



### Inside This Newsletter

JMP Webinars p. 3

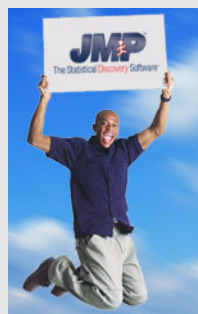
Far Out p. 6

Conferences and Trade Shows  
p. 9

Applying Benford's Law p. 9

Predicting Patterns in Child  
Genders: What's the Chance?  
p. 11

### Breaking News



JMP 5.1 will be available before 2004 begins! A few of this version's new features include:

- Linux operating system support
- A space filling design platform in Design of Experiments (DOE)
- Surface plots, parallel plots, and cell plots
- An item analysis platform
- Six Sigma tool enhancements

## Analyzing Dose-Response Curves in JMP

Jianfeng Ding, JMP Development

Mark Bailey, SAS Statistical Training and Technical Services

Bioassays are based on exposing varying amounts of an active or toxic compound to a biological entity or a surrogate system and measuring the response to this agent. The chemical reactions involved in these assays produce a response that is not linear with the dosage. JMP can fit this kind of a response with the Nonlinear platform. Several new features in JMP 5.1 make it easy to set up such an analysis and to explore the results.

### The Assay and the Model

In this example, the result of the assay,  $y$ , is the percent toxicity. A standard compound of known potency is diluted 3:1 in a series of 12 concentrations. A compound of unknown potency is similarly diluted. Each concentration is tested four times. The mean toxicity of the replicates is the response and the reciprocal of the variance of the replicates is used for a weight regression. The more the replicates vary, the less influence this dilution will have on the regression.

The toxicity is directly proportional to the concentration of the agent. The model for this assay assumes that the response increases monotonically. The asymptotes (minimum and maximum toxicity) produce a curve

with a sigmoidal shape. This response is often referred to as a logistic curve.

It is assumed that the standard and unknown compounds work by the same chemical mechanism so that the only difference should be a shift of the curve to the left or right if the unknown compound is more or less potent, respectively. The purpose of this assay is to determine the relative potency. For reference, the concentration of each compound that produces 50% toxicity (EC50) is determined.

The logarithm of concentration is the predictor variable,  $x$ . Toxicity is modeled here as a nonlinear function of the parameters  $a$ ,  $b$ ,  $c$ ,  $x_0$ , and  $y_0$ , as seen here:

$$y = y_0 + \frac{(a - y_0)}{\left(1 + \exp\left(-\left(\frac{x - x_0}{b}\right)\right)\right)^c}$$

The  $a$  parameter is the asymptotic maximum percent toxicity and  $y_0$  is the asymptotic minimum percent toxicity. The other three parameters determine the shape. The  $b$  parameter affects the steepness of the rise in response between the asymptotes. JMP calls this

(continued on page 2)



Bioassay Template		Sample	Conc	Log Conc	Toxicity 1	Toxicity 2	Toxicity 3	Toxicity 4	Toxicity 5	Toxicity 6	Mean Toxicity	SD Toxicity	Weight	Bioassay SPL Grouped
▼ Dose Response														
▼ Weighting														
▼ Fit Bioassay														
Columns (13/0)														
Sample														
Conc														
Log Conc														
Toxicity 1														
Toxicity 2														
Toxicity 3														
Toxicity 4														
Toxicity 5														
Toxicity 6														
Mean Toxicity														
SD Toxicity														
Weight														
Bioassay SPL Grouped														
Rows														
All Rows	24													
Selected	0													
Excluded	0													
Hidden	0													
Labelled	0													

Figure 1: Bioassay data table template

(continued from page 1)

the reduced model because it doesn't take treatment groups into account.

A shift parameter,  $x_0s$ , is introduced for  $x_0$  to account for the potential difference in potency. If  $x_0s$  is zero, then the two curves are identical and the two compounds have identical potency. Two curves are fit, one with the shift for the unknown data and one without the shift for the standard data. This form is called the parallel model. The Nonlinear platform handles two curves using a new grouping feature:

$$y = y_0 + \frac{(a - y_0)}{\left(1 + \exp\left(-\left(\frac{x - (x_0 + x_0s)}{b}\right)\right)\right)^c}$$

Finally, in case the shape of the standard and the unknown curve are not the same, possibly indicating that the mechanism of toxicity is not the same, shift parameters are introduced

for all of the original five parameters. This form is called the full model:

$$y = y_0 + y_0s + \frac{(a + as - (y_0 + y_0s))}{\left(1 + \exp\left(-\left(\frac{(x - (x_0 + x_0s))}{(b + bs)}\right)\right)\right)^{(c + cs)}}$$

Hypothesis tests are performed on the sequence of fitted models. First, the parallel model is compared to the reduced model. If the parallel model is not significant, then the potency of the two agents is not significantly different. Second, the assumption of parallelism is tested by determining if the full model is significantly better than the parallel model. If it is, then the assumption is violated. Finally, the full model is compared to the reduced model in case the parallel model is not significant.

The complete data table for this example is shown in Figure 1. Space is provided for up to six replicates.

## Using the Formula Library

The last column in the bioassay data table (Figure 1), Bioassay 5PL Grouped, contains a formula for the shifted five-parameter logistic curve function. This column is used instead of the predictor variable, Log Conc. You can create this formula yourself with the formula editor, but there is a easier way in JMP 5.1:

1. Select **Analyze > Modeling > Nonlinear**.
2. When the dialog appears, click the **Model Library** button under the column list, as shown in Figure 2. A dialog opens with a list of nonlinear models, as shown in Figure 3.

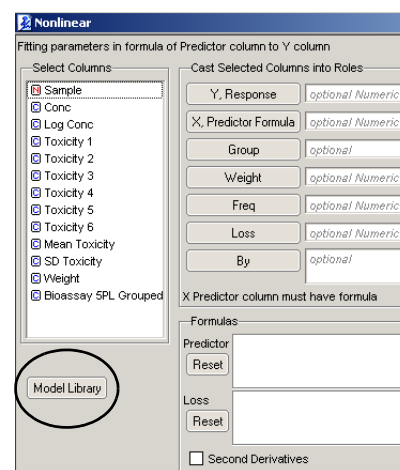


Figure 2: Nonlinear dialog

(continued on page 3)

(continued from page 2)

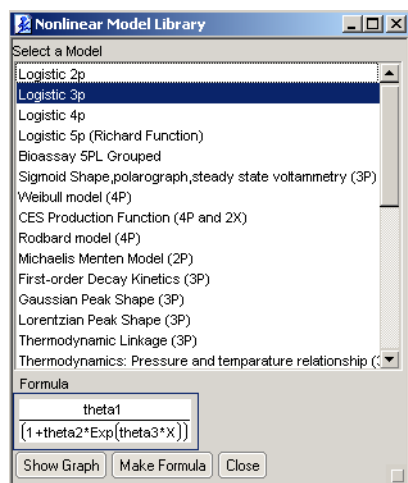


Figure 3: Nonlinear model library

3. Select Logistic 3P from the list. The formula for the model is shown below the list.
4. To see a graph of the model, click **Show Graph** at the bottom of the library window.

You can see the shape of this function based on a set of initial values. You can also use the slider to change the values or enter numbers in a box. Figure 4 shows the shape of the same model with different values for all three parameters. These models are flexible fitting tools.

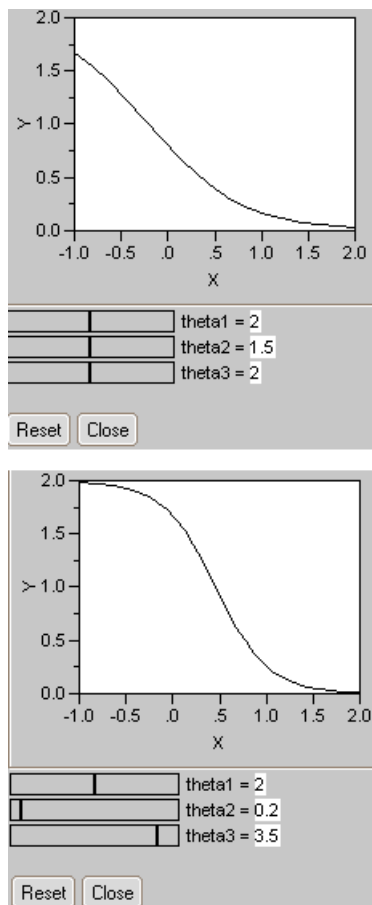


Figure 4: Graphs of nonlinear functions from library

To continue creating the formula for this example:

1. Select Bioassay 5PL Grouped from the library list.

2. Click the **Make Formula** button.
3. When the column selection dialog appears, select Log Conc, click **X**, and then **OK**.
4. When the dialog prompts for a Group variable, select Sample, click **Group** and then **OK**.
5. Enter the value that identifies the unknown agent. In this example, enter Unknown and click **OK**. The model formula dialog then prompts for additional information.
6. Select Mean Toxicity, and then click the **Response** button.
7. Select Weight, click the **Weight** button, then click **OK**.

There is now a new column in the data table containing the customized formula for the chosen model. Here, the new column is a duplicate of the Bioassay 5PL Grouped column.

Note: The model library feature is scripted. You can edit the definitions in the built-in script to add models, edit models, or delete models to suit your needs. The instructions for these changes are found in the chapter about

## JMP Webinars

Register for a free JMP Webinar at [http://www.jmp.com/news/regwebinar\\_form.shtml](http://www.jmp.com/news/regwebinar_form.shtml) or call 1-877-594-6567.

Date and Time	Title and Description
Friday, November 14, 2003 at 1:00 pm EST	<p>Accessing Data with JMP</p> <p>Have data in multiple locations and formats? Need to access data in multiple database tables using SQL? Want to easily manipulate your JMP data? JMP's powerful access and data manipulation tools will help you access and prepare your data for analysis.</p>
Monday, December 15, 2003 at 1:00 pm EST	<p>JMP for Six Sigma</p> <p>In this webinar you'll see how JMP can be customized to fit the varied needs and levels of practitioners in your organization and how JMP's graphical interactivity enables everyone to make a contribution to the productivity gains promised by Six Sigma.</p>

## Fitting the Curve

You can now complete the Nonlinear launch dialog using this constructed model column.

1. Select Mean Toxicity and click **Y, Response**.
2. Select the Bioassay 5PL Grouped column (or the duplicate you just created) and click **X, Predictor Formula**.
3. Select Sample and click **Group**.
4. Select Weight and click **Weight** for a weighted regression.

Figure 5 shows the completed nonlinear launch dialog.

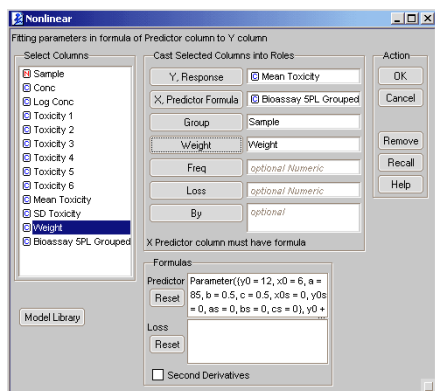


Figure 5: Completed nonlinear dialog

Click **OK** on the launch dialog to begin the fitting process, which is interactive and iterative. The initial setup is presented and it is waiting for you to initiate the fit or make changes. Note starting values for the parameters are entered for you.

You will fit the reduced model first, without any contribution to the model from the shift parameters. Check the boxes to the right of the last five parameters, as shown in Figure 6.

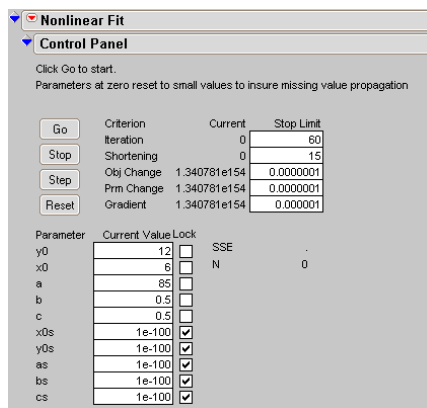


Figure 6: Initial nonlinear fit control panel

Click **Go**. Figure 7 shows that the fitting process converges to a solution. The reduced model indicates that the minimum mean toxicity is 10.657% and the maximum mean toxicity, (indicated by the value of the parameter a) is 82.859%.

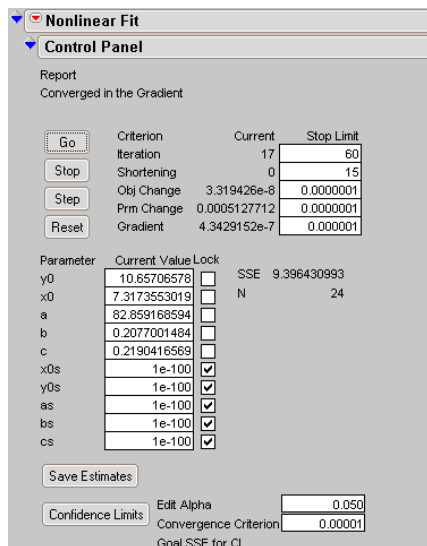


Figure 7: Results for reduced model

## Comparing Solutions

This example fits three models and compares the results, as follows:

1. Save these results by clicking the red triangle on the Nonlinear Fit title bar and selecting **Remember Solution**.

2. When prompted, type a name for this fit (name it Reduced Model) in the text box and click **OK**.
3. To fit the parallel model, uncheck the box to the right of the  $x_0s$  parameter.
4. Change the initial value, currently  $1e-100$ , to 0 for the  $x_0s$  parameter.
5. Click **Reset**, and then click **Go**.
6. When this new model converges, again save it with the **Remember Solution** command, and name it Parallel Model.
7. For the last model, uncheck the remaining four boxes next to the parameters, and change their initial values to 0.
8. Click **Reset**, then click **Go**. You will see the results as shown in Figure 8.
9. Save it with the **Remember Solution** command, and name it Full Model.

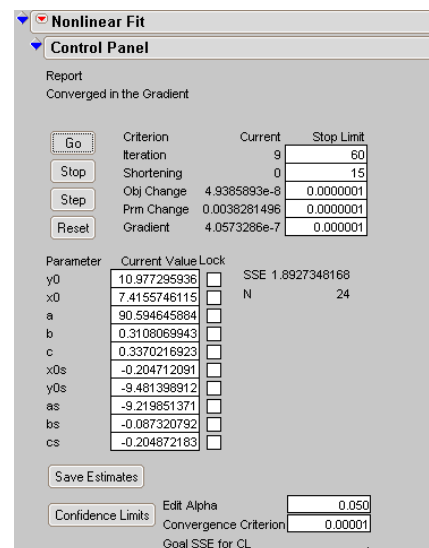


Figure 8: Results of full model nonlinear fit

Look below the Control Panel section at the plots of the standard and unknown curves of Mean Toxicity by Log Conc. These plots are based on the parameter estimates from the full model fit. Another benefit of the grouping feature (specified in the

Nonlinear Fit launch dialog) is that the curves appear on the same plot. You can see that the curves are similar but not parallel.

Look at the end of the report for the Remembered Models section, as shown in Figure 9. The table in the middle presents three hypothesis tests.

- The first line indicates that the parallel model is significant (Prob > F 0.0047).
- The second line indicates that the full model is significant (Prob > F 0.0002).
- The third line indicates that the full model is significantly better than the parallel model (Prob > F 0.0003).

Model	SSE	DFE	MSE	Restrictions
Reduced Model	9.396431	19	0.494549	$x0s=0, y0s=0, as=0, bs=0, cs=0$
Parallel Model	7.877735	18	0.437652	$y0s=0, as=0, bs=0, cs=0$
Full Model	1.8927348	14	0.135195	

Hypothesized	Alternative	Denominator	SS	NDF	DDF	F Ratio	Prob > F
Reduced Model	Parallel Model	Full Model	1.5198974	1	14	11.233	0.0047
Reduced Model	Full Model	Full Model	7.5038962	5	14	11.101	0.0002
Parallel Model	Full Model	Full Model	5.9849987	4	14	11.067	0.0003

Parameter	Reduced Model	Parallel Model	Full Model
y0	10.65706578	10.587322896	10.977256936
x0	7.3173553019	7.3394015551	7.4155746115
a	82.859168594	82.783505638	90.594645884
b	0.2077001484	0.2083287444	0.3108068943
c	0.2190416569	0.2184841767	0.3370216923
x0s	0	-0.249638509	-0.204712091
y0s	0	0	-9.481388912
as	0	0	-9.219851371
bs	0	0	-0.087320792
cs	0	0	-0.204872183

Figure 9: Remembered models report

When you created the model formula for Bioassay5PL Grouped, the Model Library script also created a table property in the data table called Fit Bioassay. Fit Bioassay contains a script that automates the entire fitting process. To see it work:

1. Close the current Nonlinear window.
2. In the data table, click the red triangle beside the Fit Bioassay table property.
3. Select **Run Script**. This script fits the reduced, parallel, and full models

and then records the remembered results for comparison.

## Custom Estimates

The  $x_0$  parameter is related to the horizontal location of the inflection point of the curve. Suppose you want to know about the difference in the concentration between the inflection point for the standard and the unknown curves. You can use a new feature, **Custom Estimate**, for this purpose:

1. Click the red triangle icon on the Nonlinear Fit title bar and select **Custom Estimate**.
2. Enter the expression  $10^{x_0}$ . This expression converts the logarithm back to the original concentration units for the standard compound.

The answer, which appears at the bottom of the window, is 26036021 with a standard error of 6067999 (see Figure 10).

3. Repeat this step for the unknown compound using  $10^{(x_0+x_{0s})}$  for the expression. The other inflection point occurs lower at 16250343 with a standard error of 11715584 (Figure 10).

Expression	Estimate	Std Error
$10^{x_0}$	26036021	6067999
$10^{(x_0 + x_{0s})}$	16250343	11715584

Figure 10: Custom estimate results

## Inverse Prediction to Find EC50

Assume for the moment that the assay curves are parallel. You want to obtain an estimate of the EC50 level for the

standard and the unknown.

To do this, you must invert both the standard and unknown parallel models (solve the prediction equation for Log Conc in terms of Mean Toxicity and the parameters). JMP can solve the equation for you:

1. Evaluate each inverted model with the associated parameter estimates to get both EC50 values.
2. Take the anti-logarithm to get the concentration corresponding to 50% toxicity.
3. The ratio of the two concentrations is the same as 10 to the power of the difference of their logarithms.
4. Use this relation to compute the ratio.

The script shown below performs all of these computations. Be sure to leave the Nonlinear window open after fitting the models and before you run this script.

```
// solve 5 PL model for x in
//terms of y.
invStdEq = Invert Expr(
    y0 + (a - y0) / (1 + Exp(-(x
- x0) / b))) ^ c,
    x,
    50
);
invUnkEq = Invert Expr(
    y0 + (a - y0) / (1 + Exp(-(x
- (x0 + x0s)) / b))) ^ c,
    x,
    50
);

// get parameter estimates.
{ y0, x0, a, b, c, x0s, y0s, as,
bs, cs } =
    (Nonlinear[1] << Report)
    ["Remembered Models"]
    [NumberColBox(10)]

<< Get;

(Nonlinear[1] << Report)["Remem-
bered Models"] << Append(
    Outline Box( "Relative
Potency",
        Table Box(
            String Col Box(
                "Parameter", {
```

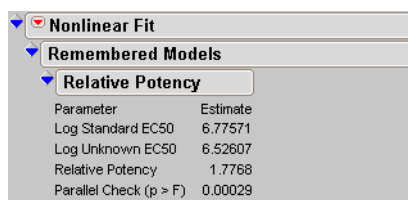


```

F)"
    "Log Standard EC50",
    "Log Unknown EC50",
    "Relative Potency",
    "Parallel Check (p >
    } ),
    Number Col Box( "Esti-
mate", {
    ( invStdEq ),
    ( invUnkEq ),
    10^(Eval( invStdEq
)-Eval( invUnkEq )),
    (Nonlinear[1] <<
Report)
    ["Remembered Models"]
    [NumberColBox(8)] <<
Get(3)
    } )
    )
);

```

After it is run, the script adds the custom report, shown in Figure 11, to the bottom of the Nonlinear window as part of the Remembered Models section.



Nonlinear Fit	
Remembered Models	
Relative Potency	
Parameter	Estimate
Log Standard EC50	6.77571
Log Unknown EC50	6.52607
Relative Potency	1.7768
Parallel Check (p > F)	0.00029

Figure 11: Custom report for inverse prediction

In this example, the unknown sample is 1.78 times more potent than the standard compound. The appropriate *p*-value from the Remembered Models report is copied here. It indicates a significant deviation from the assumption of parallelism so the estimate may not be valid.

## Conclusion

Several new features of JMP 5.1 make it easy to set up and fit nonlinear models. Special scripts that are built into JMP support specialized but common tasks, e.g., bioassay curve analysis. Scripts enable you to take an analysis even further. Stay tuned to find out when you can obtain your copy of JMP 5.1.

## From the Trainer

### Far Out

Mark Bailey, SAS Statistical Training and Technical Services

You sometimes have an idea or model in your mind about how your data should look, and you're surprised when the results don't match your expectations.

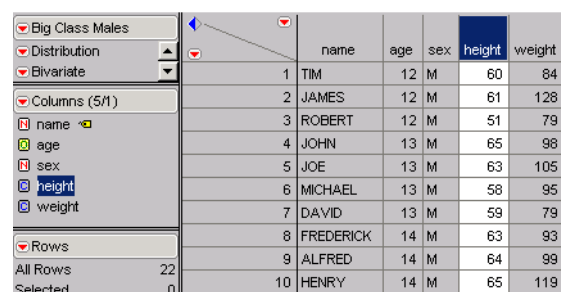
Unexpected values can be those that are much lower or higher than the other data. John Tukey called such values *outlying* data, or simply *outliers*. JMP provides several tools in the Distribution platform for you to use to explore outliers, including Tukey's schematic plot (sometimes called the outlier box plot).

In addition to plots, numerous statistics can detect outliers or assess their impact. To the uninitiated, it might seem like these tests are all the same and that one test is as good as another. However, the only thing that they share is a common purpose. The wide-ranging battery of tests result from many possible populations that you might encounter and the myriad processes by which samples can be contaminated. Each test is built for a specific population and specific type of contamination. A more complete discussion of outliers and data contamination can be found in Barnett and Lewis (1994).

### That's Normal

This article illustrates a case in which the population is known to follow the Gaussian (normal) distribution. Observations of physical characteristics of people, such as height, or data involving sums, such as sample means, are approximated well by this distribution. The following examples look at height and the calibration factor of an instrument.

This first example uses a table that is the subset of males from the Big Class data table (Figure 12), found in the JMP Sample Data folder. This folder was installed on your computer when you installed JMP.



	name	age	sex	height	weight
1	TIM	12	M	60	84
2	JAMES	12	M	61	128
3	ROBERT	12	M	51	79
4	JOHN	13	M	65	98
5	JOE	13	M	63	105
6	MICHAEL	13	M	58	95
7	DAVID	13	M	59	79
8	FREDERICK	14	M	63	93
9	ALFRED	14	M	64	99
10	HENRY	14	M	65	119

Figure 12: Males from Big Class.jmp

A distribution of the height of male students (Figure 13) shows an outlier box plot, and a normal quantile plot. Both plots suggest that one value is an outlier.

One method of detecting outliers is Grubbs' outlier test (Grubbs, 1969). The null hypothesis tests that there is no outlier in the data. The test statistic is based on the ratio of the maximum absolute deviation from the sample mean to the sample

standard deviation, as shown below:

$$G = \frac{\text{Max}(|y_i - \bar{y}|)}{s}$$

The critical  $G$  value for the hypothesis test is given by this equation:

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2}{N-2+t^2}}$$

$N$  is the sample size and  $t$  is the quantile from the Student's  $t$  distribution with probability  $\alpha/(2N)$  and  $N - 2$  degrees of freedom. The  $G$  and critical  $G$  values can be computed with a column formula or a JSL expression. The script listed at the end of this article was used in JMP to incorporate the Grubbs' test results into a distribution analysis of the male students' heights. The script can also be downloaded from [www.jmp.com](http://www.jmp.com) by clicking **Downloads > JMP Script Library**.

The script displays the dialog in Figure 14 that prompts for one numeric  $Y$  variable and the significance level. When you click **OK** in this dialog, the script performs the distribution in Figure 13, and it appends the Grubbs test after the Moments table.

As shown on the right in Figure 14, the sample  $G$ , 2.99622, is greater than the critical value, 2.75773 ( $\alpha = 0.05$ ). The observed  $G$  has a  $p$ -value of 0.01451. Therefore, the conclusion is that this low height value is an outlier.

The outlier can then be excluded and the test repeated. In this way, more than one outlier can be detected with the same method. It is generally recommended that this test be used only with a sample size greater than six.

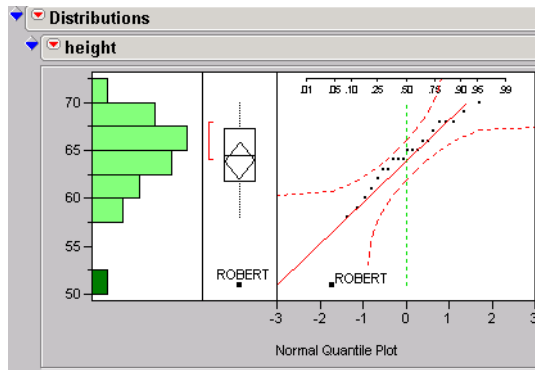


Figure 13: Distribution of male students' heights

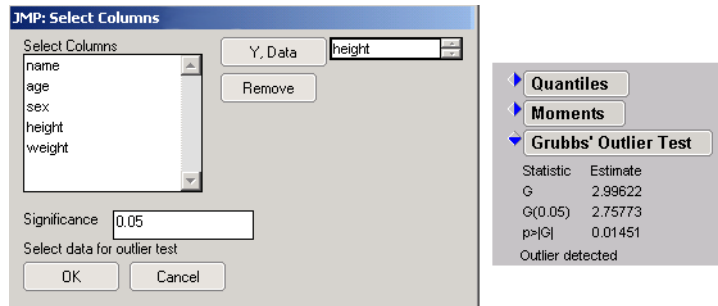


Figure 14: Dialog and results of Grubbs' outlier test

NIST ZARR13		calibration factor
Note	This data set was	
Distribution		1 9.206343
Columns (1/0)		2 9.299992
calibration factor		3 9.277895
		4 9.305795
Rows		5 9.275351
		6 9.288729
All Rows	196	7 9.287239
Selected	0	8 9.260973
Excluded	0	

Figure 15: Partial listing of calibration factor data

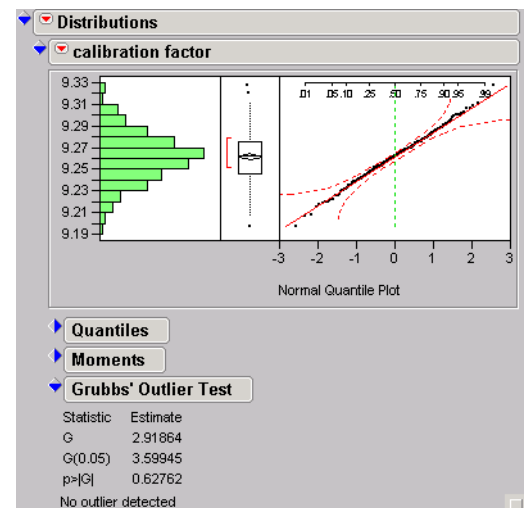


Figure 16: Outlier analysis of calibration factor data

A second example is from the electronic handbook by NIST/SEMATECH, 2003. The data includes 196 actual calibration factors from a heat flow meter calibration and stability analysis (see Figure 15).

The same distribution analysis with the Grubbs test indicates that there is no outlier in this sample (Figure 16).

## The Grubbs Test and Normality

The Grubbs test assumes that the population is normally distributed with fixed location and dispersion, which can be verified using control charts. The individual values and moving ranges in Figure 17 show points (rows 1, 2, 3, 45, 46, and 189) that indicate the calibration process might not have been stable during this

period of observation. If so, then the sample is a mixture of data from more than one population. This situation invalidates Grubbs test because shifts in the process mean or variance can mask the presence of real contaminating data.

Other tests have been developed for situations involving non-normal populations, mixtures of populations, more than one outlier beyond the same tail, outliers beyond both tails, and so on. See Barnett and Lewis (1994) for more information about these tests.

## Conclusion

This article highlighted an important statistic for finding an outlier among normally distributed data. It complements the distribution and control charts for detecting and assessing such outliers.

## References

- American Society for Testing and Materials. *Standard E 178: Standard Practice for Dealing with Outlying Observations in ASTM Standards on Precision and Bias for Various Applications*, 4<sup>th</sup> Edition, pages 274- 290, 1992.
- Barnett, Vic and Toby Lewis. *Outliers in Statistical Data*, 3rd Edition, John Wiley & Sons, 1994.
- Chambers, John M., William S. Cleveland, Beat Kleiner, and Paul A. Tukey. *Graphical Methods for Data Analysis*, Wadsworth International and Duxbury Press, 1983.
- Grubbs, Frank. "Procedures for Detecting Outlying Observations in Samples." *Technometrics*, 11(1), 1-21. February 1969.
- Iglewicz, Boris and David Hoaglin.

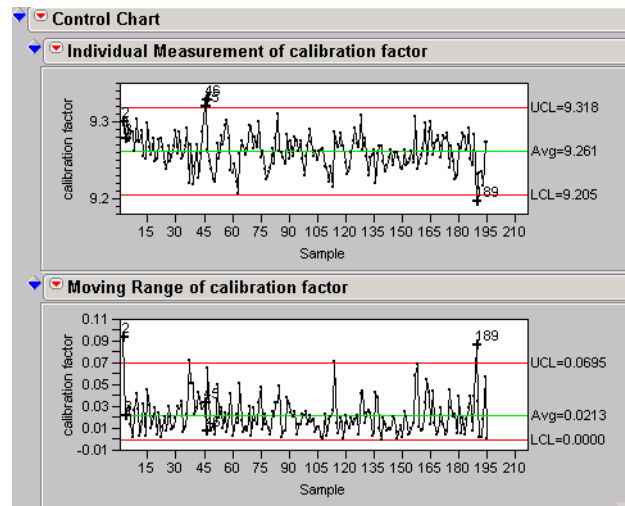


Figure 17: Control charts for calibration factor data

"Volume 16: How to Detect and Handle Outliers." *American Society for Quality*, 1993.

NIST/SEMATECH *e-Handbook of Statistical Methods*, <http://www.itl.nist.gov/div898/handbook/>, June 14, 2003.

Tukey, John W. *Exploratory Data Analysis*, Addison-Wesley, 1977.

## Script

The script used to run this analysis can be downloaded from [www.jmp.com](http://www.jmp.com) by clicking **Downloads > JMP Script Library > Grubbs Outlier Test**.

```
Clear Globals();

dIlg = Column Dialog(
    yCol = Col List( "Y, Data",
        Data Type( Numeric ),
        Min Col(1),
        Max Col(1)
    ),
    Line Up( 2,
        "Significance", a = Edit
        Number( 0.05 )
    ),
    "Select data for outlier
    test"
);
If( dIlg["Button"] == -1, Throw(
    "User cancelled" ) );
Remove From( dIlg ); Eval List(
    dIlg );

dt = current data table();
yCol = Column( yCol[1] );

dist = Distribution(
    Continuous Distribution(
        Column( yCol ),
        Quantiles(1),
        Moments(1),
        Normal Quantile Plot(1)
    )
);
yVal = yCol << Get As Matrix;
n = N Row( yVal );

// save current selection.
r0 = dt << Get Selected Rows;
// find excluded rows.
dt << Select Excluded;
r1 = dt << Get Selected Rows;
dt << Clear Select;
//deleted excluded data, in
reverse order.
r2 = J( n, 1, 1 );
r2[r1] = 0;
yVal = yVal[Loc(r2)];

// restore original selection.
If( N Row( r0 ),
    For( i=1, i<=N Row( r0 ),
        i++,
        Selected( Row State(
            r0[i] ) ) = 1;
    );
);
yRes = yVal - Mean( yVal );
g = Maximum( Abs( yRes ) ) / Std
Dev(yVal);
t0 = Abs( t Quantile( a/(2*n),
    n-2 ) );
g0 = ((n-1)/Sqrt(n)) * Sqrt(
    t0^2 / (n - 2 + t0^2) );
p = 2 * n * (1 - t Distribution(
    Sqrt( g^2*(2-n)/(2+g^2-1/n-n) ),
    n-2 ));
distr = dist << Report;
distr[Outline Box(2)] << Append(
    Outline Box( "Grubbs' Out-
    lier Test",
```



```

Table Box(
  String Col Box( "Statistic", {"G", "G("||Char(a)||")",
    "p>|G|"} ),
    Number Col Box( "Estimate", Matrix( {g, g0, p} ) )
  ),
  Text Box(
    If( g>g0,
      "Outlier detected",
      "No outlier detected"
    )
  )
);
distr["Quantiles"] << Close;
distr["Moments"] << Close;

```

## Look for JMP at these Conferences and Trade Shows

Oct. 26-28, 2003	South Central SAS User's Group (SCSUG) in Houston, TX
Nov. 5-7, 2003	Western Users of SAS Software (WUSS) in San Francisco, CA
Nov. 16-19, 2003	Southeast Regional Meeting of the American Chemical Society (SERMACS) in Atlanta, GA

## Case Study

### Applying Benford's Law

Meredith Blackwelder, JMP Development

In 1881, an astronomer named Simon Newcomb examined logarithm tables and noticed that the first pages of the book—pages containing logarithms that begin with the digit one—were worn more than the remaining pages. In 1938, a physicist at the General Electric Company named Frank Benford saw same phenomenon. He took it a step further and analyzed 20,229 sets of numbers, including baseball statistics, numbers in magazine articles, the areas of rivers, and the street addresses of the first 342 people in the book *American Men of Science*. He found that although all unrelated, the same first-digit probability pattern he found in the worn pages of logarithm tables was present in all the sets of numbers he investigated. He discovered that the number one was the first digit about 30 percent of the time, more often than any other.

Named after Frank Benford, Benford's Law states that if you randomly select a number (which must somehow be socially or naturally related to all the other numbers under investigation—Benford's Law does not apply to uniform distributions), the probability that the first digit will be a one is about 30%. Furthermore, the probability that the first digit will be a two is about 18%, a three is 12%, a four is 9%, etc.

Without knowing about Benford's Law, you might expect the first digit of laboratory data to be uniformly distributed. However, Benford's Law shows that these digits follow an unusual and asymmetric distribution (Bogomolny, 2003).

To illustrate Benford's Law in JMP, first open a data table that contains socially- or naturally-related numbers, such as those shown in Figure 17. The data table shown in Figure 18 contains house numbers of 14,399 registered voters in Wake County, North Carolina. You can find the data at <http://msweb03.co.wake.nc.us/bordelec/Waves/WavesOptions.asp>.

House Number	First Digit
106	1
1224	1
111	1
203	2
204	2
5523	5
404	4
500	5
401	4
637	6
109	1
721	7

Figure 18: House number data

(continued on page 10)

(continued from page 9)

Note that the First Digit column in Figure 18 contains a formula that automatically

retrieves the first digit of the house numbers and inserts them into the column.

The formula is:

```
Num(Substr(Char(Abs( :House  
Number)), 1, 1))
```

To graph the house number's first digits to see if they are consistent with Benford's Law, select **Analyze >**

**Distribution** and set First Digits as the Y column. The report in Figure 19 shows the distribution of first digit numbers. You can see that the distribution is indeed consistent with Benford's Law.

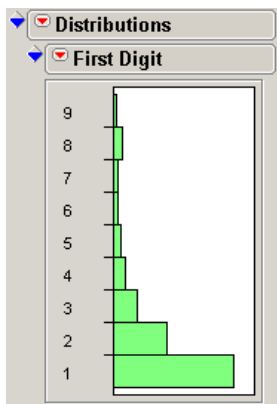


Figure 19: Distribution of first digits

However, if you run a distribution on numbers that are manually entered at random into a data table, such as those in Figure 20, the distribution pattern will be inconsistent with Benford's Law. The data and script used to run this analysis can be downloaded from [www.jmp.com](http://www.jmp.com) by clicking **Downloads**

#### > JMP Script Library > Benford's Law.

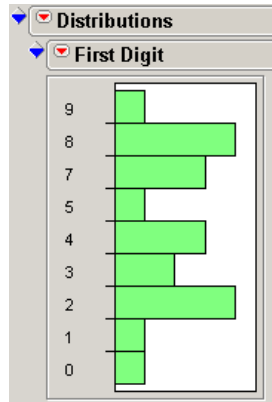


Figure 20: Random numbers entered manually don't produce a distribution consistent with Benford's Law

As a real-world example, investigators have used Benford's Law most recently for determining the strength of evidence of fraudulent data. It is used as a tool to raise red flags at potential frauds, embezzlers, tax evaders, sloppy accountants, and even computer bugs. Income tax agencies, including those for the state of California, as well as large corporations and accounting houses, use fraud detection software based on Benford's Law.

Mark Nigrini from Southern Methodist University is a pioneer of applying Benford's Law to tax evasion and other fraud detection (Nigrini and Mittermaier, 1997). He uses a system where he looks at a tax return (which, at its lowest level, contains a set of data), and he compares the numbers with the frequencies and ratios predicted by Benford's Law. If the data match Benford's Law's predictions, the data are probably honest. However, if he can graph the numbers and see

discrepancies between their patterns and those predicted by Benford's Law, he calls for further investigation. (Nigrini, 2000)

According to Malcolm Browne's article in *The New York Times* (1998), in one test alone, Nigrini correctly identified seven cases as "involving probably fraud" for Robert Burton, the chief financial investigator for the Brooklyn District Attorney, yet he warns that the fit of number sets with Benford's Law is not infallible.

## References

- Bogomolny, Alexander. 2003. *Benford's Law and Zipf's Law* [online]. Cited 7 October, 2003. Available from World Wide Web: <[http://www.cut-the-knot.org/do\\_you\\_know/zipfLaw.shtml](http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml)>
- Browne, Malcolm W. 1998. Following Benford's Law, or Looking Out for No. 1. *The New York Times*, 4 August, final edition. Also available from World Wide Web: <<http://www.rexswain.com/benford.html>>
- Nigrini, Mark. 2000. *Digital Analysis Using Benford's Law: Tests Statistics for Auditors*. Vancouver, BC, Canada: Global Audit Publications.
- Nigrini, Mark and L.I. Mittermaier. 1997. The Use of Benford's Law as an Aid in Analytical Procedures. *Auditing: A Journal of Practice and Theory* 16: 52-67.
- Robertson, Jack C. and Mark Nigrini. 1999. *Fraud Examination for Managers and Auditors: 2000 Edition*. Austin, TX, USA: Viesca Books.

## Predicting Patterns in Child Genders: What's the Chance?

Lee Creighton, JMP Development

How would you answer the following problem?

Suppose that couples who wanted children were to continue having children until a boy is born. Assuming that each newborn child is equally likely to be a boy or a girl, would this behavior change the proportion of boys in the population?

This question was posed to readers of *American Statistician*, and many answered incorrectly. You can use JMP to create a simulation to estimate the long-run proportion of boys in the population if families were to continue to have children until they have a boy. This proportion is an estimate of the probability that a randomly selected child from this population is a boy. Note that every sibling group would have exactly one boy.

To create this simulation in JMP, you can use the Random Integer function to generate numbers from either 1 or 2. The value 1 represents a male birth. If the birth is not a boy, select another random number until a boy is born.

To accomplish this simulation, we need a data table to store our results.

1. Create a new data table with a single column, called Children, and open the formula editor. The formula editor that appears is normally used to enter formulas using point-and-click menus and buttons. However,

you can actually enter scripts into the formula editor as well.

2. Double-click in the editing area where you see no formula. This changes the formula editor into text editing mode for typing formulas rather clicking to generate them.
3. Type in the following short program. However, when you enter it, place it all on one line—the separate lines here are so that the program can be easily explained.

```
t = Random Integer(2);
n = 1;
While(t == 2,
t = Random Integer(2);
n++);
n;
```

Here's what the program does:

- The first line picks a random number: either 1 or 2.
- The second line initializes a counter (named *n*) to hold the number of times we had to pick the random number, until the value 1 (male birth) appears.
- The third line begins a **while** loop. The condition it checks is that the random number is equal to 2. If the number is equal to 2, another random integer is picked and the counter (*n*) is increased by one. This process continues until the value 1 (male birth) appears.
- The last line, containing only an *n*, is the number that is placed in the data table column.

When you've returned to the data table, add 100 rows to the table by selecting **Rows > Add Rows**. Now your data table is filled with 100 instances of

the simulation. Note: Your table and graph will probably have different numbers than the one pictured here. That's randomness!

We now want to discover the proportion of these births that are male. Because of the way we set up the simulation, we know that there's only one male birth represented in each row. So, the proportion we are interested in is the number of rows in the data table (representing the number of males) divided by the total of the Children column (representing the number of total births).

To keep a running total of this proportion, add another column to the data table, with a formula:

1. Select **Cols > New Column**.
2. Name the column Proportion.
3. Click **New Property** and select **Formula**.
4. Enter the following formula.

$$\frac{\text{Row}()}{\sum_{i=1}^{\text{Row}()} \text{Children}_i}$$

Your data table should now look like the one in Figure 21.

Children	Proportion
1	1
4	0.4
4	0.33333333
1	0.4
2	0.41666667
4	0.375
1	0.41176471

Figure 21: Data table with two columns

(continued on page 12)

(continued from page 11)

You can now construct an overlay plot to show the series' long-term behavior.

1. Select **Graph > Overlay Plot**.
2. Select Proportion and click the **Y** button, then click **OK**.
3. Click the title bar's red triangle icon and select **Connect Thru Missing**.

Surprisingly, the proportion of males

and females are the same over the long term (Figure 22).

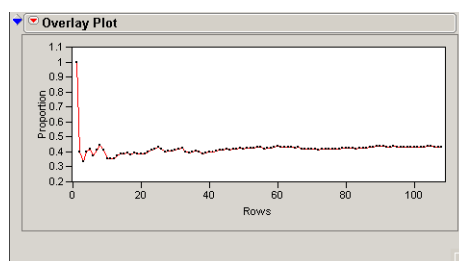


Figure 22: The overlay plot

Thus, couples who continue having children until a boy is born would not change the proportion of boys in the population.

Note: This column is inspired by a problem in Peck, Olsen, and Devore *Statistics and Data Analysis*, Duxbury Press: Belmont, CA.

#### About JMPer Cable

Issue 12 Fall 2003

JMPer Cable is mailed to JMP users who are registered users with SAS Institute. It is also available online at [www.jmp.com](http://www.jmp.com).

#### Contributors

Mark Bailey

Jianfeng Ding

Meredith Blackwelder

Lee Creighton

#### Editor

Ann Lehman

#### Designer

Meredith Blackwelder

#### Printing

SAS Institute Print Center

#### Questions or comments

[jmp@sas.com](mailto:jmp@sas.com)

#### To order JMP software

1-877-594-6567

#### For more information on JMP

1-877-594-6567

[www.jmp.com](http://www.jmp.com)

Copyright 2003, SAS Institute. All rights reserved. SAS, JMP, JMPer Cable, and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates registration. Other brand and product names are trademarks of their respective companies.