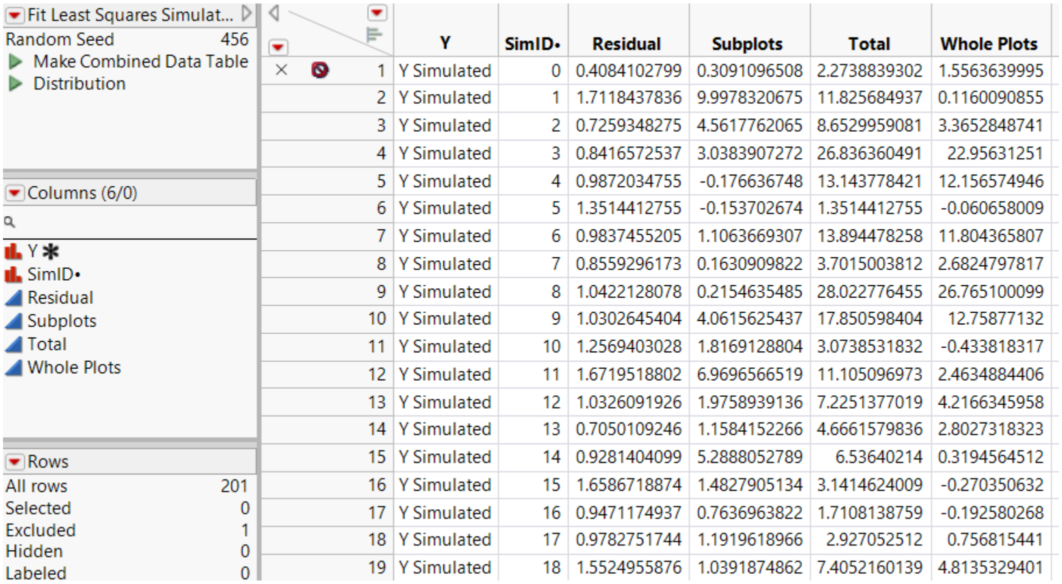
2. Next to **Number of Samples**, enter **200**.
3. (Optional) Next to **Random Seed**, enter 456.

This reproduces the values shown in Figure 10.6, except for the values in row 1.

4. Click **OK**.

The entries in your row 1 might differ from those that appear in Figure 10.6.

Figure 10.6 Table of Simulated Results for Var Component (Partial View)



Fit Least Squares Simulat...

Random Seed	456

Make Combined Data Table
Distribution

Columns (6/0)

Y
SimID•
Residual
Subplots
Total
Whole Plots

Rows

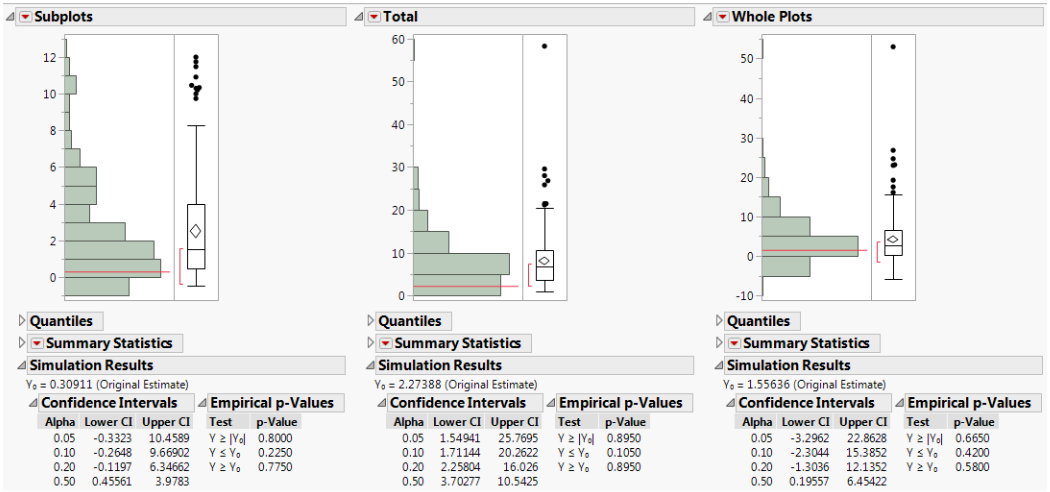
All rows	201
Selected	0
Excluded	1
Hidden	0
Labeled	0

	Y	SimID•	Residual	Subplots	Total	Whole Plots
1	Y Simulated	0	0.4084102799	0.3091096508	2.2738839302	1.5563639995
2	Y Simulated	1	1.7118437836	9.9978320675	11.825684937	0.1160090855
3	Y Simulated	2	0.7259348275	4.5617762065	8.6529959081	3.3652848741
4	Y Simulated	3	0.8416572537	3.0383907272	26.836360491	22.95631251
5	Y Simulated	4	0.9872034755	-0.176636748	13.143778421	12.156574946
6	Y Simulated	5	1.3514412755	-0.153702674	1.3514412755	-0.060658009
7	Y Simulated	6	0.9837455205	1.1063669307	13.894478258	11.804365807
8	Y Simulated	7	0.8559296173	0.1630909822	3.7015003812	2.6824797817
9	Y Simulated	8	1.0422128078	0.2154635485	28.022776455	26.765100099
10	Y Simulated	9	1.0302645404	4.0615625437	17.850598404	12.75877132
11	Y Simulated	10	1.2569403028	1.8169128804	3.0738531832	-0.433818317
12	Y Simulated	11	1.6719518802	6.9696566519	11.105096973	2.4634884406
13	Y Simulated	12	1.0326091926	1.9758939136	7.2251377019	4.2166345958
14	Y Simulated	13	0.7050109246	1.1584152266	4.6661579836	2.8027318323
15	Y Simulated	14	0.9281404099	5.2888052789	6.53640214	0.3194564512
16	Y Simulated	15	1.6586718874	1.4827905134	3.1414624009	-0.270350632
17	Y Simulated	16	0.9471174937	0.7636963822	1.7108138759	-0.192580268
18	Y Simulated	17	0.9782751744	1.1919618966	2.927052512	0.756815441
19	Y Simulated	18	1.5524955876	1.0391874862	7.4052160139	4.8135329401

The first row of the Fit Least Squares Simulate Results (Var Component) data table contains the initial values of **Var Component** and is excluded. The remaining rows contain simulated values.

5. Run the **Distribution** script.

Figure 10.7 Distribution Plots for Variance Components (Partial View)



Subplots

Y_0 = 0.30911 (Original Estimate)

Alpha	Lower CI	Upper CI
0.05	-0.3323	10.4589
0.10	-0.2648	9.66902
0.20	-0.1197	6.34662
0.50	0.45561	3.9783

Test	p-Value
$Y \geq \lvert Y_0 \rvert$	0.8000
$Y \leq Y_0$	0.2250
$Y \geq Y_0$	0.7750

Total

Y_0 = 2.27388 (Original Estimate)

Alpha	Lower CI	Upper CI
0.05	1.54941	25.7695
0.10	1.71144	20.2622
0.20	2.25804	16.026
0.50	3.70277	10.5425

Test	p-Value
$Y \geq \lvert Y_0 \rvert$	0.8950
$Y \leq Y_0$	0.1050
$Y \geq Y_0$	0.8950

Whole Plots

Y_0 = 1.55636 (Original Estimate)

Alpha	Lower CI	Upper CI
0.05	-3.2962	22.8628
0.10	-2.3044	15.3852
0.20	-1.3036	12.1352
0.50	0.19557	6.45422

Test	p-Value
$Y \geq \lvert Y_0 \rvert$	0.6650
$Y \leq Y_0$	0.4200
$Y \geq Y_0$	0.5800

For each variance component, confidence intervals at various confidence levels are shown in the Simulation Results report. Compare the 95% intervals in the Alpha=0.05 row of each table to the intervals given in the REML report (Figure 10.4):

- The simulated 95% confidence interval for the whole-plot variance component is -3.296 to 22.863. The Wald interval given in the REML report is -1.973 to 5.086.
- The simulated 95% confidence interval for the sub-plot variance component is -0.332 to 10.459. The Wald interval given in the REML report is -0.605 to 1.223.

The intervals that you obtain using simulation are considerably wider than the REML interval calculated from your single set of values. For more precise intervals, consider running a larger number of simulations.

Launch the Simulate Feature

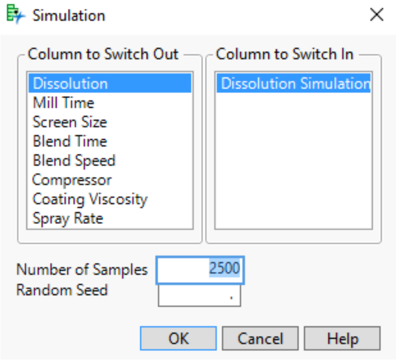
To launch the Simulate feature, right-click a column of calculated values in a JMP report window and select Simulate. To use the Simulate feature, the data table must contain a formula with a random component that simulates data. The Simulate option is available in many reports, including all reports that support bootstrapping.

Note: The Simulate feature is not available in reports that use a By variable.

The Simulate Window

Specify columns and the number of samples in the Simulate window, then run the Simulation by clicking OK. A progress bar and a Stop Early option appears. The number of the sample being simulated appears above the progress bar. If you click Stop Early, the simulated values that have been computed up to that point are presented in a Simulate Results table. The window also shows you which analyses are being run at any given time.

Figure 10.8 Simulate Window for Tablet Production.jmp



Number of Samples 2500
Random Seed .

The Simulate window contains these panels and options:

Column to Switch Out The column that is replaced by the Column to Switch In.

Column to Switch In The column that replaces the Column to Switch Out. The analysis is repeated with values simulated according to the formula in the Column to Switch In. Only columns with formulas are listed in the Column to Switch In panel.

Number of Samples Number of times that the report is re-run for a set of simulated data. The default value is 2500.

Random Seed A value that controls the simulated results. The random seed makes the results reproducible.

The Simulate Results Table

When you run a simulation, the results appear in a data table. Note the following:

- The first row of the table contains the values for the table items that appear in the report. For this reason, the first row is always excluded.
- The remaining rows give the simulation results. The number of remaining rows is equal to the Number of Samples that you specified in the Simulate launch window.

- The rows in the report are identified by the first column in the report table that contains the selected column of calculated values. A column appears in the simulated results table for each item in this first column.
- The table contains a **Distribution** script that constructs a Distribution report. This report contains histograms, quantiles, summary statistics, and simulation results for each column in the simulated results data table. In addition to the standard Distribution report, the report contains the following items:
  - A red line that denotes the original estimate appears on the histogram.
  - A Simulation Results report containing the original estimate, as well as confidence intervals and empirical *p*-values for the simulation. See "Simulation Results Report".
  - If the values in the simulated results data table have a PValue format, a Simulated Power report is also provided. See "Simulated Power Report".
- The table contains a **Power Analysis** script only if you have simulated a column of *p*-values. This script constructs a Distribution report showing histograms of *p*-values and provides a Simulated Power report. See "Simulated Power Report".

Simulation Results Report

The Simulation Results report contains the following information:

Original Estimate The value of the original estimate, also shown in the first row of the Simulate Results data table. This estimate is labeled Y_0.

Confidence Intervals Lower and upper limits for quantile-based confidence intervals at the following significance levels: 0.05, 0.10, 0.20, and 0.50.

Empirical p-Values The empirical *p*-values for a two-sided test and both one-sided tests that compare the simulated values to the original estimate. These *p*-values are computed as the proportions of the simulated values that fall in the ranges that are specified in the Test column of the report.

Simulated Power Report

The Simulated Power report contains the following information:

Alpha The significance level: 0.01, 0.05, 0.10, and 0.20.

Rejection Count The number of simulations where the test rejects at the corresponding significance level.

Rejection Rate The proportion of simulations where the test rejects at the corresponding significance level.

Lower 95% and Upper 95% Lower and upper limits for a 95% confidence interval for the simulated rejection rate. The interval is computed using the Wilson score method. See Wilson (1927).

Tip: Increase the Number of Samples for a narrower confidence interval.

Additional Examples of the Simulate Feature

This section contains additional examples using the Simulate feature.

- "Example of a Permutation Test"
- "Example of Retaining a Factor in Generalized Regression"
- "Example of a Prospective Power Analysis"

For an example that shows how to simulate a confidence interval for Ppk and the percent nonconforming for a non-normal variable, see *Quality and Process Methods*.

Example of a Permutation Test

In this example, you use Simulate to conduct a randomization or permutation test. You are studying the effects of three drugs on pain, and are interested in whether they differ in their effects. Because you have a very small sample size and are somewhat concerned about violations of the usual ANOVA assumptions, you can use Simulate to conduct a permutation test.

First, you construct a formula that randomly shuffles the pain measurements among the three drugs. Under the null hypothesis of no effect, any of these allocations is as likely as any other. It follows that the *F* ratios obtained in this manner approximate the distribution of *F* ratios under the null hypothesis. Finally, you compare the observed value of the *F* ratio to the null distribution obtained by simulation.

Define the Simulation Formula

1. Select **Help > Sample Data Folder** and open Analgesics.jmp.
2. Select **Cols > New Columns**.
3. Type Pain Shuffled for Column Name.
4. From the Column Properties list, select **Formula**.
5. In the function list, select **Row > Col Stored Value**.
6. In the Columns list, double-click pain.
7. Click the insert key (^) in the list of symbols above the editor panel.

8. From the list of functions, select **Random > Col Shuffle**.

Figure 10.9 Completed Formula

Col Stored Value (*pain*, Col Shuffle ())

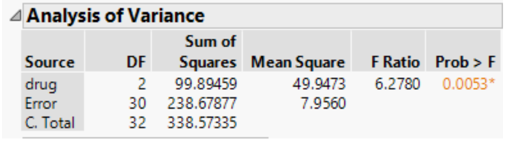
This formula randomly shuffles the entries in the pain column.

9. Click **OK** in the Formula Editor window.
10. Click **OK** in the Column Info window.

Perform the Permutation Test

1. Select **Analyze > Fit Y by X**.
2. Select pain and click **Y, Response**.
3. Select drug and click **X, Factor**.
4. Click **OK**.
5. Click the Oneway Analysis red triangle and select **Means/Anova**.

Figure 10.10 Analysis of Variance Report



Analysis of Variance

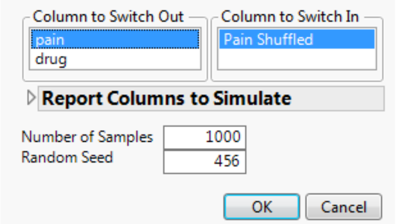
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
drug	2	99.89459	49.9473	6.2780	0.0053*
Error	30	238.67877	7.9560		
C. Total	32	338.57335			

Notice that the *F* ratio is 6.2780.

6. In the Analysis of Variance outline, right-click the F Ratio column and select **Simulate**.
7. In the Column to Switch Out list, click pain.
8. In the Column to Switch In list, click Pain Shuffled.
9. Next to **Number of Samples**, enter 1000.
10. (Optional) Next to **Random Seed**, enter 456.

This reproduces the values in this example.

Figure 10.11 Completed Simulate Window



Column to Switch Out: pain, drug (pain selected)

Column to Switch In: Pain Shuffled

Report Columns to Simulate

Number of Samples: 1000

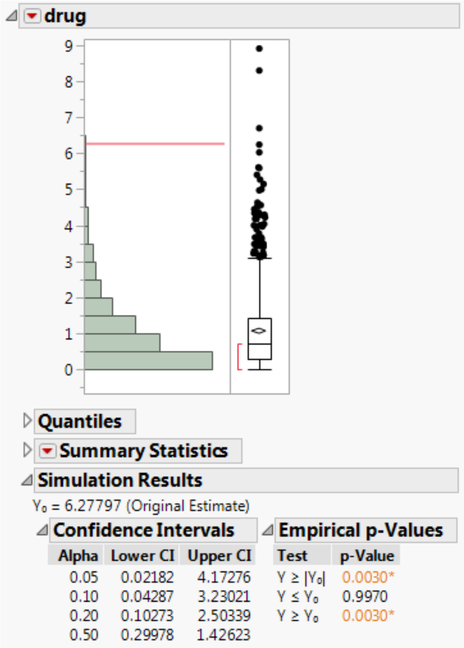
Random Seed: 456

11. Click **OK**.

    In the table of simulated results, the C. Total and Error columns are empty, since the *F* Ratio value in the Analysis of Variance table applies only to drug.

12. In the table of simulated values, run the **Distribution** script.

Figure 10.12 Simulated Distribution of *F* Ratios under the Null Distribution



Simulation Results

Y_0 = 6.27797 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	0.02182	4.17276
0.10	0.04287	3.23021
0.20	0.10273	2.50339
0.50	0.29978	1.42623

Empirical p-Values

Test	p-Value
$Y \geq \lvert Y_0 \rvert$	0.0030*
$Y \leq Y_0$	0.9970
$Y \geq Y_0$	0.0030*

The observed *F* ratio value of 6.2780 is represented with a red line in the histogram. This value falls in the upper 0.5% of the simulated null distribution of *F* ratios. This presents strong evidence that the three drugs differ in their effects on pain.

Example of Retaining a Factor in Generalized Regression

In this example, you construct a generalized regression model and use the active parameters to fit a reduced model. Based on this reduced model, you use the Simulate feature to explore the likelihood that one of the factors is included in the model.

A pharmaceutical manufacturer has historical information about the dissolution of a tablet inside the body and various factors that can affect the dissolution rate. A tablet with a dissolution rate below 70 is considered defective. You want to understand which factors affect dissolution rate.

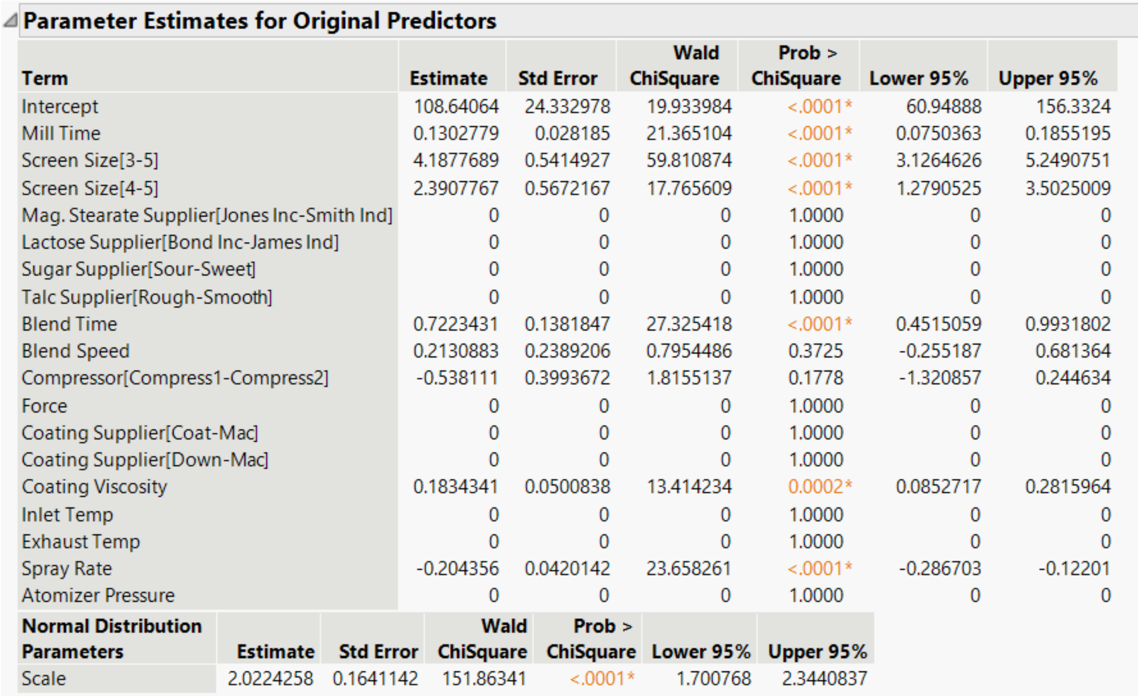
Fit the Model

In this section, you fit a model using generalized regression.

Tip: If you prefer not to work through the steps in this section, click the green triangle next to the Generalized Regression script in the Tablet Production.jmp data table to obtain the model.

1. Select **Help > Sample Data Folder** and open Tablet Production.jmp.
2. Select **Analyze > Fit Model**.
3. Click Dissolution and click **Y**.
4. Select Mill Time through Atomizer Pressure and click **Add**.
5. From the Personality list, select **Generalized Regression**.
6. Click **Run**.
7. In the Model Launch panel, select the **Adaptive** box.
8. In the Model Launch panel, click **Go**.

Figure 10.13 Model Based on Adaptive Lasso



Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	108.64064	24.332978	19.933984	<.0001*	60.94888	156.3324
Mill Time	0.1302779	0.028185	21.365104	<.0001*	0.0750363	0.1855195
Screen Size[3-5]	4.1877689	0.5414927	59.810874	<.0001*	3.1264626	5.2490751
Screen Size[4-5]	2.3907767	0.5672167	17.765609	<.0001*	1.2790525	3.5025009
Mag. Stearate Supplier[Jones Inc-Smith Ind]	0	0	0	1.0000	0	0
Lactose Supplier[Bond Inc-James Ind]	0	0	0	1.0000	0	0
Sugar Supplier[Sour-Sweet]	0	0	0	1.0000	0	0
Talc Supplier[Rough-Smooth]	0	0	0	1.0000	0	0
Blend Time	0.7223431	0.1381847	27.325418	<.0001*	0.4515059	0.9931802
Blend Speed	0.2130883	0.2389206	0.7954486	0.3725	-0.255187	0.681364
Compressor[Compress1-Compress2]	-0.538111	0.3993672	1.8155137	0.1778	-1.320857	0.244634
Force	0	0	0	1.0000	0	0
Coating Supplier[Coat-Mac]	0	0	0	1.0000	0	0
Coating Supplier[Down-Mac]	0	0	0	1.0000	0	0
Coating Viscosity	0.1834341	0.0500838	13.414234	0.0002*	0.0852717	0.2815964
Inlet Temp	0	0	0	1.0000	0	0
Exhaust Temp	0	0	0	1.0000	0	0
Spray Rate	-0.204356	0.0420142	23.658261	<.0001*	-0.286703	-0.12201
Atomizer Pressure	0	0	0	1.0000	0	0

Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	2.0224258	0.1641142	151.86341	<.0001*	1.700768	2.3440837

You are interested in the parameter estimates shown in the Normal Adaptive Lasso with AICc Validation report. Based on the nonzero parameter estimates, the model suggests that Mill Time, Screen Size, Blend Time, Blend Speed, Compressor, Coating Viscosity, and Spray Rate are related to Dissolution.

Reduce the Model

Before reducing the model, ensure that no columns are selected in the Tablet Production.jmp data table. Selected columns are not deselected in the first step below. Ensuring that no columns are selected prevents the inadvertent inclusion of columns with zeroed terms.

If you prefer not to work through the steps in this section, click the green triangle next to the **Generalized Regression Reduced Model** script in the Tablet Production.jmp data table to obtain the reduced model.

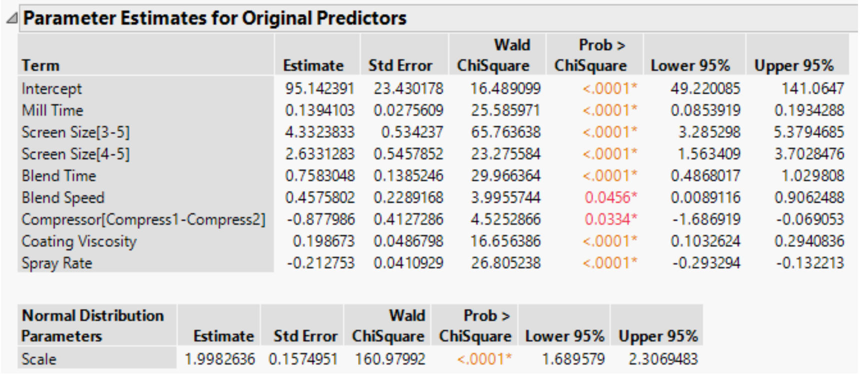
1. Click the red triangle next to Normal Adaptive Lasso with AICc Validation and select **Relaunch Active Set > Relaunch with Active Effects**.

   This opens a Fit Model window that places the terms with nonzero coefficient estimates in the Parameter Estimates reports into the Construct Model Effects list. The response is entered as Y. The Generalized Regression personality is selected.

2. Click **Run**.

3. In the Model Launch panel, select the **Adaptive** box.
4. In the Model Launch panel, click **Go**.

Figure 10.14 Reduced Model Using Adaptive Lasso



Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	95.142391	23.430178	16.489099	<.0001*	49.220085	141.0647
Mill Time	0.1394103	0.0275609	25.585971	<.0001*	0.0853919	0.1934288
Screen Size[3-5]	4.3323833	0.534237	65.763638	<.0001*	3.285298	5.3794685
Screen Size[4-5]	2.6331283	0.5457852	23.275584	<.0001*	1.563409	3.7028476
Blend Time	0.7583048	0.1385246	29.966364	<.0001*	0.4868017	1.029808
Blend Speed	0.4575802	0.2289168	3.9955744	0.0456*	0.0089116	0.9062488
Compressor[Compress1-Compress2]	-0.877986	0.4127286	4.5252866	0.0334*	-1.686919	-0.069053
Coating Viscosity	0.198673	0.0486798	16.656386	<.0001*	0.1032624	0.2940836
Spray Rate	-0.212753	0.0410929	26.805238	<.0001*	-0.293294	-0.132213

Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	1.9982636	0.1574951	160.97992	<.0001*	1.689579	2.3069483

Notice that the estimate for Blend Speed has a confidence interval (Lower 95%) that comes very close to including zero. Next, perform a simulation study to see how often Blend Speed would be included in the model if other data values from the dissolution distribution have been observed.

Explore the Inclusion of Blend Speed in the Model

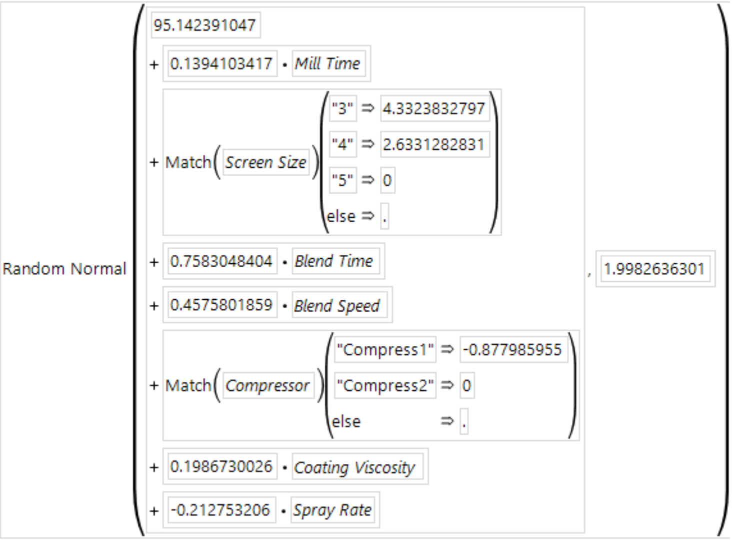
Use the report for the reduced model (Figure 10.14) in the steps below.

1. Click the red triangle next to Normal Adaptive Lasso with AICc Validation and select **Save Columns > Save Simulation Formula**.

   This adds a new column called Dissolution Simulation Formula to the Tablet Production.jmp data table.

2. (Optional) In the data table Columns panel, click the plus sign to the right of Dissolution Simulation Formula.

Figure 10.15 Simulation Formula



```
Random Normal( 95.142391047
             + 0.1394103417 • Mill Time
             + Match( Screen Size ) ( "3" ⇒ 4.3323832797
                                      "4" ⇒ 2.6331282831
                                      "5" ⇒ 0
                                      else ⇒ . )
             + 0.7583048404 • Blend Time
             + 0.4575801859 • Blend Speed
             + Match( Compressor ) ( "Compress1" ⇒ -0.877985955
                                     "Compress2" ⇒ 0
                                     else        ⇒ . )
             + 0.1986730026 • Coating Viscosity
             + -0.212753206 • Spray Rate , 1.9982636301 )
```

For each row, this formula simulates a value that could be obtained given the model and the distribution of Dissolution, which is estimated to be Normal with standard deviation about 1.998.

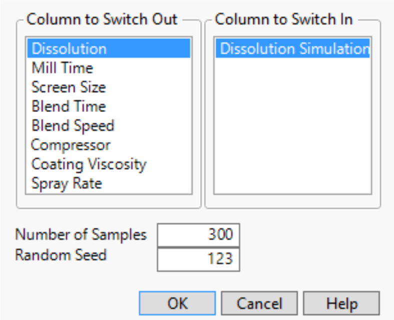
3. Click **Cancel**.
4. Go back to the reduced model report window. In the Parameter Estimates for Original Predictors report, right-click in the Estimate column and select **Simulate**.

   Make sure that Dissolution is selected in the Column to Switch Out list.
5. Next to **Number of Samples**, enter 300.

   For the simulation, you ask JMP to replace the Dissolution column in each of 300 analyses with values simulated using the Dissolution Simulation Formula column.
6. (Optional) Set the **Random Seed** to 123.

   This reproduces the values in this example.

Figure 10.16 Completed Simulation Window



Column to Switch Out: Dissolution

Column to Switch In: Dissolution Simulation

Number of Samples: 300

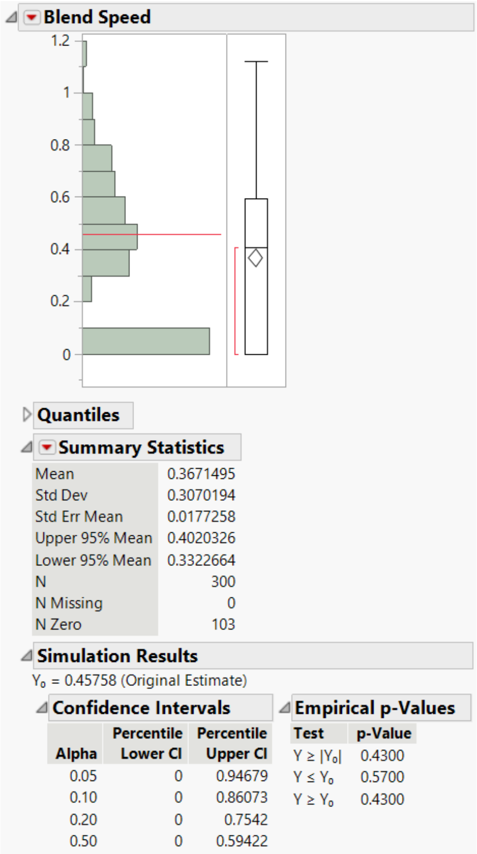
Random Seed: 123

7. Click **OK**.

   The first row of the table contains the initial values of the Estimates and is excluded. The remaining rows contain simulated values.

8. Run the **Distribution** script.
9. Press Ctrl, click the Blend Speed red triangle and select **Display Options > Customize Summary Statistics**.
10. Select **N Zero**.
11. Click **OK**.
12. Scroll to the Distribution report for Blend Speed.

Figure 10.17 Histogram of Simulated Blend Speed Coefficient Estimates



Summary Statistics

Mean	0.3671495
Std Dev	0.3070194
Std Err Mean	0.0177258
Upper 95% Mean	0.4020326
Lower 95% Mean	0.3322664
N	300
N Missing	0
N Zero	103

Simulation Results

Y_0 = 0.45758 (Original Estimate)

Confidence Intervals

Alpha	Percentile Lower CI	Percentile Upper CI
0.05	0	0.94679
0.10	0	0.86073
0.20	0	0.7542
0.50	0	0.59422

Empirical p-Values

Test	p-Value
$Y \geq \lvert Y_0 \rvert$	0.4300
$Y \leq Y_0$	0.5700
$Y \geq Y_0$	0.4300

The Summary Statistics report shows that for 103/300 = 34.3% of the simulations, the Blend Speed estimates are zero.

Example of a Prospective Power Analysis

In this example, you use the Simulate feature to perform a prospective power analysis for a nonlinear model. You are interested in the main effects of six continuous factors on whether a part passes or fails inspection. The response is binomial and you can afford a total of 60 runs.

You model the probability of a success using a generalized linear model with the logit as a link function. The logit link function fits a logistic model:

$$\pi(\mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_6 X_6)}}$$

where $\pi(\mathbf{X})$ denotes the probability that a part passes at the given design settings $\mathbf{X} = (X_1, X_2, ..., X_6)$.

Denote the linear predictor by $L(\mathbf{X})$:

$$L(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_6 X_6$$

Next, you explore power for the following values of the coefficients of the linear predictor:

Coefficient	**Value**
β_0	0
β_1	1
β_2	0.9
β_3	0.8
β_4	0.7
β_5	0.6
β_6	0.5

Because the intercept in the linear predictor is 0, when all factors are set to 0, the probability of a passing part equals 50%. The probabilities associated with the levels of the i^{th} factor, when all other factors are set to 0, are given below.

Factor	**Percent Passing at X_i = 1**	**Percent Passing at X_i = -1**	**Difference**
X_1	73.11%	26.89%	46.2%
X_2	71.09%	28.91%	42.2%
X_3	69.00%	31.00%	38.0%
X_4	66.82%	33.18%	33.6%
X_5	64.56%	35.43%	29.1%
X_6	62.25%	37.75%	24.5%

For example, when all factors other than X_1 are set to 0, the difference in pass rates that you want to detect is 46.2%. The smallest difference in pass rates that you want to detect occurs when all factors other than X_6 are set to zero and that difference is 24.5%.

Construct the Design

In this section, you construct a custom design for your experiment.

Note: If you prefer to skip the steps in this section, select **Help > Sample Data Folder** and open Design Experiment/Binomial Experiment.jmp. Click the green triangle next to the **DOE Simulate** script and then go to "Define Simulated Responses".

1. Select **DOE > Custom Design**.

   Note: Although a custom design is not optimal for a non-linear situation, in this example, for simplicity, you can use the Custom Design platform rather than the Nonlinear Design platform. For an example illustrating why a design constructed using the Nonlinear Design platform is better than an orthogonal design, see the *Design of Experiments Guide*.

2. In the Factors outline, type 6 next to **Add N Factors**.
3. Click **Add Factor > Continuous**.
4. Click **Continue**.

   You are constructing a main effects design, so do not make any changes to the Model outline.

5. Under Number of Runs, type 60 next to **User Specified**.
6. Click the Custom Design red triangle and select **Simulate Responses**.

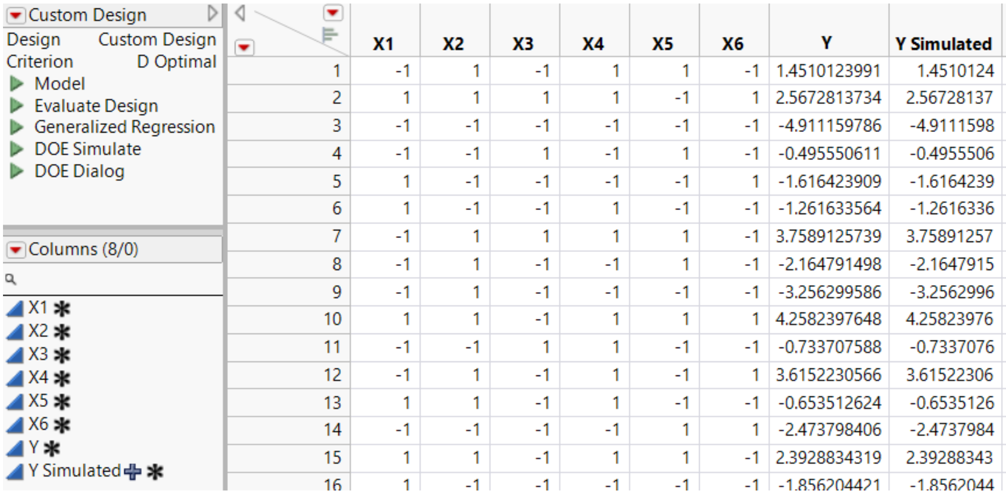
   This opens the Simulate Responses window after you select Make Table to construct the design table.

   Note: Setting the Random Seed in step 7 and Number of Starts in step 8 reproduces the same design shown in this example. In constructing a design on your own, these steps are not necessary.

7. (Optional) Click the Custom Design red triangle and select **Set Random Seed**. Type 12345 and click **OK**.
8. (Optional) Click the Custom Design red triangle and select **Number of Starts**. Type 1 and click **OK**.
9. Click **Make Design**.
10. Click **Make Table**.

Note: The entries in your Y and Y Simulated columns might differ from those that appear in Figure 10.18.

Figure 10.18 Partial View of Design Table

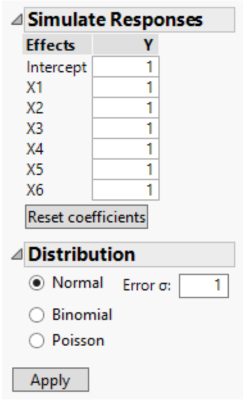


Custom Design
Design Custom Design
Criterion D Optimal
▶ Model
▶ Evaluate Design
▶ Generalized Regression
▶ DOE Simulate
▶ DOE Dialog

Columns (8/0)
X1 ✱
X2 ✱
X3 ✱
X4 ✱
X5 ✱
X6 ✱
Y ✱
Y Simulated ✚ ✱

	X1	X2	X3	X4	X5	X6	Y	Y Simulated
1	-1	1	-1	1	1	-1	1.4510123991	1.4510124
2	1	1	1	1	-1	1	2.5672813734	2.56728137
3	-1	-1	-1	-1	-1	-1	-4.911159786	-4.9111598
4	-1	-1	1	-1	1	-1	-0.495550611	-0.4955506
5	1	-1	-1	-1	-1	1	-1.616423909	-1.6164239
6	1	-1	-1	1	-1	-1	-1.261633564	-1.2616336
7	-1	1	1	1	1	-1	3.7589125739	3.75891257
8	-1	1	-1	-1	1	-1	-2.164791498	-2.1647915
9	-1	1	-1	-1	-1	-1	-3.256299586	-3.2562996
10	1	1	-1	1	1	1	4.2582397648	4.25823976
11	-1	-1	1	1	-1	-1	-0.733707588	-0.7337076
12	-1	1	-1	1	-1	1	3.6152230566	3.61522306
13	1	1	-1	1	-1	-1	-0.653512624	-0.6535126
14	-1	-1	-1	-1	1	1	-2.473798406	-2.4737984
15	1	1	-1	1	1	-1	2.3928834319	2.39288343
16	1	-1	-1	-1	-1	-1	-1.856204421	-1.8562044

Figure 10.19 Simulate Responses Window



Simulate Responses

Effects	**Y**
Intercept	1
X1	1
X2	1
X3	1
X4	1
X5	1
X6	1

Distribution

● Normal Error σ: 1
○ Binomial
○ Poisson

The design table and a Simulate Responses window appear. Two columns are added to the design table:

– Y contains a set of values simulated according to the specifications in the Simulate Responses window.

– Y Simulated contains a formula that calculates its values using the formula for the model that is specified in the Simulate Responses window. To view the formula, click the plus sign to the right of the column name in the Columns panel.

Continue to the next section, where you simulate binomial responses and fit a generalized linear model to these simulated responses.

Define Simulated Responses

Your plan is to simulate binomial response data where the probability of success is given by a logistic model. For more information about Simulate Response, see the *Design of Experiments Guide*.

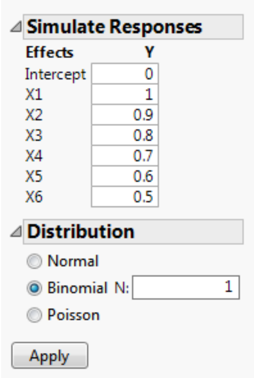
Note: If you prefer to skip the steps in this section, click the green triangle next to the **Simulate Model Responses** script. Then go to "Fit the Generalized Linear Model".

---

1. In the Simulate Responses window (Figure 10.19), enter the following values under Y:
   - Next to Intercept, type 0.
   - Next to X1, 1 is entered by default. Keep that value.
   - Next to X2, type 0.9.
   - Next to X3, type 0.8.
   - Next to X4, type 0.7.
   - Next to X5, type 0.6.
   - Next to X6, type 0.5.
2. In the Distribution outline, select **Binomial**.

   Leave the value for **N** set to 1, indicating that there is only one unit per trial.

Figure 10.20 Completed Simulate Responses Window



Effects	Y
Intercept	0
X1	1
X2	0.9
X3	0.8
X4	0.7
X5	0.6
X6	0.5

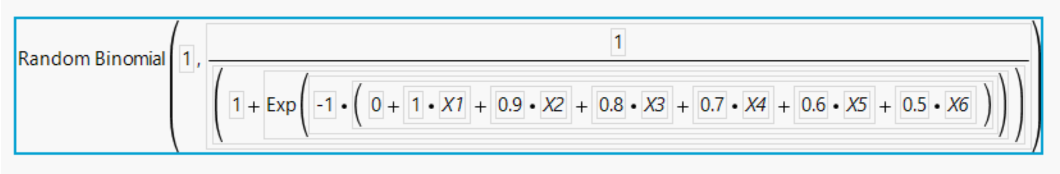
○ Normal
● Binomial N: 1
○ Poisson

3. Click **Apply**.

   In the design data table, the Y Simulated column is replaced with a formula column that generates binomial values. A column called Y N Trials indicates the number of trials for each run.

4. (Optional) Click the plus sign to the right of Y Simulated in the Columns panel.

Figure 10.21 Random Binomial Formula for Y Simulated



$$\text{Random Binomial}\left(1,\ \frac{1}{\left(1 + \text{Exp}\left(-1 \bullet \left(0 + 1 \bullet X1 + 0.9 \bullet X2 + 0.8 \bullet X3 + 0.7 \bullet X4 + 0.6 \bullet X5 + 0.5 \bullet X6\right)\right)\right)}\right)$$

5. Click **Cancel**.

Fit the Generalized Linear Model

In this section, you fit a logistic model using the Generalized Linear Model personality.

1. In the data table, click the green triangle next to the **Model** script.
2. Click the Y variable next to the **Y** button and click **Remove**.
3. Click Y Simulated and click the **Y** button.

   You are replacing Y with a column that contains randomly generated binomial values.
4. From the Personality list, select **Generalized Linear Model**.
5. From the Distribution list, select **Binomial**.

   Notice that the Logit function appears in the Link Function menu.
6. Click **Run**.

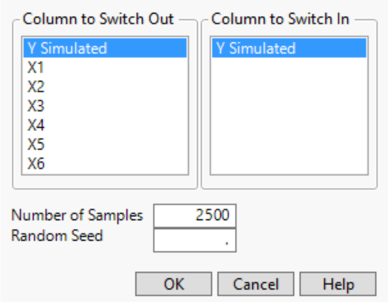
   The model that is fit is based on a single set of simulated binomial responses.

Explore Power

In this section, you simulate likelihood ratio test *p*-values to explore the power of detecting a difference over a range of probability values that is determined by the linear predictor with the coefficient values given in the example introduction.

1. In the Effect Tests outline, right-click in the **Prob>ChiSq** column and select **Simulate**.

Figure 10.22 Simulate Window



Column to Switch Out
- Y Simulated
- X1
- X2
- X3
- X4
- X5
- X6

Column to Switch In
- Y Simulated

Number of Samples: 2500
Random Seed: .

Make sure the Y Simulated column is selected in the Column to Switch Out list. This column contains the values that were used to fit the model. When you select the column Y Simulated under Column to Switch In, for each simulation, you are telling JMP to replace the values in Y Simulated with a new column of values that are simulated using the formula in the column Y Simulated.

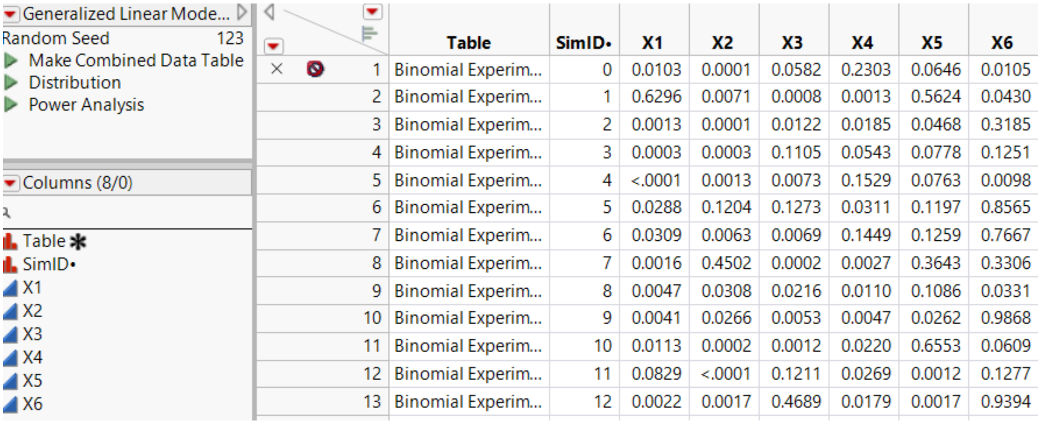
The column that you have selected in the report, **Prob>ChiSq**, is the *p*-value for a likelihood ratio test of whether the associated main effect is 0. The Prob>ChiSq value is simulated for each effect listed in the Effect Tests table.

2. Next to **Number of Samples**, type **500**.
3. Click **OK**.

   A Generalized Linear Model Simulate Results data table appears.

Note: Because response values are simulated, your simulated *p*-values might differ from those shown in Figure 10.23.

Figure 10.23 Table of Simulated Results, Partial View



Generalized Linear Mode...
Random Seed 123
- Make Combined Data Table
- Distribution
- Power Analysis

Columns (8/0)
- Table
- SimID•
- X1
- X2
- X3
- X4
- X5
- X6

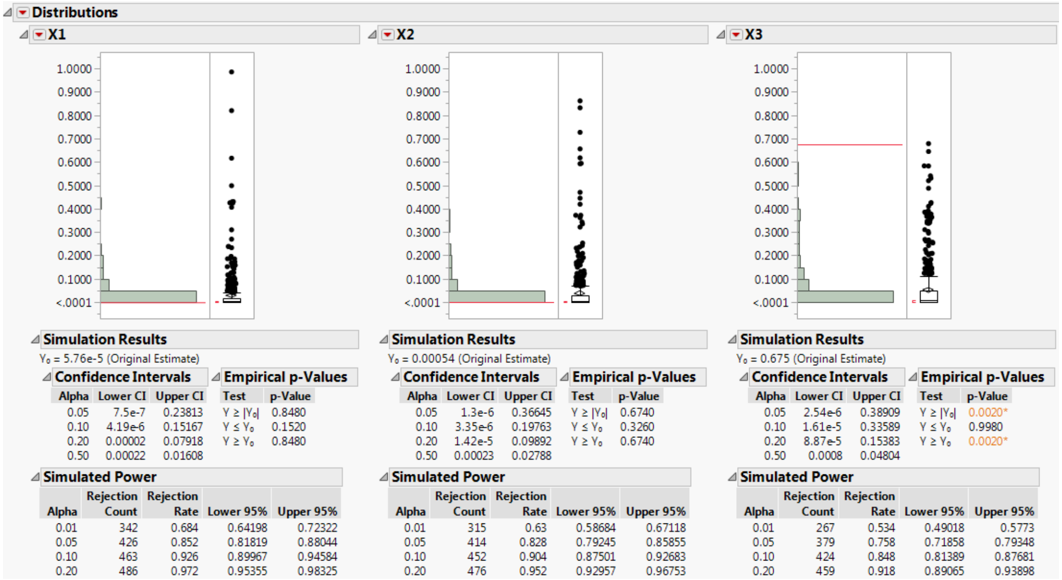
	Table	SimID•	X1	X2	X3	X4	X5	X6
1	Binomial Experim...	0	0.0103	0.0001	0.0582	0.2303	0.0646	0.0105
2	Binomial Experim...	1	0.6296	0.0071	0.0008	0.0013	0.5624	0.0430
3	Binomial Experim...	2	0.0013	0.0001	0.0122	0.0185	0.0468	0.3185
4	Binomial Experim...	3	0.0003	0.0003	0.1105	0.0543	0.0778	0.1251
5	Binomial Experim...	4	<.0001	0.0013	0.0073	0.1529	0.0763	0.0098
6	Binomial Experim...	5	0.0288	0.1204	0.1273	0.0311	0.1197	0.8565
7	Binomial Experim...	6	0.0309	0.0063	0.0069	0.1449	0.1259	0.7667
8	Binomial Experim...	7	0.0016	0.4502	0.0002	0.0027	0.3643	0.3306
9	Binomial Experim...	8	0.0047	0.0308	0.0216	0.0110	0.1086	0.0331
10	Binomial Experim...	9	0.0041	0.0266	0.0053	0.0047	0.0262	0.9868
11	Binomial Experim...	10	0.0113	0.0002	0.0012	0.0220	0.6553	0.0609
12	Binomial Experim...	11	0.0829	<.0001	0.1211	0.0269	0.0012	0.1277
13	Binomial Experim...	12	0.0022	0.0017	0.4689	0.0179	0.0017	0.9394

The first row of the table contains the initial values of **Prob>ChiSq** and is excluded. The remaining 500 rows contain simulated values.

4. Run the **Power Analysis** script.

Note: Because response values are simulated, your simulated power results might differ from those shown in Figure 10.24.

Figure 10.24 Distribution Plots for the First Two Effects



X1

Simulation Results

Y_0 = 5.76e-5 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	7.5e-7	0.23813
0.10	4.19e-6	0.15167
0.20	0.00002	0.07918
0.50	0.00022	0.01608

Empirical p-Values

Test	p-Value
$Y \geq \|Y_0\|$	0.8480
$Y \leq Y_0$	0.1520
$Y \geq Y_0$	0.8480

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	342	0.684	0.64198	0.72322
0.05	426	0.852	0.81819	0.88044
0.10	463	0.926	0.89967	0.94584
0.20	486	0.972	0.95355	0.98325

X2

Simulation Results

Y_0 = 0.00054 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	1.3e-6	0.36645
0.10	3.35e-6	0.19763
0.20	1.42e-5	0.09892
0.50	0.00023	0.02788

Empirical p-Values

Test	p-Value
$Y \geq \|Y_0\|$	0.6740
$Y \leq Y_0$	0.3260
$Y \geq Y_0$	0.6740

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	315	0.63	0.58684	0.67118
0.05	414	0.828	0.79245	0.85855
0.10	452	0.904	0.87501	0.92683
0.20	476	0.952	0.92957	0.96753

X3

Simulation Results

Y_0 = 0.675 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	2.54e-6	0.38909
0.10	1.61e-5	0.33589
0.20	8.87e-5	0.15383
0.50	0.0008	0.04804

Empirical p-Values

Test	p-Value
$Y \geq \|Y_0\|$	0.0020*
$Y \leq Y_0$	0.9980
$Y \geq Y_0$	0.0020*

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	267	0.534	0.49018	0.5773
0.05	379	0.758	0.71858	0.79348
0.10	424	0.848	0.81389	0.87681
0.20	459	0.918	0.89065	0.93898

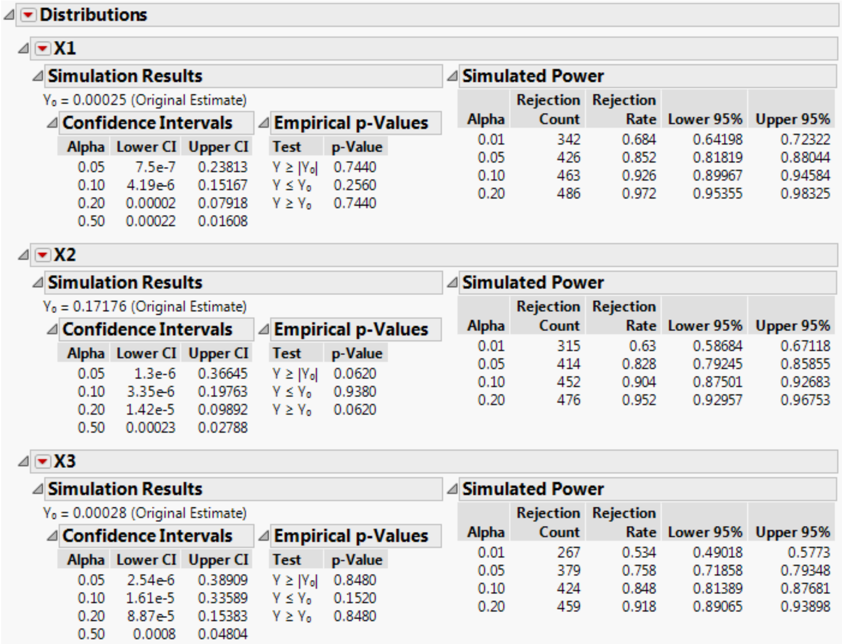
The histograms plot the 500 simulated Prob>ChiSq values for each main effect. The Simulated Power outline shows the simulated Rejection Rate in the 500 simulations.

For easier viewing, stack the reports and de-select the plots.

5. Click the Distributions red triangle and select **Stack**.
6. Press Ctrl and click the X1 red triangle, and de-select **Outlier Box Plot**.
7. Press Ctrl and click the X1 red triangle, select **Histogram Options**, and de-select **Histogram**.

Note: Because response values are simulated, your simulated power results might differ from those shown in Figure 10.25.

Figure 10.25 Power Results for the First Three Effects



Distributions

X1

Simulation Results

Y_0 = 0.00025 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	7.5e-7	0.23813
0.10	4.19e-6	0.15167
0.20	0.00002	0.07918
0.50	0.00022	0.01608

Empirical p-Values

Test	p-Value
$Y \ge \lvert Y_0 \rvert$	0.7440
$Y \le Y_0$	0.2560
$Y \ge Y_0$	0.7440

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	342	0.684	0.64198	0.72322
0.05	426	0.852	0.81819	0.88044
0.10	463	0.926	0.89967	0.94584
0.20	486	0.972	0.95355	0.98325

X2

Simulation Results

Y_0 = 0.17176 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	1.3e-6	0.36645
0.10	3.35e-6	0.19763
0.20	1.42e-5	0.09892
0.50	0.00023	0.02788

Empirical p-Values

Test	p-Value
$Y \ge \lvert Y_0 \rvert$	0.0620
$Y \le Y_0$	0.9380
$Y \ge Y_0$	0.0620

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	315	0.63	0.58684	0.67118
0.05	414	0.828	0.79245	0.85855
0.10	452	0.904	0.87501	0.92683
0.20	476	0.952	0.92957	0.96753

X3

Simulation Results

Y_0 = 0.00028 (Original Estimate)

Confidence Intervals

Alpha	Lower CI	Upper CI
0.05	2.54e-6	0.38909
0.10	1.61e-5	0.33589
0.20	8.87e-5	0.15383
0.50	0.0008	0.04804

Empirical p-Values

Test	p-Value
$Y \ge \lvert Y_0 \rvert$	0.8480
$Y \le Y_0$	0.1520
$Y \ge Y_0$	0.8480

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	267	0.534	0.49018	0.5773
0.05	379	0.758	0.71858	0.79348
0.10	424	0.848	0.81389	0.87681
0.20	459	0.918	0.89065	0.93898

In the Simulated Power outlines, the Rejection Rate for each row gives the proportion of *p*-values that are smaller than the corresponding Alpha. For example, for X3, which corresponds to a coefficient value of 0.8 and a probability difference of 38%, the simulated power for a 0.05 significance level is 379/500 = 0.758. Table 10.1 summarizes the estimated power at the 0.05 significance level for all effects. Notice how power decreases as the Difference to Detect decreases. Also notice that the power to detect an effect as large as 24.5% (X6) is only approximately 0.37.

Note: Because response values are simulated, your simulated power results might differ from those shown in Table 10.1.

Table 10.1 Simulated Power at Significance Level 0.05

Factor	**Percent Passing at X_i = 1**	**Percent Passing at X_i = -1**	**Difference to Detect**	**Simulated Power (Rejection Rate) at Alpha=0.05**
X_1	73.11%	26.89%	46.2%	0.852
X_2	71.09%	28.91%	42.2%	0.828
X_3	69.00%	31.00%	38.0%	0.758

Table 10.1 Simulated Power at Significance Level 0.05 *(Continued)*

Factor	**Percent Passing at X_i = 1**	**Percent Passing at X_i = -1**	**Difference to Detect**	**Simulated Power (Rejection Rate) at Alpha=0.05**
X_4	66.82%	33.18%	33.6%	0.654
X_5	64.56%	35.43%	29.1%	0.488
X_6	62.25%	37.75%	24.5%	0.372

Chapter 11

Bootstrapping

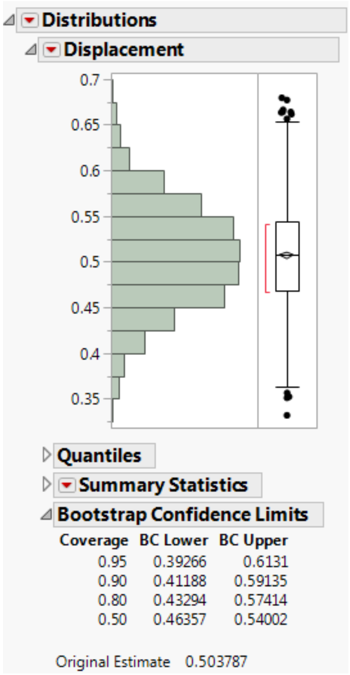
Approximate the Distribution of a Statistic through Resampling

Bootstrapping is a resampling method for approximating the sampling distribution of a statistic. You can use bootstrapping to estimate the distribution of a statistic and its properties, such as its mean, bias, standard error, and confidence intervals. Bootstrapping is especially useful in the following situations:

- The theoretical distribution of the statistic is complicated or unknown.
- Inference using parametric methods is not possible because of violations of assumptions.

Note: Bootstrap is available only from a right-click in a report. It is not a platform command.

Figure 11.1 Bootstrapping Results for a Slope Parameter



Coverage	BC Lower	BC Upper
0.95	0.39266	0.6131
0.90	0.41188	0.59135
0.80	0.43294	0.57414
0.50	0.46357	0.54002

Original Estimate 0.503787

Contents

Overview of the Bootstrapping Feature

Bootstrapping repeatedly resamples the observations that are used in your report to construct an estimate of the distribution of a statistic or statistics. The observations are assumed to be independent.

In the simple bootstrap, the n observations are resampled with replacement to produce a bootstrap sample of size n. Note that some observations might not appear in the bootstrap sample, and others might appear multiple times. The number of times that an observation occurs in the bootstrap sample is called its *bootstrap weight*. For each bootstrap iteration, the entire analysis that produced the statistic of interest is rerun with these changes:

- the bootstrap sample of n observations is the data set
- the bootstrap weight is a frequency variable in the analysis platform

This process is repeated to produce a distribution of values for the statistic or statistics of interest.

However, the simple bootstrap can sometimes be inadequate. For example, suppose your data set is small or you have a logistic regression setting where you can encounter separation issues. In such cases, JMP enables you to conduct Bayesian bootstrapping using fractional weights. When fractional weights are used, a fractional weight is associated with each observation. The fractional weights sum to n. The statistic of interest is computed by treating the fractional weights as a frequency variable in the analysis platform. For information about fractional weights, see “Fractional Weights” and “Statistical Details for Fractional Weights”.

To run a bootstrap analysis in a report, right-click in a table column that contains the statistic that you want to bootstrap and select Bootstrap.

Note: Bootstrap is available only from a right-click in a report. It is not a platform command.

JMP provides bootstrapping in many statistical platforms. See “JMP Platforms that Support Bootstrapping” for a complete list. The observations that comprise the sample are all observations that are used in the calculations for the statistics of interest. If the report uses a frequency column, the observations in that column are treated as if they were repeated the number of times indicated by the Freq variable. If the report uses a Weight variable, Bootstrap treats it as it was treated in the calculations for the report.

Tip: Bootstrap reruns the entire analysis that appears in the platform report from which Bootstrap is invoked. As a result, Bootstrap might run slowly for your selected column because of extraneous analyses in the report. If Bootstrap is running slowly, remove extraneous options from the platform report before running Bootstrap.

JMP Platforms that Support Bootstrapping

Bootstrapping is available in the following statistical platforms:

- Boosted Tree
- Bootstrap Forest
- Categorical
- Destructive Degradation
- Discriminant
- Distribution
- Fit Curve
- Fit Life by X
- Fit Parametric Survival
- Fit Proportional Hazards
- Fit Y by X
- Generalized Linear Model
- Generalized Regression
- Life Distribution
- Logistic
- Loglinear Variance
- Multiple Correspondence Analysis
- Multivariate
- Neural
- Nonlinear
- Parametric Survival
- Partial Least Squares
- Partition
- Principal Components
- Proportional Hazard
- Standard Least Squares
- Survival
- Uplift

Example of Bootstrapping

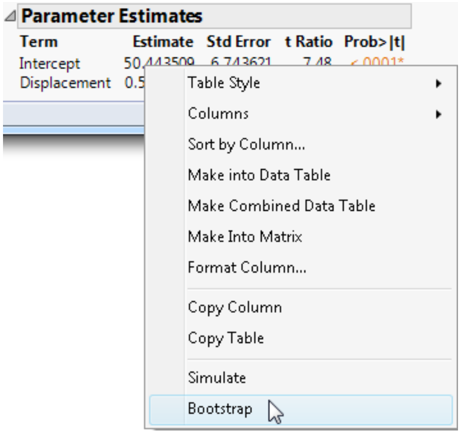
In this example, the regression assumption of homogeneity of variance is violated, so the confidence limits from a regression analysis for the slope might be misleading. For this reason, a bootstrap estimate of the confidence interval for the slope is used instead.

1. Select **Help > Sample Data Folder** and open Car Physical Data.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Horsepower and click **Y, Response**.
4. Select Displacement and click **X, Factor**.
5. Click **OK**.
6. Click the red triangle next to Bivariate Fit of Horsepower By Displacement and select **Fit Line**.

   The slope estimate is 0.503787, approximately 0.504.
7. (Optional) Right-click in the **Parameter Estimates** report and select **Columns > Lower 95%**.
8. (Optional) Right-click in the **Parameter Estimates** report and select **Columns > Upper 95%**.

   The confidence limits from the regression analysis for the slope are 0.4249038 and 0.5826711.
9. Right-click the **Estimate** column in the **Parameter Estimates** report and select **Bootstrap**.

Figure 11.2 The Bootstrap Option



The column that you right-click is relevant when the **Split Selected Column** option is selected. See “Bootstrapping Window Options”.

10. Type 1000 for the **Number of Bootstrap Samples**.
11. (Optional) To match the results in Figure 11.3, type 12345 for the **Random Seed**.
12. Click **OK**.

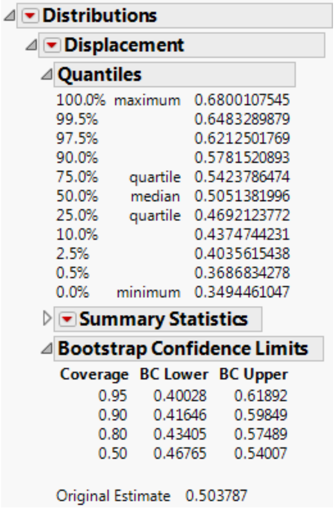
    The bootstrap process runs and produces a Bootstrap Results data table with unstacked results for the slope and intercept.

    Next, analyze the bootstrapped slope.

13. In the Bootstrap Results table, run the Distribution script.

    The Distribution report includes the Bootstrap Confidence Limits report.

Figure 11.3 Bootstrap Report



100.0%	maximum	0.6800107545
99.5%		0.6483289879
97.5%		0.6212501769
90.0%		0.5781520893
75.0%	quartile	0.5423786474
50.0%	median	0.5051381996
25.0%	quartile	0.4692123772
10.0%		0.4374744231
2.5%		0.4035615438
0.5%		0.3686834278
0.0%	minimum	0.3494461047

Coverage	BC Lower	BC Upper
0.95	0.40028	0.61892
0.90	0.41646	0.59849
0.80	0.43405	0.57489
0.50	0.46765	0.54007

Original Estimate 0.503787

The estimate of the slope (step 6) is 0.504. Based on the bootstrap results for 95% coverage, the company can estimate the slope to be between 0.40028 and 0.61892. When the displacement is changed by one unit, with 95% confidence, the horsepower changes by some amount between 0.40028 and 0.61892. The bootstrap confidence interval for the slope (0.400 to 0.619) is slightly wider than the confidence interval (0.425 to 0.583) obtained using the usual regression assumptions in step 7 and step 8.

Note: The BC Lower and BC Upper columns in the Bootstrap Confidence Limits report refer to *bias-corrected intervals*. See "Statistical Details for the Bias-Corrected Percentile Intervals".

Bootstrapping Window Options

To perform a bootstrap analysis, in a table in a JMP report, right-click a numeric column of sample statistics and select **Bootstrap**. The selected column is highlighted and the Bootstrapping window appears. After you select options and click **OK** in the Bootstrapping window, bootstrap results for every statistic in the column appear in the default results table.

Note: The Bootstrap option is not available in reports that use a By variable.

The Bootstrapping window contains the following options:

Number of Bootstrap Samples Sets the number of times that you want to resample the data and compute the statistics. A larger number results in more precise estimates of the statistics' properties. By default, the number of bootstrap samples is set to 2500.

Random Seed Sets a random seed that you can re-enter in subsequent runs of the bootstrap analysis to duplicate your current results. By default, no seed is set.

Fractional Weights Performs a Bayesian bootstrap analysis. In each bootstrap iteration, each observation is assigned a weight that is calculated as described in "Statistical Details for Fractional Weights". The weighted observations are used in computing the statistics of interest. By default, the Fractional Weights option is not selected and a simple bootstrap analysis is conducted.

Tip: Use the Fractional Weights option if the number of observations that are used in your analysis is small or if you are concerned about separation in a logistic regression setting.

Suppose that the Fractional Weights option is selected. For each bootstrap iteration, each observation that is used in the report is assigned a nonzero weight. These weights sum to *n,* the number of observations used in the calculations of the statistics of interest. For more information about how the weights are calculated and used, see "Statistical Details for Fractional Weights".

Split Selected Column Places bootstrap results for each statistic in the column that you selected for bootstrapping into a *separate* column in the Bootstrap Results table. Each row of the Bootstrap Results table (other than the first) corresponds to a single bootstrap sample.

If you deselect this option, a Stacked Bootstrap Results table appears. For each bootstrap iteration, this table contains results for the entire report table that contains the column that you selected for bootstrapping. Results for each row of the report table appear as rows in the Stacked Bootstrap Results table. Each column in the report table defines a column in the Stacked Bootstrap Results table. For an example, see "Stacked Bootstrap Results Table".

Discard Stacked Table if Split Works (Applicable only if the **Split Selected Column** option is selected.) Determines the number of results tables produced by Bootstrap.

If the Discard Stacked Table if Split Works option *is not* selected, then two Bootstrap tables are shown:

- The stacked Bootstrap Results table, which contains bootstrap results for each row of the table containing the column that you selected for bootstrapping. This table gives bootstrap results for every statistic in the report, where each column is defined by a statistic.
- The unstacked Bootstrap Results table, which is obtained by splitting the stacked table. This table provides results only for the column that is selected in the original report.

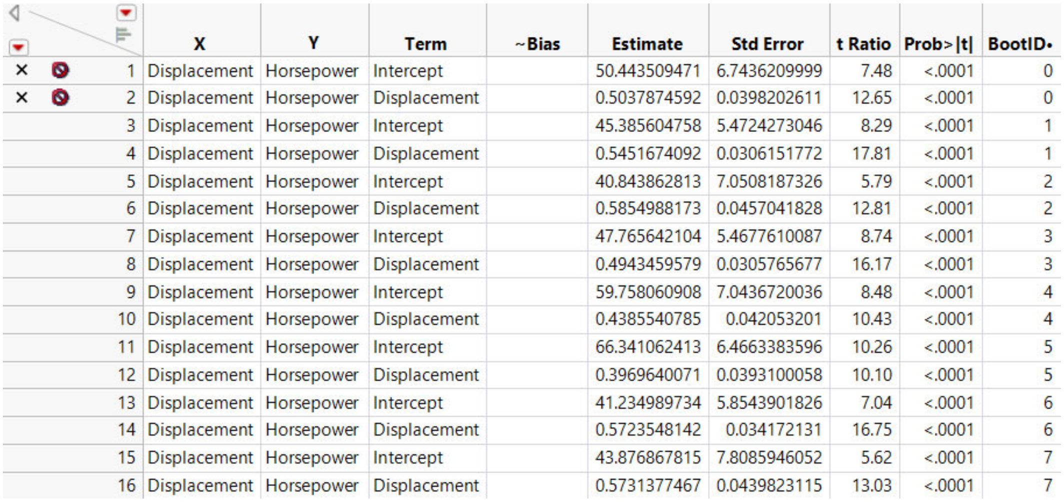
If the Discard Stacked Table if Split Works option *is* selected and if the **Split Selected Column** operation is successful, the Stacked Bootstrap Results table is not shown.

Stacked Bootstrap Results Table

A stacked bootstrap results table contains a row for each statistic and bootstrap iteration combination, with a single column for the bootstrap results. By default, the initial results of a bootstrap analysis appear in a stacked results table (Figure 11.4). This table might not appear if you have selected the Discard Stacked Table if Split Works option.

Figure 11.4 shows a bootstrap table that is based on the Parameter Estimates report obtained by fitting a Bivariate model in Fit Y by X to Car Physical Data.jmp. See “Example of Bootstrapping”.

Figure 11.4 Stacked Bootstrap Results Table



	X	Y	Term	~Bias	Estimate	Std Error	t Ratio	Prob>|t|	BootID•
× ⊘ 1	Displacement	Horsepower	Intercept		50.443509471	6.7436209999	7.48	<.0001	0
× ⊘ 2	Displacement	Horsepower	Displacement		0.5037874592	0.0398202611	12.65	<.0001	0
3	Displacement	Horsepower	Intercept		45.385604758	5.4724273046	8.29	<.0001	1
4	Displacement	Horsepower	Displacement		0.5451674092	0.0306151772	17.81	<.0001	1
5	Displacement	Horsepower	Intercept		40.843862813	7.0508187326	5.79	<.0001	2
6	Displacement	Horsepower	Displacement		0.5854988173	0.0457041828	12.81	<.0001	2
7	Displacement	Horsepower	Intercept		47.765642104	5.4677610087	8.74	<.0001	3
8	Displacement	Horsepower	Displacement		0.4943459579	0.0305765677	16.17	<.0001	3
9	Displacement	Horsepower	Intercept		59.758060908	7.0436720036	8.48	<.0001	4
10	Displacement	Horsepower	Displacement		0.4385540785	0.042053201	10.43	<.0001	4
11	Displacement	Horsepower	Intercept		66.341062413	6.4663383596	10.26	<.0001	5
12	Displacement	Horsepower	Displacement		0.3969640071	0.0393100058	10.10	<.0001	5
13	Displacement	Horsepower	Intercept		41.234989734	5.8543901826	7.04	<.0001	6
14	Displacement	Horsepower	Displacement		0.5723548142	0.034172131	16.75	<.0001	6
15	Displacement	Horsepower	Intercept		43.876867815	7.8085946052	5.62	<.0001	7
16	Displacement	Horsepower	Displacement		0.5731377467	0.0439823115	13.03	<.0001	7

Note the following about the stacked results table:

- For each bootstrap sample, there is a row for each value given in the first column of the report table. These values are shown in a column whose name is the name of the first column in the report table. In this example, for each bootstrap sample there is a row containing results for each Term: Intercept and Displacement, which appear in the Term column.
- The data table columns that are used in the analysis appear in the table. In this example, X is Displacement, and Y is Horsepower.
- There is a column for every column in the report table that you are bootstrapping. In this example, the columns are ~Bias, Estimate, Std Error, t Ratio, and Prob>|t|. Note that ~Bias is a column in the Fit Y by X report that is hidden unless one of the parameter estimates is biased.
- The BootID• column identifies the bootstrap sample. The rows where BootID• = 0 correspond to the original estimates. Those rows are marked with an X and have the excluded row state. In this example, each bootstrap sample is used to calculate results for two rows: the results for Intercept and the results for Displacement.
- The data table name begins with “Stacked Bootstrap Results”.

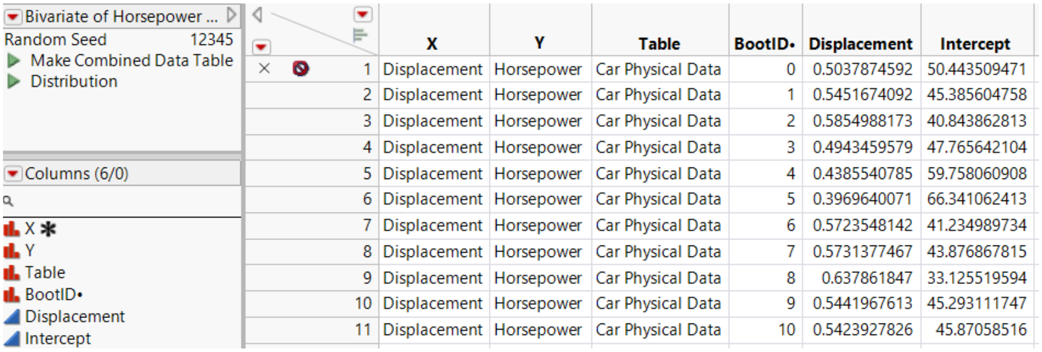
If you selected the **Split Selected Column** option, an unstacked results table might also appear. See “Unstacked Bootstrap Results Table”.

Unstacked Bootstrap Results Table

An unstacked bootstrap results table contains a single row for each bootstrap iteration, with separate columns for each bootstrap result. When bootstrapping, select **Split Selected Column** to create an unstacked bootstrap results table.

In this example, the estimates for the terms in Figure 11.4 (stacked) are split into two columns, Displacement and Intercept, in Figure 11.5 (unstacked).

Figure 11.5 Unstacked Bootstrap Results Table



	X	Y	Table	BootID•	Displacement	Intercept
1	Displacement	Horsepower	Car Physical Data	0	0.5037874592	50.443509471
2	Displacement	Horsepower	Car Physical Data	1	0.5451674092	45.385604758
3	Displacement	Horsepower	Car Physical Data	2	0.5854988173	40.843862813
4	Displacement	Horsepower	Car Physical Data	3	0.4943459579	47.765642104
5	Displacement	Horsepower	Car Physical Data	4	0.4385540785	59.758060908
6	Displacement	Horsepower	Car Physical Data	5	0.3969640071	66.341062413
7	Displacement	Horsepower	Car Physical Data	6	0.5723548142	41.234989734
8	Displacement	Horsepower	Car Physical Data	7	0.5731377467	43.876867815
9	Displacement	Horsepower	Car Physical Data	8	0.637861847	33.125519594
10	Displacement	Horsepower	Car Physical Data	9	0.5441967613	45.293111747
11	Displacement	Horsepower	Car Physical Data	10	0.5423927826	45.87058516

Note the following about the unstacked results table:

- There is a single row for each bootstrap sample.
- The data table columns used in the analysis appear in the table. In this example, X is Displacement, and Y is Horsepower.
- There is a column for each row of the report that was bootstrapped.
- If you specified a Random Seed in the Bootstrapping window, the bootstrap results table contains a table variable called Random Seed that gives its value.
- The unstacked bootstrap results table contains a Source table script and a Distribution table script. The Distribution table script enables you to quickly obtain statistics based on the bootstrap samples, including bootstrap confidence intervals.
- The BootID• column identifies the bootstrap sample. The row where BootID• = 0 corresponds to the original estimates. That row is marked with an X and has the excluded row state. In the unstacked bootstrap table, each row is calculated from a single bootstrap sample.
- The data table name ends with “Bootstrap Results (<colname>)”, where <colname> identifies the column in the report that was bootstrapped.

Analysis of Bootstrap Results

Analyze your bootstrap results using the Distribution platform:

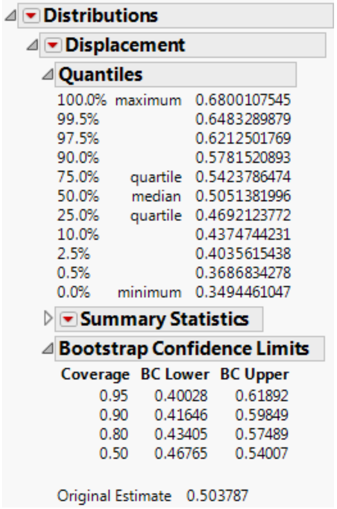
- If your analysis produced an unstacked bootstrap results table, run the Distribution script in the table.
- If your analysis produced a stacked bootstrap results table, select **Analyze > Distribution** and assign the columns of interest to the appropriate roles. In most cases, it is appropriate to assign the column that corresponds to the first column in the report table to the By role.

The Distribution platform provides summary statistics for your bootstrap results. It also produces a Bootstrap Confidence Limits report for any table that contains a BootID• column (Figure 11.6).

You can use the Distribution report to obtain two types of bootstrap confidence intervals:

- The Quantiles report provides *percentile intervals*. For example, to construct a 95% confidence interval using the percentile method, use the 2.5% and 97.5% quantiles as the interval bounds.
- The Bootstrap Confidence Limits report provides *bias-corrected percentile intervals*. The report shows intervals with 95%, 90%, 80%, and 50% coverage levels. The BC Lower and BC Upper columns show the lower and upper endpoints, respectively. For more information about the computation of the bias-corrected percentile intervals, see "Statistical Details for the Bias-Corrected Percentile Intervals".

Figure 11.6 Bootstrap Confidence Limits Report



```
Distributions
  Displacement
    Quantiles
      100.0%  maximum   0.6800107545
      99.5%             0.6483289879
      97.5%             0.6212501769
      90.0%             0.5781520893
      75.0%   quartile  0.5423786474
      50.0%   median    0.5051381996
      25.0%   quartile  0.4692123772
      10.0%             0.4374744231
      2.5%              0.4035615438
      0.5%              0.3686834278
      0.0%    minimum   0.3494461047
    Summary Statistics
    Bootstrap Confidence Limits
      Coverage  BC Lower  BC Upper
          0.95   0.40028   0.61892
          0.90   0.41646   0.59849
          0.80   0.43405   0.57489
          0.50   0.46765   0.54007

      Original Estimate  0.503787
```

The **Original Estimate** at the bottom of the Bootstrap Confidence Limits report is the estimate of the statistic using the original data.

For more information about interpreting the Bootstrap Confidence Limits report, see "Overview of the Bootstrapping Feature". Efron (1981) describes the methods for both the percentile interval and the bias-corrected percentile interval.

Additional Example of Bootstrapping

This example illustrates the benefits of the Fractional Weights (Bayesian Bootstrap) option for a small data table. The data consist of a response, Y, measured on three samples of each of seven different soil types. A scientist is interested in finding a confidence interval for the mean response for the wabash soil type.

Because each soil type has only three observations, the simple bootstrap has the potential to exclude all three of the observations for wabash from a bootstrap sample. The Fractional Weights option ensures that all observations for every soil type are represented in all bootstrap samples.

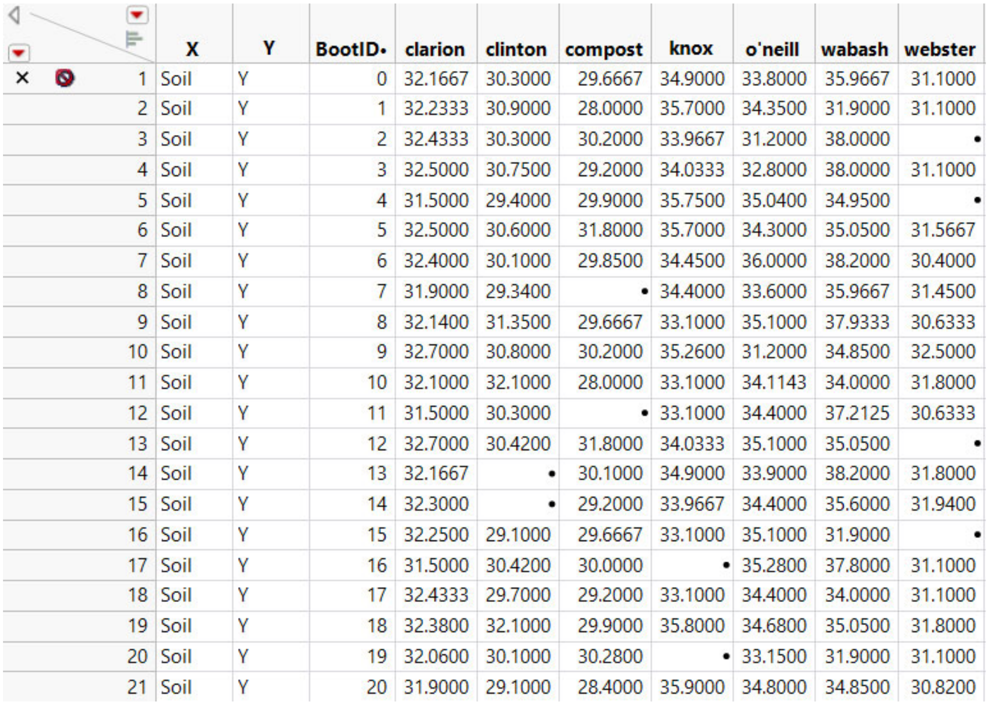
The scientist examines the distribution of wabash sample means from both bootstrap methods:

- "Simple Bootstrap Analysis"
- "Bayesian Bootstrap Analysis"

Simple Bootstrap Analysis

1. Select **Help > Sample Data Folder** and open Snapdragon.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Y and click **Y, Response**.
4. Select Soil and click **X, Factor**.
5. Click **OK**.
6. Click the red triangle next to Oneway Analysis of Y By Soil and select **Means/Anova**.
7. In the **Means for Oneway Anova** report, right-click the **Mean** column and select **Bootstrap**.
8. Type 1000 for the **Number of Bootstrap Samples**.
9. (Optional) To match the results in Figure 11.7, type 12345 for the **Random Seed**.
10. Click **OK**.

Figure 11.7 Bootstrap Results for a Simple Bootstrap



	X	Y	BootID•	clarion	clinton	compost	knox	o'neill	wabash	webster
1	Soil	Y	0	32.1667	30.3000	29.6667	34.9000	33.8000	35.9667	31.1000
2	Soil	Y	1	32.2333	30.9000	28.0000	35.7000	34.3500	31.9000	31.1000
3	Soil	Y	2	32.4333	30.3000	30.2000	33.9667	31.2000	38.0000	•
4	Soil	Y	3	32.5000	30.7500	29.2000	34.0333	32.8000	38.0000	31.1000
5	Soil	Y	4	31.5000	29.4000	29.9000	35.7500	35.0400	34.9500	•
6	Soil	Y	5	32.5000	30.6000	31.8000	35.7000	34.3000	35.0500	31.5667
7	Soil	Y	6	32.4000	30.1000	29.8500	34.4500	36.0000	38.2000	30.4000
8	Soil	Y	7	31.9000	29.3400	•	34.4000	33.6000	35.9667	31.4500
9	Soil	Y	8	32.1400	31.3500	29.6667	33.1000	35.1000	37.9333	30.6333
10	Soil	Y	9	32.7000	30.8000	30.2000	35.2600	31.2000	34.8500	32.5000
11	Soil	Y	10	32.1000	32.1000	28.0000	33.1000	34.1143	34.0000	31.8000
12	Soil	Y	11	31.5000	30.3000	•	33.1000	34.4000	37.2125	30.6333
13	Soil	Y	12	32.7000	30.4200	31.8000	34.0333	35.1000	35.0500	•
14	Soil	Y	13	32.1667	•	30.1000	34.9000	33.9000	38.2000	31.8000
15	Soil	Y	14	32.3000	•	29.2000	33.9667	34.4000	35.6000	31.9400
16	Soil	Y	15	32.2500	29.1000	29.6667	33.1000	35.1000	31.9000	•
17	Soil	Y	16	31.5000	30.4200	30.0000	•	35.2800	37.8000	31.1000
18	Soil	Y	17	32.4333	29.7000	29.2000	33.1000	34.4000	34.0000	31.1000
19	Soil	Y	18	32.3800	32.1000	29.9000	35.8000	34.6800	35.0500	31.8000
20	Soil	Y	19	32.0600	30.1000	30.2800	•	33.1500	31.9000	31.1000
21	Soil	Y	20	31.9000	29.1000	28.4000	35.9000	34.8000	34.8500	30.8200

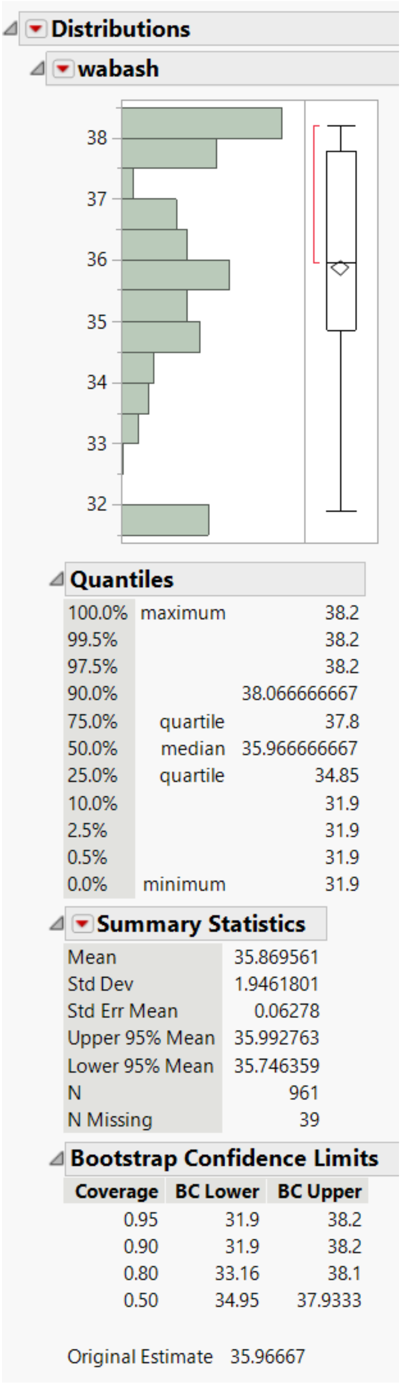
The missing values in Figure 11.7 represent bootstrap iterations in which none of the observations for a given soil type were selected for the bootstrap sample.

11. Select **Analyze > Distribution**.

12. Select wabash and click **Y, Columns**.

13. Click **OK**.

Figure 11.8 Distribution of wabash Means from a Simple Bootstrap



Distributions

wabash

Quantiles

100.0%	maximum	38.2
99.5%		38.2
97.5%		38.2
90.0%		38.066666667
75.0%	quartile	37.8
50.0%	median	35.966666667
25.0%	quartile	34.85
10.0%		31.9
2.5%		31.9
0.5%		31.9
0.0%	minimum	31.9

Summary Statistics

Mean	35.869561
Std Dev	1.9461801
Std Err Mean	0.06278
Upper 95% Mean	35.992763
Lower 95% Mean	35.746359
N	961
N Missing	39

Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	31.9	38.2
0.90	31.9	38.2
0.80	33.16	38.1
0.50	34.95	37.9333

Original Estimate 35.96667

Figure 11.8 shows the distribution of wabash means from the simple bootstrap analysis. Notice the following:

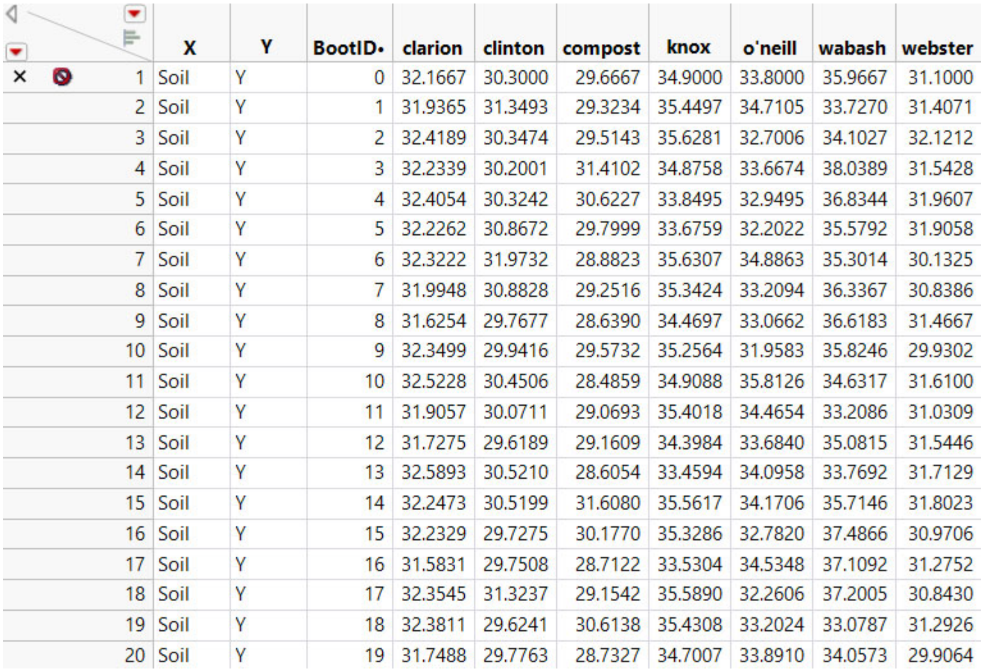
- The Summary Statistics report indicates that the number of rows containing bootstrap means for wabash is N = 961. Although you conducted 1,000 iterations, 39 bootstrap samples did not contain any of the three observations for wabash.
- The histogram of sample means is not smooth, with peaks at the two extremes. The three values for wabash are 38.2, 37.8, and 31.9. The peak at the low end of the distribution results from bootstrap samples that contain only the value 31.9. The peak at the high end results from bootstrap samples that contain one or both of the values 38.2 and 37.8.

Next, use the Fractional Weights (Bayesian Bootstrap) option to avoid obtaining missing values from the bootstrap samples and to smooth the distribution of bootstrapped means.

Bayesian Bootstrap Analysis

1. In the Oneway Analysis report, right-click the **Mean** column in the **Means for Oneway Anova** report and select **Bootstrap**.
2. Type 1000 for the **Number of Bootstrap Samples**.
3. (Optional) To match the results in Figure 11.9, type 12345 for the **Random Seed**.
4. Select the **Fractional Weights** option.
5. Click **OK**.

Figure 11.9 Bootstrap Results for a Bayesian Bootstrap

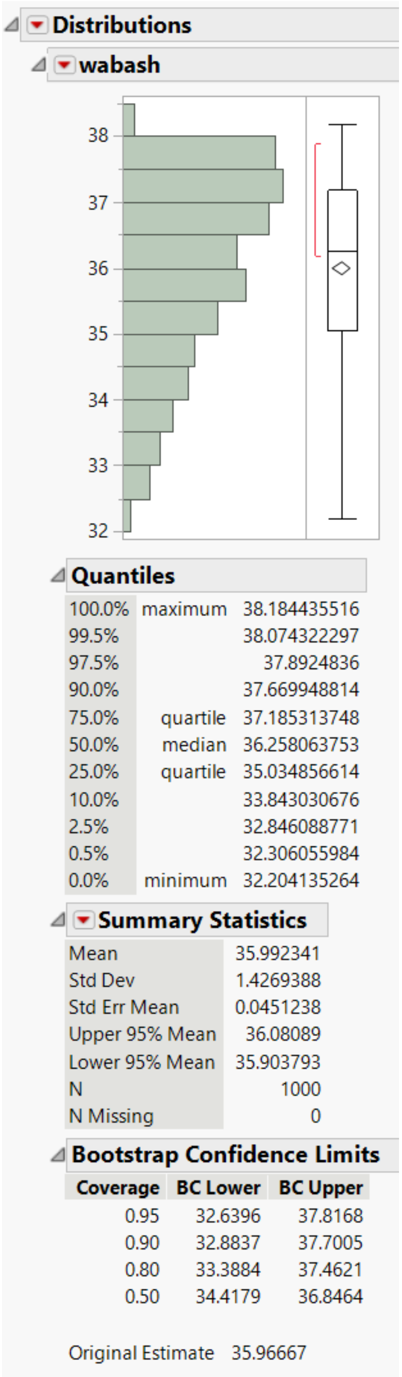


	X	Y	BootID•	clarion	clinton	compost	knox	o'neill	wabash	webster
× ⊘ 1	Soil	Y	0	32.1667	30.3000	29.6667	34.9000	33.8000	35.9667	31.1000
2	Soil	Y	1	31.9365	31.3493	29.3234	35.4497	34.7105	33.7270	31.4071
3	Soil	Y	2	32.4189	30.3474	29.5143	35.6281	32.7006	34.1027	32.1212
4	Soil	Y	3	32.2339	30.2001	31.4102	34.8758	33.6674	38.0389	31.5428
5	Soil	Y	4	32.4054	30.3242	30.6227	33.8495	32.9495	36.8344	31.9607
6	Soil	Y	5	32.2262	30.8672	29.7999	33.6759	32.2022	35.5792	31.9058
7	Soil	Y	6	32.3222	31.9732	28.8823	35.6307	34.8863	35.3014	30.1325
8	Soil	Y	7	31.9948	30.8828	29.2516	35.3424	33.2094	36.3367	30.8386
9	Soil	Y	8	31.6254	29.7677	28.6390	34.4697	33.0662	36.6183	31.4667
10	Soil	Y	9	32.3499	29.9416	29.5732	35.2564	31.9583	35.8246	29.9302
11	Soil	Y	10	32.5228	30.4506	28.4859	34.9088	35.8126	34.6317	31.6100
12	Soil	Y	11	31.9057	30.0711	29.0693	35.4018	34.4654	33.2086	31.0309
13	Soil	Y	12	31.7275	29.6189	29.1609	34.3984	33.6840	35.0815	31.5446
14	Soil	Y	13	32.5893	30.5210	28.6054	33.4594	34.0958	33.7692	31.7129
15	Soil	Y	14	32.2473	30.5199	31.6080	35.5617	34.1706	35.7146	31.8023
16	Soil	Y	15	32.2329	29.7275	30.1770	35.3286	32.7820	37.4866	30.9706
17	Soil	Y	16	31.5831	29.7508	28.7122	33.5304	34.5348	37.1092	31.2752
18	Soil	Y	17	32.3545	31.3237	29.1542	35.5890	32.2606	37.2005	30.8430
19	Soil	Y	18	32.3811	29.6241	30.6138	35.4308	33.2024	33.0787	31.2926
20	Soil	Y	19	31.7488	29.7763	28.7327	34.7007	33.8910	34.0573	29.9064

There are no missing values in the Bayesian Bootstrap results table. All 21 rows in the Snapdragon.jmp data table are included, with varying bootstrap weights, in each bootstrap sample.

6. Select **Analyze > Distribution**.
7. Select wabash and click **Y, Columns**.
8. Click **OK**.

Figure 11.10 Distribution of wabash Means from a Bayesian Bootstrap



Quantiles

100.0%	maximum	38.184435516
99.5%		38.074322297
97.5%		37.8924836
90.0%		37.669948814
75.0%	quartile	37.185313748
50.0%	median	36.258063753
25.0%	quartile	35.034856614
10.0%		33.843030676
2.5%		32.846088771
0.5%		32.306055984
0.0%	minimum	32.204135264

Summary Statistics

Mean	35.992341
Std Dev	1.4269388
Std Err Mean	0.0451238
Upper 95% Mean	36.08089
Lower 95% Mean	35.903793
N	1000
N Missing	0

Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	32.6396	37.8168
0.90	32.8837	37.7005
0.80	33.3884	37.4621
0.50	34.4179	36.8464

Original Estimate 35.96667

The Bayesian Bootstrap produces a much smoother distribution for the wabash sample means. All 1,000 bootstrap samples include the three observations for wabash. For each iteration, the wabash sample mean is calculated using different fractional weights.

The Bootstrap Confidence Limits report shows that a 95% confidence interval for the mean is 32.6396 to 37.8168.

Statistical Details for Bootstrapping

This section contains statistical details for bootstrapping.

- "Statistical Details for Fractional Weights"
- "Statistical Details for the Bias-Corrected Percentile Intervals"

Statistical Details for Fractional Weights

This section describes how the fractional weights are calculated in a bootstrap analysis. The Fractional Weights option is based on the Bayesian bootstrap (Rubin 1981). The number of times that an observation occurs in a given bootstrap sample is called its *bootstrap weight*. In the simple bootstrap, the bootstrap weights for each bootstrap sample are determined using simple random sampling with replacement.

In the Bayesian approach, sampling probabilities are treated as unknown parameters and their posterior distribution is obtained using a non-informative prior. Estimates of the probabilities are obtained by sampling from this posterior distribution. These estimates are used to construct the bootstrap weights:

- Randomly generate a vector of n values from a gamma distribution with shape parameter equal to $(n - 1)/n$ and scale parameter equal to 1.

  Note: Rubin (1981) uses 1 as the gamma shape parameter. The shape parameter that is used in JMP ensures that the mean and variance of the fractional weights are equal to the mean and variance of the simple bootstrap weights.

- Compute S as the sum of the n values.
- Compute the fractional weights by multiplying the vector of n values by N / S, where N equals the number of rows or the sum of the frequencies if a Freq variable is specified.

Note: If a Freq variable is specified for the analysis, multiply the shape parameter for the gamma distribution by the Freq values on a row-by-row basis. The sum of the values of the Freq variable must be greater than 1. Then the shape parameters are equal to $f_i(N - 1)/N$, where f_i is the Freq value for the i^{th} row and N equals the sum of the Freq values.

This procedure scales the fractional weights for each row to have mean and variance over bootstrap sampling equal to those of the simple bootstrap weights. The fractional bootstrap weights in each bootstrap sample are positive, sum to N, and have a mean of 1.

Statistical Details for the Bias-Corrected Percentile Intervals

This section describes the calculation of the bias-corrected (BC) confidence intervals that appear in the Bootstrap Confidence Limits report when you run the Distribution script in the Bootstrap Results table. Bias-corrected percentile intervals improve on the ability of percentile intervals in accounting for asymmetry in the bootstrap distribution. See Efron (1981).

Notation

- p^* is the proportion of bootstrap samples with an estimate of the statistic of interest that is less than or equal to the original estimate.
- z_0 is the p^* quantile of a standard normal distribution.
- z_α is the α quantile of a standard normal distribution.

Bias-Corrected Confidence Interval Endpoints

The endpoints of a (1 - α) bias-corrected confidence intervals are given by quantiles of the bootstrap distribution:

- The lower endpoint is the following quantile:

$$\Phi\left(2z_0 + z_{\frac{\alpha}{2}}\right)$$

- The upper endpoint is the following quantile:

$$\Phi\left(2z_0 + z_{1-\frac{\alpha}{2}}\right)$$

Chapter 12

Text Explorer

Explore Unstructured Text in Your Data

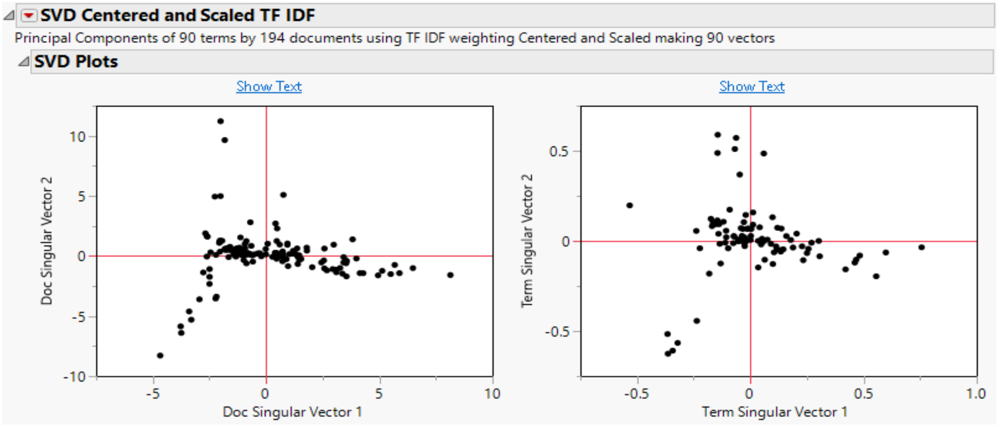
Many features in this platform are available only in JMP Pro and noted with this icon.

The Text Explorer platform enables you to analyze unstructured text, such as comment fields in surveys or incident reports. Interact with the text data by using tools to combine similar terms, recode misspecified terms, and understand the underlying patterns in your textual data.

The JMP Pro version of the platform contains analysis tools that use singular value decomposition (SVD) to group similar documents into topics. You can cluster text documents or cluster terms that are in a collection of documents. You can also cluster documents using latent class analysis.

The JMP Pro version of the platform also contains tools for identifying important terms and sentiments in your documents. Term selection enables you to identify the terms that best explain different responses. Sentiment analysis enables you to identify sentiment terms in document using lexical analysis and scores documents for positive, negative, and overall sentiment. The Sentiment Analysis feature includes basic natural language processing (NLP) support.

Figure 12.1 SVD Plots in Text Explorer



Principal Components of 90 terms by 194 documents using TF IDF weighting Centered and Scaled making 90 vectors

Contents

Overview of the Text Explorer Platform

The Text Explorer platform enables you to explore unstructured text in order to better understand its meaning. Unstructured text data are common. For example, unstructured text data could result from a free response field in a survey, product review comments, or incident reports.

Text analysis is often an iterative process, so you might alternate between curating and analyzing the list of terms.

Curating the List of Terms

Text analysis uses some unique terminology. A *term* or *token* is the smallest piece of text, similar to a word in a sentence. However, you can define terms in many ways, including through the use of regular expressions; the process of breaking the text into terms is called *tokenization*.

- A *phrase* is a short collection of terms; the platform has options to manage phrases that are specified as terms in and of themselves.
- A *document* refers to a collection of words; in a JMP data table, the unstructured text in each row of the text column corresponds to a document.
- A *corpus* refers to a collection of documents.

It is often desirable to exclude some common words from the analysis. These excluded words are called *stop words*. The platform has a default list of stop words, but you can also add specific words as stop words. Although stop words are not eligible to be terms, they can be used within phrases. However, phrases cannot start or end with a stop word.

You can also recode terms; this is useful for combining synonyms into one common term.

Stemming is the process of combining words with identical beginnings (*stems*) by removing the endings that differ. This results in “jump”, “jumped”, and “jumping” all being treated as the term “jump·”. The stemming procedure is similar to the procedure used in the Snowball string processing language. When a phrase is stemmed, each word in the phrase is stemmed as it would be stemmed as a stand-alone term.

Analyzing the List of Terms

In the Text Explorer platform, text analysis uses a *bag of words* approach. Other than in the formation of phrases, the order of terms is ignored. The analysis is based on the term counts.

After you curate the list of terms through the use of regular expressions, stop words, recoding, and stemming, you can perform analyses on the curated list of terms. The analysis options in the platform are based on the *document term matrix* (DTM). Each row in the DTM corresponds to a document (a cell in a text column of a JMP data table). Each column in the DTM corresponds to a term from the curated term list. This approach implements the bag of words approach since it ignores word ordering. In its simplest form, each cell of the DTM contains the frequency (number of occurrences) of the column's term in the row's document. There are various other weighting schemes for the DTM; these are described in "Save Options".

JMP PRO The analysis options that are available in the platform first perform a singular value decomposition (SVD) on the document term matrix. This can greatly reduce the number of columns needed to represent the term information in the data. For more information about singular value decomposition, see *Multivariate Methods*. Hierarchical clustering options are available for clustering the terms and for clustering the documents. These options enable you to group similar terms or documents together.

Text Explorer Platform Workflow

These are the expected steps for using the Text Explorer platform:

1. Specify the method for tokenizing (either built-in or customized regular expression).
2. Use the report to specify additional stop words, add phrases to the term list, perform recodes of terms, and specify exceptions to stemming rules.
3. Specify the preference for stemming.
4. Use word and phrase counts, SVD, and clustering approaches to identify important terms and phrases.

   Note: JMP PRO The SVD and clustering options are available only in JMP Pro.

5. Save results for use in further analysis: the term table, the DTM, the singular vectors, or other results.

   Note: JMP PRO The option to save the singular vectors is available only in JMP Pro.

6. Save Phrase, Recode, and Stop Words properties for use in future analyses of similar text data.

Text Processing Steps

In the Text Explorer platform, text is processed in three stages: tokenizing, phrasing, and terming.

Tokenizing Stage

The Tokenizing stage performs the following operations:

1. Convert text to lowercase.
2. Apply Tokenizing method (either Basic Words or Regex) to group characters into tokens.
3. Recode tokens based on specified recode definitions. Note that recoding occurs before stemming.

Note: Recode operations are processed internally in one pass regardless of the order that they are specified in the report window.

Phrasing Stage

The Phrasing stage collects phrases that occur in the corpus (collection of documents) and enables you to specify that individual phrases be treated as terms. Phrases cannot start or end with a stop word, but they can contain a stop word.

Terming Stage

The Terming stage creates the term list from the tokens and phrases that result from the previous stages.

For each token, the Terming stage performs the following operations:

1. Check that the minimum and maximum length requirements specified in the launch window are met. Tokens that contain only numbers are excluded from this operation.
2. Check that the token is qualified to become a term; tokens parsed by the Basic Words tokenization method must contain at least one alphabetical or Unicode character. Tokens that contain only numbers are excluded from this operation. The Regex tokenization method uses regular expressions to determine what characters are part of a token.
3. Check that the token is not a stop word.
4. Apply stemming and stem exceptions.

For each phrase that you add, the Terming stage performs the following operations:

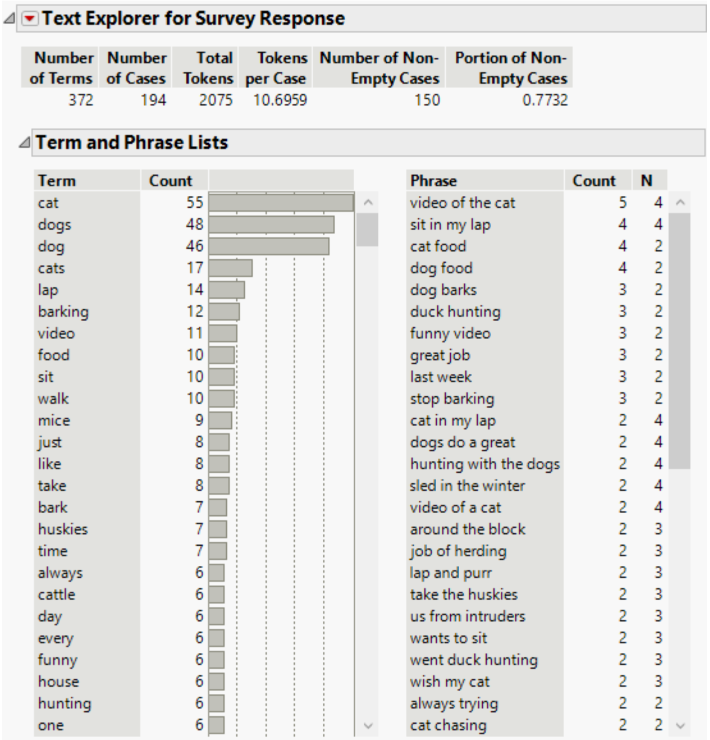
1. Add the phrase to the term list. Phrases should apply stemming to each word in the phrase that is stemmed in the term list. Phrases that have different raw tokens but the same stems are combined in the term list.
2. Remove token term occurrences that appear in the phrase.

Example of the Text Explorer Platform

Learn how to explore text responses in JMP. In this example, you want to explore the text responses from a survey about pets.

1. Select **Help > Sample Data Folder** and open Pet Survey.jmp.
2. Select **Analyze > Text Explorer**.
3. Select Survey Response and click **Text Columns**.
4. From the Language list, select **English**.
5. Click **OK**.

Figure 12.2 Example of Initial Text Explorer Report



Text Explorer for Survey Response

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
372	194	2075	10.6959	150	0.7732

Term and Phrase Lists

Term	Count
cat	55
dogs	48
dog	46
cats	17
lap	14
barking	12
video	11
food	10
sit	10
walk	10
mice	9
just	8
like	8
take	8
bark	7
huskies	7
time	7
always	6
cattle	6
day	6
every	6
funny	6
house	6
hunting	6
one	6

Phrase	Count	N
video of the cat	5	4
sit in my lap	4	4
cat food	4	2
dog food	4	2
dog barks	3	2
duck hunting	3	2
funny video	3	2
great job	3	2
last week	3	2
stop barking	3	2
cat in my lap	2	4
dogs do a great	2	4
hunting with the dogs	2	4
sled in the winter	2	4
video of a cat	2	4
around the block	2	3
job of herding	2	3
lap and purr	2	3
take the huskies	2	3
us from intruders	2	3
wants to sit	2	3
went duck hunting	2	3
wish my cat	2	3
always trying	2	2
cat chasing	2	2

At a glance, you can see that there are 372 unique terms in 194 documents. In all, there are 2075 tokenized terms. The most common term is “cat”, and it occurs 55 times.

6. Click the red triangle next to Text Explorer for Survey Response and select **Term Options > Stemming > Stem All Terms**.
7. In the Phrase List table, select **cat food** and **dog food**, right-click the selection, and select **Add Phrase**.

The terms cat food and dog food are included in the term list.

8. Scroll down in the Term List report and find the cat and dog food entries.

   You can see that there are four occurrences of each phrase.

Figure 12.3 Term List after Modifications and Scrolling

Term	Count
barn·	4
cat· food·	4
couch·	4
dog· food·	4
door·	4
duck·	4
get·	4
great·	4
guard·	4
job·	4
know·	4
made·	4
make·	4
now·	4
watch·	4
anymor·	3
ate·	3
back·	3
bath·	3
best·	3
box·	3
eat·	3

In the phrase list, cat food and dog food are gray, since they are now locally being treated as terms in this Text Explorer report.

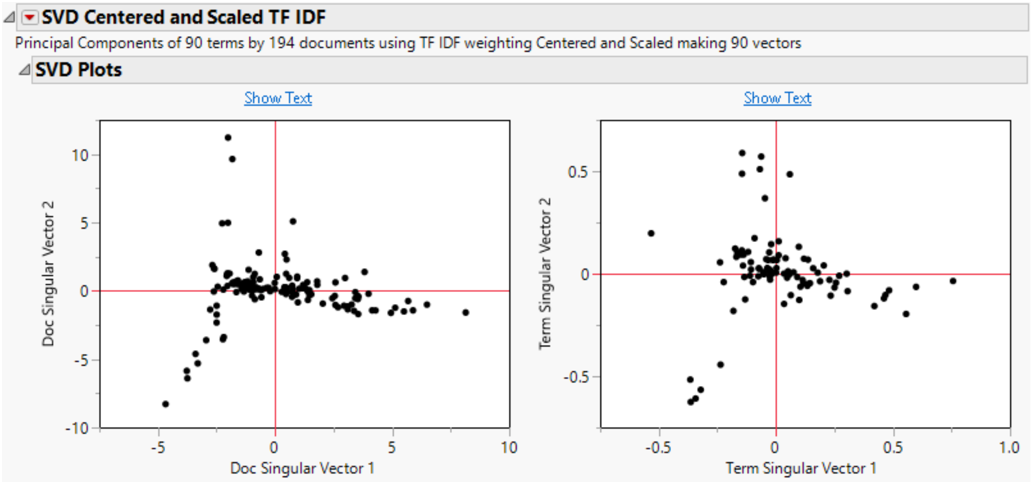
The remaining steps of this example can be completed only in JMP Pro.

9. Click the red triangle next to Text Explorer for Survey Response and select **Latent Semantic Analysis, SVD**.

10. Click **OK** to accept the default values.

    Two SVD Plots appear in the report. The one on the left shows the first two singular vectors in the document space. The one on the right shows the first two singular vectors in the term space.

Figure 12.4 SVD Plots



SVD Centered and Scaled TF IDF

Principal Components of 90 terms by 194 documents using TF IDF weighting Centered and Scaled making 90 vectors

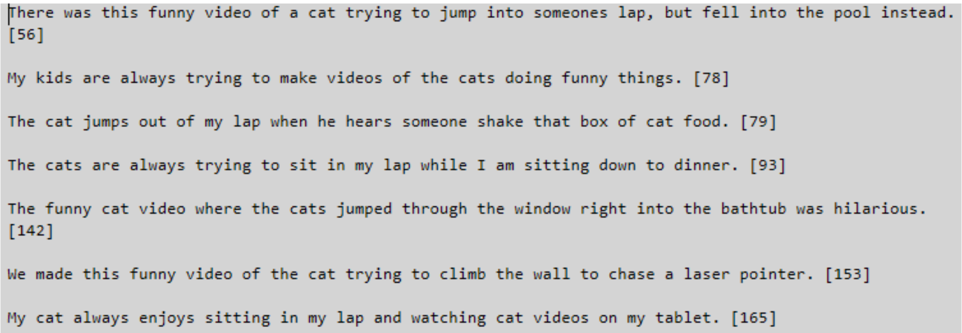
SVD Plots

11. Select the seven right-most points in the left SVD Plot.

These seven points represent survey responses that are clustered away from the rest of the points. To further investigate this cluster, you read the text of these responses.

12. Click the **Show Text** button that is above the left SVD Plot.

Figure 12.5 Text of Selected Documents



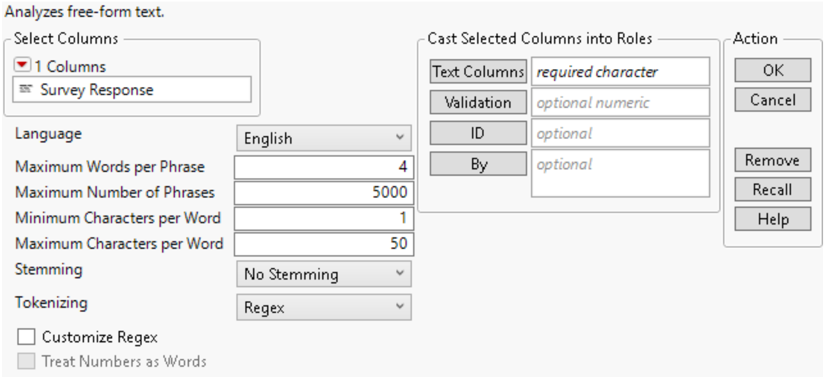
```
There was this funny video of a cat trying to jump into someones lap, but fell into the pool instead.
[56]

My kids are always trying to make videos of the cats doing funny things. [78]

The cat jumps out of my lap when he hears someone shake that box of cat food. [79]

The cats are always trying to sit in my lap while I am sitting down to dinner. [93]

The funny cat video where the cats jumped through the window right into the bathtub was hilarious.
[142]

We made this funny video of the cat trying to climb the wall to chase a laser pointer. [153]

My cat always enjoys sitting in my lap and watching cat videos on my tablet. [165]
```

A window appears that contains the text of the seven documents represented by the selected points. These survey responses are similar in that they all refer to some combination of “funny”, “cat”, and “video”. These documents have larger positive values for the first singular vector than the rest of the documents. These larger values indicate that they are different from the rest of the documents in that dimension.

Further investigation of the singular vector dimensions could lead to interpretations of what the dimensions represent. For example, many of the documents on the far right of the plot are responses that are about cats. On the far left, many of the responses are about dogs. Therefore, the first singular vector is picking up differences based on whether the response was about a cat or a dog.

Launch the Text Explorer Platform

Launch the Text Explorer platform by selecting **Analyze > Text Explorer**.



Analyzes free-form text.

Language: English
Maximum Words per Phrase: 4
Maximum Number of Phrases: 5000
Minimum Characters per Word: 1
Maximum Characters per Word: 50
Stemming: No Stemming
Tokenizing: Regex
☐ Customize Regex
☐ Treat Numbers as Words

Figure 12.6 The Text Explorer Launch Window

For more information about the options in the Select Columns red triangle menu, see *Using JMP*. The Text Explorer launch window contains the following options:

Text Columns Assigns the columns that contain text data. If you specify multiple columns, a separate analysis is created for each column.

JMP PRO **Validation** In JMP Pro, you can enter a Validation column. If you click the Validation button with no columns selected in the Select Columns list, you can add a validation column to your data table. For more information about the Make Validation Column utility, see *Predictive and Specialized Modeling*.

The specification of a Validation column does not affect the calculation of the document-term matrix. However, when a Validation column is specified, only the training set is used for the Latent Class Analysis, Latent Semantic Analysis, Topic Analysis, and Discriminant Analysis options. The Validation column is used as the Generalized Regression validation method for the Term Selection option.

ID Assigns a column used to identify separate respondents in the Save Stacked DTM for Association output data table. This output data table is suitable for association analysis. This column is also used to identify separate respondents in the Latent Class Analysis report.

By Identifies a column that creates a report consisting of separate analyses for each level of the variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Note: If you specify a By variable, the Customize Regex option and settings apply to all levels of the By variables.

Language Specifies the language used for text processing. This affects stemming and the built-in lists of stop words, recodes, and phrases. This option is independent of the language in which JMP is running. Unless the Language platform preference is set, the Language option is set according to the JMP Display Language preference.

Maximum Words per Phrase Specifies a maximum number of words that a phrase can contain to be included as a phrase in the analysis.

Maximum Number of Phrases Specifies the maximum number of phrases that appear in the phrase list.

Minimum Characters per Word Specifies the number of characters that a word must contain to be included as a term in the analysis.

Maximum Characters per Word Specifies the largest number of characters (up to 2000) that a word can contain to be included as a term in the analysis.

Stemming (Available only when the Language option is set to English, German, Spanish, French, or Italian.) Specifies a method for combining terms with similar beginning characters but different endings. The following options are available:

No Stemming No terms are combined.

Stem for Combining Stems only the terms where two or more terms stem to the same term.

Stem All Terms Stems all terms.

Note: The use of the Stemming option also affects phrases that have been added to the term list. Phrase identification occurs after terms within a phrase have been stemmed. For example, “dogs bark” and “dog barks” would both match the specified phrase “dog· bark·”. Phrases cannot be removed from the term list when a Stemming option is selected.

Tokenizing (Available only when the Language option is set to English, German, Spanish, French, or Italian.) Specifies a method for parsing the text into terms or tokens. The following tokenization options are available:

Regex Parses text using a default set of built-in regular expressions. If you want to add to, remove, or edit the set of regular expressions used to parse the text, select the **Customize Regex** option. See "Customize Regex in the Regular Expression Editor".

Basic Words Text is parsed into words based on a set of characters that typically separate words. These characters include spaces, tabs, new lines, and most punctuation marks. If you want numbers to be parsed into terms for the analysis, select the **Treat Numbers as Words** option. If you do not select this option, pieces of text between delimiters that contain only numbers are ignored in the tokenizing step.

Tip: You can view the default set of delimiters using the **Display Options > Show Delimiters** option in a Text Explorer report that uses the Basic Words Tokenizing method.

Customize Regex (Available only with the Regex Tokenizing method.) Enables you to use the Text Explorer Regular Expression Editor window to modify the Regex settings. Use this option to accommodate non-traditional words. Examples include phone numbers or words formed by a combination of characters and numbers. Using the Customize Regex option is not recommended unless the default Regex method is not giving you the results that you need. This can happen when your text contains structures that the default Regex method does not recognize. See "Customize Regex in the Regular Expression Editor".

Treat Numbers as Words (Available only with the Basic Words Tokenizing method.) Allows numbers to be tokenized as terms in the analysis. When this option is selected, the Minimum Characters per Word setting is ignored for terms that contain numeric digits.

After you click **OK** on the launch window, the Text Explorer Regular Expression Editor window appears if you selected **Customize Regex** in the launch window. Otherwise, the Text Explorer report appears.

Note: The processing of text input is not case-sensitive. All text is converted to lowercase internally prior to tokenization and all analysis steps. This conversion affects the processing of regular expressions and the aggregation of terms in the Text Explorer output.

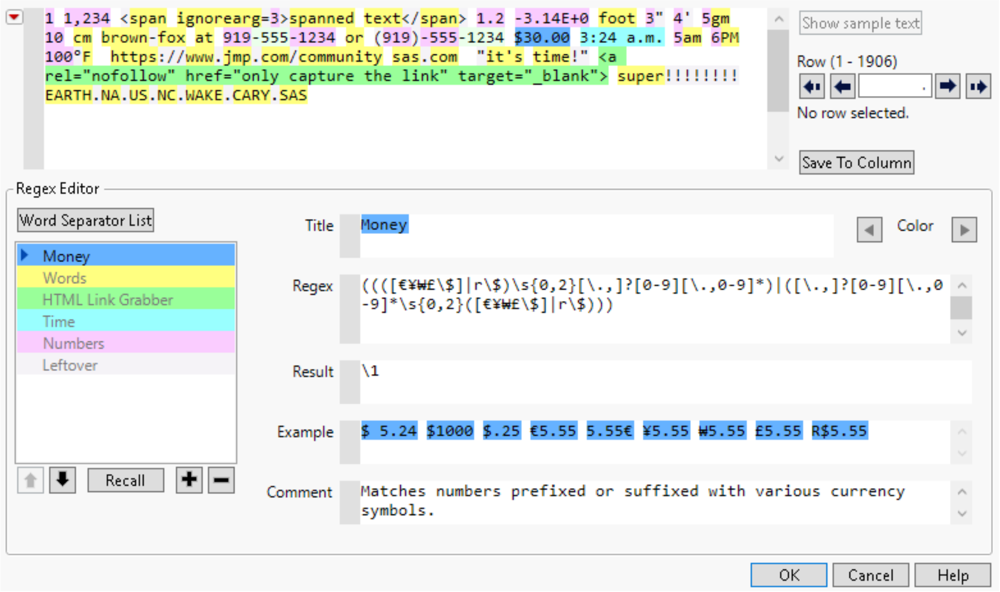
Customize Regex in the Regular Expression Editor

When you select the Customize Regex option, the Text Explorer Regular Expression Editor appears. Use this window to parse text documents using a wide variety of built-in regular expressions, such as phone numbers, times, or monetary values. You can also create your own regular expression definitions.

Note: Using the Customize Regex option is recommended only if you are not getting desired results from the default Regex method. This can happen when your text contains structures that the default Regex method does not recognize.

Tip: If Japanese, Chinese (Simplified), or Chinese (Traditional) is specified as the Language option in the launch window, the list of Regex patterns contains a single Regex for the specified language. If you want to add other Regex patterns, it is recommended that you add them after the single Regex pattern. You should avoid using the Words pattern before the language-specific Regex pattern, because the Words pattern can gather long runs of Asian language characters into single words.

Figure 12.7 Text Explorer Regular Expression Editor



```
1 1,234 <span ignorearg=3>spanned text</span> 1.2 -3.14E+0 foot 3" 4' 5gm
10 cm brown-fox at 919-555-1234 or (919)-555-1234 $30.00 3:24 a.m. 5am 6PM
100°F  https://www.jmp.com/community sas.com  "it's time!" <a
rel="nofollow" href="only capture the link" target="_blank"> super!!!!!!!!
EARTH.NA.US.NC.WAKE.CARY.SAS
```

Row (1 - 1906)

No row selected.

Title: Money

Regex:

```
((([€¥₩£\$]|r\$)\s{0,2}[\.,]?[0-9][\.,0-9]*)|([\.,]?[0-9][\.,0
-9]*\s{0,2}([€¥₩£\$]|r\$)))
```

Result: `\1`

Example: `$ 5.24 $1000 $.25 €5.55 5.55€ ¥5.55 ₩5.55 £5.55 R$5.55`

Comment: Matches numbers prefixed or suffixed with various currency symbols.

Parsing with the Script Editor Box

The script editor box at the top of the window shows you how the parsing would proceed for sample text. The results of parsing the regular expressions in the Regex Editor list are highlighted in colors that correspond to the colors in the Regex Editor list.

- Click the **First**, **Previous**, **Next**, and **Last** row buttons to populate the script editor box with text from your own data. This enables you to see how a given row of text data is parsed. You can also enter a row number in the edit box to populate the script editor box with text from a specific row in the data table.
- Click the **Save to Column** button to save a new column to the data table that contains the result of the regular expression tokenization. For more information about specifying the result of the regular expression, see "Editing the Regular Expressions". The **Save to Column** button does not appear if you access the Regular Expression Editor through **Cols >Utilities > New Column by Text Matching**.

Note: The **Save to Column** button uses only the regular expression to match text. The following settings are not used: stop words, recodes, stemming, phrases, or minimum and maximum characters per word to modify the output of the regular expression.

Adding Regular Expressions

To add a regular expression to be used in tokenization, click the plus sign below the list. The Regex Library Selections window appears. This window contains all the built-in regular expressions as well as any recently modified regular expressions that you created in previous instances of the Regular Expression Editor. Built-in regular expressions are labeled. Custom regular expressions that are saved in your library are labeled with the name that you specified. Only the most recent expression for a given name is stored in the Regex Library.

Click the **Recall** button to populate the regular expressions list with the regular expressions from the most recent instance of the Regular Expression Editor. The recalled regular expressions are ones that were present in the previous instance of the editor when either the **Save to Column** button or the **OK** button was clicked.

Select one or more regular expressions in the list and click **OK** to add the selected regular expressions to be used in tokenization. Use the **Delete Selected Item** button to remove one or more custom regular expressions from the Regex Library. The Regex Library for each user is stored as a JSL file in a directory called TextExplorer. The location of this directory is based on your computer's operating system:

- Windows: "C:/Users/<username>/AppData/Roaming/SAS/JMP/TextExplorer/"
- macOS: "/Users/<username>/Library/Application Support/JMP/TextExplorer/"

These files can be shared with other users, but you should not edit the file directly. Use the Regular Expression Editor instead.

Editing the Regular Expressions

Terms are tokenized by processing the regular expressions in the order specified in the Regex Editor panel. To change the order of the regular expressions, select a regular expression in the list and click the up or down arrow buttons below the list. You can also drag and drop items in the regular expression list to change the order of execution. A blue triangle represents the currently selected regular expression. To remove a regular expression and exclude it from the tokenization, select it in the list and click the minus sign below the list. The "Leftover" regular expression cannot be removed and must appear last in the sequence of regular expressions.

When you select a regular expression in the list, the editable fields in the Regex Editor panel refer to the selected regular expression. Click and type in any of these fields to edit them.

Each regular expression has the following attributes:

Title Specifies a name used to identify the regular expression in the current window (as well as in the Regex Library later).

Regex Specifies the regular expression definition. The regular expression must have at least one set of parentheses to designate the regular expression capture.

Result Specifies what replaces the text matched by the regular expression. This value can be static text, blank, or the value of the regular expression capture. The regular expression capture is defined as the result of the Regex definition:

- To replace the matched text with static text, specify the static text in the Result field.
- To ignore the matched text, leave the Result field blank.
- To keep the text that results from the outer-most parentheses in the regular expression, use “\1” (without quotation marks) in the Result field.
- To keep the entire result of the regular expression, use “\0” (without quotation marks) in the Result field.

Example (Optional) Specifies an example text string with colors indicating the behavior of the regular expression.

Comment (Optional) Specifies a comment to explain the regular expression and its behavior.

Color Specifies the color used to identify matches of the regular expression in the text in the Script Editor box and in the Example field. Use the arrow buttons to change the color.

Note: If the regular expression definition in the Regex field is invalid, a red X appears next to the name of the regular expression in the list of regular expressions.

Creating a Custom Regular Expression

Follow these steps to create your own custom regular expression:

1. Click the plus sign below the list.
2. In the Regex Library Selections window, note that the Blank regular expression is selected.
3. Click **OK**.
4. Edit the Regex definition in the Regex Editor panel.
5. Give your custom regular expression a unique name in the Title field.

Tip: When editing the Regex definition field, it is helpful to have the Log window open and visible. Some error messages appear only in the Log window. To open the Log window, select **View > Log**. There are many internet resources available for troubleshooting regular expressions, such as https://regexr.com/.

The Word Separator List

The **Word Separator List** button enables you to specify a list of characters that occur between words in the tokenization process. The *between-word characters* cannot begin a word, but they can appear inside a word if one of the regular expressions allows it. You can add or remove characters from the list in the window that appears when you click the button. By default, the only character in the list is a whitespace character. In the Separator Characters window, click the **Reset** button to undo any modifications to the list of separator characters. Modifications to the list of separator characters are applied only to the current regular expression tokenization.

The following steps describe the processing of the specified regular expressions and the required “Leftover” regular expression:

1. Compare the current character in the text stream to the list of separator characters.
   – If the character is in the list of separator characters, ignore the character, process any accumulated characters in the “Leftover” temporary string, move to the next character, and repeat step 1.
   – If the character is not in the list of separator characters, go to step 2.
2. Compare the string starting at the current character to each regular expression (one at a time, up to, but not including, the “Leftover” regular expression).
   – If the string starting at the current character matches one of the regular expressions, the following events occur. Any accumulated characters in the “Leftover” temporary string are processed. The value of the Result field is saved as a term. The current character in the text stream becomes the character following the matched string. The processing returns to step 1.
   – If the string starting at the current character does not match any of the regular expressions up to the “Leftover” regular expression, go to step 3.
3. Collect characters into the “Leftover” temporary string by appending the current character and setting the current character to the next character in the text stream. Return to step 1.
   – The “Leftover” temporary string is accumulated one character at a time, until one of the other regular expressions produces a match.
   – The default Result of the “Leftover” regular expression is to discard the accumulated “Leftover” temporary string.

Tips:

- If you set the Result of the “Leftover” regular expression to \1, you might want to add more separator characters, such as punctuation marks. This ensures that your results do not include the specified punctuation marks.
- Instead of changing the Result of the “Leftover” regular expression to \1, you might want to consider one or more of the following actions to capture terms of interest:

  - Add more regular expressions from the Regex Library.
  - Create custom regular expressions.

The processing follows the above steps until reaching the end of the text string for each row in the data table.

Saving the Results to a Column in the Data Table

Click the **Save to Column** button to save to the data table a new column that contains the results of the regular expression tokenization. The new column is a character column with the same name as the text column specified in the Text Explorer launch window; a number is appended to the name so that the column names are unique. You can also use a stand-alone Regular Expression utility under Cols > Utilities > New Column by Text Matching. See *Using JMP*.

Note: When you save the results of the custom regular expression tokenization to a column in the data table, the regular expression process is run on the original text in each row of the data table. It is not run on the version of the text string that was converted to lowercase.

Closing the Text Explorer Regular Expression Editor

After you click **OK** in the Text Explorer Regular Expression Editor window, the following events occur:

1. The custom regular expressions defined in the Text Explorer Regular Expression Editor window are saved to the Regex Library.

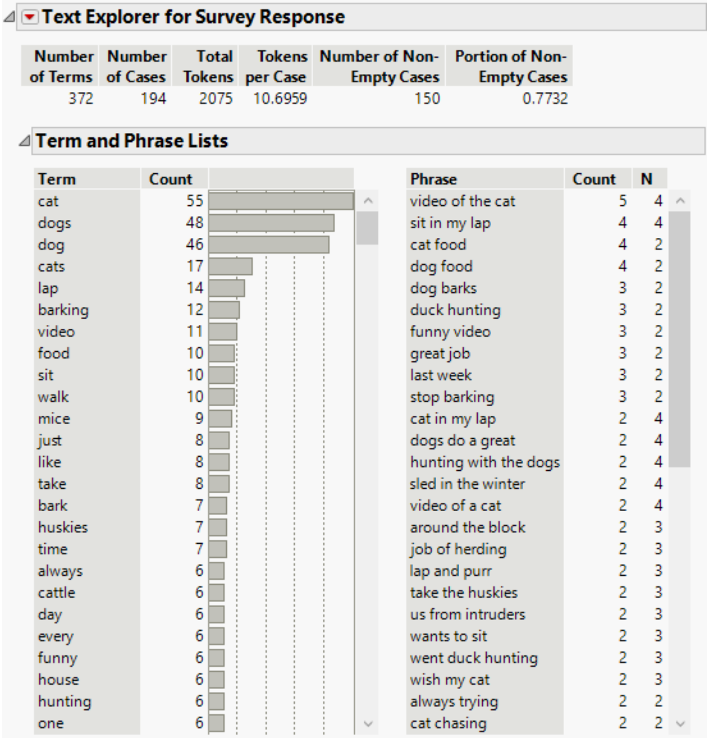
   Caution: The custom Regex Library is saved only when you click **OK** and there are customized regular expressions. The most recently saved regular expressions will be available next time. Use unique names to keep additional regular expressions in the Regex Library. To ensure that a regular expression is available later, you can save a script from the Text Explorer report window.

2. The Text Explorer report appears. The report shows the result of using the specified regular expression settings to tokenize the text.

The Text Explorer Report

The Text Explorer report contains the Summary Counts report and the Term and Phrase Lists report.

Figure 12.8 Example of a Text Explorer Report



Text Explorer for Survey Response

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
372	194	2075	10.6959	150	0.7732

Term and Phrase Lists

Term	Count
cat	55
dogs	48
dog	46
cats	17
lap	14
barking	12
video	11
food	10
sit	10
walk	10
mice	9
just	8
like	8
take	8
bark	7
huskies	7
time	7
always	6
cattle	6
day	6
every	6
funny	6
house	6
hunting	6
one	6

Phrase	Count	N
video of the cat	5	4
sit in my lap	4	4
cat food	4	2
dog food	4	2
dog barks	3	2
duck hunting	3	2
funny video	3	2
great job	3	2
last week	3	2
stop barking	3	2
cat in my lap	2	4
dogs do a great	2	4
hunting with the dogs	2	4
sled in the winter	2	4
video of a cat	2	4
around the block	2	3
job of herding	2	3
lap and purr	2	3
take the huskies	2	3
us from intruders	2	3
wants to sit	2	3
went duck hunting	2	3
wish my cat	2	3
always trying	2	2
cat chasing	2	2

Summary Counts Report

The first table in the Text Explorer report contains the following summary statistics:

Number of Terms The number of terms in the term list.

Number of Cases The number of documents in the corpus.

Total Tokens The total number of terms in the corpus.

Tokens per Case The number of tokens divided by the number of cases.

Number of Non-Empty Cases The number of documents in the corpus that contain at least one term.

Portion of Non-Empty Cases The proportion of documents in the corpus that contain at least one term.

Term and Phrase Lists

The Term and Phrase Lists section of the Text Explorer report contains tables of terms and phrases found in the text after tokenization has occurred. See Figure 12.8 for an example of the Term and Phrase Lists report. The Count column in the term list indicates the number of occurrences of the term in the corpus. The Count column in the phrase list indicates the number of occurrences of the phrase in the corpus; the N column indicates the number of words in the phrase.

By default, the Terms List is sorted in descending count order; terms that are tied in count are sorted alphabetically. The Phrases List is sorted in descending count order; phrases that are tied in count are then sorted in descending length (N) order. Further ties in the Phrases List are sorted alphabetically. The sort order of each list can be changed to alphabetical sorting using the options in each list.

The phrases that appear in the phrase list are determined by the settings of the **Maximum Words per Phrase** and **Maximum Number of Phrases** options in the launch window. Phrases that occur only one time in the data table do not appear in the phrase list.

Phrases can be specified as terms at various scopes. Phrases in the phrase list that have been specified as terms are colored based on the scope of the phrase specification (Table 12.1). For more information about specifying phrases in different scopes, see "Term Options Management Windows".

Table 12.1 Colors for Specified Phrases

Scope	Color
Built-in	Red
User Library	Green
Project	Blue
Column Property	Orange
Local	Gray

Actions for Terms and Phrases

You can access options in the Term List and Phrase List tables by selecting items and then right-clicking in the left-most column of each table. You can save each table as a data table by right-clicking in the Count column of each table and selecting Make into Data Table.

Term List Pop-up Menu Options

When you right-click in the Term column of the Term List table, a pop-up menu appears with the following options:

Select Rows Selects rows in the data table that contain the selected terms.

Show Text Shows the documents that contain the selected terms.

> **Note:** By default, only the first 10,000 documents are shown. If there are more than 10,000 documents that contain a selected term, a window appears that enables you to increase this limit.

Alphabetical Order Specifies the sort order of the term list. When this option is selected, the terms are sorted in alphabetical order. When this option is not selected, the terms are sorted in descending Count order.

Numerical Order (Available only when the Alphabetical Order option is selected.) Specifies the sort order of the term list. When this option is selected, the items are split into string and numeric segments, and the numeric segments are then sorted in numerical order. For more information about the sorting rules used by the Numerical Order option, see *Using JMP*.

Copy Places the selected terms onto the clipboard.

Color Enables you to assign a color to the selected terms.

Label Places labels on the corresponding points in the Term SVD Plot for the selected terms.

Containing Phrases Selects the phrases in the Phrase List table that contain the selected terms.

Save Indicators Saves an indicator column to the data table for each term selected in the term list. The value of the indicator column for each row is 1 if the document in that row contains the term and 0 otherwise.

Save Formula Saves a column formula to the data table for each term selected in the term list. The column formula for each row evaluates to 1 if the document in that row contains the term and 0 otherwise. This is useful for new documents.

Recode Enables you to change the values for one or more terms. Select the terms in the list before selecting this option. After you select this option, the Recode window appears. See *Using JMP*.

Add Stop Word Adds the selected terms to the list of stop words and removes those terms from the term list. This action also updates the phrase list.

Note: If you add a stemmed word as a stop word, all of the tokens that correspond to that stem are added as stop words.

Add Stem Exception (Available only when the Language option is set to English, German, Spanish, French, or Italian.) Adds the selected terms to the list of terms that are excluded from stemming.

Remove Phrase (Available only when a specified phrase is selected in the term list and the selected Stemming method is No Stemming.) Removes the selected phrase from the set of specified phrases and updates the Term Counts accordingly.

Note: If a phrase as been added as a Sentiment Phrase, the Remove Phrase option also removes the phrase from the list of sentiment terms in the current Sentiment Analysis report.

JMP PRO **Add Sentiment** (Available only when a Sentiment Analysis report is open in the current report window.) Adds the selected terms to the list of sentiment terms in the current Sentiment Analysis report.

Note: If you add a stemmed word as a sentiment term, all of the tokens that correspond to that stem are added as sentiment terms.

Show Filter Shows or hides a search filter above the term list. See "Search Filter Options".

Make into Data Table Creates a JMP data table from the report table.

Make Combined Data Table Searches the report for other tables like the one you selected and combines them into a single JMP data table.

Phrase List Pop-up Menu Options

When you right-click in the Phrase column of the Phrase List table, a pop-up menu appears with the following options:

Select Rows Selects rows in the data table that contain the selected phrases.

Show Text Shows the documents that contain the selected phrases.

Save Indicators Saves an indicator column to the data table for each phrase selected in the phrase list. The value of the indicator column for each row is 1 if the document in that row contains the phrase and 0 otherwise.

Alphabetical Order Specifies the sort order of the phrase list. When this option is selected, the terms are sorted in alphabetical order. When this option is not selected, the terms are sorted in descending Count order.

Numerical Order (Available only when the Alphabetical Order option is selected.) Specifies the sort order of the phrase list. When this option is selected, the items are split into string and numeric segments, and the numeric segments are then sorted in numerical order. For more information about the sorting rules used by the Numerical Order option, see *Using JMP*.

Copy Places the selected phrases onto the clipboard.

Select Contains Selects larger phrases in the phrase list that contain the selected phrase.

Select Contained Selects smaller phrases in the phrase list and terms in the term list that are contained by the selected phrase.

Add Phrase Adds the selected phrases to the term list and updates the Term Counts accordingly.

Add Stop Word Adds the selected phrases to the list of stop words. This action also updates the term list.

JMP PRO **Add Sentiment Phrase** (Available only when a Sentiment Analysis report is open in the current report window.) Adds the selected phrases to the term list and to the list of sentiment terms in the current Sentiment Analysis report.

Show Filter Shows or hides a search filter above the phrase list. See “Search Filter Options”.

Make into Data Table Creates a JMP data table from the report table.

Make Combined Data Table Searches the report for other tables like the one you selected and combines them into a single JMP data table.

Search Filter Options

Click the down arrow button next to the search box to refine your search.

Contains Terms Returns items that contain a part of the search criteria. A search for “ease oom” returns messages such as “Release Zoom”.

Contains Phrase Returns items that contain the exact search criteria. A search for “text box” returns entries that contain “text” followed directly by “box” (for example, “Context Box” and “Text Box”).

Starts With Phrase Returns items that start with the search criteria.

Ends With Phrase Returns items that end with the search criteria.

Whole Phrase Returns items that consist of the entire string. A search for “text box” returns entries that contain only “text box”.

Regular Expression Enables you to use the wildcard (*) and period (.) in the search box. Searching for “get.*name” looks for items that contain “get” followed by one or more words. It returns “Get Color Theme Names”, “Get Name Info”, and “Get Effect Names”, and so on.

Invert Result Returns items that do not match the search criteria.

Match All Terms Returns items that contain both strings. A search for “t test” returns elements that contain either or both of the search strings: “Pat Test”, “Shortest Edit Script” and “Paired t test”.

Ignore Case Ignores the case in the search criteria.

Match Whole Words Returns items that contain each word in the string based on the Match All Terms setting. If you search for “data filter”, and Match All Terms is selected, entries that contain both “data” and “filter” are returned.

Text Explorer Platform Options

This section describes the options available in the Text Explorer platform.

- “Text Preparation Options”
- “Text Analysis Options”
- “Save Options”
- “Report Options in Text Explorer”

Text Preparation Options

The Text Explorer red triangle menu contains the following options for text preparation:

Display Options Shows a submenu of options to control the report display.

Show Word Cloud Shows or hides the Word Cloud report. The Word Cloud red triangle menu enables you to change the layout and font for the word cloud. See “Word Cloud Options”.

The word cloud can be interactively resized by changing the width. The height is then determined automatically. The rows in the term list are linked to the terms in the word cloud.

Show Term List Shows or hides the Term List report.

Show Phrase List Shows or hides the Phrase List report.

Show Term and Phrase Options Shows or hides buttons in the Term and Phrase Lists report corresponding to the options available in the pop-up menus for each list. See "Term and Phrase Lists".

Show Summary Counts Shows or hides the Summary Counts table. See "Summary Counts Report".

Show Stop Words Shows or hides a list of the stop words used in the analysis. A built-in list of stop words is used initially. To add a stop word, right-click it in the Term List report and select **Add Stop Word** from the pop-up menu. See "Term Options Management Windows".

Show Recodes Shows or hides a list of the recoded terms. See "Term Options Management Windows".

Show Specified Phrases Shows or hides a list of the phrases that have been specified by the user to be treated as terms. See "Term Options Management Windows".

Show Stem Exceptions (Available only when the Language option is set to English, German, Spanish, French, or Italian.) Shows or hides the terms that are excluded from stemming. See "Term Options Management Windows".

Show Delimiters (Available only when the Language option is set to English, German, Spanish, French, or Italian and the selected Tokenizing method is Basic Words.) Shows or hides the delimiters used by the Basic Words Tokenizing method. To modify the set of delimiters used, you must use the `Add Delimiters()` or `Set Delimiters()` messages in JSL.

Show Stem Report (Available only when the Language option is set to English, German, Spanish, French, or Italian and the selected Stemming method is not No Stemming.) Shows or hides the Stemming report that contains two tables of stemming results. The table on the left maps each stem to the corresponding terms. The table on the right maps each term to its corresponding stem.

Show Selected Rows Opens a window that contains the text of the documents that are in the currently selected rows.

Show Filters for All Tables Shows or hides filters that can be used for searching tables in the report. This option applies to the following tables: Stop Words, Specified Phrases, Stem Exceptions, Term List, Phrase List, and the Stem Report. For more information about the filter tool, see "Search Filter Options".

Term Options Shows a submenu of options that apply to the Term List report.

Stemming (Available only when the Language option is set to English, German, Spanish, French, or Italian.) See the description of stemming options in "Launch the Text Explorer Platform".

Include Builtin Stop Words Specifies if the stop words used in the tokenizing process include built-in stop words or not.

Include Builtin Phrases Specifies if the phrases used in the tokenizing process include built-in phrases or not.

Manage Stop Words Shows a window that enables you to add or remove stop words. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude stop words that are specified in any of the other levels. See "Term Options Management Windows".

Manage Recodes Shows a window that enables you to add or remove recodes. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude recodes that are specified in any of the other levels. See "Term Options Management Windows".

Manage Phrases Shows a window that enables you to add or remove the phrases that are treated as terms. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude phrases that are specified in any of the other levels. See "Term Options Management Windows".

Manage Stem Exceptions (Available only when the Language option is set to English, German, Spanish, French, or Italian.) Shows a window that enables you to add or remove exceptions to stemming. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude stem exceptions that are specified in any of the other levels. See "Term Options Management Windows".

Parsing Options Shows a submenu of options that apply to parsing and tokenization.

Tokenizing (Available only when the Language option is set to English, German, Spanish, French, or Italian.) See the description of tokenizing options in "Launch the Text Explorer Platform".

Customize Regex (Available only with the Regex Tokenizing method.) Shows the Customize Regex window. This option enables you to modify the Regex settings for the current Text Explorer report.

Note: If you specified a By variable in the platform launch window, the Customize Regex option automatically broadcasts to all level of the By variables.

Treat Numbers as Words (Available only when the Language option is set to English, German, Spanish, French, or Italian and Basic Words is the selected Tokenizing method.) Allows numbers to be tokenized as terms in the analysis. Note that this option is affected by the setting for Minimum characters per word.

Word Cloud Options

The Word Cloud red triangle menu contains the following options:

Layout Specifies the arrangement of the terms in the word cloud. By default, the Layout is set to Ordered.

- **Ordered** Presents the terms in horizontal lines ordered from most to least frequent.
- **Alphabetical** Presents the terms in horizontal lines sorted in ascending alphabetical order.
- **Centered** Presents the terms in a cloud and sized by frequency.

Coloring Specifies the coloring of the terms in the word cloud. By default, the Coloring is set to None.

- **None** Colors each term the same color as it is colored in the term list.
- **Uniform Color** Colors each term the same color. You can change this color in the Legend.
- **Arbitrary Grays** Colors each term in varying shades of gray.
- **Arbitrary Colors** Colors each term in various colors. You can adjust the colors in the Legend.
- **By column values** Colors each term on a gradient color scale. The scale is based on the score for a term generated by the Score Terms by Column option. You can adjust the colors and gradient in the Legend.

Font Specifies the font, style, and size of the terms in the word cloud.

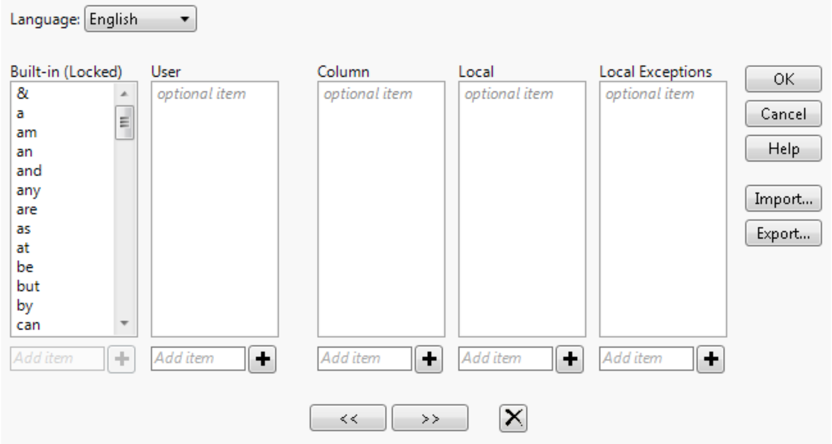
Show Legend Shows or hides the legend for the word cloud.

Term Options Management Windows

Stop word, recode, phrase, and stem exception information can be specified for many different scopes. They can be stored in the following locations: the Text Explorer user library (User scope), the current project, a column property for the analysis column (Column scope), or in a platform script (Local scope). You can save the local specifications and local exceptions for a specific instance of Text Explorer by saving the script for the Text Explorer report.

The Term Options management windows are four similar windows that enable you to manage the collections of stop words, recodes, phrases, and stem exceptions. Figure 12.9 shows the Manage Stop Words window. The Manage Phrases and Manage Stem Exceptions are identical to the Manage Stop Words window. The Manage Recodes window differs slightly. See “Manage Recodes”.

Figure 12.9 Manage Stop Words Window



Language: English

Manage Stop Words

The Manage Stop Words window contains multiple lists of stop words that represent the different scopes (or locations) of specified stop words. Below each list is a text edit box and an add button. These controls enable you to add custom stop words to each scope. You can move stop words from one scope to another by dragging them. You can copy and paste items from one list to another list. Two buttons at the bottom of the window move the selected items from one scope to the next, either left or right. The X button removes the selected items from their current scope. You can edit existing items in the lists by double-clicking on an item and changing the text.

Language Specifies the list of built-in stop words and to which language the user library selections are saved. If you select Apply Items for Language, the changes are saved to the main user library. The Language setting applies only to the Built-in, User, and Project scopes.

Built-in (Locked) Lists the built-in list of stop words for the specified language. You can exclude a built-in stop word by placing it in the Local Exceptions list.

User Lists the stop words in the user library for the specified language.

Project (Available only when Text Explorer is launched within a project.) Lists the stop words in the current project for the specified language.

Column Lists the stop words in the “Stop Words” column property for the text column.

Local Lists the stop words in the local scope. You can specify them when Text Explorer is launched via JSL. These stop words are used only in the current Text Explorer platform report.

Local Exceptions Lists words that are not treated as stop words in the current Text Explorer platform. You can specify them when Text Explorer is launched via JSL. The words listed in Local Exceptions override words listed in all of the other scopes.

Import Enables you to import stop words from a text file. The stop words are copied to the clipboard. You can paste them into any of the lists other than Built-in.

Export Enables you to export stop words to the clipboard or to a text file. An Export window appears that enables you to select the scopes for which you would like to export stop words and the location of the export.

The user library files are located in a TextExplorer directory. The location of this directory is based on your computer's operating system:

- Windows: "C:/Users/<username>/AppData/Roaming/SAS/JMP/TextExplorer/<lang>/"
- macOS: "/Users/<username>/Library/Application Support/JMP/TextExplorer/<lang>/"

The main user library files are located in the TextExplorer directory itself. These files are not language-specific.

When you click OK, changes to the User, Project, and Column lists are saved to the user library, the project, and the column properties, respectively. Anything specified in the Local and Local Exceptions lists is saved only when you save the script of the Text Explorer report.

If you are saving stop words to the user library, the file is named stopwords.txt. If you are saving to a column property, the property is called "Stop Words".

Manage Recodes

The Manage Recodes window differs slightly from the Manage Stop Words window. Instead of one text edit box below each list, there are two text edit boxes. The old value (specified in the top box) is recoded to the new value (specified in the bottom box).

If you are saving recodes to the user library, the file is named recodes.txt. If you are saving to a column property, the property is called "Recodes".

Manage Phrases

If you are saving phrases to the user library, the file is named phrases.txt. If you are saving to a column property, the property is called "Phrases".

Manage Stem Exceptions

If you are saving stem exceptions to the user library, the file is named stemExceptions.txt. If you are saving to a column property, the property is called "Stem Exceptions".

Note: The Local Exceptions list in the Manage Stem Exceptions window lists stem exceptions that are excluded from the stem exception list. The words in this list are involved in the stemming operation.

Manage Negation Terms

The Manage Negation Terms window is available in the Sentiment Analysis report. See "Sentiment Analysis".

If you are saving negation terms to the user library, the file is named negations.txt. If you are saving to a column property, the property is called "Negation Terms".

Note: Terms that appear in the Local or Local Exceptions lists apply only to the current Sentiment Analysis report.

Manage Intensifier Terms

The Manage Intensifier Terms window, which is available in the Sentiment Analysis report, differs slightly from the Manage Stop Words window. In addition to the text edit box below each list, there is a Multiplier control. The Multiplier control enables you specify an intensification multiplier for a term when you add it to the set of intensifier terms. See "Sentiment Analysis".

If you are saving intensifier terms to the user library, the file is named intensifiers.txt. If you are saving to a column property, the property is called "Intensifier Terms".

Note: Terms that appear in the Local or Local Exceptions lists apply only to the current Sentiment Analysis report.

Manage Sentiment Terms

The Manage Sentiment Terms window, which is available in the Sentiment Analysis report, differs slightly from the Manage Stop Words window. In addition to the text edit box below each list, there is a Score control. The Score control enables you specify a sentiment score for a term when you add it to the set of sentiment terms. See "Sentiment Analysis".

If you are saving sentiment terms to the user library, the file is named sentiments.txt. If you are saving to a column property, the property is called "Sentiment Terms".

Note: Terms that appear in the Local or Local Exceptions lists apply only to the current Sentiment Analysis report.

Text Analysis Options

The Text Explorer red triangle menu contains the following analysis options:

Latent Class Analysis Performs a latent class analysis on the binary weighted document term matrix using sparse matrix routines. See "Latent Class Analysis".

When you select Latent Class Analysis from the Text Explorer red triangle menu, a Specifications window appears with the following options:

Maximum Number of Terms The maximum number of terms included in the latent class analysis.

Minimum Term Frequency The minimum number of occurrences a term must have to be included in the latent class analysis.

Number of Clusters The number of clusters in the latent class analysis.

Latent Semantic Analysis, SVD Performs a partial singular value decomposition of the document term matrix. See "Latent Semantic Analysis (SVD)".

Discriminant Analysis Predicts membership of each document in a group or category based on the document term matrix. See "Discriminant Analysis".

Term Selection Analyzes which terms best explain different responses. Term Selection can also be useful for sentiment analysis when the responses are ratings. See "Term Selection".

Sentiment Analysis (Available only when the Language option is set to English.) Identifies sentiment terms in documents using lexical analysis and scores documents for positive, negative, and overall sentiment. See "Sentiment Analysis".

Singular Value Decomposition Specifications Windows

In the Text Explorer platform, the analysis options are based on the Document Term Matrix (DTM). The DTM is formed by creating a column for each term in the term list (up to a specified Maximum Number of Terms). Each text document (equivalent to a row in the data table) corresponds to a row of the DTM. The values in the cells of the DTM depend on the type of weighting specified by the user in the Specifications window.

Figure 12.10 shows the Singular Value Decomposition Specifications window. When you select options from the Text Explorer red triangle menu that perform a singular value decomposition on the document term matrix, the Specifications window appears with the following options:

Maximum Number of Terms The maximum number of terms included in the singular value decomposition.

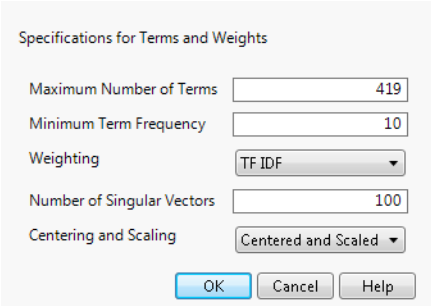
Minimum Term Frequency The minimum number of occurrences a term must have to be included in the singular value decomposition.

Weighting The weighting scheme that determines the values that go into the cells of the document term matrix. The weighting scheme options are described in "Document Term Matrix Specifications Window".

Number of Singular Vectors The number of singular vectors in the singular value decomposition. The default value is the minimum of the number of documents, the number of terms, or 100.

Centering and Scaling Options for centering and scaling of the document term matrix. You can choose between **Centered and Scaled**, **Centered**, and **Uncentered**. By default, the document term matrix is both centered and scaled.

Figure 12.10 SVD Specification Window



Specifications for Terms and Weights

Maximum Number of Terms: 419
Minimum Term Frequency: 10
Weighting: TF IDF
Number of Singular Vectors: 100
Centering and Scaling: Centered and Scaled

Save Options

The Text Explorer red triangle menu contains the following options to save information to data tables, table columns, and column properties:

Save Document Term Matrix Saves columns to the data table for each column of the document term matrix (up to a specified Maximum Number of Terms).

JMP PRO **Save Stacked DTM for Association** Saves a stacked version of the document-term matrix to a JMP data table. The stacked format is appropriate for analysis in the Association Analysis platform. See *Predictive and Specialized Modeling*. If you specify an ID variable in the Text Explorer launch window, the ID variable is used to identify the rows that each term came from in the original text data table. The stacked table also contains a table script to launch Association Analysis.

Save DTM Formula Saves a formula column with the Vector modeling type to the data table. The length of the vector depends on user-specified options for the maximum number of terms, the minimum term frequency, and the weighting. The resulting column uses the

`Text Score()` JSL function. For more information about this function, see Help > Scripting Index.

Save Term Table Creates a data table that contains each term from the term list, the number of occurrences, and the number of documents that contain each term. If you select the Score Terms by Column option after selecting Save Term Table, a column containing scores for each term is added to the data table created by the Save Term Table option.

Score Terms by Column Saves scores based on values in a specified column to the JMP data table created by the Save Term Table option. The scores for each term are the mean value of the specified column weighted by the number of occurrences of the term in each row. If you have already selected the Save Term Table option, the Score Terms by Column option adds a column containing scores to the data table created by the Save Term Table option. Otherwise, the JMP data table for the term table is created. When the specified column is not Continuous, columns containing scores for each level in the specified column are created.

Document Term Matrix Specifications Window

When you select the Save Document Term Matrix and Save DTM Formula options from the Text Explorer red triangle menu, the Document Term Matrix Specifications window appears with the following options:

Maximum Number of Terms The maximum number of terms included in the document term matrix.

Minimum Term Frequency The minimum number of occurrences a term must have to be included in the document term matrix.

Weighting The weighting scheme that determines the values that go into the cells of the document term matrix.

The following options are available for Weighting:

Binary Assigns 1 if a term occurs in each document and 0 otherwise. This is the default weighting, unless an SVD analysis has previously been run.

Ternary Assigns 2 if a term occurs more than once in each document, 1 if it occurs only once and 0 otherwise.

Frequency Assigns the count of a term’s occurrence in each document.

Log Freq Assigns $log10(1 + x)$, where x is the count of a term’s occurrence in each document.

TF IDF Assigns *TF * log10(nDoc / nDocTerm)*. Abbreviation for *term frequency - inverse document frequency*. This is the default weighting for an SVD analysis. The terms in the formula are defined as follows:

TF = frequency of the term in the document

nDoc = number of documents in the corpus

nDocTerm = number of documents that contain the term

Note: If you select Save Document Term Matrix or Save DTM Formula after you have run an SVD analysis, the Specifications window contains the specifications from the most recent SVD analysis.

Report Options in Text Explorer

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Latent Class Analysis

In the Text Explorer platform, latent class analysis enables you to group the documents from the corpus into clusters of similar documents. The Latent Class Analysis report contains the model specifications, the Bayesian Information Criterion (BIC) value for the model and a Show Text button. If one or more clusters in the Cluster Mixture Probabilities table is selected, the Show Text button opens a window that contains the text of the documents that are deemed most likely to belong to the selected cluster.

The Latent Class Analysis red triangle menu contains the following options:

Display Options Specifies the contents of the Latent Class Analysis report. By default, all of the report options are shown except for the word clouds for each cluster.

Cluster Mixture Probabilities Shows or hides a table of the probability of an observation belonging to each cluster.

Tip: You can select one or more rows in the Mixture Probabilities by Cluster table to select the observations assigned to the corresponding clusters.

Term Probabilities by Cluster Shows or hides a table of terms with an estimate for each cluster of the conditional probability that a document contains the term, given that the document belongs to a particular cluster. By default, the terms in this table are sorted by descending frequency in the corpus.

The Cluster Most Characteristic column shows the cluster that the term occurs in at the highest rate.

The Cluster Most Probable column shows the cluster in which a randomly chosen document that contains the term is most likely to be found.

Top Terms by Cluster Show or hides a table of the ten terms with the highest scores in each cluster. The score $S_{t,c}$ for term t in cluster c is calculated as follows:

$$S_{t,\,c} = 100 \bullet \mathrm{mean}(p_t) \bullet \log 10\left(\frac{p_{t,\,c}}{\mathrm{mean}(p_t)}\right)$$

where $\mathrm{mean}(p_t)$ is the mean of the term probabilities by cluster for term t and $p_{t,c}$ is the term probability by cluster for term t in cluster c.

MDS Plot Shows or hides a multidimensional scaling plot, which is a two-dimensional representation of the proximity of the clusters. For more information about MDS plots, see *Multivariate Methods*. The Show Text button opens a window that contains the text of the selected documents.

Cluster Probabilities by Row Shows or hides the Mixture Probabilities table, which displays probabilities of cluster membership for each row. The Most Likely Cluster column indicates the cluster with the highest probability of membership for each row.

Word Clouds by Cluster Shows or hides a matrix of word clouds, one for each cluster.

Rename Clusters Enables you to add descriptive names for one or more of the clusters.

Save Probabilities Saves the values in the Mixture Probabilities table to the corresponding rows in the data table.

Save Probability Formulas Saves a formula column to the data table for each cluster as well as a formula column for the most likely cluster.

The score formula that is saved uses the `Text Score()` JSL function with the weighting argument set to "`LCA`".

Color by Cluster Colors each row in the data table according to its most likely cluster.

Remove Removes the Latent Class Analysis report from the Text Explorer report.

For more information about latent class analysis, see *Multivariate Methods*.

Note: The LCA algorithm that is used in the Text Explorer platform takes advantage of the sparsity of the document term matrix. For this reason, the LCA results in the Text Explorer platform do not exactly match the results in the Latent Class Analysis platform.

Latent Semantic Analysis (SVD)

In the Text Explorer platform, latent semantic analysis is centered around computing a partial singular value decomposition (SVD) of the document term matrix (DTM). This decomposition reduces the text data into a manageable number of dimensions for analysis. Latent semantic analysis is equivalent to performing principal components analysis (PCA).

The partial singular value decomposition approximates the DTM using three matrices: $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}'$. The relationship between these matrices is defined as follows:

$$DTM \approx \mathbf{U} * \mathbf{S} * \mathbf{V}'$$

Define $nDoc$ as the number of documents (rows) in the DTM, $nTerm$ as the number of terms (columns) in the DTM, and $nVec$ as the specified number of singular vectors. Note that $nVec$ must be less than or equal to min($nDoc$, $nTerm$). It follows that $\mathbf{U}$ is an $nDoc$ by $nVec$ matrix that contains the left singular vectors of the DTM. $\mathbf{S}$ is a diagonal matrix of dimension $nVec$. The diagonal entries in $\mathbf{S}$ are the singular values of the DTM. $\mathbf{V}'$ is an $nVec$ by $nTerm$ matrix. The rows in $\mathbf{V}'$ (or columns in $\mathbf{V}$) are the right singular vectors.

The right singular vectors capture connections among different terms with similar meanings or topic areas. If three terms tend to appear in the same documents, the SVD is likely to produce a singular vector in **V′** with large values for those three terms. The **U** singular vectors represent the documents projected into this new term space.

Latent semantic analysis also captures indirect connections. If two words never appear together in the same document, but they generally appear in documents with another third word, the SVD is able to capture some of that connection. If two documents have no words in common but contain words that are connected in the dimension-reduced space, they map to similar vectors in the SVD output.

The SVD transforms text data into a fixed-dimensional vector space, making it amenable to all types of clustering, classification, and regression techniques. The Save options enable you to export this vector space to be analyzed in other JMP platforms.

The DTM, by default, is centered, scaled, and divided by *nDoc* minus 1 before the singular value decomposition is carried out. This analysis is equivalent to a PCA of the correlation matrix of the DTM.

You can also specify Centered or Uncentered in the Specifications window.

- If you specify Centered, the DTM is centered and divided by *nDoc* minus 1 before the singular value decomposition. This analysis is equivalent to a PCA of the covariance matrix of the DTM.
- If you specify Uncentered, the DTM is divided by *nDoc* before the singular value decomposition. This analysis is equivalent to a PCA of the unscaled DTM.

The SVD implementation takes advantage of the sparsity of the DTM even when the DTM is centered.

SVD Report

In the Text Explorer platform, the Latent Semantic Analysis option produces two SVD plots and a table of the singular values from the singular value decomposition.

SVD Plots

The first plot contains a point for each document. For a given document, the point that is plotted is defined by the document’s values in the first two singular vectors (the first two columns of the **U** matrix) multiplied by the diagonal singular values matrix (**S**). This plot is equivalent to the Score Plot in the Principal Components platform. Each point in this plot represents a document (row of the data table). You can select the points in this plot to select the corresponding rows in the data table.

The second plot contains a point for each term. For a given term, the point that is plotted is defined by the term's values in the first two singular vectors (the first two rows of the **V'** matrix) multiplied by the diagonal singular values matrix (**S**). This plot is equivalent to the Loadings Plot in the Principal Components platform. In this plot, the points correspond to rows in the Term List table.

Above each of the SVD Plots, you can click a Show Text button to open a window that contains the text of the selected points in the plot.

Singular Values

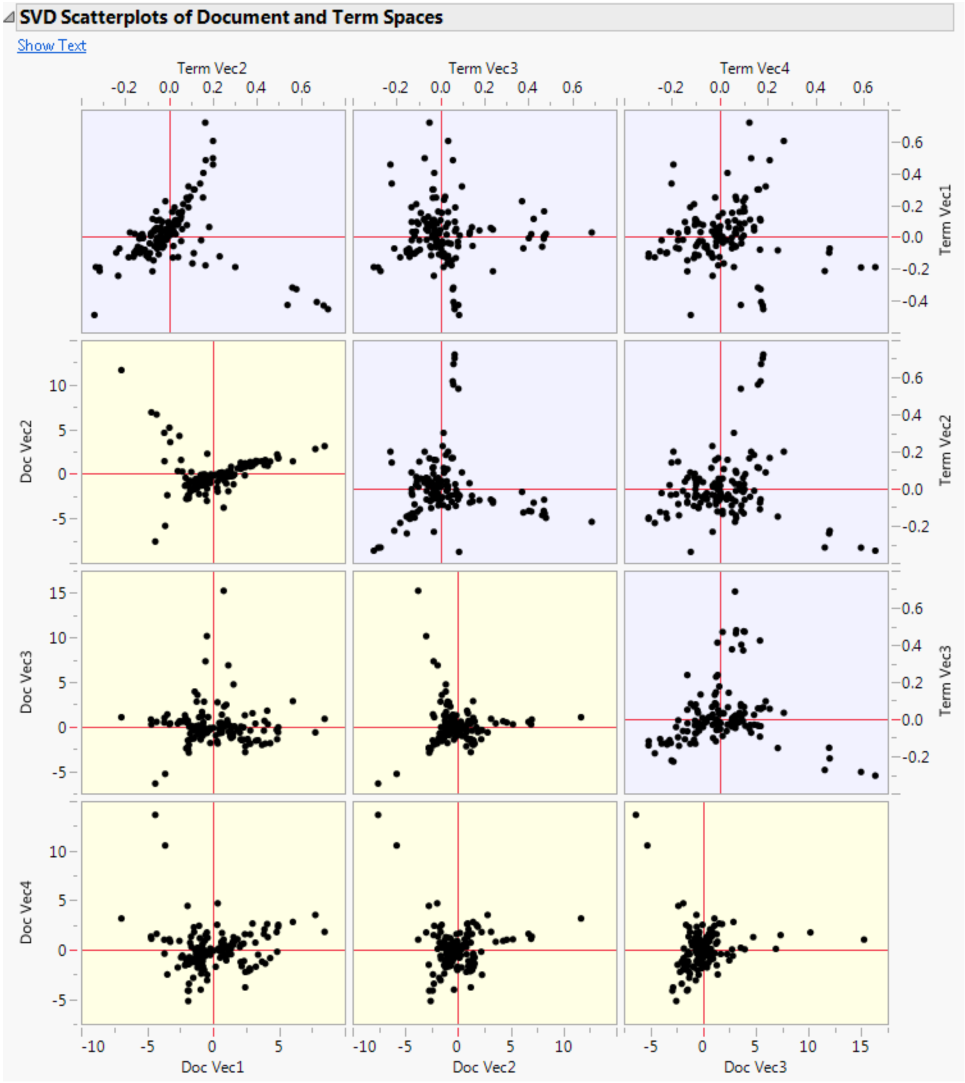
Below the document and term SVD plots, a table of the singular values appears. These are the diagonal entries of the **S** matrix in the singular value decomposition of the document term matrix. The Singular Values table also contains a column of corresponding eigenvalues for the equivalent principal components analysis. Like in the Principal Components platform, there are columns for the percent and cumulative percent of variation explained by each eigenvalue (or singular value). You can use the Cum Percent column to decide what percent of variance from the DTM you want to preserve, and then use the corresponding number of singular vectors.

SVD Report Options

In the Text Explorer platform, the SVD red triangle menu contains the following options:

SVD Scatterplot Matrix Shows or hides a scatterplot matrix of the term and document singular value decomposition vectors. You are prompted to select the size of the scatterplot matrix when you select this option. This scatterplot matrix enables you to visualize more than the first two dimensions of the singular value decomposition. The Show Text button opens a window that contains the text of the selected documents.

Figure 12.11 SVD Scatterplots of Document and Term Spaces



Topic Analysis, Rotated SVD Performs a varimax rotated partial singular value decomposition of the document term matrix to produce groups of terms called topics. You can select this option multiple times to find different numbers of topics. See "Topic Analysis".

Cluster Terms Shows or hides a hierarchical clustering analysis of the terms in the data. To the right of the dendrogram, there are options to set the number of clusters and save the clusters to a data table. For each term, this data table contains its frequency, the number of documents that contain it, and its assigned cluster. For more information about hierarchical clustering and dendrograms, see *Multivariate Methods*.

Cluster Documents Shows or hides a hierarchical clustering analysis of the documents in the data. To the right of the dendrogram, there are options to do the following: set the number of clusters, save the clusters to a column in the data table, and show the documents in a selected branch of the dendrogram plot.

Select Near Neighbors Finds and selects the nearest neighbors of the selected points in the document SVD plot. The algorithm uses the **U** matrix from the singular value decomposition to find the nearest neighbors. When you select this option, you must specify the number of nearest neighbors to select. By default, the option selects the 10 nearest neighbors.

Save Document Singular Vectors Saves a user-specified number of singular vectors from the document singular value decomposition as columns to the data table. The first two saved columns represent the points plotted in the document SVD plot. See “Latent Semantic Analysis (SVD)”.

Save Singular Vector Formula Saves a formula column with the Vector modeling type that contains the document singular value decomposition to the data table. The resulting column uses the `Text Score()` JSL function. For more information about this function, see Help > Scripting Index.

Save Term Singular Vectors Saves a user-specified number of singular vectors from the terms singular value decomposition as columns to a new data table where each row corresponds to a term. If a Term Table data table is already open, this option saves the columns to that data table. The first two saved columns represent the points plotted in the term SVD plot. See “Latent Semantic Analysis (SVD)”.

Remove Removes the SVD report from the Text Explorer report window.

Topic Analysis

In the Text Explorer platform, the Topic Analysis, Rotated SVD option performs a varimax rotation on the partial singular value decomposition (SVD) of the document term matrix (DTM). You must specify a number of rotated singular vectors, which corresponds to the number of topics that you want to retain from the DTM. After you specify a number of topics, the Topic Analysis report appears.

Topic analysis is equivalent to a rotated principal component analysis (PCA). The varimax rotation takes a set of singular vectors and rotates them to make them point more directly in the coordinate directions (toward the terms). This rotation makes the vectors help explain the text as each rotated vector orients toward a set of terms. Negative values indicate a repulsion force. The terms with negative values occur in a topic less frequently compared to the terms with positive values.

Topic Analysis Report

In the Text Explorer platform, the Topic Analysis report shows the terms that have the largest loadings in each topic after rotation. There are additional reports that show the components of the rotated singular value decomposition.

The Top Loadings by Topic report shows a table of terms for each topic. The terms in each table are the ones that have the largest loadings in absolute value for each topic. Each table is sorted in descending order by the absolute value of the loading. These tables can be used to determine conceptual themes that correspond to each topic.

The Topic Analysis report also contains the following reports:

Topic Loadings Contains a matrix of the loadings across topics for each term. This matrix is equivalent to the factor loading matrix in a rotated PCA.

Word Clouds by Topic Contains a matrix of word clouds, one for each topic.

Topic Scores Contains a matrix of document scores for each topic. Documents with higher scores in a topic are more likely to be associated with that topic.

Topic Scores Plots Contains a Show Text button and a plot of topic scores for each document. The Show Text button opens a window that contains the text of the selected documents.

The Topic Scores Plots report is a visual representation of the matrix in the Topic Scores report. Each panel in the plot corresponds to one of the topics, or one of the columns of the Topic Scores matrix. Within each panel, each point corresponds to one of the documents in the corpus, or one of the rows of the Topic Scores matrix.

Variance Explained by Each Topic Contains a table of the variance explained by each topic. The table also contains columns for the percent and cumulative percent of the variation explained by each topic.

Rotation Matrix Contains the rotation matrix for the varimax rotation.

Topic Analysis Report Options

In the Text Explorer platform, the Topic Analysis red triangle menu contains the following options:

Topic Scatterplot Matrix Shows or hides a scatterplot matrix of the rotated singular value decomposition vectors. The Show Text button opens a window that contains the text of the selected documents.

Display Options Contains options to show or hide content that appears in the Topic Analysis report. See "Topic Analysis Report".

Rename Topics Enables you to add descriptive names for one or more of the topics.

Save Document Topic Vectors Saves a user-specified number of singular vectors from the rotated singular value decomposition as columns to the data table.

Save Topic Vector Formula Saves a formula column with the Vector modeling type that contains the rotated singular value decomposition to the data table. The resulting column uses the `Text Score()` JSL function. For more information about this function, see Help > Scripting Index.

Save Term Topic Vectors Saves the topic vectors as columns to the data table created by the Save Term Table option.

Remove Removes the Topic Analysis report from the SVD report.

Discriminant Analysis

In the Text Explorer platform, discriminant analysis predicts membership of each document in a group or category based on the columns in the document term matrix (DTM). Specifically, discriminant analysis predicts a classification of each document into a category of a response column. When you select the Discriminant Analysis option, you must select a response column that contains categories or groups. Group membership is predicted by the columns of the DTM. For more information about discriminant analysis, see *Multivariate Methods*.

The discriminant analysis method in the Text Explorer platform is based on a singular value decomposition of the centered DTM. Each group of the response column has its own group mean that is used to center the DTM. The discriminant analysis method in the Text Explorer platform is faster than the Discriminant Analysis platform because it takes advantage of the sparsity of the DTM.

Discriminant Analysis Specifications Window

In the Text Explorer platform, the Discriminant Analysis option is based on the Document Term Matrix (DTM). The DTM is formed by creating a column for each term in the term list (up to a specified Maximum Number of Terms). Each text document (equivalent to a row in the data table) corresponds to a row of the DTM. The values in the cells of the DTM depend on the type of weighting specified by the user in the Specifications window.

When you select the Discriminant Analysis option from the Text Explorer red triangle menu, the Specifications window appears with the following options:

Maximum Number of Terms The maximum number of terms included in the discriminant analysis.

Minimum Term Frequency The minimum number of occurrences a term must have to be included in the discriminant analysis.

Weighting The weighting scheme that determines the values that go into the cells of the document term matrix. The weighting scheme options are described in "Document Term Matrix Specifications Window".

Number of Singular Vectors The number of singular vectors in the discriminant analysis. The default value is the minimum of the number of documents, the number of terms, or 100.

Discriminant Analysis Report

By default, the Discriminant Analysis report in the Text Explorer platform contains two open reports: the Classification Summary and the Discriminant Scores. The other reports are initially closed.

The Discriminant Analysis report contains the following reports:

Term Means Provides a table of the terms used in the discriminant analysis. The terms correspond to the columns of the DTM. The table contains the means in each group for each term, as well as the overall mean and weighted standard deviation for each term.

Squared Distances to Each Group Provides a table that contains the squared Mahalanobis distances to each group for each document. For more information about Mahalanobis distances, see *Multivariate Methods*.

Probabilities to Each Group Provides a table that contains the probability that a document belongs to each group.

Classification Summary Provides a report that summarizes the discriminant scores. This report corresponds to the Score Summaries report in the Discriminant Analysis platform report.

Discriminant Scores Provides a table of the predicted classification of each document and other supporting information. This table corresponds to the Discriminant Scores table in the Discriminant Analysis platform report.

Discriminant Analysis Report Options

In the Text Explorer platform, the Discriminant Analysis red triangle menu contains the following options:

Canonical Plot Shows or hides a plot of the documents and group means in canonical space. Canonical space is the space that most separates the groups. If there are more than two levels of the response variable, you must specify the number of canonical coordinates. If you specify more than two canonical coordinates, this option produces a matrix of canonical plots.

Save Probabilities Saves a probability column to the data table for each response level as well as a column that contains the most likely response. The Most Likely response column contains the level with the highest probability based on the model.

Each probability column gives the posterior probability of an observation's membership in that level of the response. The Response Probability column property is saved to each probability column. For more information about the Response Probability column property, see *Using JMP*.

Save Probability Formulas Saves formula columns to the data table for the prediction of the most likely response. The first saved column contains a formula that uses the `Text Score()` function to calculate the probability for each response level. There are also columns that contain probabilities for each response level as well as a column that contains the predicted response.

Save Canonical Scores Saves columns to the data table that contain the scores from canonical space for each observation. Canonical space is the space that most separates the groups. The column for the k^{th} canonical score is named Canonical<k>.

Remove Removes the Discriminant Analysis report from the Text Explorer report window.

Term Selection

In the Text Explorer platform, term selection identifies which terms best explain different responses. The analysis uses the Generalized Regression platform to perform variable selection on the document term matrix (DTM) and to identify terms that most impact the response. Term selection can be used with binary responses, similar to sentiment analysis, as well as other types of responses. The fitted model uses an appropriate response distribution for the specified response column.

Tip: For an example of Term Selection, select **Help > Sample Data Folder**, open Chips.jmp, and run the Text Explorer - Term Selection table script.

Term Selection Settings

The Settings section of the Term Selection report enables you to select a response column, specify the target level of the response, and adjust the settings for the model. When you have specified the model settings, click the Run button to run the model. The fitted model then appears in the Summary report. See "Term Selection Summary Report".

Target Level

After you choose a response column, the Target Level outline appears.

- For nominal responses, choose one level of the response to be the target level in a logistic regression model; the response in the logistic regression model is the target level versus all of the other levels combined.
- For ordinal responses, all response levels are initially included in the model. Using the local data filter, you can select levels of the response to be excluded from the model; the underlying numeric values of the included levels are modeled with a normal response distribution.

  Note: For ordinal responses, the term selection model can be fit only when the data type of the response column is numeric.

- For continuous responses, use the local data filter histogram to select values of the response to be excluded from the model; the included values are modeled with a normal response distribution.
- For response columns with the Multiple Response modeling type, choose one or more levels of the response to be the target level in a binary logistic regression model. If you choose more than one level, a document belongs to the target level if any of the levels are present in the response column for that document. Select the **Combine with AND** option to require that all selected levels are present in a document's response column for that document to be included in the target level.

Model Settings

By default, the Generalized Regression model uses the Elastic Net estimation method with early stopping and the AICc validation method. You can change these settings in the Model Settings outline. See *Fitting Linear Models*.

Note: If a Validation column is specified in the Text Explorer launch window, the Generalized Regression platform in the Term Selection report uses the Validation column as the Validation method.

Term Settings

The Term Settings define the document term matrix (DTM) that is used in the regression model. You can change the weighting technique as well as the maximum number of terms included in the DTM; each term corresponds to a column of the DTM. Note that terms that have fewer than 10 occurrences in the corpus are not included in the DTM used by the model. For more information about the DTM options, see "Document Term Matrix Specifications Window".

Term Selection Report

After you run an analysis in the Text Explorer platform, the Term Selection report consists of three sections. The Settings report contains controls for specifying an analysis. See "Term Selection Settings". Below the Settings report, there are initially closed Generalized Regression reports for each analysis that you have run. See *Fitting Linear Models*. The last section of the report is the Summary report.

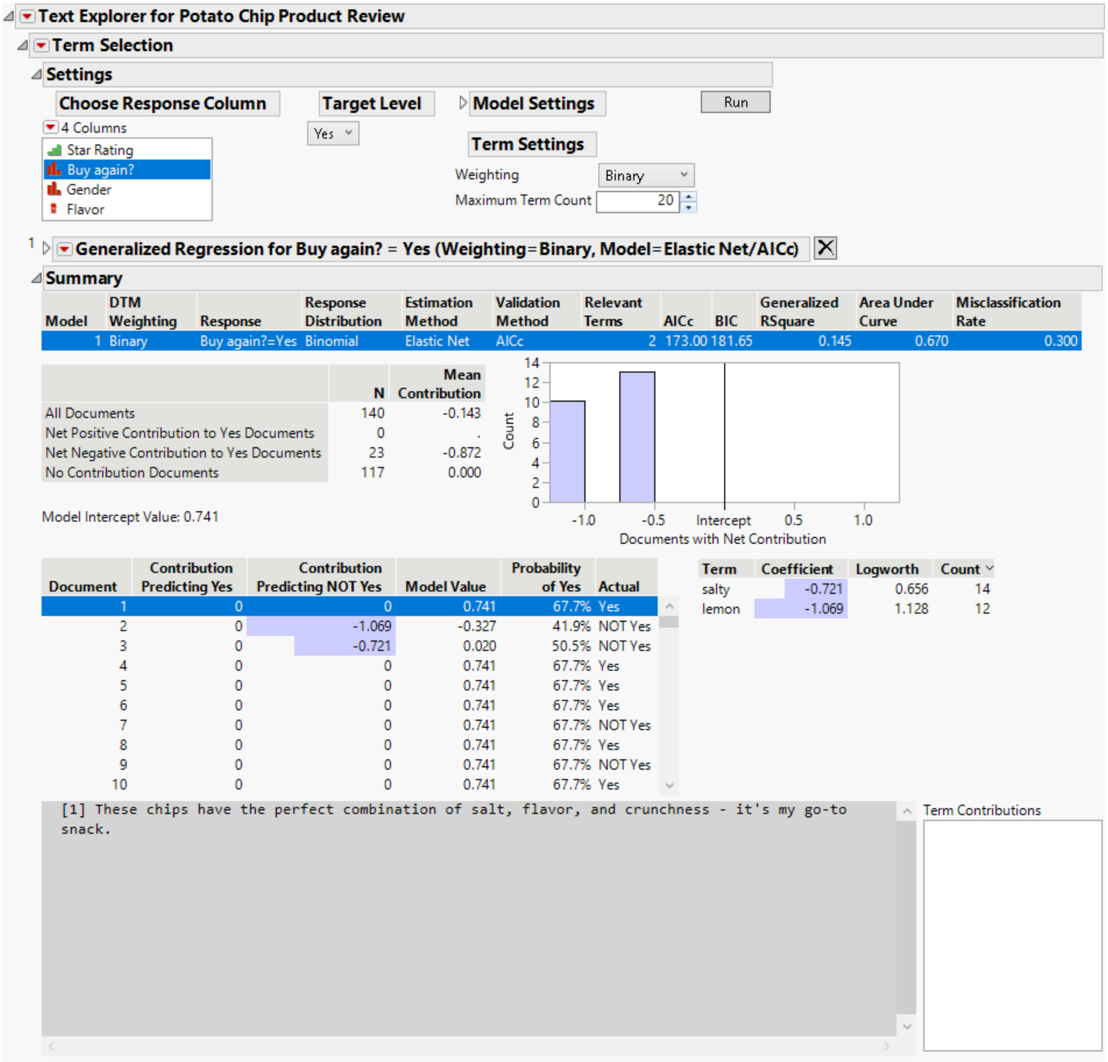


Figure 12.12 Term Selection Report

4 Columns

- Star Rating
- Buy again?
- Gender
- Flavor

Target Level: Yes

Weighting: Binary

Maximum Term Count: 20

Model	DTM Weighting	Response	Response Distribution	Estimation Method	Validation Method	Relevant Terms	AICc	BIC	Generalized RSquare	Area Under Curve	Misclassification Rate
1	Binary	Buy again?=Yes	Binomial	Elastic Net	AICc	2	173.00	181.65	0.145	0.670	0.300

	N	Mean Contribution
All Documents	140	-0.143
Net Positive Contribution to Yes Documents	0	.
Net Negative Contribution to Yes Documents	23	-0.872
No Contribution Documents	117	0.000

Model Intercept Value: 0.741

Document	Contribution Predicting Yes	Contribution Predicting NOT Yes	Model Value	Probability of Yes	Actual
1	0	0	0.741	67.7%	Yes
2	0	-1.069	-0.327	41.9%	NOT Yes
3	0	-0.721	0.020	50.5%	NOT Yes
4	0	0	0.741	67.7%	Yes
5	0	0	0.741	67.7%	Yes
6	0	0	0.741	67.7%	Yes
7	0	0	0.741	67.7%	NOT Yes
8	0	0	0.741	67.7%	Yes
9	0	0	0.741	67.7%	NOT Yes
10	0	0	0.741	67.7%	Yes

Term	Coefficient	Logworth	Count
salty	-0.721	0.656	14
lemon	-1.069	1.128	12

```
[1] These chips have the perfect combination of salt, flavor, and crunchness - it's my go-to
snack.
```

Term Selection Summary Report

The Summary report contains a Model Comparison table, a Summary table and histogram, a Document Scores table, a Term Scores table, and a text box.

The Model Comparison table contains a row for each fitted model. The rest of the Summary report shows results from the currently selected model in this table.

The Summary table shows counts and mean scores for the documents, overall and by the predicted value of the response from the model. The Mean Contribution is the average of the contribution values in the Document Scores table. The Summary histogram shows the distribution of the overall contribution values of the documents. The histogram is interactive, so you can click on a bar to highlight the corresponding documents in the Document Scores table.

The Document Scores table shows the positive and negative contribution values for each document, as well as predicted and actual values for each document. For binomial response models, the predicted values are probabilities of the document being in the target level; for normal response models, the predicted values are the predictions from the fitted model for each document. If you select a row of the table, the text of the corresponding document appears in the text box below the table.

The Term Scores table lists each term that was selected by the fitted model, its coefficient from the model, its Logworth, and the count of occurrences of the term in the corpus. If you select a row of the table, the text of the corresponding document appears in the text box below the table.

The text box shows the text of documents that are selected in the Document Scores table or the context of terms that are selected in the Term Scores table.

Term Selection Report Options

In the Text Explorer platform, the Term Selection red triangle menu contains the following options:

Save Document Scores (Available only when an analysis is selected in the Summary table.) Saves the columns from the Document Scores table to new columns in the data table. The new columns contain the positive and negative contributions, as well as the predicted value for each document.

Save Term Score DTM (Available only when an analysis is selected in the Summary table.) Saves columns to the data table for each relevant term in the currently selected analysis. The columns contain the term scores for each document, using the Weighting specified in the Term Selection Term Settings.

Save Prediction Formulas (Available only when an analysis is selected in the Summary table.) Saves columns to the data table that contain the prediction formulas for the currently selected analysis.

Show Term Cloud Shows or hides a word cloud in the Summary report. The word cloud shows the coefficient terms in the currently selected analysis. The words are sized by the absolute value of their coefficients and colored by the sign of their coefficients.

Remove Removes the Term Selection report from the Text Explorer report window.

Sentiment Analysis

In the Text Explorer platform, sentiment analysis identifies sentiment terms in documents using lexical analysis and scores documents for positive, negative, and overall sentiment. The analysis assumes that each document is free text with binary sentiment on a single topic. Sentiment Analysis incorporates basic natural language processing (NLP) into the results. For more information about natural language processing, see https://opennlp.apache.org/. If you prefer to not use NLP, deselect the Parse Documents option.

Tip: For an example of Sentiment Analysis, select **Help > Sample Data Folder**, open Chips.jmp, and run the Text Explorer - Sentiment Analysis table script.

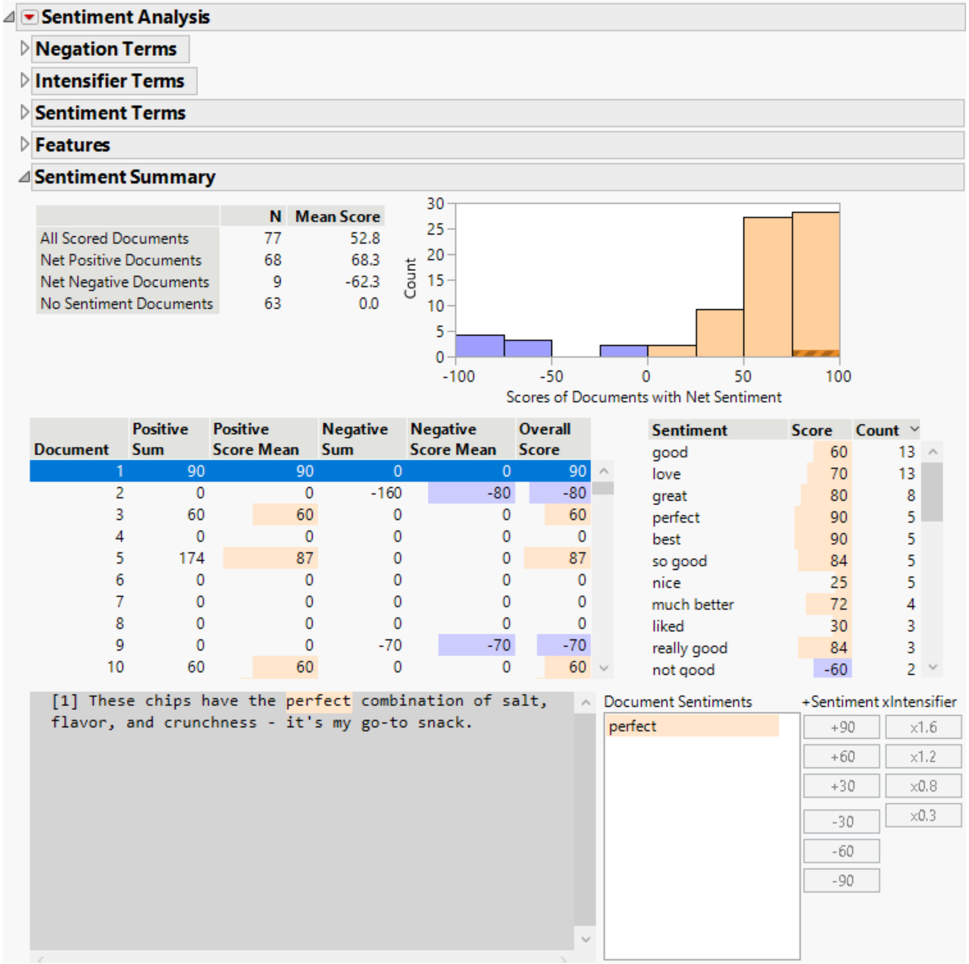
Notes:

- In Sentiment Analysis, a word can take effect in only one class of terms: negation, intensifier, or sentiment.
- Sentiment Analysis recognizes some emoticons, or sequences of characters that are treated as a single unit. You can see the built-in emoticons and their default sentiment scores in the Sentiment Terms report or in the Manage Sentiment Terms window.
- If you specify a word as a negation, intensifier, or sentiment term and it is already specified as a stop word, the word is temporarily removed as a stop word as long as the Sentiment Analysis report is open. This temporary removal affects the entire Text Explorer report. It is restored as a stop word when the Sentiment Analysis report closes.

Sentiment Analysis Report

By default, the Sentiment Analysis report in the Text Explorer platform contains one open report: the Sentiment Summary. The other reports are initially closed.

Figure 12.13 Sentiment Analysis Report



	N	Mean Score
All Scored Documents	77	52.8
Net Positive Documents	68	68.3
Net Negative Documents	9	-62.3
No Sentiment Documents	63	0.0

Document	Positive Sum	Positive Score Mean	Negative Sum	Negative Score Mean	Overall Score
1	90	90	0	0	90
2	0	0	-160	-80	-80
3	60	60	0	0	60
4	0	0	0	0	0
5	174	87	0	0	87
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	-70	-70	-70
10	60	60	0	0	60

Sentiment	Score	Count
good	60	13
love	70	13
great	80	8
perfect	90	5
best	90	5
so good	84	5
nice	25	5
much better	72	4
liked	30	3
really good	84	3
not good	-60	2

```
[1] These chips have the perfect combination of salt,
flavor, and crunchness - it's my go-to snack.
```

Document Sentiments: perfect

The Sentiment Analysis report contains the following reports:

Negation Terms

Contains a list of the negation terms in the current sentiment analysis. Right-click the list to see a menu of additional options. You can select terms in the list to remove them.

Intensifier Terms

Contains a list of the intensifier terms and their corresponding multiplier values. Right-click the list to see a menu of additional options. You can select terms in the list to remove them.

Sentiment Terms

Contains a list of the sentiment terms and their corresponding score values. This report also enables you to add new sentiment terms. The Possible Sentiment table contains counts of terms that you might consider adding as sentiment terms. To add a term as a sentiment term, select it in the Possible Sentiment table and click one of the buttons below +Sentiment. To choose a sentiment score value that is not listed, you can edit the score value after you have added it to the list of sentiment terms.

When you select a term in the Possible Sentiment table, documents that contain that term appear on the right side of the Sentiment Terms report. This provides context for how the term is used in the corpus.

Features

Contains options to score features in the corpus. A feature is something that is being described by the sentiment terms. Click the **Search** button to generate a list of possible feature terms. When you select one or more terms from the Possible Feature table, excerpts of the documents that contain those terms appear in the text box to the right of the table. Click the **Score Selected Features** button to update the Sentiment Summary report to show the results of scoring the selected feature terms.

Note: If you select the Parse Documents option, the Features report scores words when they occur within the same umbrella clause as a sentiment.

Sentiment Summary

Contains the results of the sentiment analysis based on the current settings. This report contains a Summary table and histogram, a Document Scores table, a Sentiment Terms table, a text box, and a control panel that enables you to add more sentiment and intensifier terms.

The Summary table shows counts and mean scores for the documents, broken down by how the documents were scored. The Mean Score is determined by the setting of the Scoring option. See “Sentiment Analysis Report Options”. The Summary histogram shows the distribution of the overall sentiment scores of the documents. The histogram is interactive, so you can click on a bar to highlight the corresponding documents in the Document Scores table.

The Document Scores table shows the positive and negative sentiment score sums and means, as well as the overall sentiment score for each document. If you select a row of the table, the text of the corresponding document appears in the text box below the table. If you specify a Score Column, the table contains the values from the scoring column.

Tip: You can hover over the cells in the Document Scores table to see the scoring calculations that were used to produce the table results.

The Sentiment Terms table lists each sentiment term, its score value, and the count of occurrences of the term in the corpus.

Tip: For sentiment terms with multiple words, you can hover over the cells in the Score column to see the calculations that were used to produce the score.

The text box shows the text of documents that are selected in the Document Scores table or the context of terms that are selected in the Sentiment Terms table. When you select a document in the Document Scores table, a list of the sentiments in that document appears to the right of the text box.

Tip: When you hover over a term in the text box that is classified as a negation, intensifier, or sentiment term, a box appears that shows the classification and contains a Remove button. Click the **Remove** button to quickly remove that term from the list of negation, intensifier, or sentiment terms.

The control panel activates when you select a term in the text box. To add a term as a sentiment term, select it in the text box and click one of the buttons below +Sentiment. To add a term as an intensifier term, select it in the text box and click one of the buttons below ×Intensifier.

Sentiment Analysis Report Options

In the Text Explorer platform, the Sentiment Analysis red triangle menu contains the following options:

Scoring Contains the following options for calculating the Overall Scores for documents:

- **Scaled** Scores of positive and negative phrases are summed. The sum is then divided by the number of phrases in the document to determine the Overall Score.
- **Min Max** The Overall Score is calculated as the sum of the maximum positive score and the minimum negative score.

Score Column Specifies a data table column that contains known information that can be compared with the calculated sentiment. The score column is added to the Document Scores table.

Tip: You can visually compare the Overall Score column and the score column to assess the sentiment scoring.

Parse Documents Specifies if natural language processing (NLP) is used to parse the documents. For more information about natural language processing, see https://opennlp.apache.org/.

Save Document Scores Saves the columns from the Document Scores table to new columns in the data table. The new columns contain the positive and negative sentiment score sums and means, as well as the overall sentiment score for each document.

Save Count of Sentiment Scores by Document Saves a column to the data table for each sentiment term. Each column contains counts of the occurrences of each sentiment term in each document.

Show Negation Terms Shows or hides the Negation Terms report.

Show Intensifier Terms Shows or hides the Intensifier Terms report.

Show Sentiment Terms Shows or hides the Sentiment Terms report.

Show Feature Finder Shows or hides the Features report.

Show Sentiment Cloud Shows or hides a word cloud of the sentiment terms in the Sentiment Summary report.

Include Builtin Negation Terms Specifies if the negation terms used in the sentiment analysis include built-in negation terms or not.

Include Builtin Intensifier Terms Specifies if the intensifier terms used in the sentiment analysis include built-in intensifier terms or not.

Include Builtin Sentiment Terms Specifies if the sentiment terms used in the sentiment analysis include built-in sentiment terms or not.

Manage Negation Terms Shows a window that enables you to add or remove negation terms. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude negation terms that are specified in any of the other levels. See "Term Options Management Windows".

Manage Intensifier Terms Shows a window that enables you to add or remove intensifier terms. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude intensifier terms that are specified in any of the other levels. See "Term Options Management Windows".

Manage Sentiment Terms Shows a window that enables you to add or remove sentiment terms. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude sentiment terms that are specified in any of the other levels. See “Term Options Management Windows”.

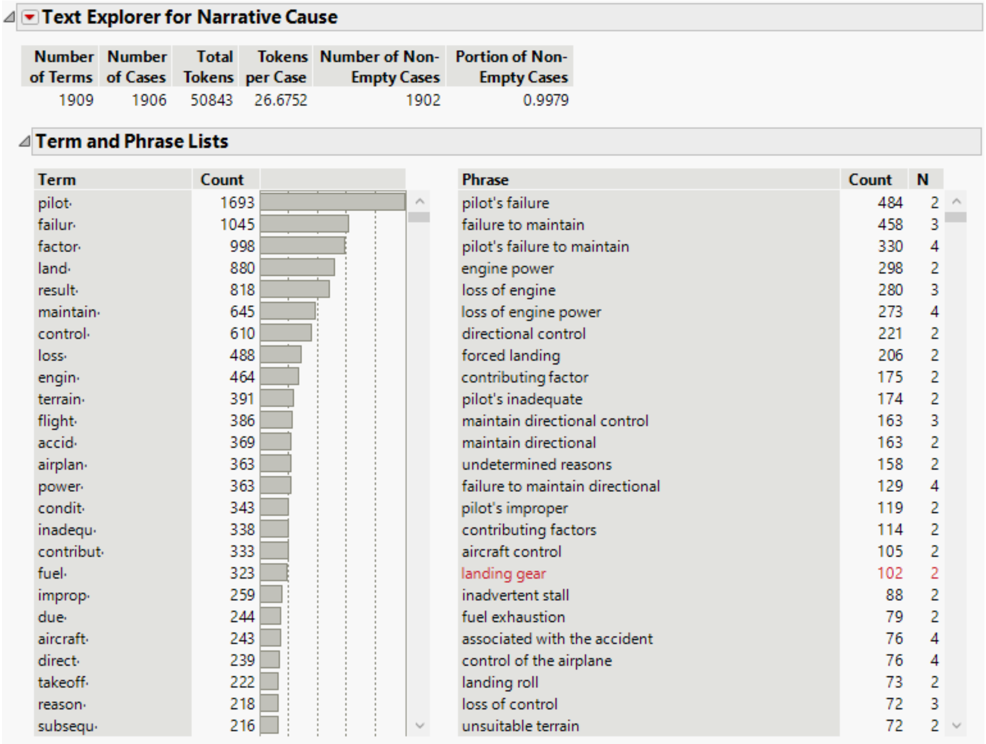
Additional Example of the Text Explorer Platform

This example examines aircraft incident reports from the National Transportation Safety Board for events occurring in 2001 in the United States. You want to explore the text that contains a description of the results of the investigation into the cause of each incident. You also want to find themes in the collection of incident reports.

1. Select **Help > Sample Data Folder** and open Aircraft Incidents.jmp.
2. Select **Rows > Color or Mark by Column**.
3. Select Fatal from the columns list and click **OK**.

   The rows that contain accidents involving fatalities are colored red.
4. Select **Analyze > Text Explorer**.
5. Select Narrative Cause from the Select Columns list and click **Text Columns**.
6. From the Language list, select **English**.
7. From the Stemming list, select **Stem All Terms**.
8. From the Tokenizing list, select **Basic Words**.
9. Click **OK**.

Figure 12.14 Text Explorer Report for Narrative Cause



Text Explorer for Narrative Cause

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
1909	1906	50843	26.6752	1902	0.9979

Term and Phrase Lists

Term	Count
pilot·	1693
failur·	1045
factor·	998
land·	880
result·	818
maintain·	645
control·	610
loss·	488
engin·	464
terrain·	391
flight·	386
accid·	369
airplan·	363
power·	363
condit·	343
inadequ·	338
contribut·	333
fuel·	323
improp·	259
due·	244
aircraft·	243
direct·	239
takeoff·	222
reason·	218
subsequ·	216

Phrase	Count	N
pilot's failure	484	2
failure to maintain	458	3
pilot's failure to maintain	330	4
engine power	298	2
loss of engine	280	3
loss of engine power	273	4
directional control	221	2
forced landing	206	2
contributing factor	175	2
pilot's inadequate	174	2
maintain directional control	163	3
maintain directional	163	2
undetermined reasons	158	2
failure to maintain directional	129	4
pilot's improper	119	2
contributing factors	114	2
aircraft control	105	2
landing gear	102	2
inadvertent stall	88	2
fuel exhaustion	79	2
associated with the accident	76	4
control of the airplane	76	4
landing roll	73	2
loss of control	72	3
unsuitable terrain	72	2

From the report, you see that there are almost 51,000 tokens and about 1,900 unique terms.

10. Right-click pilot· in the term list and select **Select Rows**.

    From the number of selected rows in the data table, you see that some form of the word "pilot" occurs in more than 1,300 of the incident reports.

11. Right-click pilot· and select **Add Stop Word**.

    Because some form of the word "pilot" occurs frequently compared to other terms, these terms do not provide much information to differentiate among documents. All of the terms that stem to pilot· are added to the stop word list.

The remaining steps of this example can be completed only in JMP Pro.

12. Click the red triangle next to Text Explorer for Narrative Cause and select **Latent Semantic Analysis, SVD**.

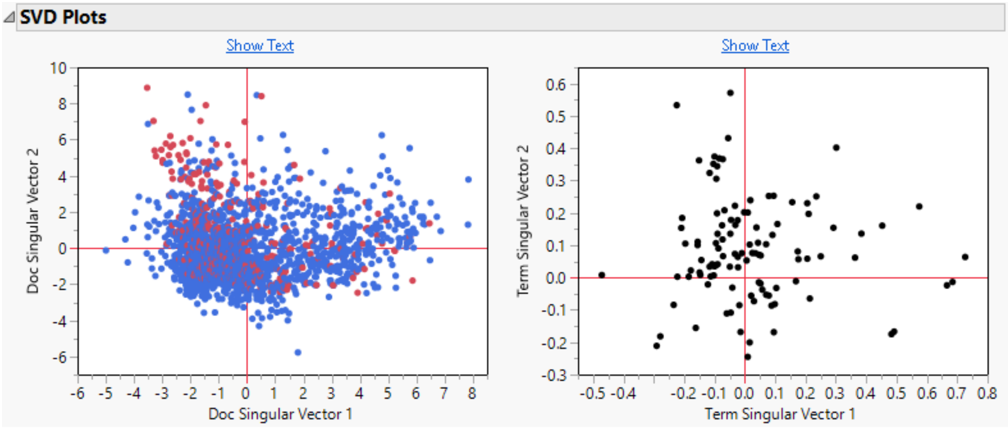
    This is the first analysis step toward topic analysis, which performs a rotation of the SVD.

13. In the Specifications window, type 50 for Minimum Term Frequency.

    Because there are approximately 51,000 tokens, this frequency is equivalent to a term that represents at least 0.1% of all the terms.

14. Click **OK**.

Figure 12.15 SVD Plots for Narrative Cause



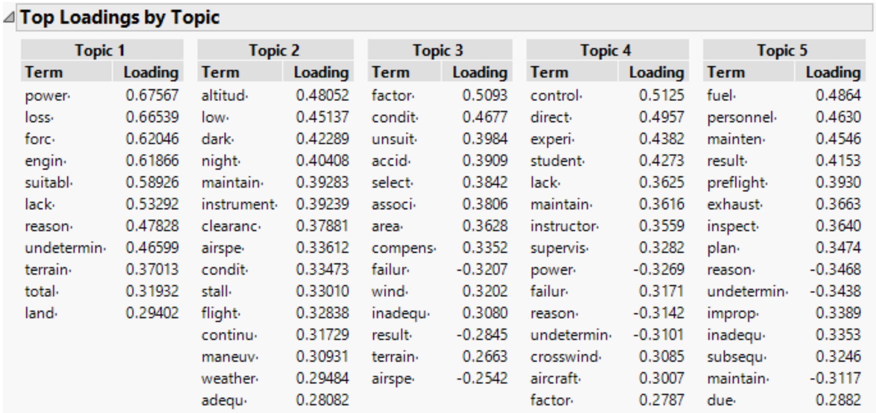
There is not a lot of difference in the document SVD plot between fatal and non-fatal incidents.

15. JMP PRO Click the red triangle next to SVD Centered and Scaled TF IDF and select **Topic Analysis, Rotated SVD**.

    You want to look for groups of terms that form topics.

16. JMP PRO Type 5 for Number of Topics.
17. JMP PRO Click **OK**.

Figure 12.16 Top Loadings by Topics for Narrative Cause



Top Loadings by Topic

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Term	**Loading**	**Term**	**Loading**	**Term**	**Loading**	**Term**	**Loading**	**Term**	**Loading**
power·	0.67567	altitud·	0.48052	factor·	0.5093	control·	0.5125	fuel·	0.4864
loss·	0.66539	low·	0.45137	condit·	0.4677	direct·	0.4957	personnel·	0.4630
forc·	0.62046	dark·	0.42289	unsuit·	0.3984	experi·	0.4382	mainten·	0.4546
engin·	0.61866	night·	0.40408	accid·	0.3909	student·	0.4273	result·	0.4153
suitabl·	0.58926	maintain·	0.39283	select·	0.3842	lack·	0.3625	preflight·	0.3930
lack·	0.53292	instrument·	0.39239	associ·	0.3806	maintain·	0.3616	exhaust·	0.3663
reason·	0.47828	clearanc·	0.37881	area·	0.3628	instructor·	0.3559	inspect·	0.3640
undetermin·	0.46599	airspe·	0.33612	compens·	0.3352	supervis·	0.3282	plan·	0.3474
terrain·	0.37013	condit·	0.33473	failur·	-0.3207	power·	-0.3269	reason·	-0.3468
total·	0.31932	stall·	0.33010	wind·	0.3202	failur·	0.3171	undetermin·	-0.3438
land·	0.29402	flight·	0.32838	inadequ·	0.3080	reason·	-0.3142	improp·	0.3389
		continu·	0.31729	result·	-0.2845	undetermin·	-0.3101	inadequ·	0.3353
		maneuv·	0.30931	terrain·	0.2663	crosswind·	0.3085	subsequ·	0.3246
		weather·	0.29484	airspe·	-0.2542	aircraft·	0.3007	maintain·	-0.3117
		adequ·	0.28082			factor·	0.2787	due·	0.2882

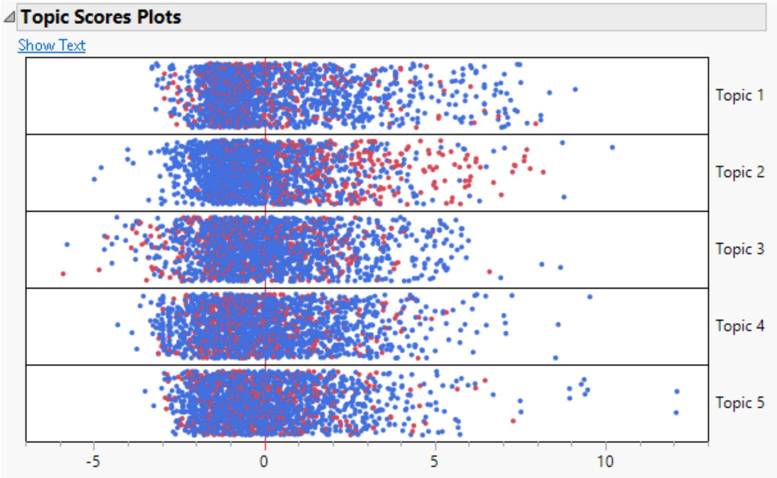
The terms for each topic with the highest loadings enable you to interpret whether the topic is capturing a theme in the incident reports.

For example, Topic 1 has high loadings for power, loss, and engine, indicating a theme of losing power to the engine as a cause of the incident. This corresponds to the phrase "loss of engine power" occurring 273 times in the set of incident reports.

Based on the words with high loadings in Topic 2, it can be described as being related to incidents that involved darkness or low altitude.

18. Click the gray disclosure icon next to Topic Scores Plots.

Figure 12.17 Topic Scores Plots for Narrative Cause



Each topic score plot contains a point for each document in the corpus. You can select points in these plots to further examine the text of specific documents.

You want to further explore the subject matter of Topic 2.

19. Select the three right-most points in the Topic 2 plot and click the **Show Text** button at the top left of the graph.

The text of the three documents with the highest scores in Topic 2 appear in a new window. From these, you can confirm that Topic 2 relates to low altitude.

At this stage of the text analysis, you have many choices in how to proceed. Text analysis is an iterative process, so you might use topic information to further curate your term list by adding stop words or specifying phrases. You might save the weighted document-term matrix, the vectors from the SVD or rotated SVD as numeric columns in your data table and use them in other JMP analysis platforms. When you use these columns in other platforms, you can also include other columns from your data table in further analyses.

Appendix A

References

The following sources are referenced in *Basic Analysis*.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Agresti, A., and Coull, B. A. (1998). “Approximate is Better Than ‘Exact’ for Interval Estimation of Binomial Proportions.” *American Statistician* 52:119–126.

Asiribo, O., and Gurland, J. (1990). “Coping with Variance Heterogeneity.” *Communication in Statistics: Theory and Methods* 19:4029–4048.

Bablok, W., Passing, H., Bender, R., and Schneider, B. (1988). “A General Regression Procedure for Method Transformation.” *Journal of Clinical Chemistry Clinical Biochemistry* 26:783–90.

Bartlett, M. S., and Kendall, D. G. (1946). “The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation.” *Supplement to the Journal of the Royal Statistical Society* 8:128–138.

Bissell, A. F. (1990). “How Reliable is Your Capability Index?” *Applied Statistics* 30:331–340.

Boos, D. D., and Duan, S. (2021). “Pairwise Comparisons Using Ranks in the One-Way Model.”*American Statistician* 75:414–423.

Brown, M. B., and Benedetti, J. K. (1977). “Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables.” *Journal of the American Statistical Association* 72:305–315.

Brown, M. B., and Forsythe, A. B. (1974). “Robust Tests for Equality of Variances.” *Journal of the American Statistical Association* 69:364–367.

Chen, S.-X., and Hall, P. (1993). “Empirical Likelihood Confidence Intervals for Quantiles.” *The Annals of Statistics* 21:1166–1181.

Chou, Y.-M., Owen, D. B., and Borrego, S. A. (1990). “Lower Confidence Limits on Process Capability Indices.” *Journal of Quality Technology* 22:223–229.

Cleveland, W. S. (1979). “Robust Locally Weighted Regression and Smoothing Scatterplots.” *Journal of the American Statistical Association* 74:829–836.

Cohen, J. (1960). “A Coefficient of Agreement for Nominal Scales.” *Education Psychological Measurement* 20:37–46.

Conover, W. J. (1972). “A Kolmogorov Goodness-of-fit Test for Discontinuous Distributions.” *Journal of the American Statistical Association* 67:591–596.

Conover, W. J. (1980). *Practical Nonparametric Statistics*. New York: John Wiley & Sons.

Conover, W. J. (1999). *Practical Nonparametric Statistics*. 3rd ed. New York: John Wiley & Sons.

Cureton, E. E. (1967). “The Normal Approximation to the Signed-Rank Sampling Distribution when Zero Differences are Present.” *Journal of the American Statistical Association* 62:1068–1069.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics* 44:837–845.

Devore, J. L. (1995). *Probability and Statistics for Engineering and the Sciences*. Pacific Grove, CA: Duxbury Press.

Dunn, O. J. (1964). “Multiple Comparisons Using Rank Sums.” *Technometrics* 6:241–252.

Dunnett, C. W. (1955). “A Multiple Comparisons Procedure for Comparing Several Treatments with a Control.” *Journal of the American Statistical Association* 50:1096–1121.

Efron, B. (1981). “Nonparametric Standard Errors and Confidence Intervals.” *The Canadian Journal of Statistics* 9:139–158.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. 2nd ed. Boca Raton, Florida: CRC.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). “Large-Sample Standard Errors of Kappa and Weighted Kappa.” *Psychological Bulletin* 72:323–327.

Friendly, M. (1994). “Mosaic Displays for Multi-Way Contingency Tables.” *Journal of the American Statistical Association* 89:190–200.

Goodman, L. A., and Kruskal, W. H. (1979). *Measures of Association for Cross Classification*. New York: Springer-Verlag.

Gupta, S. S. (1965). “On Some Multiple Decision (Selection and Ranking) Rules. *Technometrics* 7:225–245.

Hajek, J. (1969). *A Course in Nonparametric Statistics*. San Francisco: Holden-Day.

Hartigan, J. A., and Kleiner, B. (1981). “Mosaics for Contingency Tables.” In *Computer Science and Statistics: Proceedings of the Thirteenth Symposium on the Interface*, edited by W. F. Eddy, 268–273. New York: Springer-Verlag.

Hayter, A. J. (1984). “A Proof of the Conjecture That the Tukey-Kramer Method Is Conservative.” *Annals of Mathematical Statistics* 12: 61–75.

Hosmer, D. W., and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

Howell, D. C. (2013). *Statistical Methods for Psychology*. 8th ed. Belmont, California: Wadsworth.

Hsu, J. (1981). “Simultaneous Confidence Intervals for All Distances from the ‘Best’.” *Annals of Statistics* 9:1026–1034.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. London: Chapman and Hall.

Huber, P. J. (1973). “Robust Regression: Asymptotics, Conjecture, and Monte Carlo.” *Annals of Statistics* 1:799–821.

Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics.* 2nd ed. New York: John Wiley & Sons.

Iman, R. L. (1974). “Use of a t-statistic as an Approximation to the Exact Distribution of Wilcoxon Signed Ranks Test Statistic.” *Communications in Statistics—Simulation and Computation* 3:795–806.

Jones, M. C., and Pewsey, A. (2009). “Sinh-Arcsinh Distributions.” *Biometrika* 96:761–780.

Kendall, M., and Stuart, A. (1979). *The Advanced Theory of Statistics*. 4th ed. Vol. 2. New York: Macmillan.

Keuls, M. (1952). “The Use of the ‘Studentized Range’ in Connection with an Analysis of Variance.” *Euphytica* 1.2:112–122.

Kramer, C. Y. (1956). “Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications.” *Biometrics* 12:307–310.

Lehmann, E. L., and D’Abrera, H. J. M. (2006). *Nonparametrics: Statistical Methods Based on Ranks*. Rev. ed. San Francisco: Holden-Day.

Levene, H. (1960). “Robust Tests for the Equality of Variance.” In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, edited by I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann. Palo Alto, CA: Stanford University Press.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

Meeker, W. Q., and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley & Sons.

Meeker, W. Q., Hahn, G. J., and Escobar, L. A. (2017). *Statistical Intervals: A Guide for Practitioners and Researchers*. 2nd ed. New York: John Wiley & Sons.

Miller, A. J. (1972). “Letter to the Editor.” *Technometrics* 14:507.

Nagelkerke, N. J. D. (1991). “A Note on a General Definition of the Coefficient of Determination.” *Biometrika* 78:691–692.

Nelson, P. R., Wludyka, P. S., and Copeland, K. A. F. (2005). *The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions*. Philadelphia: Society for Industrial and Applied Mathematics.

Neter, J., Wasserman, W., and Kutner, M. H. (1996). *Applied Linear Statistical Models*. 4th ed. Boston: Irwin.

O’Brien, R. G. (1979). “A General ANOVA Method for Robust Tests of Additive Models for Variances.” *Journal of the American Statistical Association* 74:877–880.

O'Brien, R., and Lohr, V. (1984). “Power Analysis For Linear Models: The Time Has Come.” *Proceedings of the Ninth Annual SAS User’s Group International Conference*, 840–846. Cary, NC: SAS Institute Inc.

Olejnik, S. F., and Algina, J. (1987). “Type I Error Rates and Power Estimates of Selected Parametric and Nonparametric Tests of Scale.” *Journal of Educational Statistics* 12:45–61.

Passing, H., Bablok, W. (1983). “A New Biometrical Procedure for Testing the Equality of Measurements from Two Different Analytical Methods”. *Journal of Clinical Chemistry Clinical Biochemistry*, 21:709–20.

Passing, H., Bablok, W. (1984). “Comparison of Several Regression Procedures for Method Comparison Studies and Determination of Sample Sizes”. *Journal of Clinical Chemistry Clinical Biochemistry*, 22:431–45.

Pratt, J. W. (1959). “Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures.” *Journal of the American Statistical Association* 54:655–667.

Reinsch, C. H. (1967). “Smoothing by Spline Functions.” *Numerische Mathematik* 10:177–183.

Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Rubin, D. (1981). “The Bayesian Bootstrap.” *The Annals of Statistics* 9:130–134.

SAS Institute Inc. (2020a). “Introduction to Nonparametric Analysis.” In *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/intronpar.pdf.

SAS Institute Inc. (2020b). “The FREQ Procedure.” In *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/freq.pdf.

Schuirmann, D. J. (1987). “A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability.” *Journal of Pharmacokinetics and Biopharmaceutics* 15:657–680.

Slifker, J. F., and Shapiro, S. S. (1980). “The Johnson System: Selection and Parameter Estimation.” *Technometrics* 22:239–246.

Snedecor, G. W., and Cochran, W. G. (1980). *Statistical Methods*. 7th ed. Ames, Iowa: Iowa State University Press.

Somers, R. H. (1962). “A New Asymmetric Measure of Association for Ordinal Variables.” *American Sociological Review* 27:799–811.

Stephens, M. A. (1974). “EDF Statistics for Goodness of Fit and Some Comparisons.” *Journal of the American Statistical Association* 69:730–737.

Tan, C. Y., and Iglewicz, B. (1999). “Measurement-Methods Comparisons and Linear Statistical Relationship.” *Technometrics* 41:192–201.

Tamhane, A. C., and Dunlop, D. D. (2000). *Statistics and Data Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Tukey, J. W. (1953). “The Problem of Multiple Comparisons.” In *Multiple Comparisons, 1948–1983*, edited by H. I. Braun, vol. 8 of *The Collected Works of John W. Tukey* (published 1994), 1–300. London: Chapman & Hall. Unpublished manuscript.

Welch, B. L. (1951). “On the Comparison of Several Mean Values: An Alternative Approach.” *Biometrika* 38:330–336.

Wheeler, D. J. (2003). *Range Based Analysis of Means*. Knoxville, TN: SPC Press.

Wilson, E. B. (1927). “Probable Inference, the Law of Succession, and Statistical Inference.” *Journal of the American Statistical Association* 22:209–212.

Wludyka, P. S., and Nelson, P. R. (1997). “An Analysis-of-Means-Type Test for Variances From Normal Populations.” *Technometrics* 39:274–285.

Appendix B

Technology License Notices

- Scintilla is Copyright © 1998–2017 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

  NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.
- Progressë® Telerik® UI for WPF: Copyright © 2008-2019 Progress Software Corporation. All rights reserved. Usage of the included Progress® Telerik® UI for WPF outside of JMP is not permitted.
- ZLIB Compression Library is Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.
- Packages is Copyright © 2009–2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  Neither the name of the WhiteBox nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED

WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- iODBC software is Copyright © 1995–2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  – Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL OPENLINK OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- This program, “bzip2”, the associated library “libbzip2”, and all documentation, are Copyright © 1996–2019 Julian R Seward. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.

3. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

4. The name of the author may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Julian Seward, jseward@acm.org

bzip2/libbzip2 version 1.0.8 of 13 July 2019

- R software is Copyright © 1999–2012, R Foundation for Statistical Computing.
- MATLAB software is Copyright © 1984-2012, The MathWorks, Inc. Protected by U.S. and international patents. See www.mathworks.com/patents. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.
- libopc is Copyright © 2011, Florian Reuter. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  - Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.
  - Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR

ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- libxml2 - Except where otherwise noted in the source code (e.g. the files hash.c, list.c and the trio files, which are covered by a similar license but with different Copyright notices) all the files are:

  Copyright © 1998–2003 Daniel Veillard. All Rights Reserved.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

  THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

- Regarding the decompression algorithm used for UNIX files:

  Copyright © 1985, 1986, 1992, 1993

  The Regents of the University of California. All rights reserved.

  THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3. Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

- Snowball is Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton.

  All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS \”AS IS\” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- Pako is Copyright © 2014–2017 by Vitaly Puzrin and Andrei Tuputcyn.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

- HDF5 (Hierarchical Data Format 5) Software Library and Utilities are Copyright 2006–2015 by The HDF Group. NCSA HDF5 (Hierarchical Data Format 5) Software Library and Utilities Copyright 1998-2006 by the Board of Trustees of the University of Illinois. All rights reserved. DISCLAIMER: THIS SOFTWARE IS PROVIDED BY THE HDF GROUP AND THE CONTRIBUTORS "AS IS" WITH NO WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED. In no event shall The HDF Group or the Contributors be liable for any damages suffered by the users arising out of the use of this software, even if advised of the possibility of such damage.
- agl-aglfn technology is Copyright © 2002, 2010, 2015 by Adobe Systems Incorporated. All Rights Reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  – Neither the name of Adobe Systems Incorporated nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT

(INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- dmlc/xgboost is Copyright © 2019 SAS Institute.

  Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libzip is Copyright © 1999–2019 Dieter Baron and Thomas Klausner.

  This file is part of libzip, a library to manipulate ZIP archives. The authors can be contacted at <libzip@nih.at>.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3. The names of the authors may not be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE AUTHORS ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- OpenNLP 1.5.3, the pre-trained model (version 1.5 of en-parser-chunking.bin), and dmlc/xgboost Version .90 are licensed under the Apache License 2.0 are Copyright © January 2004 by Apache.org.

You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

- You must give any other recipients of the Work or Derivative Works a copy of this License; and
- You must cause any modified files to carry prominent notices stating that You changed the files; and
- You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
- If the Work includes a "NOTICE" text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.
- You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

- LLVM is Copyright © 2003–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- clang is Copyright © 2007–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS”, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- lld is Copyright © 2011–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS”, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libcurl is Copyright © 1996–2021, Daniel Stenberg, daniel@haxx.se, and many contributors, see the THANKS file. All rights reserved.

  Permission to use, copy, modify, and distribute this software for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

  THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

- On the Windows operating system, JMP utilizes the OpenBLAS library. OpenBLAS is licensed under the 3-clause BSD license. Full license text follows: Copyright © 2011-2015, The OpenBLAS Project

  All rights reserved. Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. Neither the name of the OpenBLAS project nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE OPENBLAS PROJECT OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.