SAS INSTITUTE INC.
FALL 1997
ISSUE 5

JMPer Cable®

NEWSLETTER FOR JMP® USERS

In This Issue

A NEW ANGLE ON MATCHED PAIRS

John Sall
Senior Vice President
SAS Institute Inc.

As we wrote up the paired t-test for the *JMP Start Statistics* book we noticed that the paired t-test was very rich in opportunities for alternate interpretations. In this article we will review those alternate views and add a new view, to be implemented in Version 4 of JMP.

Different Views

Here are four ways to view a paired t-test:

- Distribution of difference: the paired t-test examines the distribution of the difference of two variables, and tests if the mean of those differences is different from zero. The mean of the differences is the same number as the difference of the means, as in the 2-group t-test, but the variance of the distribution is not the same as if it were two groups.
- Restricted regression: We adopted the scatterplot in the Fit Y-by-X (continuous by continuous) as the best place to show the paired t-test graphically. The scatterplot shows the distribution of each variable, and the correlation between them. A 45 degree line through the origin shows the boundary where the values are equal. The test for the mean difference is a tug of war across that boundary. If you fit a regression line that is constrained to have a slope of 1, the vertical distance between the fitted line and the line through the origin, will be the mean difference, and the test for it is equivalent to the paired t-test (**Figure A** left plot).
- MANOVA: If you regard both variables as responses in a MANOVA model that has no effects other than an intercept, and then you test a contrast across the two intercepts, this will be equivalent

➡

to the paired t-test. The paired t-test is the simplest MANOVA model.

- Two-way ANOVA: You stack the data, mapping the values from the two columns into one column with twice as many observations. You now run a two-way model without interaction, where one term identifies the original column from which the value came, and the other term identifies the original observation that the value came from. You might call them the within-subject and between-subjects effects. The resulting F test on the first term is equivalent to the original paired t-test. The square root of the F value will match the t-value.

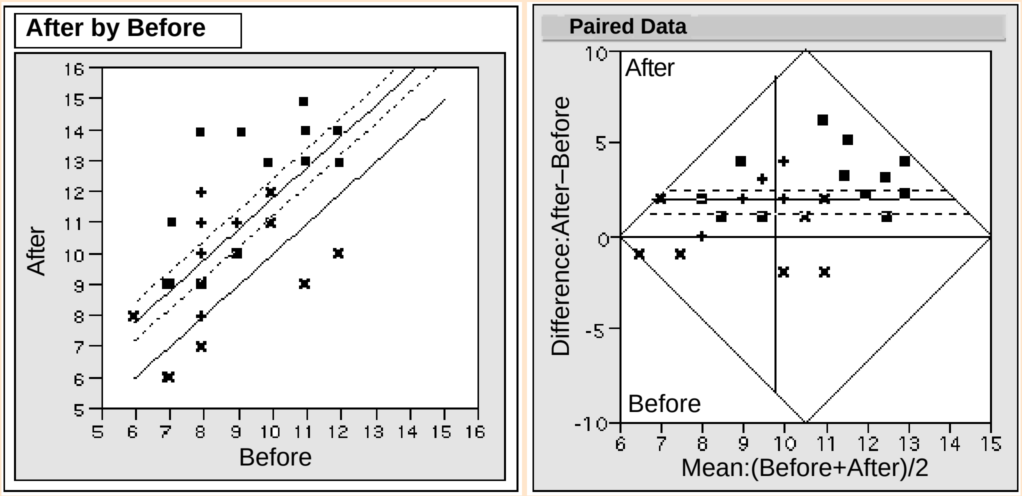
Before and After, Version 3 and Version 4

There are at least these four ways to do a paired t-test in version 3 of JMP. The problem was that users had a hard time finding these ways. The preferred way was in continuous-by-continuous Fit Y-by-X platform, which is the same platform that does simple regression. With version 4 of JMP, we resolved to find a more prominent place for the paired t-test that users could easily find, a new platform called "Matched Pairs."

In version 3 of JMP, you have to tilt your head to visualize the distribution of the difference between paired values. With version 4 of JMP, we rotate the plot by 45 degrees so that the difference is now vertical. **Figure A** shows the before-and-after pictures. Remember that these are the same graphs, except for the rotation. Pick out some points and features on the left, and note that the graph on the right is the same, tilted. The immediate benefits are:

- You will find Matched Pairs on the Analyze menu, rather than having to dig for it inside Fit Y by X.
- The new platform scales the axes for X and Y, so that you don't have to manually scale plots to get axes.

Graphical t test for Version 3 (left) and Version 4 (right)



- You don’t have to turn your head. The distribution of interest is now vertical. You are testing that it’s mean is not equal to zero, the horizontal line across the middle. The mean difference line is the line above the middle, with two dotted 95% confidence limits shown above and below it. The test is significant at 5% if the confidence interval does not contain the zero line.

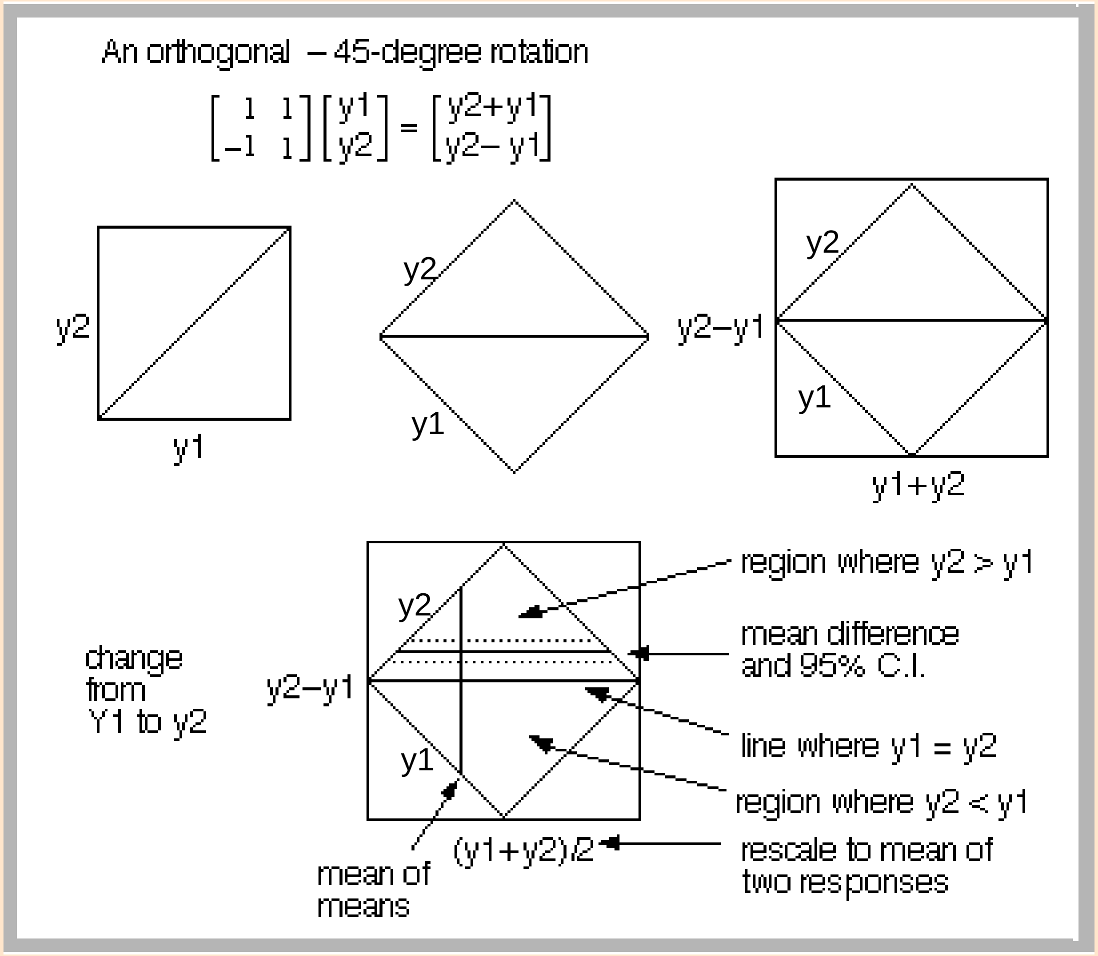
Notice that the variance of the difference is heavily dependent on the correlation between the two variables. You have positively correlated data, it will tend to be close to the line of fit, and the variance of the difference, which is the distance away from the line, will be relatively small. If you have negatively correlated data, it will tend to vary in the other direction away from the line of fit, and the variance of the difference will be relatively large. On the new rotated plot, these patterns will now be on the vertical and horizontal directions directly, rather than on the 45 and –45 degree directions.

The New Angle

Now we can notice that the new angle helps unify the different views of the analysis. A 45-degree negative rotation

Figure B

Illustration of a Negative 45-Degree Rotation


An orthogonal –45-degree rotation
$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} y1 \\ y2 \end{bmatrix} = \begin{bmatrix} y2+y1 \\ y2-y1 \end{bmatrix}$
y2
y1
y2
y1
y2
y2–y1
y1
y1+y2
change from Y1 to y2
y2
y2–y1
y1
(y1+y2)/2
mean of means
region where y2 > y1
mean difference and 95% C.I.
line where y1 = y2
region where y2 < y1
rescale to mean of two responses

is the same as multiplying the data by a certain “orthogonal” matrix.

The first coordinate of the rotation is the sum Y1+Y2, which becomes the new horizontal axis. The second coordinate is the difference Y2-Y1, which becomes the new vertical axis. So the 45 degree rotation is equivalent to changing two variables into the sum and difference (see **Figure B)**.

Where have we seen this before?

If you consider this as a MANOVA, then the first coordinate is the SUM response function, and the second coordinate the CONTRAST response function. This is the classical pattern of a repeated measures analysis. The alternate representation of that is as a split plot. The between-subjects is the whole plot effect. The within-subjects is the sub plot effect.

You can still see a plot of the original data by rotating your head 45 degrees. You can see which points are low in Y1 and high in Y2 and so on. But the focus is now on sum and difference, and even a relationship between them.

There are many possibilities for making statements regarding the patterns to be discovered in the new coordinates. For the six following examples (see **Figure C**), we have:

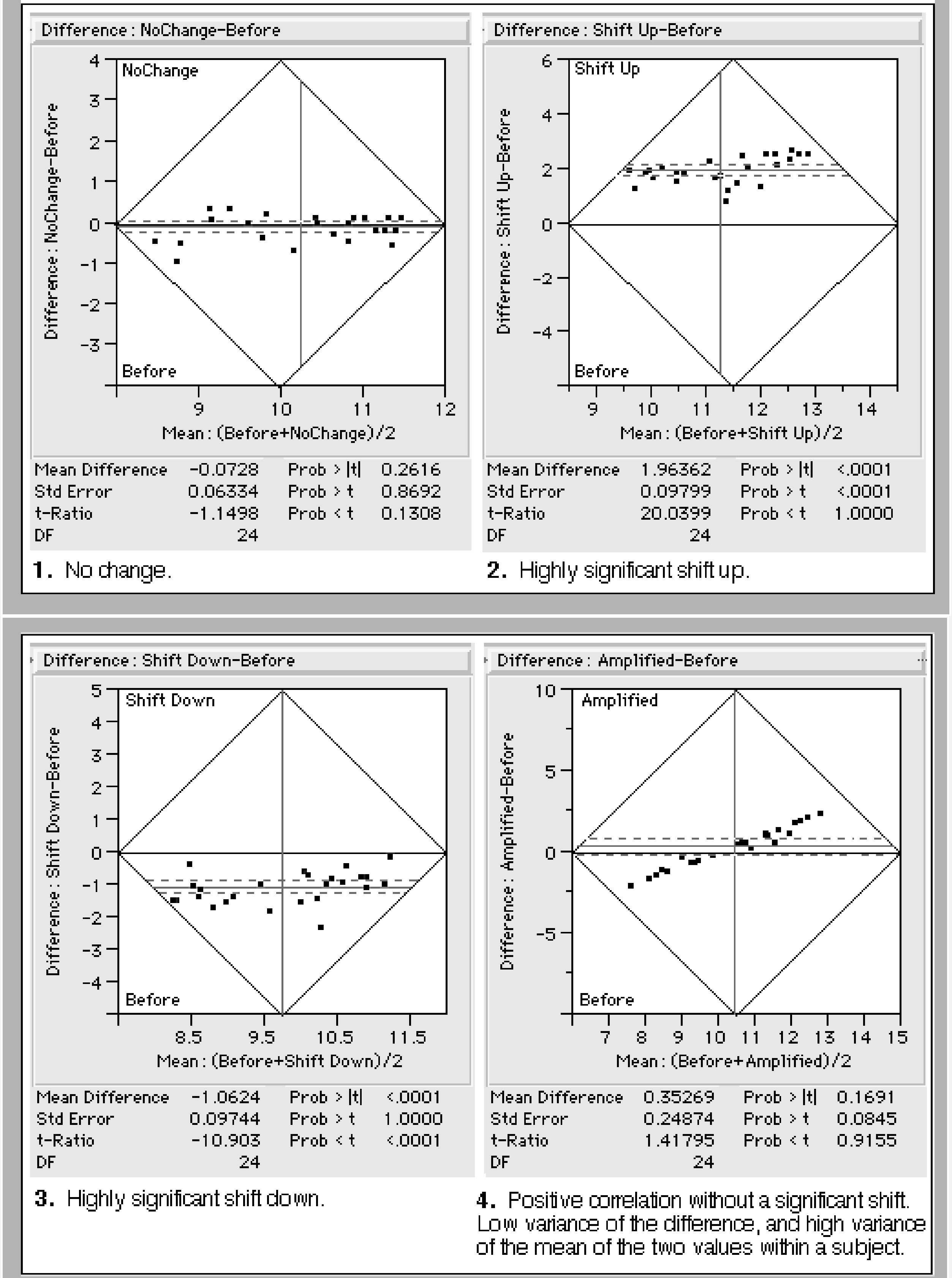
1. No Change. The distribution vertically is small, and centered at zero. The change from Y1 to Y2 is not significant. This was the high-positive-correlation pattern of the original scatterplot, and is typical.
2. Shift Up. The Y2 score is consistently higher than Y1 across all subjects.
3. Shift Down. The Y2 score is consistently lower than Y1 across all subjects.
4. No average shift, but amplified relationship. Overall the mean is the same from Y1 to Y2, but individually the high scores got higher, and the low scores got lower.
5. No average shift, but reverse relationship. Overall the mean was the same from Y1 to Y2, but the high Y1s got low Y2s and vice-versa. This was the high-negative-correlation pattern of the original scatterplot and is unusual.
6. No average shift, but damped relationship. Overall the mean was the same from Y1 to Y2, but the high scores went down a little, and the low scores went up a little.

It is interesting that rotation turns correlation phenomena into variance phenomena and vice versa.

And of course the real value of the analysis might be to discover individuals that don’t fit the pattern such as individuals that got better when others got worse.

So this is one more small chapter in the continuing quest for JMP to provide illumination and innovation in the discovery and understanding process.

Figure C Comparison of Six Matched Pairs (see previous list)


Difference : NoChange-Before
NoChange
Before
Difference : NoChange-Before
Mean : (Before+NoChange)/2
4
3
2
1
0
-1
-2
-3
9
10
11
12
Difference : Shift Up-Before
Shift Up
Before
Difference : Shift Up-Before
Mean : (Before+Shift Up)/2
6
4
2
0
-2
-4
9
10
11
12
13
14
Difference : Shift Down-Before
Shift Down
Before
Difference : Shift Down-Before
Mean : (Before+Shift Down)/2
5
4
3
2
1
0
-1
-2
-3
-4
8.5
9.5
10.5
11.5
Difference : Amplified-Before
Amplified
Before
Difference : Amplified-Before
Mean : (Before+Amplified)/2
10
5
0
-5
7
8
9
10
11
12
13
14
15

```
Mean Difference   -0.0728   Prob > |t|   0.2616
Std Error         0.06334   Prob > t     0.8692
t-Ratio           -1.1498   Prob < t     0.1308
DF                     24
```

1. No change.

```
Mean Difference   1.96362   Prob > |t|   <.0001
Std Error         0.09799   Prob > t     <.0001
t-Ratio           20.0399   Prob < t     1.0000
DF                     24
```

2. Highly significant shift up.

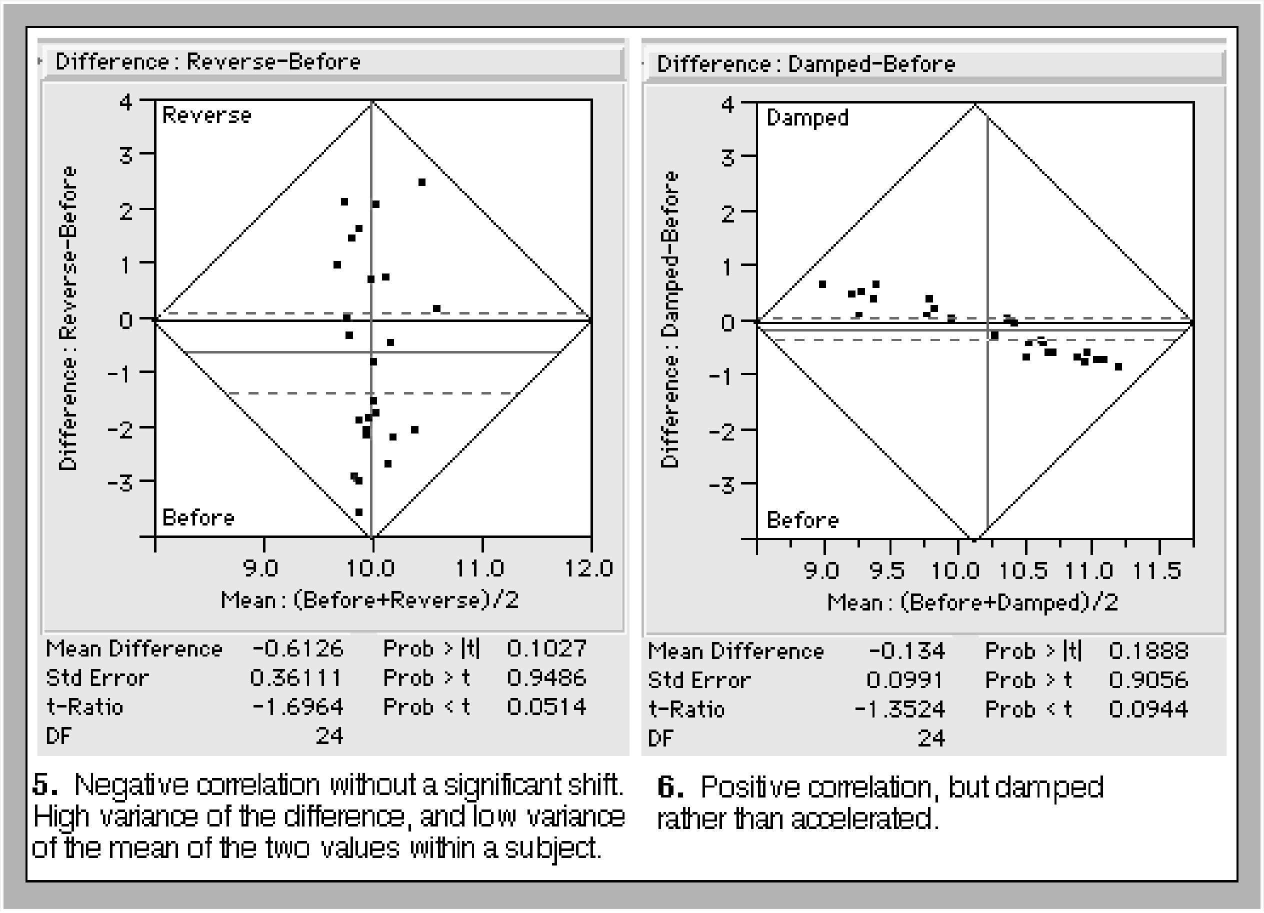
```
Mean Difference   -1.0624   Prob > |t|   <.0001
Std Error         0.09744   Prob > t     1.0000
t-Ratio           -10.903   Prob < t     <.0001
DF                     24
```

3. Highly significant shift down.

```
Mean Difference   0.35269   Prob > |t|   0.1691
Std Error         0.24874   Prob > t     0.0845
t-Ratio           1.41795   Prob < t     0.9155
DF                     24
```

4. Positive correlation without a significant shift. Low variance of the difference, and high variance of the mean of the two values within a subject.

Figure C (continued)


Difference : Reverse-Before
Reverse
Before
Difference : Reverse-Before
4
3
2
1
0
-1
-2
-3
9.0
10.0
11.0
12.0
Mean : (Before+Reverse)/2
Difference : Damped-Before
Damped
Before
Difference : Damped-Before
4
3
2
1
0
-1
-2
-3
9.0
9.5
10.0
10.5
11.0
11.5
Mean : (Before+Damped)/2

Mean Difference	-0.6126	Prob > |t|	0.1027
Std Error	0.36111	Prob > t	0.9486
t-Ratio	-1.6964	Prob < t	0.0514
DF	24		

Mean Difference	-0.134	Prob > |t|	0.1888
Std Error	0.0991	Prob > t	0.9056
t-Ratio	-1.3524	Prob < t	0.0944
DF	24		

5. Negative correlation without a significant shift. High variance of the difference, and low variance of the mean of the two values within a subject.

6. Positive correlation, but damped rather than accelerated.

Notes: The tilt idea has been done for other situations. You can also tilt for comparing means in groups. For more information see the following references:

Chambers, Cleveland, Kleiner and Tukey (1983) *Graphical Methods for Data Analysis*, page 304, Duxbury Press,CA.

Hsu and Peruggia (1994) “Graphical Representations of Tukey’s Multiple Comparison Method,” *Journal of Computational and Graphical Statistics*,
June, v3 p 156.

Congratulations to ASA Winners

These students were finalists in the 1997 Mu Sigma Rho College Bowl at the Anaheim, California American Statistical Association conference, and each won a copy of the professional version of JMP:

A. Oedekoven *University of Nebraska*	*George Capuano* *University of Nebraska*
John Castelloe *University of Iowa*	*Tiare Stone* *Brigham Young University*

"without reservation, the clinical trials that we did, and the advances that we made for the health of people, never would have been possible had it not been for JMP"

JMP DATA DISCOVERY CONFERENCE

Professional Services Division
SAS Institute Inc.

JMP software users from across the country met July 15th-18th at SAS Institute Inc.'s Training Center in Cary for the second JMP Data Discovery Conference. The stimulating, four-day interactive training conference gave attendees the opportunity to come together and share ideas about their experiences with JMP Statistical Discovery Software, and interact one-on-one with John Sall, Senior Vice President and Co-founder of SAS Institute Inc. and the JMP software development team.

Highlights of the conference included:

- a new JMP Statistical Discovery Software training curriculum
- Version 4 preview led by John Sall and the JMP development team
- keynote speakers Drs. Lenore and Leonard Herzenberg, world-renowned geneticists who use JMP software in their research efforts.

During her presentation, Dr. Lenore Herzenberg stated that "without reservation, the clinical trials that we did, and the advances that we made for the health of people, never would have been possible had it not been for JMP (software)." She also said that "discovery was a very important part of statistics."

Complimenting the technological aspects of the conference were an informal question-and-answer lunch with John Sall, a SAS campus tour led by Koka Booth of the Public Affairs Department, and an evening dinner event which featured a talk about the Triangle by Teresa Tesh, Director of North American Sales Operations.

Mark your calendar for the next exciting JMP Data Discovery Conference to be held at SAS Institute in Cary from October 28 to October 31, 1997. Bob Stuart, from Motorola University in the College of Technology, will be the Invited Speaker. His areas of expertise are Six Sigma Quality, Robust Design, Process

Improvement, Optimization and Total Process Control. Bob will address the conference on "Integrating Statistical Thinking in the Culture."

For more information about the upcoming conferences or to register, call

919-677-8000 X5005

or send a FAX to

919-677-8225

SAS INSTITUTE ACQUIRES ABACUS CONCEPTS

Cary, NC - (September 26, 1997) SAS Institute announced that it has acquired a market-leading desktop statistical analysis package for the life sciences market, StatView® software from Abacus Concepts of Berkeley, CA.

Abacus Concepts has been a leading developer of easy-to-use statistical software for more than ten years. It is often used by researchers who do statistical analysis as just one part of their jobs such as scientists, business analysts, educators, and so on. First developed for Macintosh computers,

StatView has received more awards from the Macintosh press than any other desktop statistics package. StatView is now available for both Macintosh and Windows.

John Sall, cofounder of SAS Institute and Senior Vice President, said, “We’ve had a long, amicable relationship with Abacus. Statview has a graphical user interface that makes the product extremely easy to use for researchers who are not trained statisticians. It also has phenomenal presentation graphics.”

This technology acquisition complements SAS Institute's statistical offerings, SAS® software, for analysts who need a tool to use in a client/server environment and integrated with an enterprise data warehouse, and JMP® software for desktop data discovery.

SAS Institute will distribute StatView with direct sales from nearly 100 offices in 60 countries, through mail order, and internationally via a network of distributors. For more information about StatView, please visit the following Web site:

`<http://www.abacus.com>`

FINDING THE AREA UNDER A CURVE USING JMP AND A TRAPEZOIDAL RULE

by Nicole Hill Jones
SAS Institute Inc.

Researchers who conduct clinical trials often have to measure the concentration of a new investigational drug in a subject's blood and compare it to baseline values. To do this, data are measured at discrete time points over a specified interval. The measured data represent the body's reaction as evidenced by an increase or decrease in the concentration of a particular substance. If you give a drug by IV injection, collect blood samples at various times, and measure the plasma concentrations of the drug, you might expect to see a steady decrease in concentration as the drug dissolves. **Figure A** shows data and the plasma concentration time curve from D.W.A. Borne (1995).

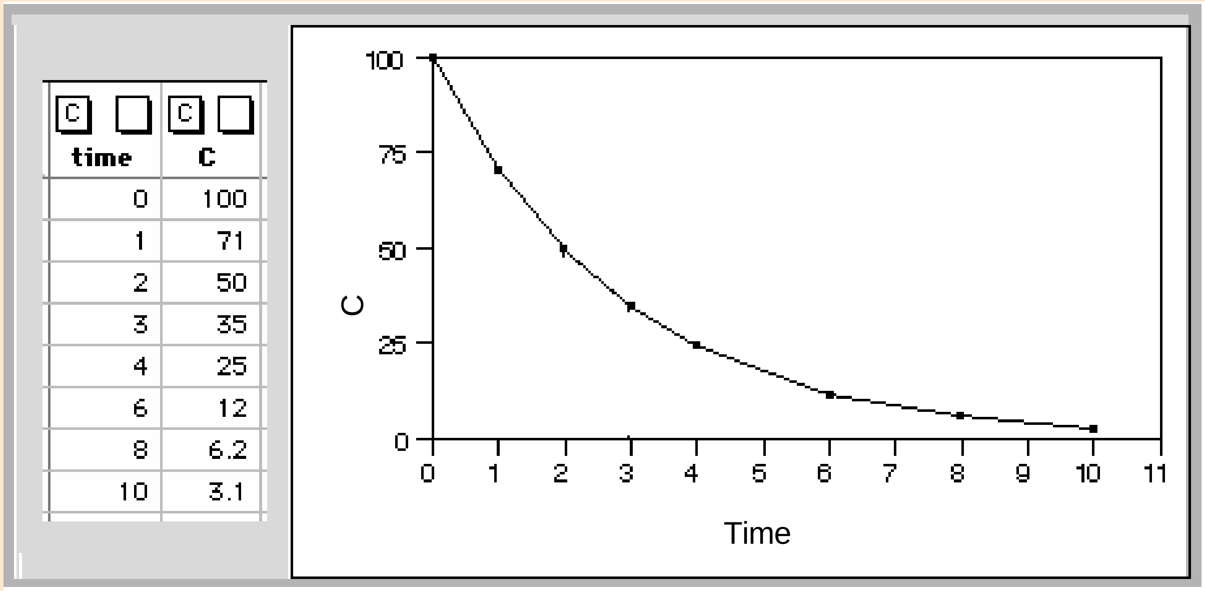
The following example illustrates how to find the area under the plasma concentration time curve. This method uses a variation of the Trapezoidal Rule. The function of the time curve is divided into segments or periods between each time point, the area in each segment is calculated and then the pieces are added together to obtain the total area under the curve.

If each time segment is considered a trapezoid, its area is given by the segment width and the average concentration within the segment width. Let's define t_i as the *i*th time point where time is given in hours, and C_i as the drug concentration at the *i*th time point. So the concentration between the 2nd and 3rd time period, $AUC_{2,3}$, is computed:

$$AUC_{2,3} = \frac{C_2 + C_3}{2} \cdot (t_3 - t_2)$$

as illustrated in the **Figure B**.

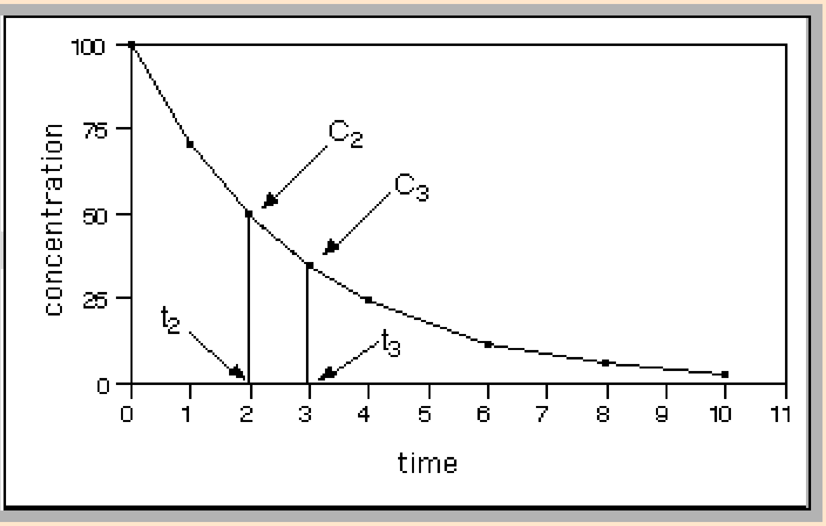
Figure A Plot of Concentration vs. Time



time	C
0	100
1	71
2	50
3	35
4	25
6	12
8	6.2
10	3.1

Figure B

Concentration (ug/ml) By Time In Segments



This works for all segments except the last, which is the time segment from 10 hours (in this example) until the drug has completely dissolved. The area under the last segment C_n can be calculated using the *elimination rate constant* (kel), which is computed from the sample data.

The kel constant is the negative slope of the relationship of the log of concentration and time. This can be computed in JMP using the standard formula for the slope in a simple linear regression of Y on X:

$$\frac{\sum (x_i - \bar{x})\ (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

You do this for the concentration example as follows:

1) Create a new column in the data table (call it ln(C)) and use the Natural Log function in the calculator Transcendental functions to compute the log of concentration.

2) Create another new column (call it beta) and use the calculator to compute the slope parameter of the regression of time on ln(C) as shown below.

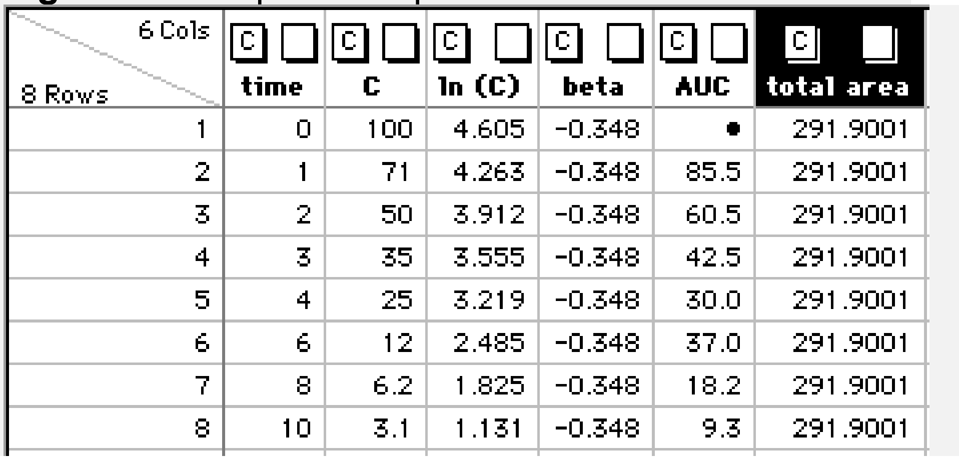
This calculator formula to compute beta uses an assignment function to create a temporary variable for the numerator (XY) and for the denominator (x2) of the slope computation. It then computes and assigns XY/x2 to the result.

$$\begin{aligned} &XY \Leftarrow \sum_{i=1}^{\blacksquare} \left(\left(time_i - \overline{time}\right) \cdot \left(ln\ (C)_i - \overline{ln\ (C)}\right)\right) \\ &x2 \Leftarrow \sum_{i=1}^{\blacksquare} \left(time_i - \overline{time}\right)^2 \\ &\text{results } \frac{XY}{x2} \end{aligned}$$

Now you can find the total area under the curve by summing the areas of the individual segments and the computed final segment:

$$\sum_{i=1}^{\blacksquare} AUC_i + \left(\frac{C_{\blacksquare}}{-\,beta}\right)$$

Figure C Computed Slope Parameter and Total AUC



6 Cols / 8 Rows	time	C	ln (C)	beta	AUC	total area
1	0	100	4.605	-0.348	•	291.9001
2	1	71	4.263	-0.348	85.5	291.9001
3	2	50	3.912	-0.348	60.5	291.9001
4	3	35	3.555	-0.348	42.5	291.9001
5	4	25	3.219	-0.348	30.0	291.9001
6	6	12	2.485	-0.348	37.0	291.9001
7	8	6.2	1.825	-0.348	18.2	291.9001
8	10	3.1	1.131	-0.348	9.3	291.9001

You can reuse this data table by saving it as a template—just delete all the rows and save the empty table with its formulas. To use the template, open it and paste your data into the time and C columns. You can rename the columns to fit your data, and the formulas automatically use the new column names and compute the AUC you are looking for.

References

Bourne, D.W.A. (1995), "A First Course in Pharmacokinetics & Biopharmaceutics", University of Oklahoma College of Pharmacy.

`http://157.142.72.143/gaps/pkbio/Ch04/Ch0403.html`

JMP *Statistics and Graphics Guide*, Version 3.1, (1995), SAS Institute, Inc.

Calculator Corner

$$\sum_{j=1}^{i}\begin{cases}\dfrac{height_i}{height_j}, & \text{if } (generation_j = \text{“father”}) \text{ and } (family\ id_i = family\ id_j)\\ \square, & \text{otherwise}\end{cases}$$

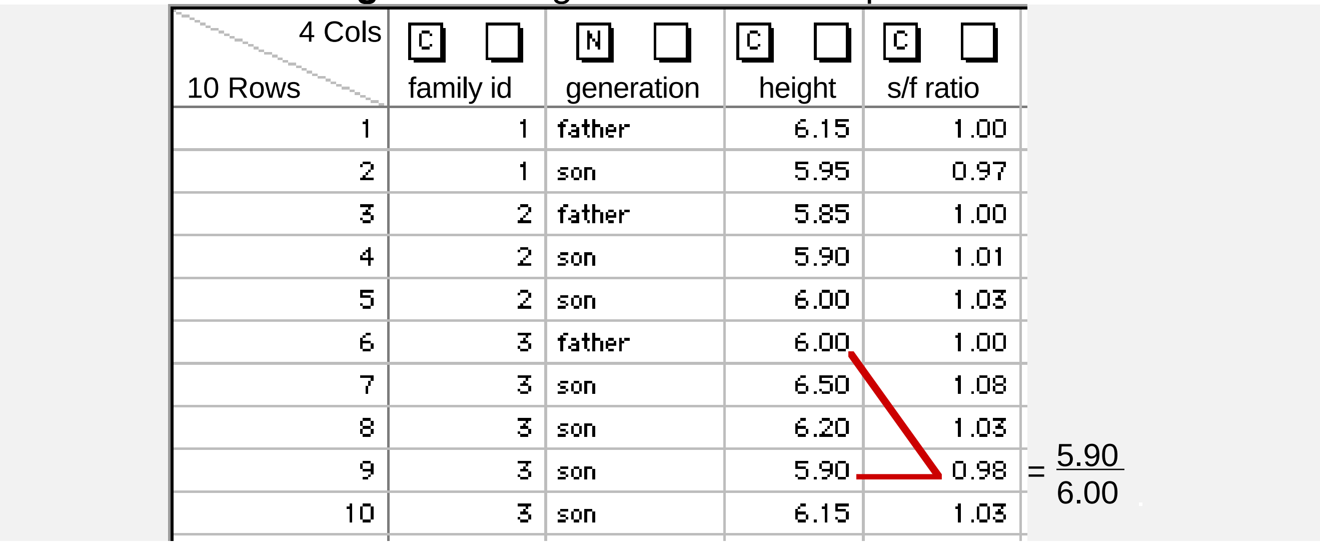
by Michael Hecht
SAS Institute Inc.

SUMMATION SIDE EFFECTS

The formula shown above is deceptively complex looking. Its purpose is simply to scan columns of values. When it finds what it is looking for it computes a ratio and assigns it to the column it is building (s/f ratio shown in **Figure A**).

This table has rows for a father followed by a row for each of his sons, with a family id and height measurement for each row. Suppose you want to compute the ratio of each son to his father (possibly for the purpose of testing whether height has increased over a generation.)

Figure A Diagram of Ratio Computations



4 Cols / 10 Rows	family id	generation	height	s/f ratio
1	1	father	6.15	1.00
2	1	son	5.95	0.97
3	2	father	5.85	1.00
4	2	son	5.90	1.01
5	2	son	6.00	1.03
6	3	father	6.00	1.00
7	3	son	6.50	1.08
8	3	son	6.20	1.03
9	3	son	5.90	0.98
10	3	son	6.15	1.03

One easy way to do this is to create a column (s/f ratio) and use the formula shown at the top of this article. This formula makes use of the summation function to scan columns of values. Scanning for a value is a logical side effect of the summation function when its argument is a condition rather than one or more simple subscripted variables. The process for finding s/f ratio goes like this:

For each row the summation function begins at the first row of the table and looks at the values of generation and family id. It continues until it locates the row where the value of generation is "father," and the family id matches the current row's family id. It then stops and computes the ratio of the current son's height (height_i) to his father's height (height_j).

As the formula stands it simply divides each father's height into itself giving 1, which could be eliminated by including an if or match conditional statement to exclude computing any value for fathers. This would be desirable if you wanted to compare the mean ratio to 1 and draw an inference about the change in heights over the father-son generation.

The next article, *Just Mooning Around,* uses a more elaborate form of this logic to compute the ratio of satellites to their respective parent planets, and planets to the sun.

JUST MOONING AROUND

by Ann Lehman
SAS Institute Inc.

The name of this article, *Just Mooning Around* is the same as an article published by Dr. Isaac Asimov in 1975, in which he conjectures that our moon is not a satellite of the earth, but rather “a planet in its own right, moving about the Sun in careful step with the earth.” He supports this conjecture admirably with a dozen or so pages of explanation, formulae, and things to visualize (if you can). The central support is a computation he called the tug-of-war (TOW) value, which is the ratio of the gravitational pull by a satellite’s parent, F_P, to the gravitational force between the satellite and the sun, F_S.

You can picture this ratio (F_P/F_S) as a tug of war going on for a satellite with its planet on one side of the gravitational rope and the sun on other. If this ratio is greater than 1 the parent planet is winning, if less than 1 the sun is winning. You expect all satellites to have a TOW value greater than 1. If a TOW value is less than 1, the moon wannabe falls away from its parent planet and into the sun or into its own orbit around the sun. Dr. Asimov shows that the formula to compute a satellite’s TOW value is:

$$\left(\frac{m_p}{m_s}\right)\cdot\left(\frac{d_c}{d_p}\right)^2$$

where $\mathbf{m_p}$ is the mass of the parent, $\mathbf{m_s}$ is the mass of the sun, $\mathbf{d_c}$ is the distance from the satellite to the sun, and $\mathbf{d_p}$ is the distance from the parent planet to its satellite child.

Asimov computed this TOW value for all the planetary satellites for which distance and mass data were available at that time—there were only 22 satellites that had both a name and physical information. (These did not include Pluto’s moon Charon, which was not discovered until 1978.) We were able to match these computations within reasonable limits and also compute a TOW for Charon and several other newer satellites.

Another factor Asimov used to examine satellite properties is the *Roche limit*. The Roche limit was first described by Edouard Roche in 1848.

It is the closest distance an object can come to another object without being pulled apart by tidal forces. If a planet and a moon have identical densities, then the Roche limit is 2.446 times the radius of the planet. The Roche limit for the Earth is 11,470 miles. If our Moon ever ventured this close to the earth, it would be pulled apart by tidal forces and the Earth could end up having rings. The four gaseous outer planets do have systems of rings inside of their respective Roche limits. On July 7, 1992, comet Shoemaker-Levy 9 broke into 21 pieces when it approached to within Jupiter's Roche limit. On subsequent passes, each of the pieces of the comet impacted Jupiter.

The Roche limit defines the closest a moon can exist to a planet, and the TOW can be used to compute a maximum distance a moon can be from its parent before it is captured by the sun. For example, Mars' Roche limit is 5,160 miles above the surface and the distance defined by its TOW value is 15,000 miles. Both Diemos with a distance of 14,569 miles from Mars and Phobos at 5,824 miles fall within this interval and can be classified as true satellites:

5,160 <
(both 5,824 and 14,569)
< 15,000

Another interesting example is Mercury, which has a Roche limit of 3,706 and a maximum distance defined by its TOW value of 2,675, which falls within its Roche limit. This means it is highly improbable for Mercury to ever have a satellite.

What about Earth and Luna?

For starters, the TOW value computed for the Earth and Luna is .455, which implies that Earth is losing its tug of war with the sun to keep Luna as a satellite. Next, for earth a minimum distance defined by the Roche limit is 9,734 miles, and the maximum distance defined by the TOW is 29,000. An earth satellite should lie in the small band between 9,734 and 29,000 miles from the earth's center. But as we all know the distance of the moon from earth is 237,000 miles!

So what's going on here? According to Asimov, if you draw a picture of the orbits of the Earth and Moon exactly to scale as they orbit the Sun, you would see that the Moon's orbit is everywhere concave toward the Sun; it is always falling toward the sun. All the other satellites fall away from the sun through part of their orbits and are contained by the superior pull of their parent. So it seems Luna does not orbit the earth as its satellite, but rather the earth and Luna orbit the sun together as a binary system .

You can find the details of these observations and a couple of other arguments Dr. Asimov uses to question the cosmic relationship between the earth and its moon in his article.

What about Pluto and Charon?

Although Dr. Asimov didn't have data on Pluto and Charon he would have been pleased when his neat logic supported the moonness of Charon; Charon is nearly half the size of Pluto, and its TOW value computes as a respectable 601.66. Further, it is 12,196 miles from Pluto, nicely outside Pluto's Roche limit of 1,784 miles and inside the maximum allowable distance of 54,619 miles. Nevertheless, astronomers today prefer to view Charon and Pluto as a binary system because:

- They are so close together and so nearly the same size that they both orbit the center of mass that lies between them, and that center of mass orbits the sun. Although this is true for all planets and their moons, the relative sizes of the bodies and the distances between them places their joint center of mass near the center of the planet.
- Further, they orbit each other with the same sides permanently facing each other, which is also like double-planet (binary) systems.

A JMP Look at Moons and Planets

This brings us to a JMP view of the solar system. Exploring the moons data and plotting the ratio of a child mass to its parent reveals a simple relationship that puts both the Earth/Luna and and Pluto/Charon pairs into the same category, which seems to support the astronomers and might give Dr. Asimov food for thought (most everything did).

It's easy to create a JMP table that has one row for each moon and each planet in the solar system, and using readily available information you include its mass, giving the top table shown in **Figure A**.

Note: In this example mass is transformed by ln(cube root), which makes for a more readable plot later in the discussion.

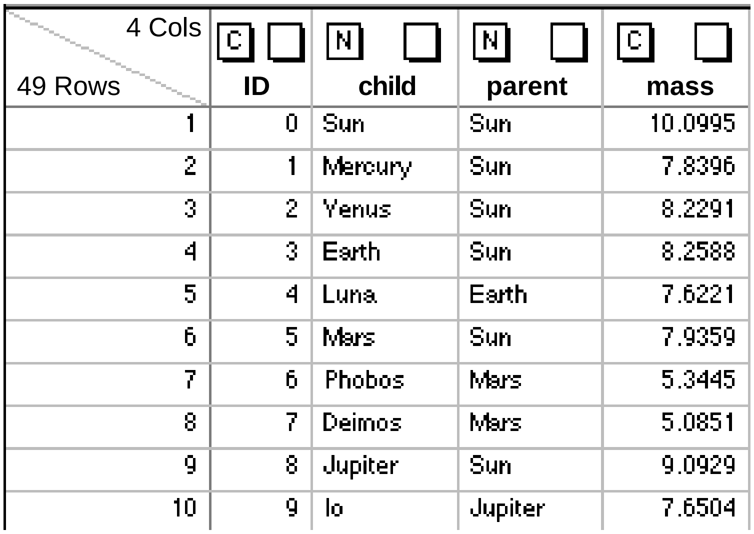
We want to compare two groups of mass ratios: the ratios of all planets and moons to their respective parent body (call this the Child group) to the group containing only the ratios of the planets to the sun (called the Parent group). Note that planetary satellites all fall only into the Child group and the sun is only in the Parent group. But the planets fall into both categories and need to be included in both groups—as children of the sun and parents to their children.

To compare these two groups of mass ratios we first stacked the child and parent columns using Tables→Stack. The new stacked column name is body and the new ID column name is type, as shown in the bottom table in **Figure A**. There are now a pair of rows (a child row and a parent row) for each single row in the top table in **Figure A**. However, the mass row duplicates the mass of the "child" for each pair. This is fine for all the “child” rows but is incorrect for the “parent” sun, and also whenever a planet is in the “parent” role. So, the computation of the correct mass ratio for each row can be found with a (not for the faint hearted) calculator formula, which does the following:

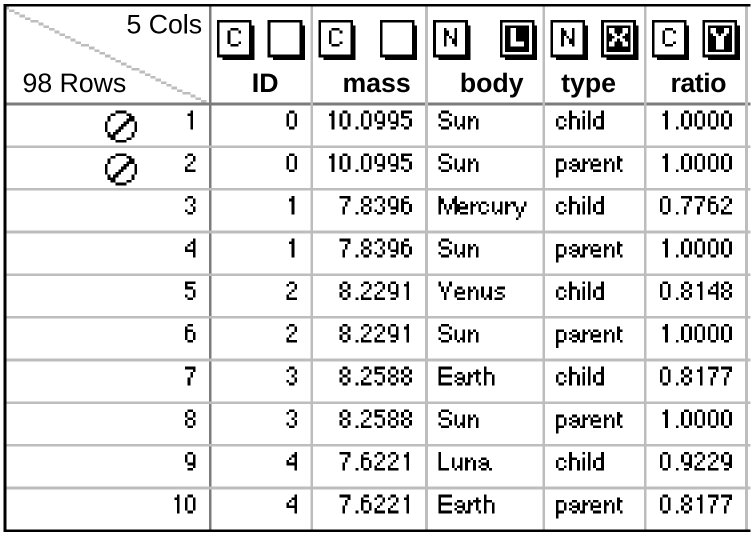
- For the purpose of the example the sun is considered its own parent and has mass ratio of one (1).
- When body is a planet, compute its mass ratio to the sun.
- When body is a moon, compute the mass ratio to its parent planet.

This appears to be a lot to ask of this data table, which doesn’t tell you whether a body (other than the Sun) is a planet or a moon. However, a calculator formula can figure that out by using the summation operator to scan the values of body and type, and thus find the correct denominator to compute the ratio column shown in the bottom table of **Figure A**.

Figure A Table of Mass Values (top) and Stacked With Computed Mass Ratios (bottom)



4 Cols 49 Rows	C ID	N child	N parent	C mass
1	0	Sun	Sun	10.0995
2	1	Mercury	Sun	7.8396
3	2	Venus	Sun	8.2291
4	3	Earth	Sun	8.2588
5	4	Luna	Earth	7.6221
6	5	Mars	Sun	7.9359
7	6	Phobos	Mars	5.3445
8	7	Deimos	Mars	5.0851
9	8	Jupiter	Sun	9.0929
10	9	Io	Jupiter	7.6504



5 Cols 98 Rows	C ID	C mass	N body	N type	C ratio
⊘ 1	0	10.0995	Sun	child	1.0000
⊘ 2	0	10.0995	Sun	parent	1.0000
3	1	7.8396	Mercury	child	0.7762
4	1	7.8396	Sun	parent	1.0000
5	2	8.2291	Venus	child	0.8148
6	2	8.2291	Sun	parent	1.0000
7	3	8.2588	Earth	child	0.8177
8	3	8.2588	Sun	parent	1.0000
9	4	7.6221	Luna	child	0.9229
10	4	7.6221	Earth	parent	0.8177

The *Calculator Corner* article in this news-letter explains how the summation operator scans a column to find a specific value.

The formula used to compute ratio is shown in **Figure C**.

The final step is to plot the child-parent pairs of ratios. To do this the we used Analyze→Fit Y by X with ratio as Y and type as X (the one-way Anova platform). The Matching Variable option from the Analysis popup menu used the ID column to match the children to their parents, giving the plot in **Figure B**.

This plot clearly shows that all the children look up to their parents except Charon and Luna.

Reference

Asimov, Isaac (1975), *Of Time, Space, and Other Things*, pgs 87-98, Avon Books, New

Figure B Two groups of Mass Ratios

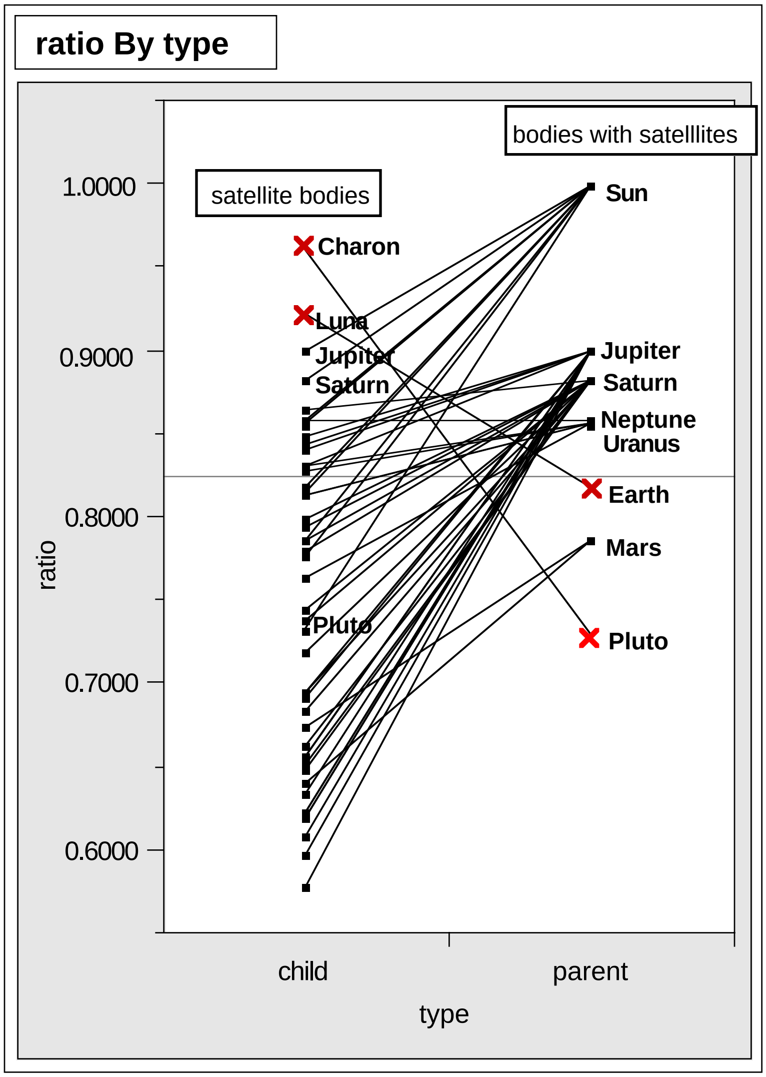
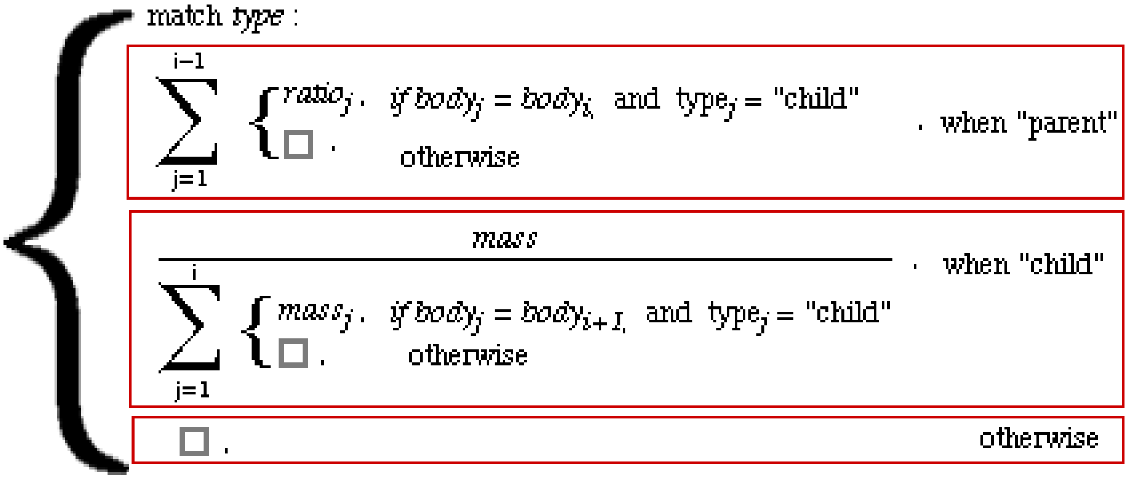


Figure C Formula to Compute Mass Ratio



match type :

$$\begin{cases} \displaystyle\sum_{j=1}^{i-1} \begin{cases} ratio_j, & if\ body_j = body_i \text{ and } type_j = \text{"child"} \\ \square, & \text{otherwise} \end{cases} & \text{when "parent"} \\ \dfrac{mass}{\displaystyle\sum_{j=1}^{i} \begin{cases} mass_j, & if\ body_j = body_{i+1} \text{ and } type_j = \text{"child"} \\ \square, & \text{otherwise} \end{cases}} & \text{when "child"} \\ \square, & \text{otherwise} \end{cases}$$

CHECKING THE PROPORTIONAL HAZARD ASSUMPTION WITH KAPLAN-MEIER SURVIVAL ESTIMATES

by Rolf E. Taffs
Food and Drug Administration

The Cox proportional hazards model (Cox, 1972) evaluates treatment, diagnostic, or prognostic factors to determine the magnitude and significance of their effects on population survival or failure time. For each factor (e.g. an exposure level or a prognostic index) a predictor variable is used to describe the subjects according to the factor level to which each subject or unit belongs. The model can be used in multivariate analysis to test the hypothesis that survival does not depend on the level of a treatment or risk factor (Kalbfleish and Prentice, 1980; Christensen, 1987).

Unlike parametric methods, which compare regression coefficients under the constraints of a given mathematical model, the proportional hazards model does not require that the precise nature of the survival function be known nor that it be constrained by the assumption of a particular mathematical form. However, the model assumes that the hazard functions for the levels of a given factor or treatment are proportional.

This assumption of proportionality must be met for the analysis to be valid. The proportional hazard can be written as the ratio of cumulative hazard functions:

$$\frac{\Lambda_A(t)}{\Lambda_B(t)}=e^{b\Sigma d}$$

where, for the factor being analyzed, **b** is the regression coefficient and **d** is the difference between subjects receiving treatment **A** or **B**. Under the assumption of proportional hazards, the corresponding proportion between the cumulative integrated hazards would be the same, and plots of the logarithm of the cumulative hazards corresponding to values differing by **d** should be parallel and approximately **b•d** apart vertically. This allows the assumption to be examined graphically when a formal statistical test is unavailable.

The hazard functions needed to perform the graphic comparison can be calculated from the Kaplan-Meier (product-limit) survival estimates S(t) obtained from the data according to the formula:

$$\Lambda(t) = \pm\ln(S(t))$$

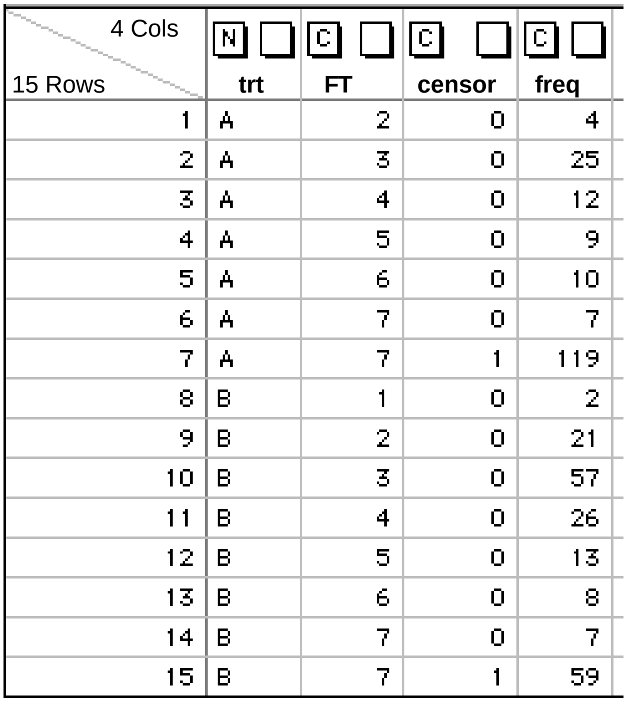
The Kaplan-Meier Method in the Survival platform in JMP generates the product-limit survival estimates needed to examine the proportional hazards assumption.

The sample data shown in **Figure A** show frequencies (Freq) of time to failure (FT) for 379 units observed for a period of seven days. Units that did not fail during the observation period are censored. In this example, the assumption of proportionality for treatments **A** and **B** are graphically examined.

Begin by choosing Analyze→Survival. In the Kaplan-Meier Method window, select FT as the Time variable, trt (treatment) as the Grouping variable, and the Censor and Freq columns as the Censor and Frequency variables. Click OK for survival rate calculations and a survival plot.

Next, select the Save Estimates option in the $ border popup window to generate a new data table that lists the cumulative survival probabilities for each treatment and for the combined data, as shown in **Figure B**.

Figure A Example Survival Data



4 Cols 15 Rows	trt	FT	censor	freq
1	A	2	0	4
2	A	3	0	25
3	A	4	0	12
4	A	5	0	9
5	A	6	0	10
6	A	7	0	7
7	A	7	1	119
8	B	1	0	2
9	B	2	0	21
10	B	3	0	57
11	B	4	0	26
12	B	5	0	13
13	B	6	0	8
14	B	7	0	7
15	B	7	1	59

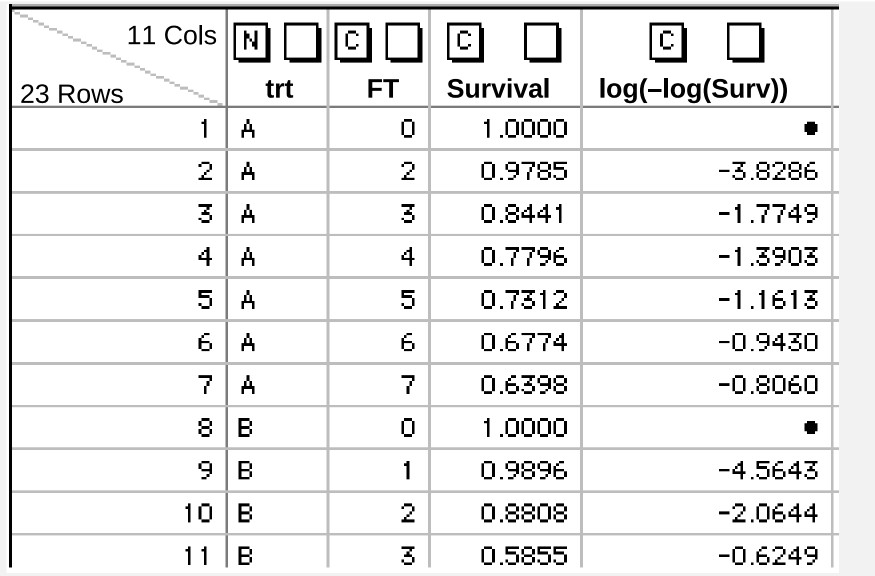
Now you overlay the values in each treatment group to graphically compare them. To do this you need first to split the table by the column you want to plot, log(–log(Survival)).

To split the table, choose Tables→Split Columns and complete the Split Columns dialog as follows.

- log(–log(Survival)) as the Split variable,
- trt as Col ID
- FT as the Group variable.

Using FT as a Group variable causes the split operation to look at FT values within treatment levels and create a row for each unique value found.

Figure B New Data Table With Survival Estimates



11 Cols 23 Rows	trt	FT	Survival	log(–log(Surv))
1	A	0	1.0000	●
2	A	2	0.9785	-3.8286
3	A	3	0.8441	-1.7749
4	A	4	0.7796	-1.3903
5	A	5	0.7312	-1.1613
6	A	6	0.6774	-0.9430
7	A	7	0.6398	-0.8060
8	B	0	1.0000	●
9	B	1	0.9896	-4.5643
10	B	2	0.8808	-2.0644
11	B	3	0.5855	-0.6249

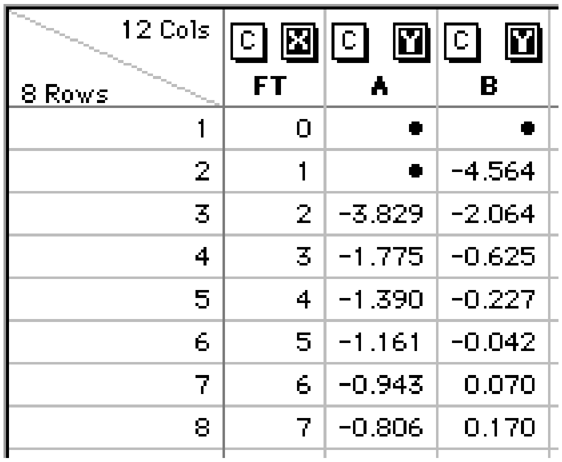
Also, click the Drop All radio button on the dialog to keep only the variables needed for the overlay plot.

When you click OK JMP splits the log(–log(Survival)) column into three columns as defined by the levels of the trt variable, and creates the new untitled table in **Figure C**. For the overlay plot, assign X and Y roles as shown.

Now choose Graph→Overlay Plot and click OK to see the plot in **Figure D**. By default, the Connect option for the Overlay Plot is in effect and the treatment groups have the markers identified in the plot legend.

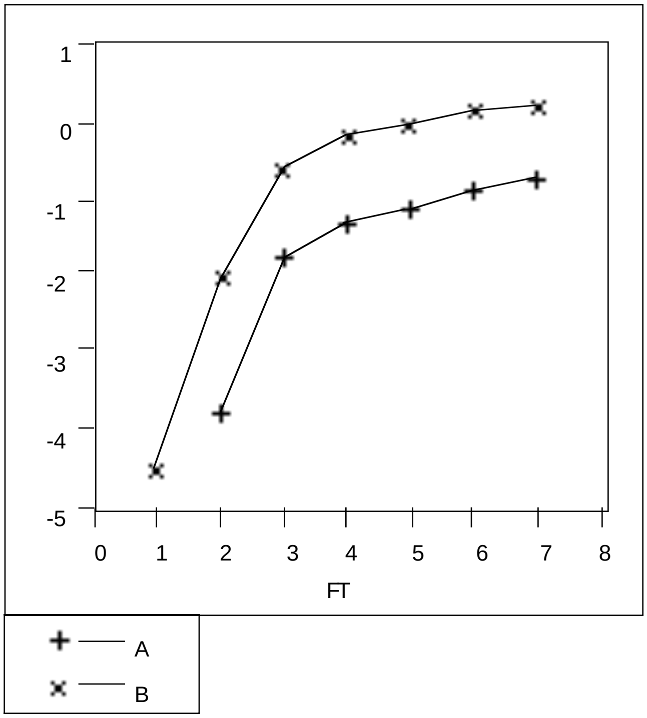
In **Figure D** you can visually examine the relationship between the treatment levels as defined by the treatment groups. The group curves should be parallel if the proportional hazards assumption is met. Curves that intersect or are decidedly nonparallel indicate that the assumption may not be met, in which case the proportional hazards analysis may be invalid. This example shows the curves to be nearly parallel (equidistant vertically at each point on the FT axis), indicating further analysis using a proportional hazards model is appropriate.

Figure C Data Table Reshaped to Overlay Survival Estimates



12 Cols 8 Rows	FT	A	B
1	0	•	•
2	1	•	-4.564
3	2	-3.829	-2.064
4	3	-1.775	-0.625
5	4	-1.390	-0.227
6	5	-1.161	-0.042
7	6	-0.943	0.070
8	7	-0.806	0.170

Figure D Comparison of Survival Curves for Two Groups



References

Cox, D.R. (1972, Regression Models in Life tables (with discussion), J.R. Statistical Society B 34:187-220

Kalbfleisch, J.D. and Prentice, R.L. (1980), The Statistical Analysis of Failure Time Data, New York: John Wiley and Sons.

Christensen, E. (1987), Multivariate Survival Analysis Using Cox's Regression Model, Hepatology 7:1346-1358.

Tips and Techniques

AN ALTERNATIVE TO JITTERING

The Jitter option was recently added to the Outlier Box plot in the Distribution platform and to the Nominal by Continuous scatterplot in the Fit Y by X platform. The Jitter option adds random horizontal jitter so that you can see points that overlay on the same Y value. The horizontal adjustment of points varies from .375 to .625 with a 4*(Uniform–.5)**5 distribution. This is helpful when you are plotting a large number of values. For smaller numbers of points, simply labeling a point with the number of values it represents is an alternative to jittering.

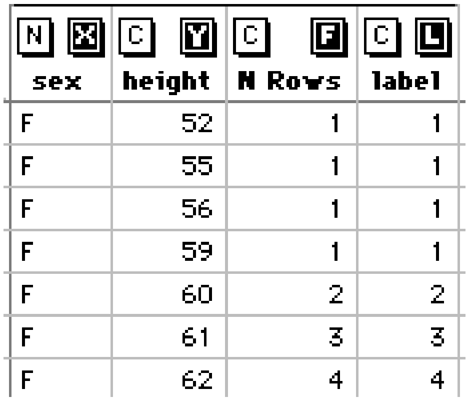
The following example uses the Height and sex variables in the BIG CLASS sample data table. Use Analyze→Fit Y by X with height as Y and sex as X to see the left-hand plot in **Figure A**.

To see the second plot in **Figure A**, choose the Jitter option from the Display popup menu for that plot. The jittered points indicate that the same height value occurs multiple times (for both males and females).

To label points instead of jittering requires several steps:

1) First, use Tables→Group/Summary to summarize the data by the X and Y variables you are using (height and sex in this example). The values in the N Rows summary column tells the number of rows in the source table with the same sex and height values (**Figure B**).

Figure B Summary Table With Label Column



sex	height	N Rows	label
F	52	1	1
F	55	1	1
F	56	1	1
F	59	1	1
F	60	2	2
F	61	3	3
F	62	4	4

2) Before doing anything else, use Tables→Subset to make a duplicate of the summary table. You do this to have a copy of the summary information that is not linked to the source table.

3) Next, duplicate the N Rows column. The easiest way to do that is to highlight N Rows and copy it to the clipboard. Then create a new column and paste the values into it. We named this new column label (**Figure B**).

4) Use the modeling type popup menu at the top of the columns to assign roles as follows:
   - height to Y (Y)
   - N Rows to Freq (F)
   - label to Label (L).

   Now the same values (N Rows and label) act as both frequencies and labels.

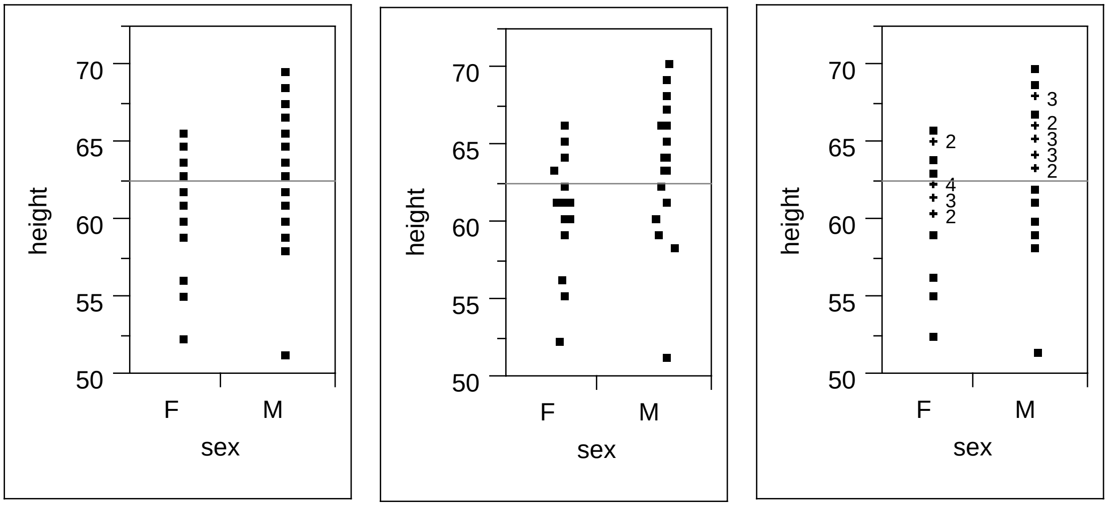
5) The last step (almost) is to create a column that labels and marks all the rows whose value of N Rows is greater than 1. To do this create a new column and specify Row State as its Data Type and Formula as its Data Source. Then use the following calculator formula to assign row state constants:

   « label(*N Rows* > 1), marker(*N Rows* > 1) »

   Use the Copy to Row State command in the popup menu at the top of the row state column to activate the row states, as shown in **Figure C**.

With this table active select Analyze→Fit Y by X to see the right-hand table in **Figure A**.

Figure A Comparison of default points (left-hand plot) with jittered points (middle plot) and labeled points (right-hand plot).



EDITOR
Ann Lehman

CONTRIBUTORS
Michael Hecht
Colleen Jenkins
Nicole Hill Jones
Ann Lehman
John Sall
Rolf E. Taffs

If you have questions or comments about JMPer Cable write to

JMPer Cable
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513

JMPer Cable is sent only to JMP users who are registered with SAS Institute.

For more information on JMP, or to order a copy, contact
SAS Institute, JMP Sales
phone: 919-677-8000 x 5071
FAX: 919-677-8224

You can also browse our web site at
`http://www.jmpdiscovery.com`