SAS INSTITUTE INC.
FALL 2000
ISSUE 7

JMPer Cable®

NEWSLETTER FOR JMP® USERS

Transition from JMP 3 to JMP 4: Questions and Answers

John Sall, Executive Vice President
SAS Institute Inc.

Version 4 is here and this is the first issue of JMPer Cable to discuss it. But rather than cover new features, we would like to cover some transition issues in a question and answer format.

Q: Is JMP 4 much different than JMP 3?

A: Yes. Internally, it is completely rewritten in a new language. However, in functionality it is compatible with JMP 3. The surface details are different; it will take a little time to become accustomed to the new interface.

Q: When should I upgrade to JMP 4?

A: Now. We shipped an early-adopter release 4.0 in April, but more recently started shipping release 4.0.2, which had a number of fixes and improvements.

Q: Is there going to be another release?

A: The next release will be 4.0.4, which is due out in a few months. The main change will be to the formula scheduling system. There was a performance problem in earlier Version 4 releases when you used subscripting or the Lag function for tables with a large number of rows. We completely reworked the formula scheduling system and it is now faster than it was in JMP 3.

Q: Why did JMP 4 take so long to complete?

A: With Version 3.2 we decided that JMP had evolved to do most things it needed to and we had the luxury to spend a longer time on Version 4, redesigning the internals to be more flexible for supporting further growth. One big change was to go from C to C++. Another change was to create a new host operating system interface so that JMP would be more comfortable under Windows without sacrificing any comfort on the Macintosh. The display system was reworked to support larger reports and to create a new journaling system. A huge effort went into making JMP scriptable, and this investment has paid off well.

Q: Are there eature differences or incompatibilities between JMP 3 and JMP 4?

A: Veteran users have pointed out differences, but notice that the list is small:

- Some accelerator keys have changed. This might be a temporary annoyance until you get used to JMP 4.
- The product is bigger on disk. This is largely because of the growth in the Help system, which now encompasses most of the JMP documentation.
- The Ternary plot doesn't have the same functionality—it offers contours for expressions but not contours for data.
- The Fit Model contour facility for response surface models has been removed because there is a better facility available in the Contour Profiler.
- The candidate-set D-Optimal feature was removed from Fit Model because there is a much better custom designer in the new DOE facility.
- Formula editing works differently in order to facilitate the scripting language in formulas.

However, there are a multitude of reasons to move to JMP 4, many of which we will present here and future newsletters.

Q: What about incompatibilities in results between JMP 3 and JMP 4?

A: Categorical parameter naming and continuous polynomial centering in the model fitting are major changes. We used to name parameters after the coding arithmetic. Now we name parameters according to their interpretation. For example Age[12-17] was the version 3 name to represent the coefficient in the linear model for Age=12 and the negative coefficient for age[17]. The design column was formed by subtracting the two indicator values. In JMP 4, we name the parameter Age[12] because that is the way to interpret it. Age[17] for the last level is the negative sum of the others, which you can now see if you ask for the Expanded Estimates.

For continuous factors that participate in compound effects, i.e. those that are multiplied by other factors, we now subtract the mean to center the values before the multiplication.

The big reason for doing this is that now the lower order (e.g. main effect) parameter estimates are interpretable and have a more appropriate test. This feature, called *polynomial centering*, can be turned off in the model dialog if you want to get answers that agree

(continued next page) ➡

IN THIS ISSUE

(continued from previous page)

with JMP 3. Release 3 of JMP had several features, such as Effect Screening, that attempted to compensate for non-centered polynomials, but JMP 4 addresses the problem directly. You can also assign a *coding property* for columns if you want to completely specify how the factor is parameterized. This improves the interpretability of the estimates as effect sizes. The DOE facility automatically makes use of this feature.

Q: What's the deal with formula editing?

A: The formula editor in JMP 3 was smart and beautiful, but there were other issues:

1) There were certain cases where the order of evaluation was not clear by looking at the formula. To fix this, JMP 4 draws a box around each layer of the formula, so that you can immediately see the tree structure of the formula evaluation hierarchy. Boxing is an option that can be turned off if not needed.

2) Version 3 used operator precedence when you composed expressions. This means that instead of applying the operator on the selected expression, the formula editor moved up the expression tree until it found a lower-priority operator, and then applied the operator. This made Version 3 easy for entering an addition or multiplication, but sometimes caused frustration when you applied an operator and discovered that you had to start over and add parentheses first. In JMP 4 there is a simpler rule: when you enter an operator, it is always applied to the selected expression.

Q: What non-obvious features should I look for in JMP 4?

A: Many new things are visible and obvious so you won't need any hints. But there are a few features that are not visible directly, and these are some of the best features in Version 4. For example, context clicking (right-mouse click or CONTROL-click on the Macintosh) is supported everywhere, giving a new richness in functionality. Context-click on a graph, a table, an axis, a plot label, and you see a menu of useful commands for these report items. Also, we improved the way the cursor changes as you move the mouse into different areas of the interface surface. Point identification by just hovering over a point displays the row ID for the point so you can find out about the point without changing the selection. **Important:** Context-click on tables to see if there are hidden columns available. In particular, the Parameter Estimates table has extra columns you can unhide.

Q: Why is the Custom Design command the first command of the DOE features?

A: We are making a statement here, that D-Optimal design is not just for very special situations, but can be used by everyone to produce experimental designs appropriate to any situation. It is still easy to get a classical design if you want one. But there are some major breakthroughs in Custom Design that make it easy, and extraordinarily general. When an engineer wants an experimental design, the best approach is to define the factors and responses, describe their properties, and specify the kinds of effects that need to be estimable. Then, continue with a dialog that accepts information about sample size. This process is much more natural than trying to shoehorn a situation into a table of predefined classical design.

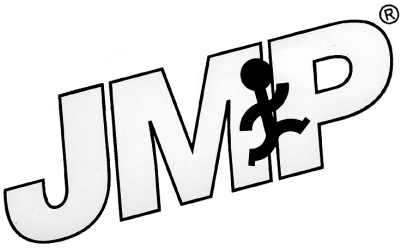
Q: Will JMP 3 still work if I install JMP 4?

A: Yes. And JMP 3 reads JMP 4 data tables. However, JMP 4 formulas are not supported in JMP 3.

Q: How do I upgrade to JMP 4?

A: All registered JMP customers who are running JMP 3.0 or above may upgrade to JMP 4 for $295 (corporate customers) or $195 (academic customers).

Note: All registered users who purchased JMP 3.2.6 on or after February 1, 2000 are entitled to receive JMP 4 free of charge. To place an upgrade order within the U.S., or if you qualify for a free upgrade, contact Fulfillment Services at 1-800-727-3228. Please have your registration number available. To qualify for the free upgrade, you must show proof of purchase. Volume discounts and leasing arrangements are available. For more information, contact JMP Sales at 919-677-8000.



Bulletin

Robert Mee, Professor and Head, Department of Statistics,
University of Tennessee, Knoxville, TN 37996-0532
rmee@utk.edu ph (423) 974-1640

The University of Tennessee has approved a new certificate program offered by the University of Tennessee Statistics department. It involves four courses that are available either locally or for at-distance students. The courses are almost exclusively based on the use of JMP. The courses consist of the applied portion of the first year of an M.S. in statistics, which includes statistical methods, applied linear models, design of experiments, and industrial statistics.

THE SUBSET COMMAND

Ann Lehman
SAS Institute

Creating a subset of a JMP table has always been very simple—select the rows and columns you want included in the subset and choose Tables→Subset. If no rows are selected, the subset contains all rows; if no columns are selected, the subset contains all columns.

Subsetting a table continues to be easy in JMP 4, but the Subset dialog now offers some new and useful choices.

Link To Original Data Table

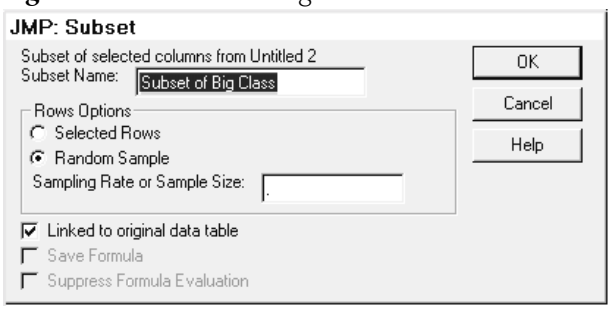
When you choose the Subset command, the Subset dialog first appears with the Linked to original data table check box checked, as shown in **Figure A**. If columns in the data table have formulas, there are also Save Formula and Suppress Formula Evaluation boxes. These check boxes are dim because linked subsets inherit the source table's column formulas. Computed columns in linked subsets have these characteristics:

- altering a formula in the original table changes values in the subset table as well as the original table
- the corresponding subset column is locked even though it does not have a formula editor of its own.

In general, the rows in a linked subset behave the same as their corresponding rows in the original table. If you select rows in the original table, any rows included in the subset are also selected. Likewise, if you select rows in the subset, the corresponding rows are selected in the original table. If you change cell values in the original table, the same cells in the subset table also change. If you change a cell value in the subset, the corresponding cell changes in the original table.

Columns behave differently than rows. Selected columns do not automatically become selected to and from the original and subset tables. However, you can select a column in either table and use its Column Info dialog to change column characteristics. The characteristics of the same column in the other table change to match.

Figure A *The Subset Dialog*



Note: You can bypass the subset dialog by using SHIFT-Subset. This action automatically produces a linked subset of the selected rows and columns.

Column Formulas in Subsets

If there are any columns with formulas in the original table and you uncheck the Link to original data table check box, the formula check boxes become active. If Save Formulas is checked, a subset contains the same column formulas as those in the original table.

However, keep in mind that formulas in a subset can be tricky. If a formula uses the row number function, Row(), the result in the subset might not be what you expect. For example, suppose a table contains baseline values in the first row, and another column compares values in each subsequent row to those in the first row. If you subset the table without including the first row, the resulting column would incorrectly make the comparison to the first row in the subset table.

That is why there is a Suppress Formula Evaluation check box. If you take an unlinked subset and know that a formula does exactly what you want, then leave the Suppress Formula Evaluation unchecked. Use the check box if you want to examine the subset before having a formula reevaluate. Then, if needed, change the formula with the Formula Editor, or delete the formula if you don't want it.

Random Sampling

The Subset facility can also create a subset that is an exact-size (to the nearest integer) random sample without replacement. When you check the Random Sample radio button and enter either a sample size or a sample rate (Sampling Rate or Sample Size), the sampling process begins in the first row, generates a random number, compares it to the sample rate probability, and accepts or rejects that observation. The number of rows left to choose from is decremented and, if the row was chosen, the sample number still needed is decremented. The sampling rate is recomputed and this process continues until the sample is complete.

The following JSL program performs the same sampling procedure as the Random Sample function of the Subset command:

```
//Random Sample of k rows;
k=20;  n=nrow();
for (i=1, i<=nrow(),i++, p=k/n;
    if (random Uniform() < p,
       k--;
       Selected (Rowstate(i))=1;
       );
    n--;
    );
```

ALL MIXED UP

Mark Bailey, Statistical Training & Technical Service
SAS Institute Inc.

This article presents a new way to design optimal experiments that were previously impractical using classical design schemes. For example, problems that encompass both process factors, such as temperature and pressure, and mixture components, such as blends of chemical solvents, are not uncommon. Traditionally, experimenters divide such problems into two parallel tracks because the available classical designs for process factors (e.g., factorial and central composite designs) do not address mixture components, and vice versa.

The design and model for the factors must assume a nominal mixture. Likewise, the mixture design has to assume nominal factor levels. Eventually, runs will vary factors and components together in a late optimization or confirmation stage. One scheme to study the factors and components together is to replicate the design for the factors at each point of the mixture design (or the other way around). These studies become very large for only a few variables.

For example, to study just 4 factors and 3 components you might use a Box-Behnken design with 27 runs and a simplex centroid design with 7 runs. Note that these choices are individually minimal designs in order to control the overall size of the combined experiment. In this case, the total size is 7x27=189 runs.

The classical designs for these cases further restrain design choices because they limit blocking structures and exclude categorical factors and fixed covariates. These important factors must be incorporated into the design in a less-than-optimal way or ignored completely.

In this hypothetical case, pretend that you are a chemist who wants to find the best conditions for purifying the reaction product for a new drug on a pilot plant scale using a recrystallization method. The goals for this process are

1) recover at least 90% of the pure compound,
2) achieve at least 98% overall purity, and
3) reduce a particular toxic impurity to no more than 1%.

These goals are listed in increasing relative importance (1-3). Note that these goals are important during the search for optimum conditions but they do not participate directly in the design of the experiment itself.

To define this design, use DOE→Custom Design (or click the DOE tab of the JMP Starter window). Open the Responses panel and use the Add Response button to display three responses. Enter the specifications for each response to match **Figure A**.

Figure A *Multiple Responses with Goals and Limits*

Custom Design

Responses

Add Response | Remove | N Responses

Response Name	Goal	Lower Limit	Upper Limit	Importance
Yield	Maximize	90	.	1
Purity	Maximize	98	.	2
Impurity	Minimize	.	1	3

Next, consider the factors. The recrystallization involves completely dissolving the crude reaction product and then selectively crystallizing only the desired drug out of solution while leaving all of the impurities behind in solution. The material and the solvent are heated and then held at an elevated temperature (70-90° C) for some period of time (1-2 hours). The amount of material (1-2 grams) can vary from one reaction to the next. Better results are obtained for some drugs if a small amount of pure crystals (*seeds*) are added after the solution cools.

The solvent must dissolve everything at higher temperatures (at least 20% water, 15% alcohol, and 20% ether) but only the impurities at room temperature (no more than 60% ether or 70% alcohol and ether combined). The whole process takes about half a day, but three reactors are available so that 6 runs per day are possible.

Use the Add Factors button to create three *continuous* factors, one 2-level *categorical* factor, three *mixture* factors, and one *blocking* factor for six runs per block. Enter the specifications for each of the factors to match **Figure B**.

Figure B *Factors Panel Showing Mixed Factors*

Custom Design

Responses

Factors

Add Factor | Remove | Add N Factors 1

Name	Role	Values	
Temperature (C)	Continuous	70	90
Time (hrs)	Continuous	1	2
Amount (g)	Continuous	1	2
Seeding	Categorical	no	yes
Water	Mixture	0.2	1
Alcohol	Mixture	0.15	1
Ether	Mixture	0.2	0.6
day	Blocking	1	2

The drug is soluble in ether and somewhat in alcohol so together these chemicals must not make up more than 70% of the solvent or the crystals will not form. The restriction on the combined proportion of alcohol and ether is created with the Add Constraint button, as shown to the right.

Define Factor Constraints

Add Constraint

Factor	Constraint
Tempurature (C)	0
Time (hrs)	0
Amount (g)	0
Water	0
Alcohol	1
Ether	1

less than or equal

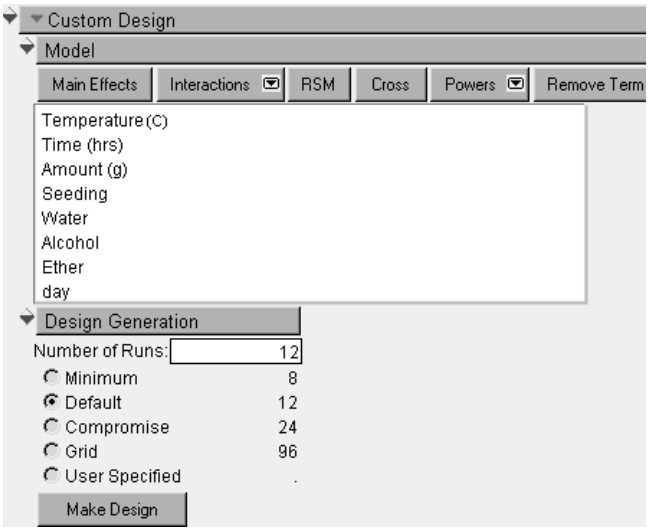
0.7

Note that constraints can be defined in terms of any of the continuous factors, not just mixture components. For example, lower temperature and shorter holding time might not work or higher temperature and longer holding time might not be necessary. You could exclude these regions from the design with additional constraints.

The optimal design collects data at points to estimate all of your important effects (model adequacy) with sufficient precision (prediction variance) in the fewest number of runs. You will assess the precision in a moment before you accept the final design.

You decide to investigate only the main effects of these factors in this initial study and determine if there is an opportunity in this region. You can add terms to your model for additional effects after screening, and augment your design with new optimal runs later. Inspect the model section of the design platform as shown in **Figure C**.

Figure C Model and Design Generation Panels



All of the information about your problem is entered. The last decision is about the sample size. The minimum number of runs is sufficient for estimating all of the effects in the model but it provides no degrees of freedom for estimating the residual error. The default choice in this case provides 4 additional degrees of freedom and will require 2 days of testing. You decide to begin this way. Use the Make Design button to continue.

You can see in this table that you are sampling the low and high levels of each factor in a generally balanced and complete way. (Please note that you may not obtain exactly the same 12 runs as shown in **Figure D** due to the random nature of the D-optimal design.)

Figure D Customized D-Optimal Design for Mixed Factors

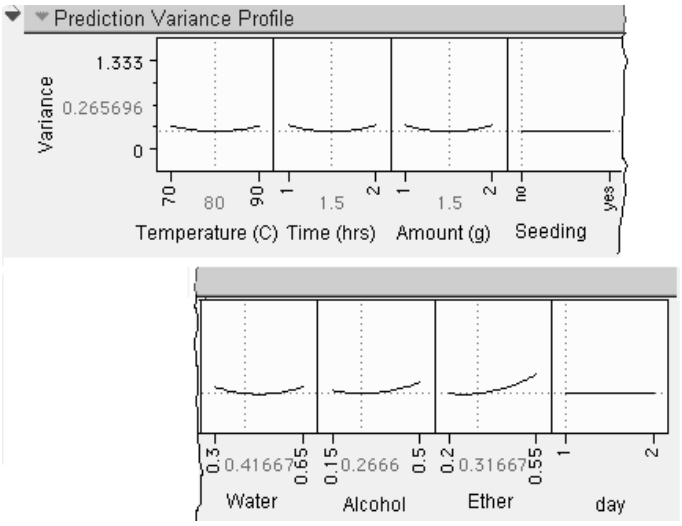
Design

Run	Temperature (C)	Time (hrs)	Amount (g)	Seeding	Water	Alcohol	Ether	day
1	70	2	2	yes	0.3	0.5	0.2	1
2	70	1	1	no	0.3	0.5	0.2	2
3	70	2	1	no	0.65	0.15	0.2	1
4	90	1	2	no	0.65	0.15	0.2	2
5	90	2	1	no	0.3	0.15	0.55	1
6	90	2	1	yes	0.3	0.5	0.2	2
7	70	1	2	yes	0.3	0.15	0.55	1
8	70	2	2	no	0.3	0.15	0.55	2
9	90	1	2	no	0.3	0.5	0.2	1
10	90	2	2	yes	0.65	0.15	0.2	2
11	90	1	1	yes	0.3	0.15	0.55	1
12	70	1	1	yes	0.65	0.15	0.2	2

How well do these 12 design points estimate the main effects that begin your investigation? Open the Prediction Variance Profile to see the plots shown in **Figure E**.

This display shows the variance of the model prediction for any point in this region. It establishes the confidence interval (bands) for the model prediction later when it is combined with the error variance (i.e., RMSE) obtained in the data analysis.

Figure E Prediction Variance Profiler



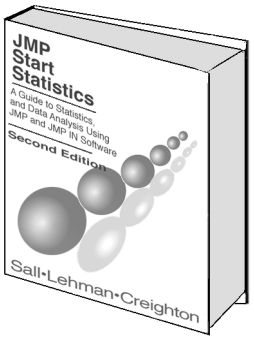
When you proceed with Make Table, you can also use the graphical exploration tools in the Distribution, Multivariate, and Spinning Plot platforms to review the balance and arrangement of the design points.

Consider: What classical design approach would handle every aspect of this problem for you as easily and as well as this custom design?

Please see the *JMP Design of Experiments* guide for more details about these procedures and design tools.

JMP Start Statistics: Second Edition

A Guide to Statistics and Data Analysis Using JMP and JMP IN Software



The Second Edition of *JMP Start Statistics* continues to fill the demand for a book that focuses on using JMP to learn about statistics. *JMP Start Statistics* is a friendly, comprehensive approach to statistics, and a guide for learning to use JMP software. The second edition has been updated to reflect the new capabilities in JMP 4.

This book is also the official reference document that accompanies release 4 of JMP IN, the student edition of JMP. The JMP IN package (software and book) is available to students from Duxbury Press and also through many campus bookstores. JMP IN provides students with a complete data analysis program useful in any statistics course, elementary to advanced. Students can handle an unlimited number of data points, so there is no artificial limit placed on the size of problems that can be solved. JMP IN also includes the new JMP Scripting Language that allows simulations and automated analyses.

JMP Start Statistics is written by John Sall, the principal developer of JMP, Ann Lehman, and Lee Creighton. It has over 500 pages of statistical discussion and explanation (interspersed with occasional informal thoughts).

Each topic is supported with hands-on examples of varying levels of complexity. JMP's interactive statistical platforms help clarify and simplify what are often perceived as difficult and tangled-up statistical concepts. Chapters conclude with a set of lab or homework exercises.

The interactive and graphical nature of the JMP examples promotes the idea that statistics is a discovery process:

- Graphics that accompany each analysis make it easier to understand results.
- Interactivity leads to further analysis, refines results, and can result in further discovery.

JMP Start Statistics begins with the basics—one variable, one sample—to look at univariate distributions. Topics progress through simple regression, *t* tests, analysis of variance, analysis of categorical variables, multiple regression, correlations, multivariate relationships, and fitting general linear models. Also, specialty chapters discuss statistical quality control, design of experiments, and elementary time series analysis.

Complete documentation for JMP IN software is included on the JMP IN CD, in the JMP IN help system, and in pdf files.

To order a copy of *JMP Start Statistics* (without software), contact SAS Institute Inc. at 1-800-727-3228 (9-5 p.m. EST, Monday through Friday), or send Email to SASBOOK@sas.com. ISBN 0-534-35967-1

For information about using JMP IN software in your classes, contact Duxbury Press at (800) 425-0563. You can visit Duxbury online at www.duxbury.com.

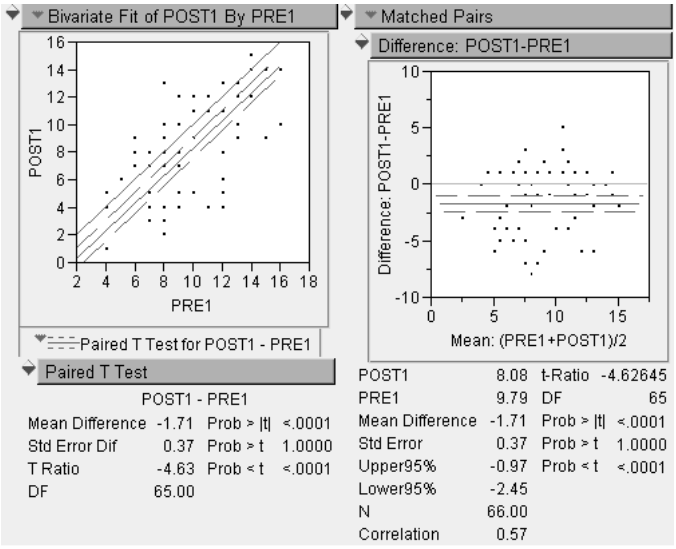
PAIRED t-TEST UPDATE

In JMP 3, the paired t-test is part of the Bivariate platform. A 45 degree line through the origin represents the reference for when there is no difference between the paired variables. The difference line and 95% confidence lines are plotted parallel to the reference line. If the reference line is outside the confidence limits, as shown on the left in **Figure A**, the t test is significant at $p < 0.05$.

JMP 4 enhances this analysis and its plot, and promotes it into its own platform, Matched Pairs in the Analyze menu. The difference of the scores is plotted on the vertical axis, as shown on the right in **Figure A**. You can tailor the axis with a reference line at zero, and see if it lies outside 95% confidence limits. JMP 4 gives a more detailed report, and also lets you use a grouping variable.

Note: If you prefer the JMP 3 paired t-test analysis, it can be accessed from the Bivariate platform. To do this, analyze the paired variables with Fit Y by X. Then use the SHIFT key when clicking the popup menu on the Bivariate title bar. The Paired t Test command then appears in the list of fitting commands.

Figure A *t-test as in JMP 3 (left) and in JMP 4 (right)*



WHAT'S NEW ALERT!

A difference between JMP 3 and JMP 4

You might have noticed that there is no longer a Paste At End command in the Edit menu. JMP 4 can perform the same functions as Paste at End, but in simpler ways.

In JMP 3, the Paste at End command creates new rows or columns as needed to hold the clipboard contents. It pastes data into new rows at the end of selected columns, or into new columns at the end of selected rows. If both rows and columns are selected, the Paste at End command is dimmed.

If no rows or columns are selected, Paste at End appends the clipboard data array to the bottom of the table, beginning in the first column, creating new rows and columns as needed. You cannot undo a Paste at End action.

In an existing data table, the Paste command is dimmed unless there is a selection of rows, columns, or both. Paste writes over the cell content of the selected area. Paste actions are undoable both JMP 3 and JMP 4.

In JMP 4, the Paste command has been enhanced to perform the duties of the JMP 3 Paste at End command. When pasting, the results depend on what section of the data table is selected to be the paste destination, or whether anything in the data table is selected at all.

Pasting Without an Explicit Destination

If no columns or rows are selected, the default paste area begins in the first column beneath the last row. The JMP 4 Paste command appends the clipboard data to the bottom of the table, creating new rows and columns as needed for the data.

If you paste into an empty table, JMP 4 creates the rows and columns needed to accommodate the clipboard data. The Paste at End command in V3 performs the same function. This action provides a simple way to transfer data from another application into JMP, or from one JMP table to another.

Designating an Explicit Paste Destination

The most familiar way to designate a paste destination is to select rows, columns, or both. As in JMP 3, Paste writes over the cell contents of the selected area.

- If the number of selected rows is less than the number of rows on the clipboard, and the selection includes the last row in the table, additional rows are created, to hold all the clipboard contents
- If the number of selected rows is less than the number of rows on the clipboard, and the selection does not include the last row in the table, Paste overwrites the selected rows with only as much data as fits the destination.
- When columns are explicitly selected for pasting, JMP 4 expects the number of selected columns to match the number of columns on the clipboard.
- If the designated number of rows is greater than the number of rows on the clipboard, values are repeated until the area is full.

You can click in any cell beneath the existing columns to define the upper-left corner of a rectangular paste destination. For example, suppose you select the rows and columns shown on the left in **Figure A**, and copy them to the clipboard.

Then click to identify a destination area in the data table, as in the middle table in **Figure A**. When you paste, the results appear as shown in the right-hand table, with the receiving rows and column automatically selected.

In this situation, if there are more columns of data on the clipboard than in the data table (as in the example below), the additional columns on the clipboard are ignored. When there are fewer columns of data on the clipboard than there are in the data table, the cells for the additional columns in the data table fill with missing values. These actions are the same as those for JMP 3 when Paste at End is used with a selected destination.

(Continued next page) ➔

Figure A *Example of Pasting Data into a Designated Paste Destination*

	name	age	sex	height
1	KATIE	12	F	59
2	LOUISE	12	F	61
3	JANE	12	F	55
4	JACLYN	12	F	66
5	LILLIE	12	F	52

Clipboard Viewer

File Edit Display Help

KATIE 12 F 59
LOUISE 12 F 61
JANE 12 F 55

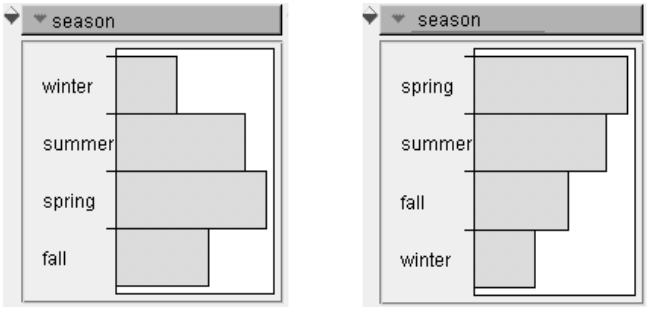
	name	age	sex	height
1	KATIE	12	F	59
2	LOUISE	12	F	61
3	JANE	12	F	55
4	JACLYN	12	F	66
5	LILLIE	12	F	52

	name	age	sex	height
1	KATIE	12	F	59
2	LOUISE	12	F	61
3	JANE	12	F	55
4	JACLYN	12	F	66
5	LILLIE	12	F	52
6		•		•
7		•	KATIE	12
8		•	LOUISE	12
9		•	JANE	12

Tips and Techniques

A very frequently asked question is, "How do I change the order of the levels in a graph or analysis?" For example, the distribution platform orders the values 'winter', 'summer,' 'spring', 'fall', as shown on the left in **Figure A**. But suppose you want the order to be 'spring', 'summer', 'fall', 'winter', as shown on the right.

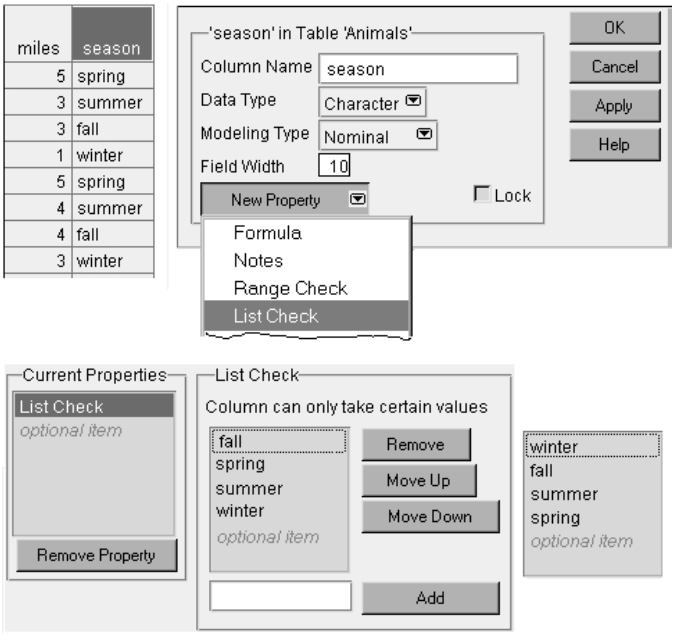
Figure A *Default Order* *Rearranged Order*



Rearranging the order of levels is easily accomplished using the List Check column property. To do this:

- Select the column whose values you want reordered.
- Choose Cols→Column Info and create the List Check property as shown below at the top in **Figure B**.
- Select a value in the List Check box and use the Move Up or Move Down buttons to reorder.

Figure B *Reorder Levels with List Check Column Property*



(continued from the previous page)

What's New Alert! Discontiguous Copy and Paste

To make a discontiguous selection of rows and columns, use CTRL-click (COMMAND-click on the Macintosh). When you copy a discontiguous selection, the clipboard knows only to arrange the information in rows and columns (**Figure B**). Thus the discontiguous nature of the selection is lost.

Figure B *Copy Discontiguous Rows and Columns*

	name	age	sex	height
1	KATIE	12	F	59
2	LOUISE	12	F	61
3	JANE	12	F	55
4	JACLYN	12	F	66
5	LILLIE	12	F	52

Clipboard Viewer
File Edit Displa
LOUISE 61
JACLYN 66

When you then paste, the default as described previously—the information fills a rectangular area starting at the bottom of the table in the first column.

	name	age	sex	height
1	KATIE	12	F	59
2	LOUISE	12	F	61
3	JANE	12	F	55
4	JACLYN	12	F	66
5	LILLIE	12	F	52
6	LOUISE	61		•
7	JACLYN	66		•

If you create two additional rows in the table and select those rows but don't select columns, Paste notes that the number of columns on the clipboard isn't the same as the number of columns in the data table and gives an alert. Nothing is pasted.

If you only select columns, the clipboard contents fill the columns completely, cycling through the values to fill the columns, as in the top table to the right.

	name	age	sex	height
1	LOUISE	12	F	61
2	JACLYN	12	F	66
3	LOUISE	12	F	61
4	JACLYN	12	F	66
5	LOUISE	12	F	61
6	JACLYN	•		66
7	LOUISE	•		61

To preserve the discontiguous paste selection, both rows and columns have to be selected, as shown in the bottom table.

	name	age	sex	height
1	KATIE	12	F	59
2	LOUISE	12	F	61
3	JANE	12	F	55
4	JACLYN	12	F	66
5	LILLIE	12	F	52
6	LOUISE	•		61
7	JACLYN	•		66

You can also use drag-and-drop action to move or copy sections of a JMP table anywhere in the same table or to another JMP table, but that's another story.

Note: Use the Undo command in the Edit menu to undo the effects of a less-than-satisfactory Paste result.

POWER LINES: ADDING A GRAPHICS SCRIPT

John Sall
SAS Institute Inc.

There are many situations where you have a graph and you want to overlay a curve corresponding to some expression on the graph. With JMP Version 4, you can context click on most graph frames and select the Add Graphics Script command, which lets you enter a JSL script that executes and draws its results. The operator in JSL to graph a function has this form

```
YFunction (expression, x variable,...)
```

One situation we recently encountered was a cDNA micro array gene expression study on fruit fly genes, done by Greg Gibson (Department of Genetics, North Carolina State University). There were 384 genes studied on the array. Unlike most micro array experiments, this study had replicates, so we could study the variation in expression for each gene, as well as the response itself. It turned out that different genes had widely varying variances, and thus we recommend replication for future experiments. This study had 6 replicates; the question is how many replicates are recommended. There is no right answer here, but it would be nice to see a graph of what combinations of mean and standard deviation would have, say, a 90% power of being detected. The response, represented by the mean of the reps, is the log ratio of the treated fly compared to a standard, as in typical cDNA micro array studies.

Figure A shows a graph of the standard deviation by the mean for each of 384 genes. How many responses like this would likely be significant if you had 2, 3, 4, 5, or 6 replicates? The points in the middle (near zero) tend to not be significant because of their small effect size (mean). The points lying in the upper portion of the plot tend to not be significant because of their large standard deviation.

To find the answer, we first find F statistics as a function of the true mean, standard deviation, and sample size. Then we can draw the contours of the sample size as a function of the F statistics.

To do this, context click (right click, or CONTROL click on the Mac) in the plot frame and select Add Graphics Script from the menu. When the script editor displays, enter the following script:

Context Menu


```
for(n=2, n<=6, n++,
   Contour Function(
      F Sample Size(0.05, 1, 1,
         XMean^2/(YStdDev^2/n), 0.9),
      XMean, YStdDev, n)
   )
```

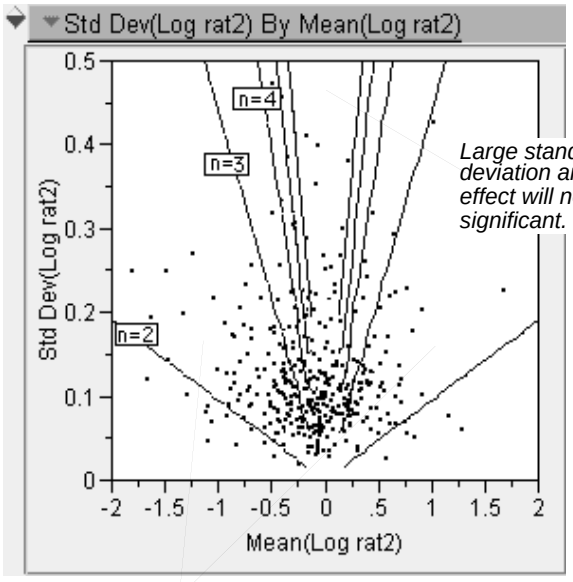
Working from the inside, the script is interpreted like this:

- The F Sample Size function has five arguments; it calculates the sample size needed to achieve a power of 90% (its fifth argument) at a 0.05 alpha-level (its first argument), using an F test with 1 numerator and 1 denominator degrees of freedom (arguments 2 and 3). This F test is just the square of a t-test—the square of the mean divided by the standard error of the mean (its fourth argument).
- The Contour Function has four arguments. It evaluates its first argument (the F Sample Size function described above) for a grid of values identified by the second and third arguments, then draws the contour line specified in the fourth argument.

Note: This JSL formula is computationally intensive and might take a few moments to complete—be patient.

In the result shown in **Figure A**, the lower outer line is the contour for 2 observations per gene; the next lines are for 3, 4, and so on, up to 6 replicates. You can see that as you move to larger samples sizes, you can detect more and more of the combinations of mean and standard deviation represented by the points in the actual experiment. Judging from this, it looks like going to 4 replicates gives you a lot of sensitivity, but going to more reps gains less and less territory for the 90% power.

Figure A *Computed Lines Show Areas of Significant Power for Various Numbers of Replicates*



Mean and standard deviation are 90% likely to be significant when replicates are increased from 2 to 3.

TRIGONOMETRIC REGRESSIONS

Lee Creighton
SAS Institute Inc.

Regression is a method of fitting curves through data points. It is often thought of as a linear technique, used for models that are linear in their parameters.

A unique set of problems require data to be fit to periodic functions that use trigonometric functions as a model. Some examples of these phenomena include

- measurements of average temperature in a city over time
- analyzing time series data from the *frequency domain* point of view
- tidal ebbs and flows
- carbon dioxide or ozone concentrations in the atmosphere.

The General Model

All of these situations involve a generic sine equation with four parameters:

$$y = \mathbf{A} * sin(\mathbf{B} * (x + \mathbf{C})) + \mathbf{D} + error$$

You probably recall from basic trigonometry that the coefficients in this model have standard names and definitions:

A is amplitude—the vertical distance from the mean of a sine wave to its peak (either positive or negative).

B reflects period—the range of X values for a complete cycle

C represents horizontal phase shift, meaning where does a given sine wave begin relative to its mean.

D represents vertical shift from zero to the mean of a given sine wave.

It is important to know the meanings of these parameters before investigating a model because initial guesses are needed for each of them. Knowing how the parameters affect the model allows for more informed guesses.

The Data

For this article, let's examine average temperature readings for Raleigh, North Carolina over a two-year period. The goal is to provide a model that can be used to predict certain temperature-dependent events. For example, farmers want to plant certain crops only when the average temperature is above 40 degrees. The temperature data for this example is shown in **Figure A**. It has a numeric month variable (beginning with 1 as January) and a corresponding temperature variable—clearly a nonlinear cyclical model; sort of a sinusoidal wave.

Nonlinear Fitting for Sine Regression

The Nonlinear Platform is a perfect tool to investigate these data. The nonlinear fitting process uses an iterative algorithm to determine the values of **A**, **B**, **C,** and **D** appropriate for the data. The Nonlinear Platform requires a general model to be entered in a separate column in the data table. So, the first step is to create a third, empty column in the data table to contain the nonlinear formula shown in **Figure A** for the generic sine equation.

Use Cols→New Column to create a new column for the nonlinear formula. In this example, the column name is X Formula f(month), to make it easy to remember that it is the X formula. Select Formula from the list of New Properties that show in the New Column dialog, then click Edit Formula to open the Formula Editor.

By default, the upper-left list in the Formula Editor lists the table columns. To create the four parameters for the sine formula, select Parameters from the popup menu as shown in **Figure B**, then click New Parameter. Name the parameter and click OK. Do this for each of the parameters, **A**, **B**, **C,** and **D**. Leave the initial values blank; you can enter them later in the Nonlinear Fitting Control Panel.

Next, enter the general sine formula into the Formula Editor. You can build the formula by highlighting terms in the formula and selecting variables, parameters, and operators as needed. Alternatively, you can double click on the formula to show a text editing box and type the formula as a scripting command.

```
A * Sine(B * ( :Month + C)) + D
```

(continued next page) ➔

Figure A *Data Table with Monthly Temperature Values: Use the Formula Editor to create parameters (Figure B) and build the sine formula*

	Month	Temp (F)	X Formula f(month)
1	1	38.9	0
2	2	42.0	0
3	3	50.4	0
4	4	59.0	0
5	5	67.0	0

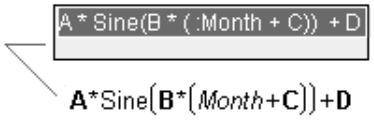
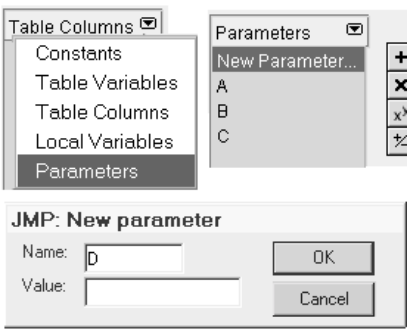


Figure B *Create Parameters in the Formula Editor*



(continued from previous page)

The next step is to choose Analyze→Nonlinear Fit. Select X Formula f(month) as **X, Predictor** in the Nonlinear Role dialog, and Temp (F) as **Y, Response**.

Initial Values for Parameters

The Nonlinear Platform attempts to find values for the parameters that give a good fit for the data points. This usually requires giving parameter starting values to the nonlinear platform. Since the guess for the temperature pattern is that it is a sine wave, the data can give hints about reasonable starting values. For example, a good way to start any analysis is to do a distribution of the response and look at summary statistics. **Figure C** shows the Moments and Quantiles from a Distribution report for the temperature variable.

Figure C *Summary Statistics and Parameter Values*

he Distribution of emp(F) gives clues bout starting values or the parameters.

Moments

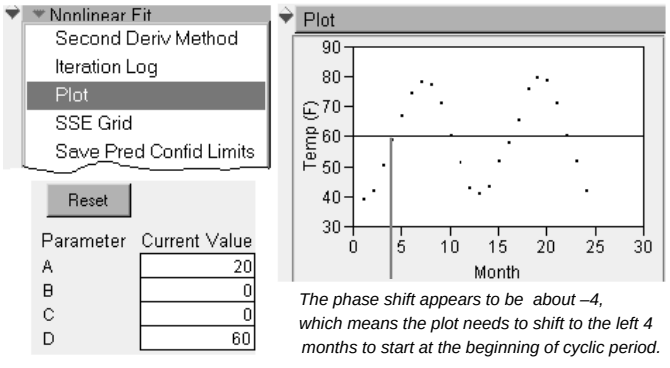
Mean	59.60417
Std Dev	14.01935
Std Err Mean	2.86169
upper 95% Mean	65.52397
lower 95% Mean	53.68436
N	24.00000

Quantiles

100.0%	maximum	79.800
99.5%		79.800
97.5%		79.800
90.0%		78.500
75.0%	quartile	73.550
50.0%	median	59.550
25.0%	quartile	45.075
10.0%		41.200
2.5%		38.900
0.5%		38.900
0.0%	minimum	38.900

The Mean is about 60, which is an estimate of the **D** parameter described above—the shift from zero to the mean of the data. The maximum is about 80. The difference between the mean and the maximum is 20, which is an estimate of the **A** parameter, amplitude. Enter those values for **A** and **D** into the control panel and click Reset to see the plot in **Figure D**.

Figure D *Summary Statistics and Parameter Values*



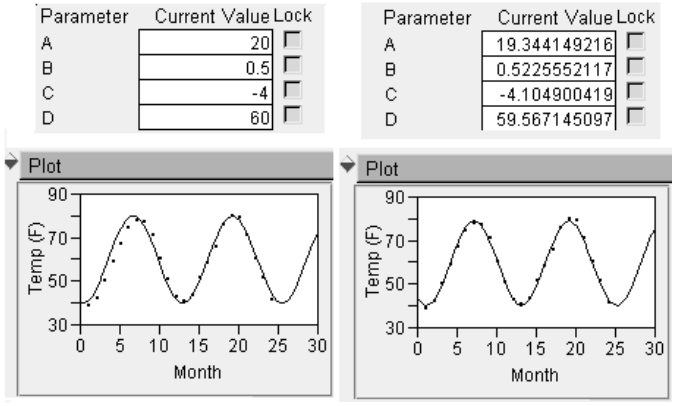
The phase shift appears to be about –4, which means the plot needs to shift to the left 4 months to start at the beginning of cyclic period.

The plot of the data gives heuristic information about the other two parameters. The beginning of the curve appears to be between 4 and 5; if you shift the plot to the left on the X axis by about 4, the curve begins at the mean. This gives an initial value for the **C** parameter (phase shift) of –4.

The period of the example data is 12 months. An estimate for the **B** parameter is found by computing $2\pi/12$, which is about 0.5. To continue with the nonlinear fitting process, enter these best guesses for **B** and **C** into the control panel and again click Reset. The results are shown on the left in **Figure E**. These common-sense starting values for the parameters give a very good starting place for nonlinear parameter estimation.

To finish, click Go on the Nonlinear Fitting Control Panel. The final parameter estimates given by the nonlinear fitting process are shown on the right in **Figure E**.

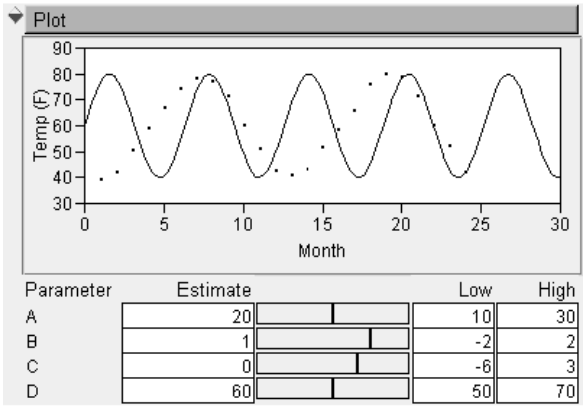
Figure E *Starting Values and Final Parameter Estimates*



A Further Note on Initial Parameter Values

Not all models are as neat as a trigonometric function, where parameters are easy to understand and readily provide starting values. The Nonlinear Fitting Control Panel has parameter sliders beneath the plot for tuning initial values until there is an acceptable fit showing in the plot. **Figure E** shows the plot for the temperature example when **B** is set to 1 and **C** is 0. Moving the **B** and **C** sliders to the values described above gives the plot in **Figure F**. The parameter sliders are useful and fun—give them a try!

Figure F *Parameter Sliders to Set Values*



ISSUE 7 FALL 2000

EDITOR
Ann Lehman

CONTRIBUTORS

Mark Bailey
Lee Creighton
Ann Lehman
John Sall

PRINTING
SAS Institute Print Center

JMPer Cable is sent only to JMP users who are registered with SAS Institute. If you know of JMP users who are not registered, pass them a copy of JMPer Cable and let them see what they are missing!

If you have questions or comments about JMPer Cable, or want to order more copies, write to

JMPer Cable
SAS Institute
SAS Campus Drive
Cary, NC 27513

To order a copy of JMP, call 1-800-727-3228. For more information on JMP, contact SAS Institute, JMP Sales
phone: 919-677-8000
FAX: 919-677-8224

FLASH

JMPer Cable is on the Web

You can now see JMPer Cable at the JMP Web site:

`http://www.jmpdiscovery.com`

If you don't keep JMPer Cable for reference, please recycle!

SAS Institute
SAS Campus Drive
Cary, NC 27513 USA
Tel: (919) 677 8000

Bulk Rate
U.S. Postage
PAID
SAS Institute Inc.