Does It Matter Who Did It?

How R&D Labs Can Learn More From Their Positive Controls

Tom Treynor

WHITE PAPER

Table of Contents

Introduction

Whether we are attempting to do profitable manufacturing or reputable science, our goal is the same: to predict the future. As the outputs of a company's manufacturing operations become more and more predictable, it gets better and better at writing contracts that grow the balance of revenues and costs in a competitive marketplace. Likewise, as the initial predictions of an R&D team's latest model are validated by subsequent experimentation, they become more willing to bet more of their time, money and reputation on the model's more far-ranging predictions.

Among both scientists and engineers, there is one prediction that is so obvious it is frequently unspoken and so foundational it is frequently untested: the prediction that two people independently executing the same procedure should get the same result. To be sure, scientists will often try to reproduce the results of a new journal article before attempting to extend it, and corporate product managers will usually budget for time-consuming comparability testing when doing tech transfer or scaling up a manufacturing process to a new site. Yet these same people will implicitly assume that Joe's and Jane's results are comparable when they do R&D in the same building – without sufficient evidence to do so.

There are three distinct ways in which Joe and Jane can get different results in the lab despite receiving the same training and using the same equipment to execute the same protocols:

1) They could execute the same experiments with different means.
2) They could execute the same experiments with different standard deviations.
3) They could generate outlying results with different frequencies.

Although most managers – and the most frequently used statistical methods – are focused on measuring differences in means, a statistically significant difference in any of these dimensions can present an opportunity to learn something valuable. To be clear, there is generally little value (and usually negative value) in thinking we've learned, "This researcher is better than that one." Rather, the valuable thing we learn is, "There must exist some specific root cause that was too easily forgotten from our protocol or our training – or completely overlooked by it." By investigating and successfully identifying the root cause of any statistically significant difference among researchers, at minimum we increase the likelihood that the protocols we share with the scientific community or our manufacturing partners will readily reproduce the outcomes we have observed in our own workplaces. These insights can also accelerate our R&D progress by increasing the signal-to-noise ratios of all subsequent experiments we do.

Although the specific root causes and the ways they are investigated will vary from lab to lab and from project to project, the statistical methodologies we might use to determine if there are researcher-related root causes worth investigating are fundamentally the same. The rest of this white paper describes a variety of ways we can learn more from our positive controls by testing for statistically significant differences in either the mean, standard deviation or outlier frequency among two or more researchers doing experiments in an industrial or academic R&D setting. This article assumes the reader has a Wikipedia-level conceptual understanding of statistical hypothesis testing, confidence intervals, multivariate linear regression, design of experiment methodologies and statistical process control.

Requirements

Although our ultimate goal is to build models that predict a future that can be precisely reproduced by anyone, anywhere and at any time, the discovery that today's models would predict a persistent difference between Joe's and Jane's outcomes can serve as an important step on the path to achieving that objective. Of course, we can't test the hypothesis that their data is comparable – or the hypothesis that we might derive a valuable insight from their lack of comparability – until we create the opportunity to do so. The two most important requirements for performing such tests are:

(1) There must be a column in our data table that documents which researcher generated which outputs.

(2) We must have access to statistical software such as JMP® to analyze the data.

When there is no record in our data tables of which researchers executed which specific aspects of each experiment, we are implicitly assuming that everyone's data is comparable and that there is nothing of value to learn by testing for differences among them. Not only is this assumption frequently wrong, but it opposes everything else we do as researchers to determine which factors influence our experiments and by how much. Perhaps more importantly, as a practical matter, whenever the "Researcher" column explains a statistically significant fraction of the variation in our data tables, including that term in our multivariate models improves the precision with which we can measure its other coefficients (for examples, see Parts 2 and 3 to follow).

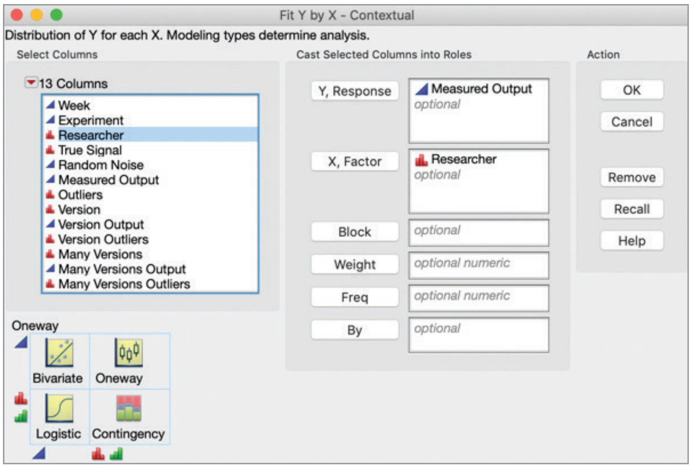
When we have access to good statistical software, we can even detect a persistent offset between researchers without replicating a single condition in our experiment. Nevertheless, many of us have made a habit of running "positive controls" alongside the truly novel conditions we test in each experiment. We even take pride in our commitment to carefully designing and executing the positive controls for each experiment we do. However, if we are not compiling the data from weeks and months of these positive controls to look for statistically significant correlations among the researchers who ran them, we are not extracting nearly as much value from them as we could.

Part 1 of this paper uses the example of a single, oft-repeated positive control to review foundational concepts and procedures for measuring the effect of the researcher on the mean, standard deviation and outlier frequency of an experiment. Parts 2 and 3 demonstrate how the same tests can be run even when we change what we run as our positive control from time to time. In addition to highlighting some underutilized features of JMP, these sections highlight some underappreciated trade-offs we make whenever we choose to update which conditions we run as positive controls.

Guides and Commentary

Part 1: How to use JMP to test for differences between researchers repeatedly running the same positive control

Imagine that over a three-month period, Joe and Jane independently run the same positive control a total of 50 times. If the data for these repeated controls is compiled into a single table, such as the sample data table provided with this article, we can test for evidence of a statistically significant difference in Joe's and Jane's mean outputs using the Fit Y by X platform in JMP:

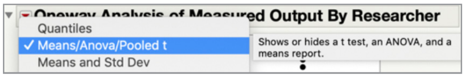


Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

Y, Response: Measured Output

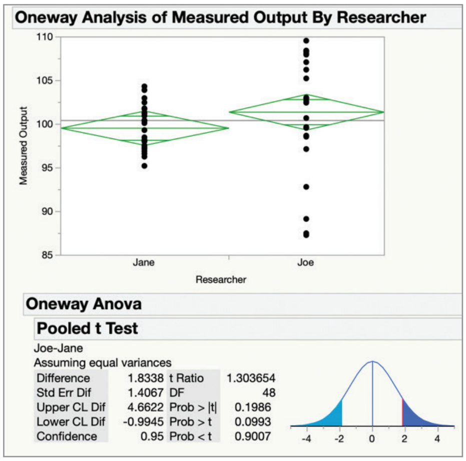
X, Factor: Researcher



Oneway Analysis of Measured Output By Researcher

✓ Means/Anova/Pooled t

Shows or hides a t test, an ANOVA, and a means report.



Oneway Analysis of Measured Output By Researcher

Oneway Anova

Pooled t Test

Joe-Jane

Assuming equal variances

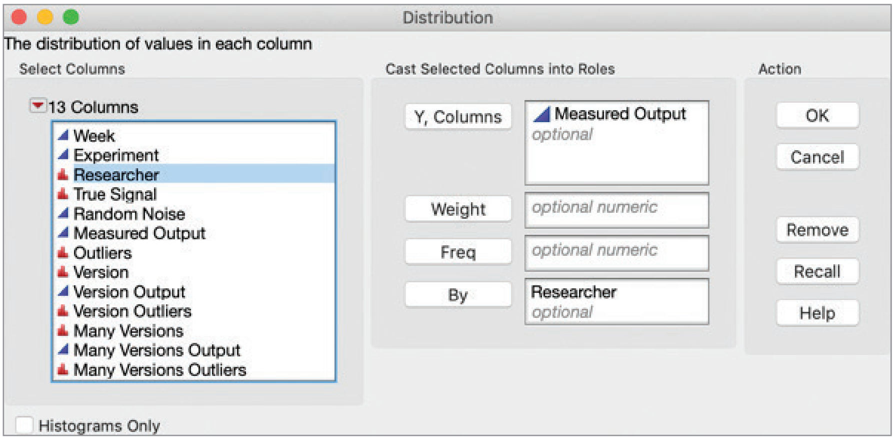
Difference	1.8338	t Ratio	1.303654
Std Err Dif	1.4067	DF	48
Upper CL Dif	4.6622	Prob > |t|	0.1986
Lower CL Dif	-0.9945	Prob > t	0.0993
Confidence	0.95	Prob < t	0.9007

If we take this analysis at face value, we might conclude from its two-tailed p-value (0.1986) that Joe and Jane are generating comparable data with their independent experiments. However, if this is the only way that we analyze the data, we will overlook many valuable stories that these positive controls have to tell.

A fundamental assumption of this one-way analysis is that the process variation around Joe's and Jane's means is normally distributed – i.e., random. However, it appears in the data visualization above that Joe's data might not be normally distributed.

Specifically, it appears as though there may be three or four outliers (measured output < 95) that are not representative of the positive controls Joe ran on 21 other occasions. If indeed there is evidence for nonrandom variation in anyone's data, we should tag those outliers and then revisit the test for means using only the normally distributed subsets of their data.

We can use the Distribution platform to test for both normality and outliers:



Distribution

The distribution of values in each column

Select Columns

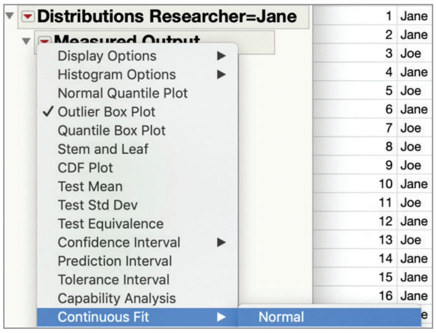
13 Columns

- Week
- Experiment
- Researcher
- True Signal
- Random Noise
- Measured Output
- Outliers
- Version
- Version Output
- Version Outliers
- Many Versions
- Many Versions Output
- Many Versions Outliers

Cast Selected Columns into Roles

Y, Columns	Measured Output *optional*
Weight	*optional numeric*
Freq	*optional numeric*
By	Researcher *optional*

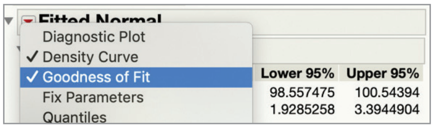
☐ Histograms Only



Distributions Researcher=Jane

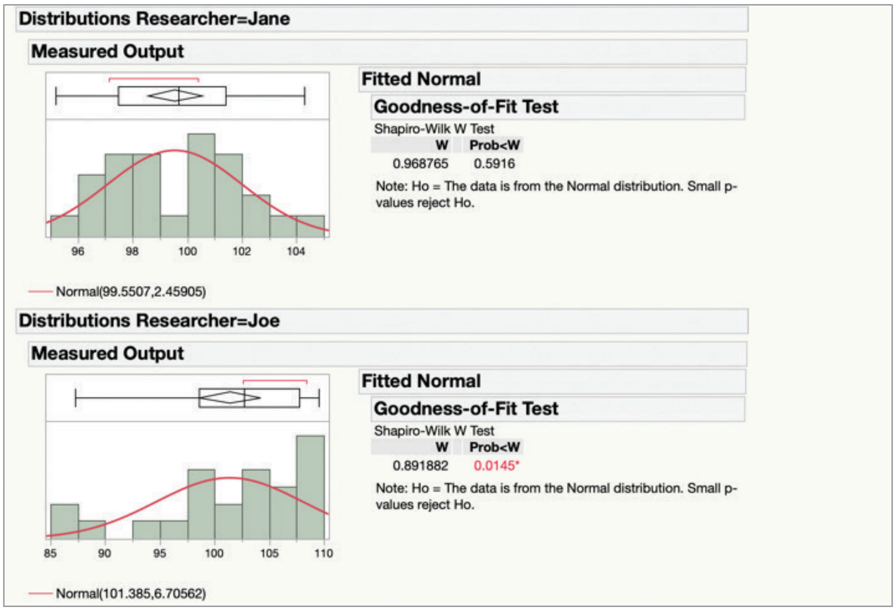
Measured Output

1	Jane
2	Jane
3	Joe
4	Jane
5	Joe
6	Jane
7	Joe
8	Joe
9	Joe
10	Jane
11	Joe
12	Jane
13	Joe
14	Jane
15	Jane
16	Jane



Fitted Normal

Lower 95%	Upper 95%
98.557475	100.54394
1.9285258	3.3944904



Distributions Researcher=Jane

Measured Output

```
Fitted Normal
Goodness-of-Fit Test
Shapiro-Wilk W Test
         W       Prob<W
   0.968765      0.5916
Note: Ho = The data is from the Normal distribution. Small p-
values reject Ho.
```

Distributions Researcher=Joe

Measured Output

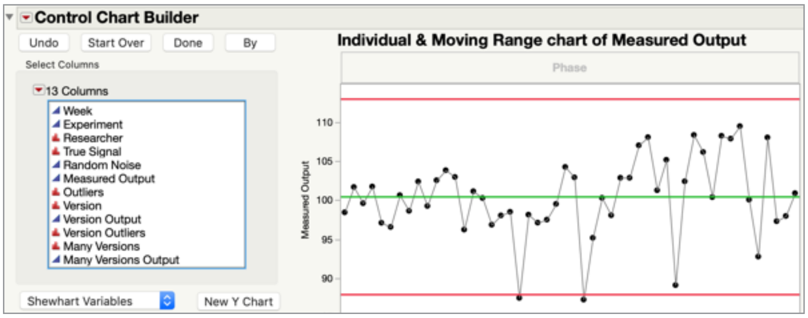
```
Fitted Normal
Goodness-of-Fit Test
Shapiro-Wilk W Test
         W       Prob<W
   0.891882      0.0145*
Note: Ho = The data is from the Normal distribution. Small p-
values reject Ho.
```

Whereas the p-value for Jane's data is >0.05 (0.5916), the p-value for Joe's data is <0.05 (0.0145). In other words, there is evidence that Joe sometimes runs his positive controls in a fundamentally (and measurably) different way than he usually runs his positive controls.

The box-and-whiskers plots adjoining the histograms above would have automatically flagged as outliers any data that is either < 1st quartile – 1.5*(interquartile range) or > 3rd quartile + 1.5*(interquartile range). In this case, none of Joe's or Jane's data is flagged as outliers, because none of their data meets these criteria.

An alternative test for outliers can be performed using control chart logic. Control charts are constructed by coupling a plot of repeated data versus time (i.e., a run chart) to an algorithm for testing each individual replicate for evidence of nonrandom variation. A run chart becomes a control chart when the outputs of this algorithm are superimposed as two red lines called control limits, each equidistant from a third line representing the output mean. Each individual point that falls outside the control limits flags evidence of nonrandom variation in the illustrated data.

We can plot Joe's and Jane's time-ordered data as a control chart using JMP software's Control Chart Builder by clicking and dragging Measured Output into the Y region of the initially empty canvas:



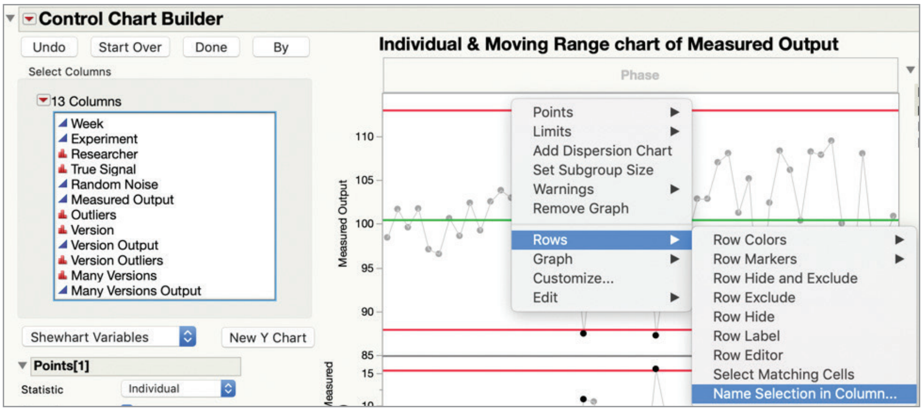
In this case, the algorithm that is used to calculate control limits flags two of 50 positive controls as statistically significant outliers.

Although control chart logic is more sensitive than box-and-whiskers logic for detecting outliers in this data set, it is notable that we are still detecting only two of the three or four points that we had hypothesized as outliers above. Have our eyes deceived us? Or do our statistical tests for nonrandom variation have a substantial false negative rate?

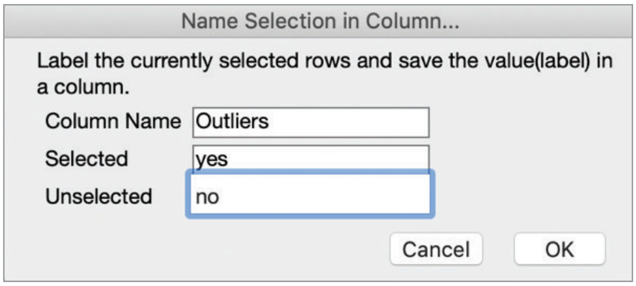
It's important to recognize that most outlier tests do, in fact, suffer a substantial false negative rate, because the presence of nonrandom variation in a data set generally inflates whatever estimate of the normal component of variation is used to test for statistical significance. To mitigate these false negative risks, it is thus wholly appropriate, after detecting the first statistically significant outliers, to iteratively filter and re-test our data for additional outliers.[1]

[1] To be clear, any approach to reduce false negative risk increases false positive risk to some degree. Since those risks are generally asymmetric, it is important to determine on a case-by-case basis which risk is more costly in a given circumstance. In this case, the false negative risk is that we infect future data with preventable errors for a potentially unbounded amount of time, whereas the false positive risk is that we initiate an unnecessary – but time-bounded – investigation into "observed" differences between researchers. Reasonable people can disagree about how to weigh these risks, but this author has observed that most researchers substantially underweight the first risk and substantially overweight the latter.

We can do this iteration without leaving the Control Chart Builder in the following manner. First, we select the two outliers in the Individual chart, right-click and select "Name Selection in Column…":



We can create a new column called "Outliers" with values of "yes" for the two selected points and "no" for the 48 unselected points:


Name Selection in Column...

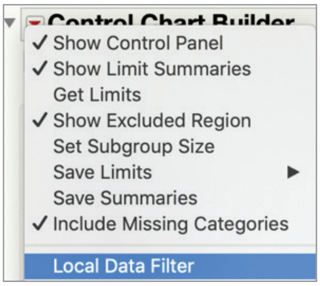
Label the currently selected rows and save the value(label) in a column.

Column Name: Outliers

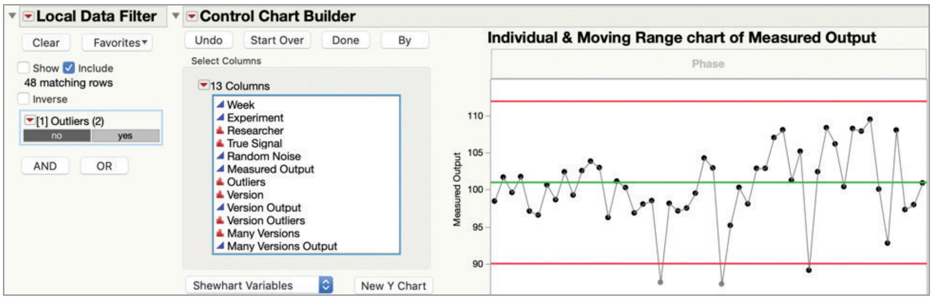
Selected: yes

Unselected: no

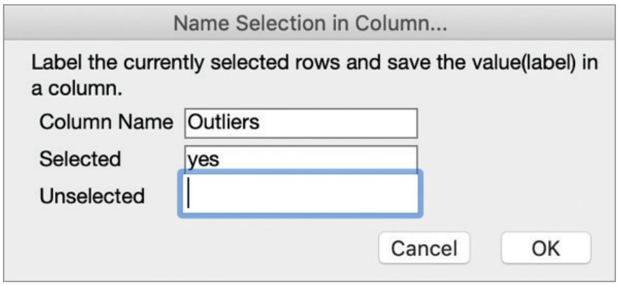
Next, we add a Local Data Filter to our Control Chart Builder:



By unchecking the “Show” box and checking the “Include” box, the Individual Chart will continue to display all 50 measurements while its control limits are recalculated using only the subset of those data where Outliers = “no”:



Since the first two outliers are no longer contributing to this control chart’s estimate of the normal component of its variation, a third measurement is now flagged as an outlier. We can again use the “Name Selection in Column…” feature to document this row of our data table as an outlier. However, if we have selected only this third measurement as we do so, it is critical that this time we do not write any values to the unselected data:


Name Selection in Column...

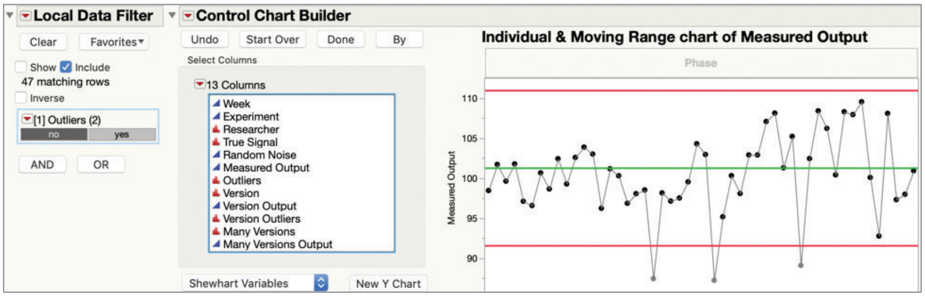
Label the currently selected rows and save the value(label) in a column.

Column Name: Outliers

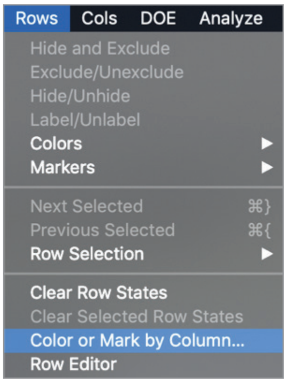
Selected: yes

Unselected:

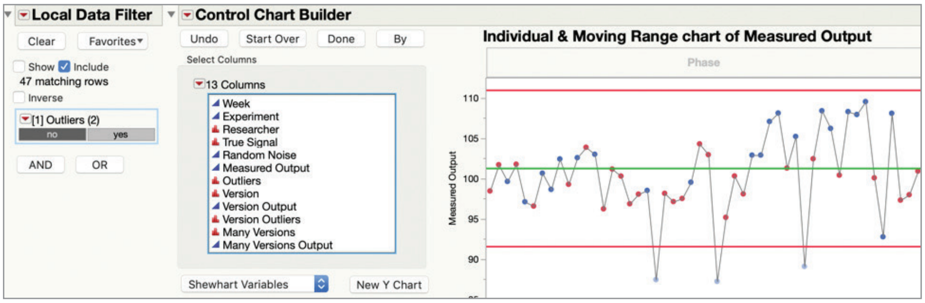
If we write “no” in the “Unselected” field when only one row is selected, this action will overwrite the values of “yes” that had been written to the two rows that are no longer selected. (Alternatively, we must select all three rows before writing “no” in the “Unselected” field.) When we click “OK,” the Individual chart will update automatically, as it now removes all three “yes” rows from the calculation of its control limits:



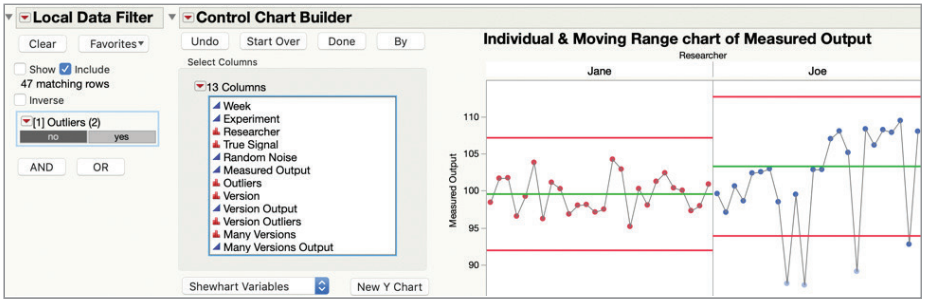
At this point no other rows of our data table are flagged as outliers. If we color the data in our control chart by researcher…



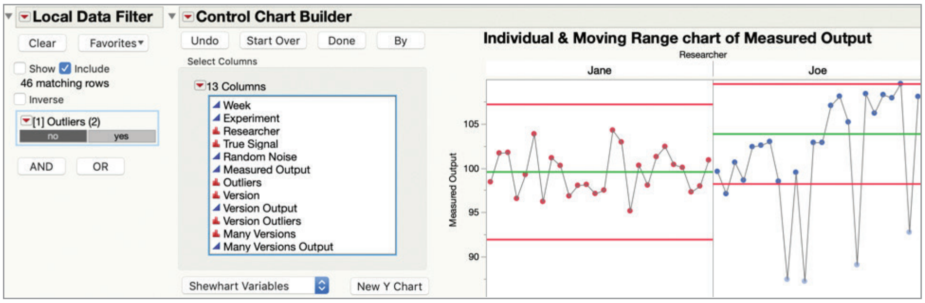
we see that all three outliers are associated with Joe:



We also see that the fourth suspected outlier, while perhaps not an outlier relative to all the positive controls, may be outlying relative to Joe's positive controls. We can test this hypothesis by dragging Researcher into the Phase region of the Control Chart Builder:

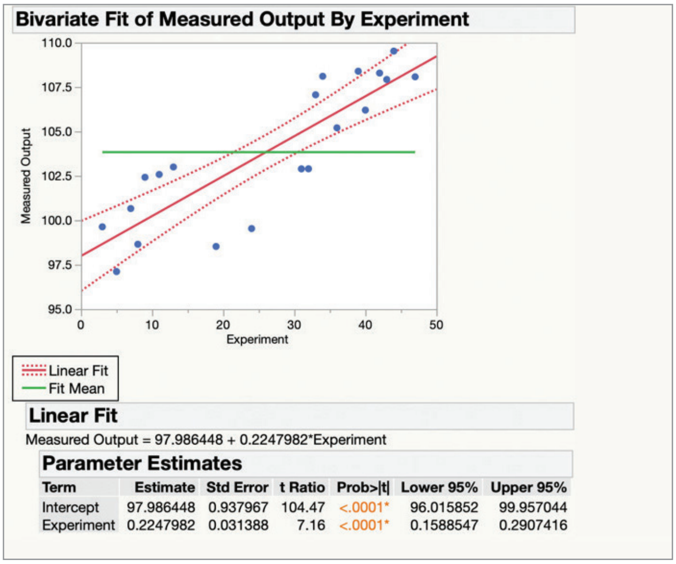


When we tag this fourth point as an outlier, yet again some of Joe's positive controls fall outside the recalculated control limits (one above the upper control limit, one below the lower control limit):



However, it would seem inappropriate to call these fifth and sixth rows of the data table "outliers," as both seem so similar in magnitude to so many more of Joe's positive controls. This important observation illustrates that the fundamental purpose of control chart logic is to detect statistically significant evidence of nonrandom ("assignable cause") variation, not "outliers" per se. In this case, we can see that the reason some data fall outside the recalculated control limits is that Joe's positive controls seem to have been trending (nonrandomly) higher over time.

We can test this hypothesis using the Fit Y by X platform by filtering to the more representative subset of Joe's data (Researcher = "Joe" and Outliers = "no") and fitting a line to a plot of "Measured Output" versus "Experiment":



Linear Fit

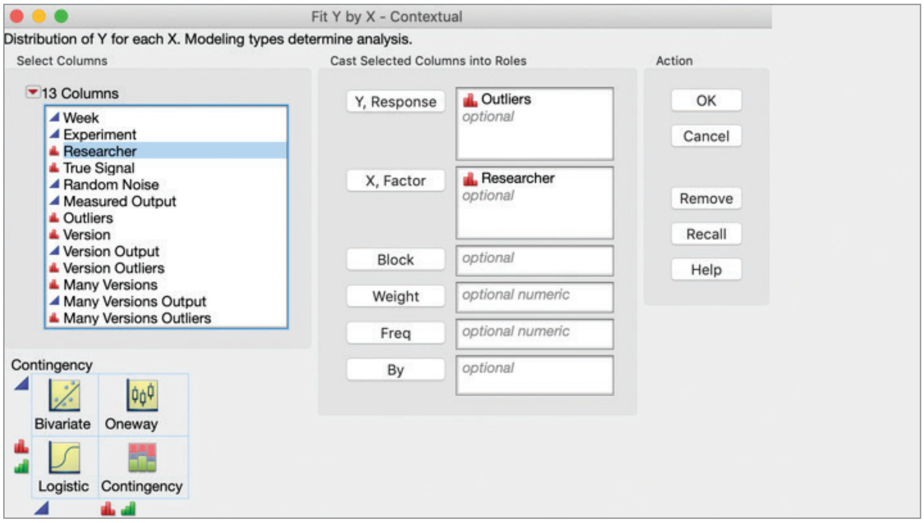
Measured Output = 97.986448 + 0.2247982*Experiment

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	97.986448	0.937967	104.47	<.0001*	96.015852	99.957044
Experiment	0.2247982	0.031388	7.16	<.0001*	0.1588547	0.2907416

Sure enough, the trend in Joe's positive controls is quite statistically significant ($p < 0.0001$), with an estimated rate of 0.16 to 0.29 units per experiment. For comparison, the equivalent fit to Jane's positive controls has p = 0.9675, with an estimated rate of -0.07 to 0.07 units per experiment (not shown).

Are Joe's four outliers revealing a failure mode of our standard operating procedure that is specific to Joe? Or might we expect Jane's data to exhibit the same failure mode with a similar frequency in a larger data set? We can use the Fit Y by X platform to test for statistically significant differences in Joe's and Jane's outlier frequencies by attempting to explain the nominal variation in the Outliers column as a function of the nominal variation in the Researcher column:



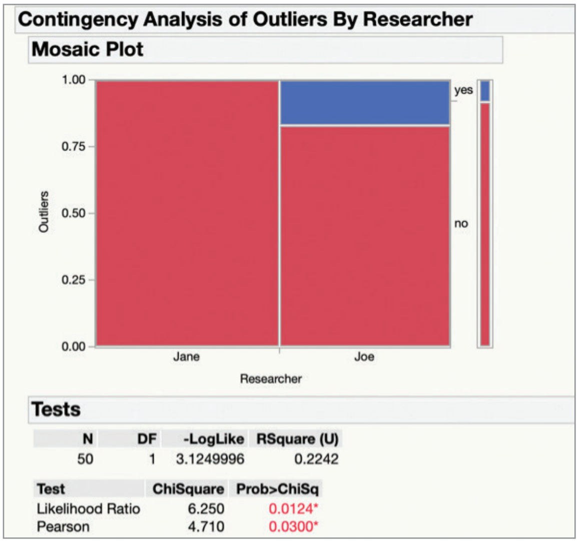
Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

Y, Response: Outliers

X, Factor: Researcher

By default, JMP runs two different chi-square tests, denoted "Likelihood Ratio" and "Pearson":



Contingency Analysis of Outliers By Researcher

Mosaic Plot

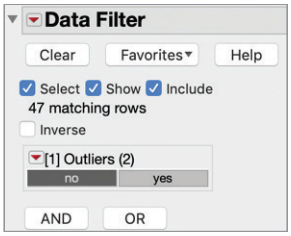
Tests

```
 N   DF   -LogLike   RSquare (U)
50    1   3.1249996       0.2242

Test              ChiSquare  Prob>ChiSq
Likelihood Ratio      6.250     0.0124*
Pearson               4.710     0.0300*
```

In this case, both p-values are less than 0.05, but greater than 0.01. Although there is enough evidence to conclude that Joe generates outliers with greater frequency than Jane, we should remain open to the possibility that there is no actual difference between them.

Before we conclude Part 1, let's finally return to the question we asked near its beginning: Is there a statistically significant difference in means between the ways Joe and Jane usually run their positive controls? First we use the Data Filter in JMP to hide and exclude the rows of our data table that have been tagged as outliers:



☑ Select ☑ Show ☑ Include

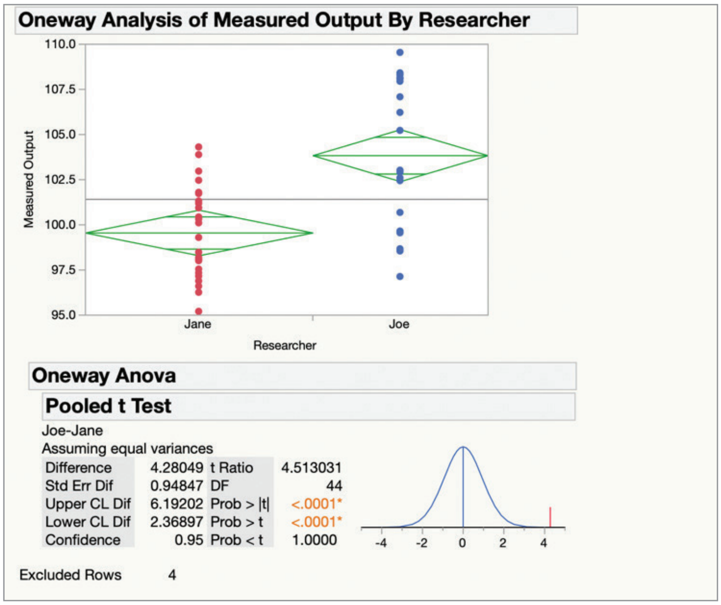
47 matching rows

□ Inverse

[1] Outliers (2)

no | yes

Now when we repeat the one-way/ANOVA without these four outliers, we get a different result: We see evidence ($p < 0.05$) that, on average, Joe measures higher outputs than Jane:



Oneway Anova

Pooled t Test

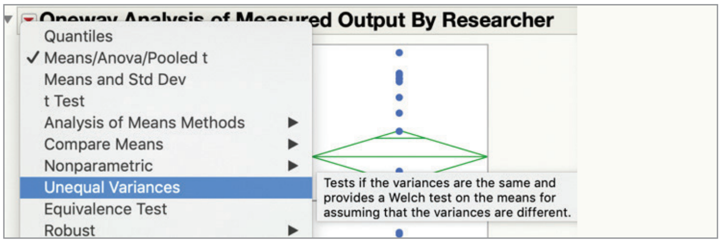
Joe-Jane

Assuming equal variances

Difference	4.28049	t Ratio	4.513031
Std Err Dif	0.94847	DF	44
Upper CL Dif	6.19202	Prob > |t|	<.0001*
Lower CL Dif	2.36897	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

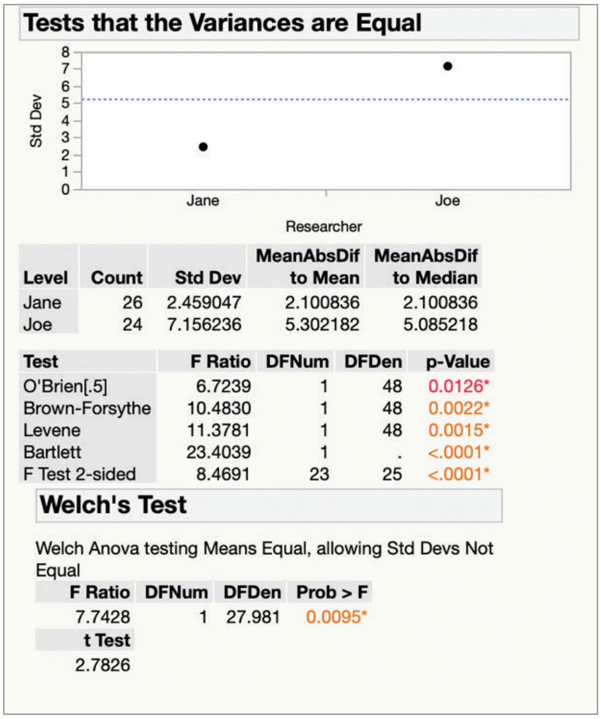
Excluded Rows 4

Moreover, even after excluding those four outliers, we observe that Joe's 20 remaining positive controls exhibit a wider range of outcomes than Jane's 26 positive controls, despite being fewer in number. Although that difference in ranges seems like evidence that Joe's experiments are more variable than Jane's, we can do a more rigorous test within the Fit Y by X platform by selecting the test for "Unequal Variances" from its red triangle menu:



Tests if the variances are the same and provides a Welch test on the means for assuming that the variances are different.

In this case, each of five distinct tests for unequal variances indicates that there is sufficient evidence ($p < 0.05$) that Joe's control experiments are still more variable than Jane's, even after excluding his four outliers:



Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Jane	26	2.459047	2.100836	2.100836
Joe	24	7.156236	5.302182	5.085218

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	6.7239	1	48	0.0126*
Brown-Forsythe	10.4830	1	48	0.0022*
Levene	11.3781	1	48	0.0015*
Bartlett	23.4039	1	.	<.0001*
F Test 2-sided	8.4691	23	25	<.0001*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
7.7428	1	27.981	0.0095*

t Test
2.7826

It is important to note that when we request the test for unequal variances, JMP will also perform Welch's Test, which is not a test for unequal variances. Instead, it's a test for a difference in means that relaxes the usual assumption of homoscedasticity (i.e., the assumption that the data associated with each level is pulled from normal distributions with identical standard deviations).

We can summarize the analyses we've done above in two ways. First, with a focus on the mechanics of data analysis:

1) We tested for a statistically significant difference in outlier frequency.
2) We tested for a statistically significant difference in means for the more representative subsets of Joe's and Jane's positive controls.
3) We tested for a statistically significant difference in variance for the more representative subsets of Joe's and Jane's positive controls. (We also tested for linear effect of time on Joe's and Jane's positive controls.)

Alternatively, we can focus on the interpretation and predictions of the data analysis:

1) There is evidence that, for some reason, Joe sometimes runs his positive control in a fundamentally different way that is not representative of the rest of his positive controls. We predict that having Joe and Jane partner to observe the different ways that they each execute their control experiments could help to identify a root cause that explains the observed difference in outlier frequency.
2) There is evidence that, for some reason, Joe tends to measure larger values than Jane for the same positive control. We predict that having Joe and Jane partner to observe the different ways that they each execute their control experiments could help to identify a root cause that explains the observed difference in their means.
3) There is strong evidence that, for some reason, Joe's data exhibits more variability than Jane's, even within the more representative subset of Joe's positive controls. We predict that having Joe and Jane partner to observe the different ways that they each execute their control experiments could help to identify a root cause that explains the observed difference in their standard deviations. (The correct hypothesis should be able to explain why Joe's positive controls have been drifting steadily upward over time.)

Although it is possible that Joe's failure to comply with one specific aspect of a documented standard operating procedure explains all six signals we have observed (four outliers, one difference in means and one difference in standard deviations), it is also possible that investigating Joe's variation will lead us to at least one previously overlooked factor that we can design into future experiments to improve our output mean (not just its stability). In this case, since all four outliers are of a similar magnitude, we might reasonably suspect that they share a single root cause. Also, since the observed difference in means and standard deviations can be traced to the same upward drift in Joe's positive controls, it seems likely that those two observations share a single root cause as well.

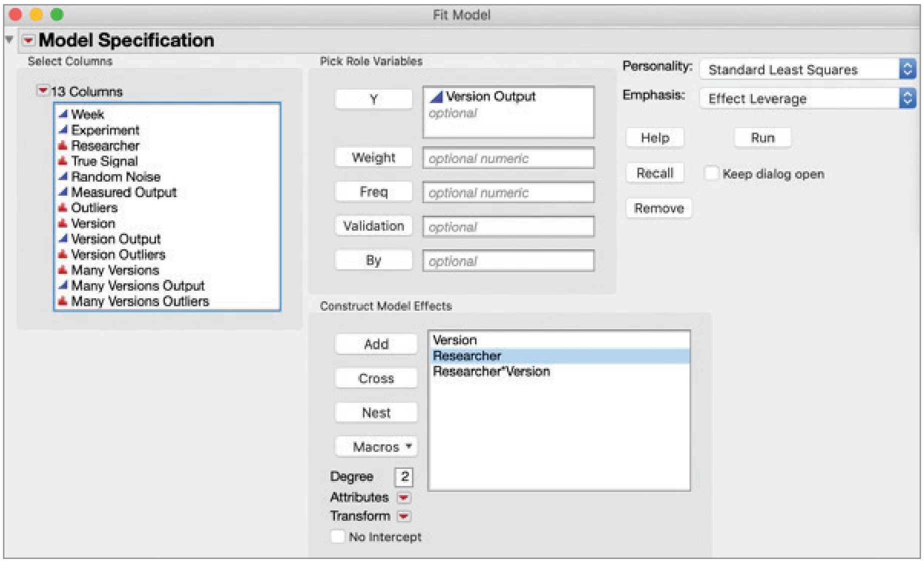
Whatever the root causes happen to be, they would be valuable to know. At minimum, improving the precision and reducing the failure rate of these experiments will reduce the time it takes our team to reach our shared goals for R&D or product development.

Part 2: How to use JMP to test for differences between researchers repeatedly running two related positive controls

As our R&D team progresses toward its goals for technology development, a performance gap widens between our latest, greatest results and our initial positive control. Never mind all the specific innovations we've made to the materials and equipment we use and how we use them – that performance gap alone is enough to generate suspicion that the old positive control may no longer be relevant to today's work. It's certainly possible that the old control experiment is still sensitive to all the same sources of variation as our latest test conditions, but it's reasonable to suspect it might not be.

If we update the conditions we run as positive controls from time to time, we ought to update how we use our statistical software as well, so we can extract as much insight as possible from our R&D team's limited budget for replication. Often the hardest part about using these features of our statistical software is simply knowing where to look for them.

The sample data table provided with this paper includes columns documenting the positive control "Version" corresponding to each "Version Output" value. In this case we have assumed the table's first 25 rows were recorded using Version A and the last 25 rows were recorded with Version B. Although we could apply the univariate and bivariate statistical methods described above to only the Version B subset of the data table, we could do a multivariate analysis of the full data set instead. For example, we can fit Version Output as function of Version, Researcher and an additional Researcher* Version interaction term using the Fit Model platform:



Fit Model

Model Specification

Select Columns

13 Columns
- Week
- Experiment
- Researcher
- True Signal
- Random Noise
- Measured Output
- Outliers
- Version
- Version Output
- Version Outliers
- Many Versions
- Many Versions Output
- Many Versions Outliers

Pick Role Variables

Y: Version Output
optional

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

Personality: Standard Least Squares

Emphasis: Effect Leverage

□ Keep dialog open

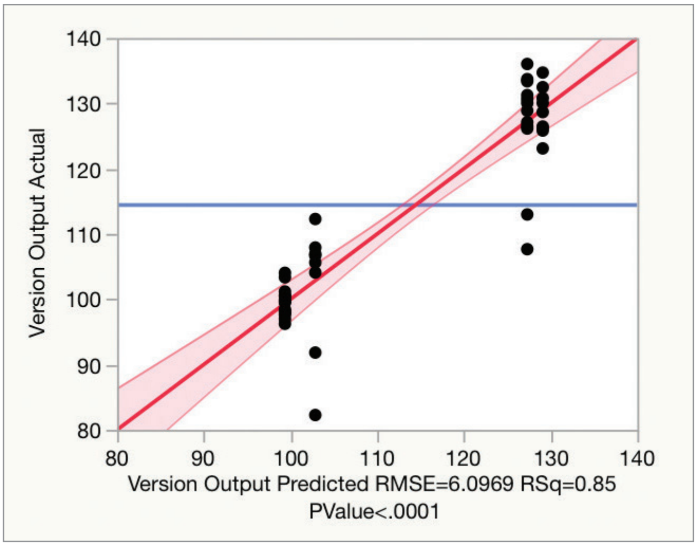
Construct Model Effects

Version
Researcher
Researcher*Version

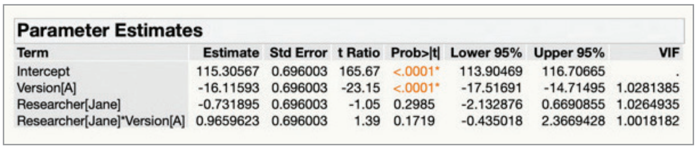
Degree: 2

□ No Intercept

The result is a very statistically significant model ($p < 0.0001$):



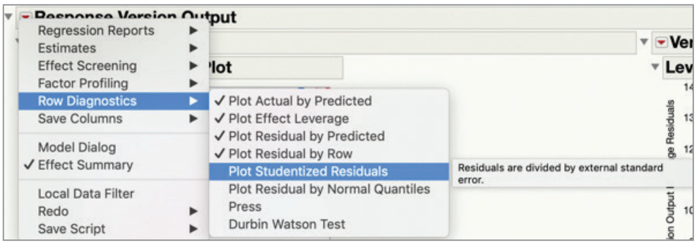
However, Version is the only statistically significant term in this model:



Parameter Estimates

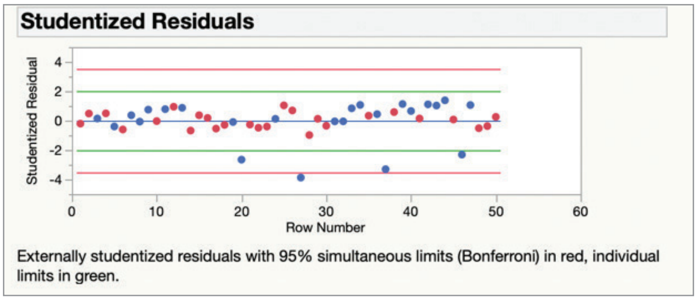
Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	115.30567	0.696003	165.67	<.0001*	113.90469	116.70665	.
Version[A]	-16.11593	0.696003	-23.15	<.0001*	-17.51691	-14.71495	1.0281385
Researcher[Jane]	-0.731895	0.696003	-1.05	0.2985	-2.132876	0.6690855	1.0264935
Researcher[Jane]*Version[A]	0.9659623	0.696003	1.39	0.1719	-0.435018	2.3669428	1.0018182

In the plot of actual versus predicted values above, we can see that there are four rows of our data table that are not fit as well by our model as the majority of our data. We can perform a rigorous test for outliers without leaving the Fit Model platform by clicking on its red triangle and selecting “Plot Studentized Residuals” from the “Row Diagnostics” menu:



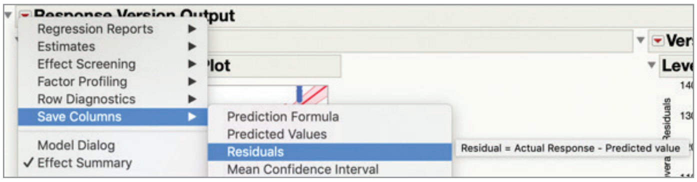
Residuals are divided by external standard error.

Our Fit Model report now includes a data visualization that, like the Control Chart example above, identifies statistically significant outliers as that data which lies outside the illustrated limits:

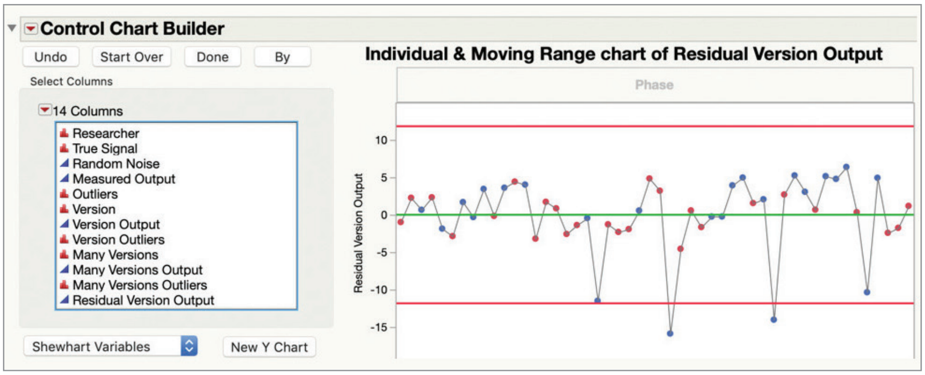


Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

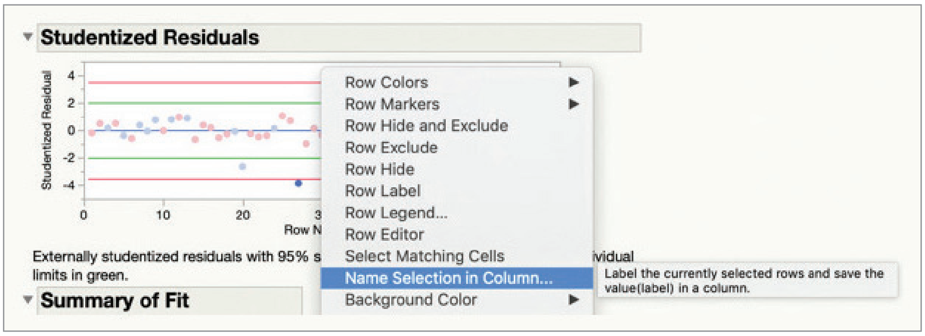
Although the specific statistical test that JMP is performing in this case is distinct from control chart logic, the data that gets flagged as outliers by each test is frequently the same. Although it is more convenient to test for outliers within the Fit Model report, if desired we can test for outliers using control chart logic by saving the residuals of our model as a column in our data table:



If we then drag the new Residual Version Output column into the Y position of the Control Chart Builder, we can verify that the same row of our data table is flagged as a statistically significant outlier by this alternate test:



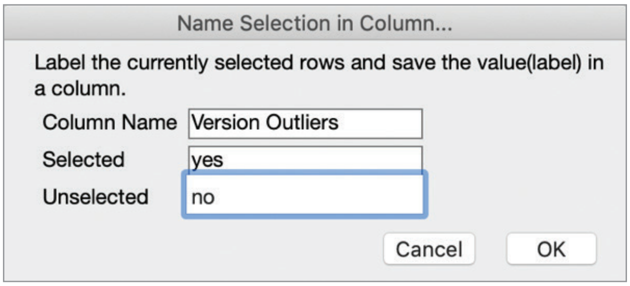
Although the Control Chart of the residuals flags one more of our four suspected outliers than the Studentized Residual chart, it seems plausible that both tests are suffering substantial false negative rates. As discussed in Part 1, to mitigate the false negative risks of our outlier tests it is wholly appropriate, after tagging the first statistically significant outliers, to iteratively filter, re-fit and re-test our data for additional outliers. We can do this iteration without leaving the Fit Model platform. First, we select the one outlier in the Studentized Residuals chart, right-click and select "Name Selection in Column…":



Externally studentized residuals with 95% s ividual limits in green.

Label the currently selected rows and save the value(label) in a column.

We can create a new column called "Version Outliers" with values of "yes" for the one selected point and "no" for the 49 unselected points:



Name Selection in Column...

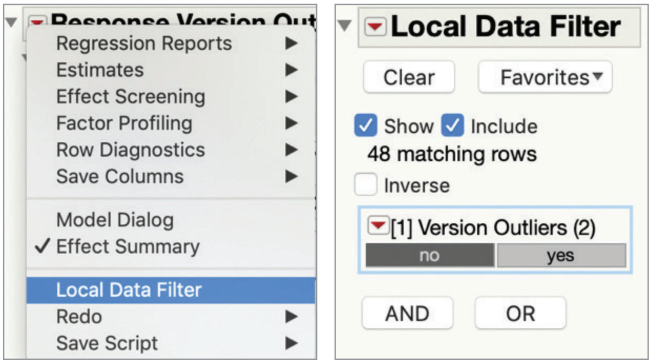
Label the currently selected rows and save the value(label) in a column.

Column Name: Version Outliers

Selected: yes

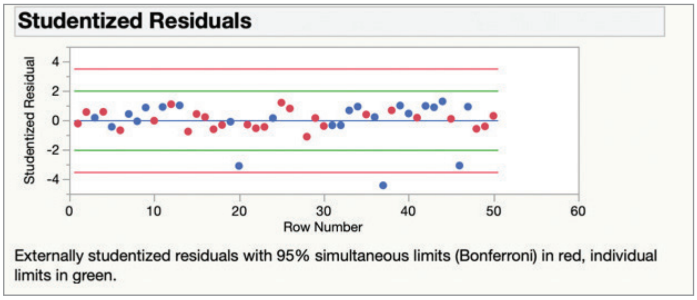
Unselected: no

Next, we add a Local Data Filter to our Fit Model report and filter to "Version Outliers" = "no":



48 matching rows

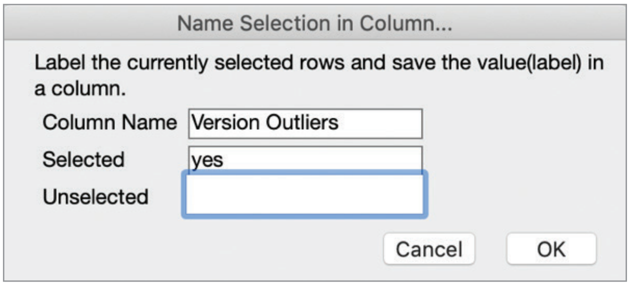
When we revisit the Studentized Residuals chart, we see that one previously untagged row of our data table is now flagged as an outlier:



Studentized Residuals

Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

As described above, the second time we utilize the "Name Selection in Column…" feature, it is critical that we do not write any values to the unselected data:



Name Selection in Column...

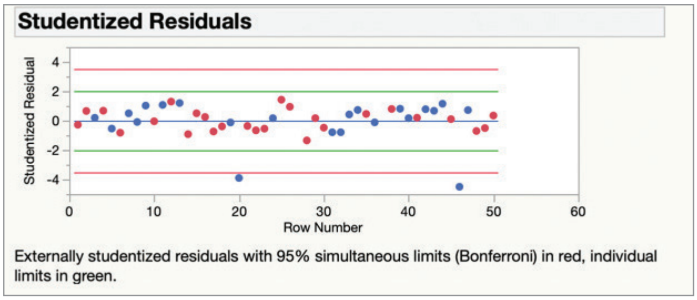
Label the currently selected rows and save the value(label) in a column.

Column Name: Version Outliers

Selected: yes

Unselected:

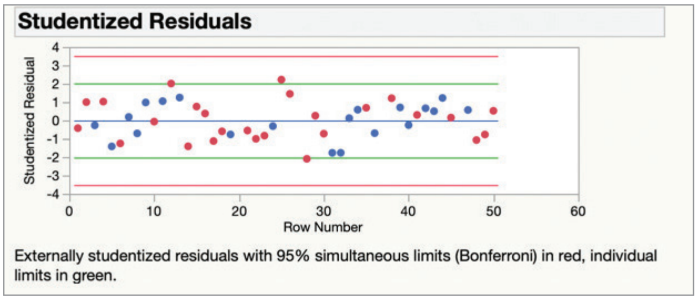
When we click OK this second time, our Studentized Residuals chart will update automatically as the Local Data Filter now removes both "yes" rows from the analysis:



Studentized Residuals

Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

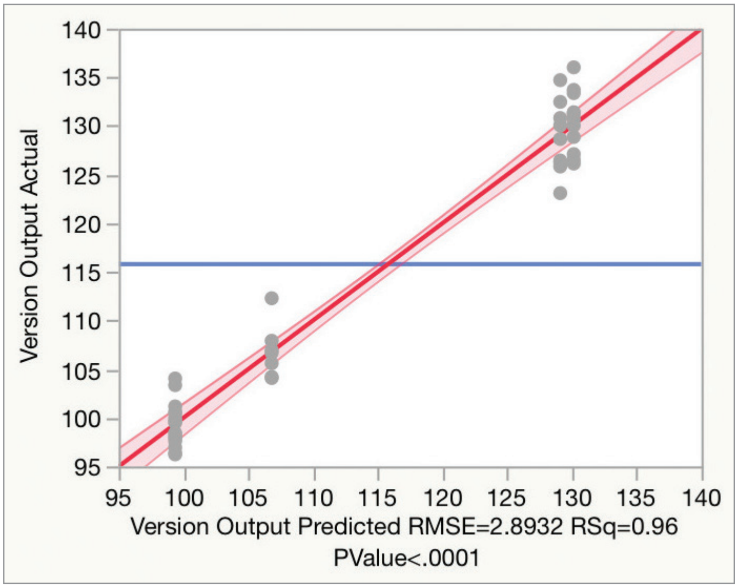
Two additional rows of our data table are now flagged as outliers. As soon as they are tagged, our Studentized Residuals chart is updated once again:



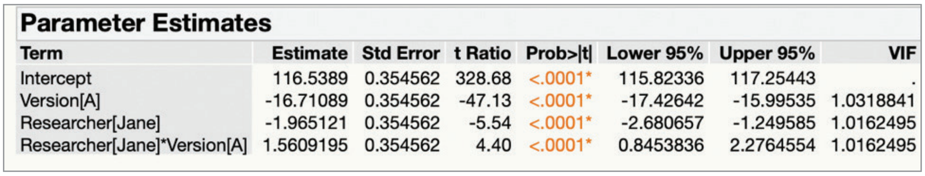
Studentized Residuals

Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

Since there are no additional rows of our data table flagged as outliers, we are finally ready to re-interpret our multivariate model:



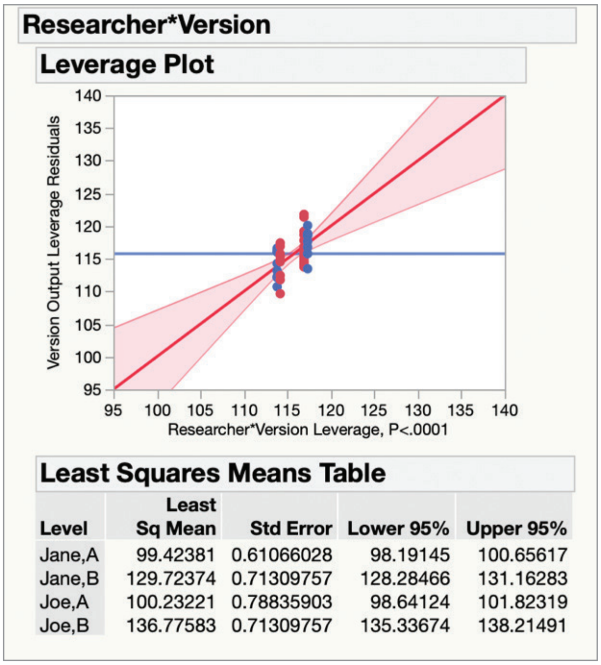
When we direct our modeling toward this more representative subset of our data (46 of 50 rows), we find the Researcher and Researcher*Version terms are now quite statistically significant (p < 0.0001), both having been insignificant (p > 0.05) before filtering the four outliers:



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	116.5389	0.354562	328.68	<.0001*	115.82336	117.25443	.
Version[A]	-16.71089	0.354562	-47.13	<.0001*	-17.42642	-15.99535	1.0318841
Researcher[Jane]	-1.965121	0.354562	-5.54	<.0001*	-2.680657	-1.249585	1.0162495
Researcher[Jane]*Version[A]	1.5609195	0.354562	4.40	<.0001*	0.8453836	2.2764554	1.0162495

We can complement what we learn from the p-values of our parameter estimates by looking at their 95% confidence intervals. For example, when we look at the Least Squares Means Table for the Researcher*Version interaction term, we see that the confidence intervals for the "Jane,A" and "Joe,A" levels overlap, but the confidence intervals for the "Jane,B" and "Joe,B" levels do not:

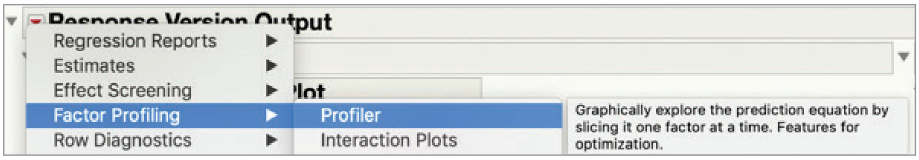

Researcher*Version

Leverage Plot

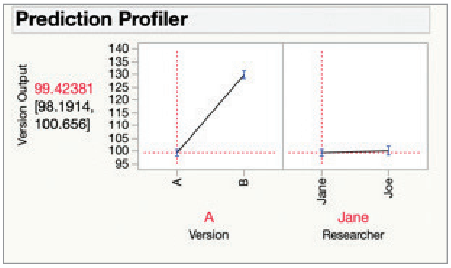
Least Squares Means Table

Level	Least Sq Mean	Std Error	Lower 95%	Upper 95%
Jane,A	99.42381	0.61066028	98.19145	100.65617
Jane,B	129.72374	0.71309757	128.28466	131.16283
Joe,A	100.23221	0.78835903	98.64124	101.82319
Joe,B	136.77583	0.71309757	135.33674	138.21491

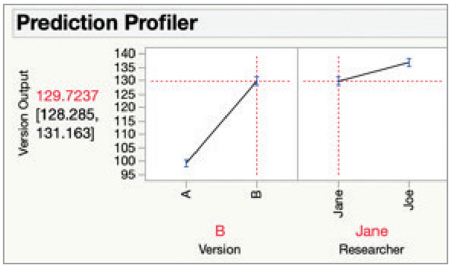
In other words, although there was no statistically significant difference between Researchers for Version A, Joe's values increased relative to Jane's when we started to run Version B. We can illustrate the same conclusions another way by turning on and interacting with the JMP Profiler:



Graphically explore the prediction equation by slicing it one factor at a time. Features for optimization.

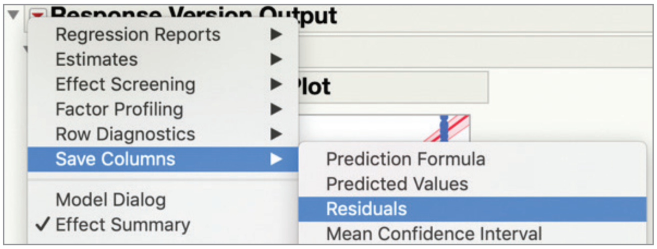

Prediction Profiler

Version Output 99.42381 [98.1914, 100.656]

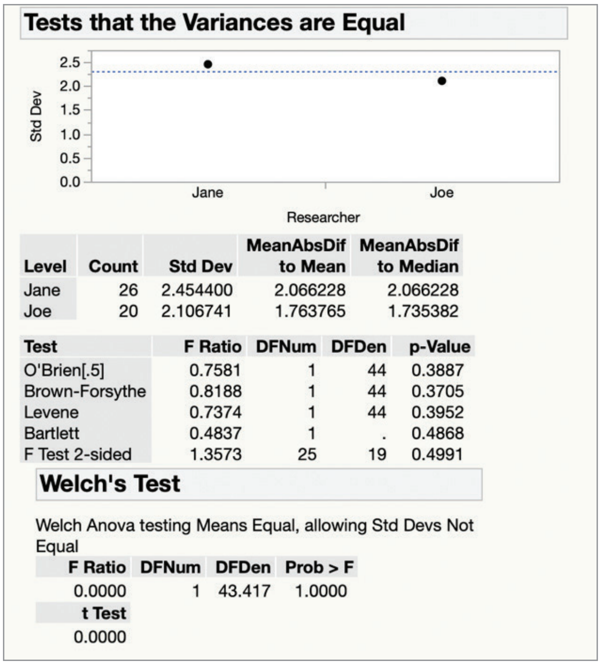

Prediction Profiler

Version Output 129.7237 [128.285, 131.163]

Having fit this more representative subset of our data table, we can next test the hypothesis that different researchers have run their positive controls with different degrees of random variation. However, since the Fit Model platform does not have its own feature to test for unequal variances in its residuals, we must first save those residuals as a new column in our data table:



We can then use the Fit Y by X platform in the same way as illustrated above to test for unequal variances when Y = Residual Version Output and X = Researcher:



Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Jane	26	2.454400	2.066228	2.066228
Joe	20	2.106741	1.763765	1.735382

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.7581	1	44	0.3887
Brown-Forsythe	0.8188	1	44	0.3705
Levene	0.7374	1	44	0.3952
Bartlett	0.4837	1	.	0.4868
F Test 2-sided	1.3573	25	19	0.4991

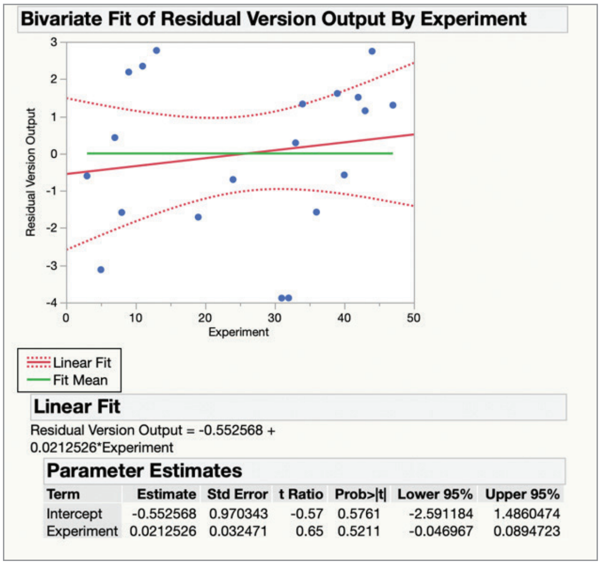
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
0.0000	1	43.417	1.0000

t Test
0.0000

In this case, all the p-values are now greater than 0.05, indicating that there is no longer evidence that the different researchers have run their positive controls with different degrees of random normal variation. Similarly, we can no longer detect a linear relationship when we filter to Joe's representative subset of the data table and fit Residual Version Output as a function of Experiment:



Bivariate Fit of Residual Version Output By Experiment

Linear Fit

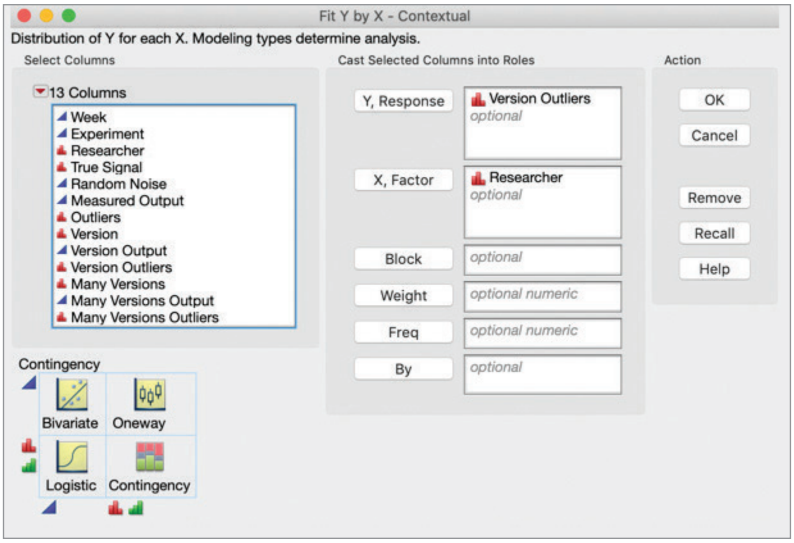
Residual Version Output = -0.552568 + 0.0212526*Experiment

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-0.552568	0.970343	-0.57	0.5761	-2.591184	1.4860474
Experiment	0.0212526	0.032471	0.65	0.5211	-0.046967	0.0894723

Although these tests are no longer detecting statistically significant signals, the upward trend in Joe's results over time hasn't actually disappeared: It now appears as the statistically significant interaction term in our model of the means.

Having tagged four of our 50 runs as outliers, we can also test for statistically significant differences between researchers with respect to outlier frequency. If we assume that these outlier frequencies are independent of which version Joe and Jane are executing, we can simply run the same Fit Y by X platform as described above:



Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

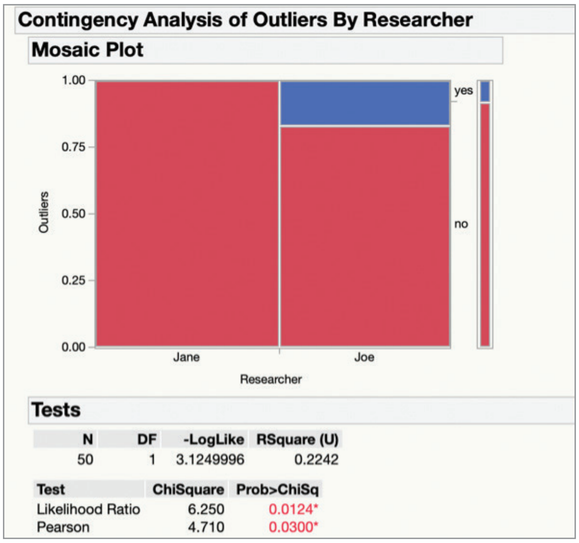
Select Columns

13 Columns

- Week
- Experiment
- Researcher
- True Signal
- Random Noise
- Measured Output
- Outliers
- Version
- Version Output
- Version Outliers
- Many Versions
- Many Versions Output
- Many Versions Outliers

Cast Selected Columns into Roles

Y, Response	Version Outliers optional
X, Factor	Researcher optional
Block	optional
Weight	optional numeric
Freq	optional numeric
By	optional



Contingency Analysis of Outliers By Researcher

Mosaic Plot

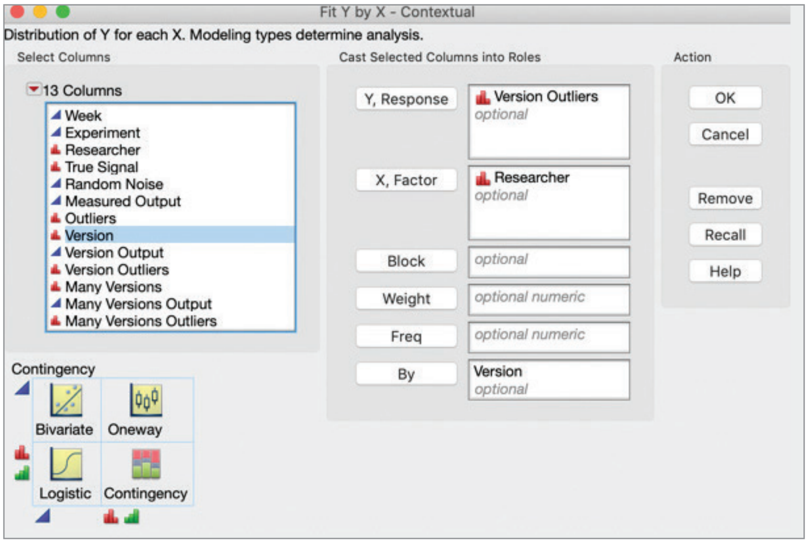
Tests

N	DF	-LogLike	RSquare (U)
50	1	3.1249996	0.2242

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	6.250	0.0124*
Pearson	4.710	0.0300*

Since each of the p-values in these tests is less than 0.05, we can conclude that there is something fundamentally different between Jane's and Joe's approaches to running their positive controls (both Versions A and B), some root cause (at least one) that leads Joe to generate outlying results more often than Jane.

We can relax the assumption that Joe and Jane have the same outlier frequencies for both Versions A and B a couple different ways. One is to introduce "Version" as a "By" variable in the Fit Y by X platform:


Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

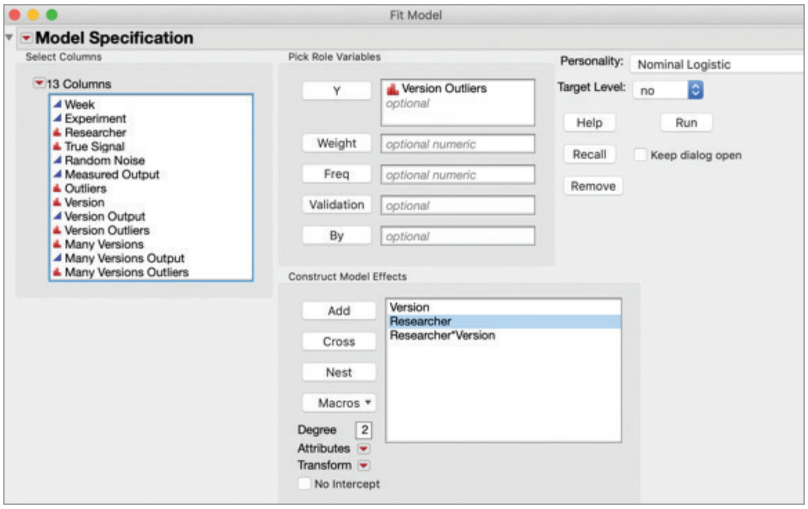
Select Columns

13 Columns
- Week
- Experiment
- Researcher
- True Signal
- Random Noise
- Measured Output
- Outliers
- Version
- Version Output
- Version Outliers
- Many Versions
- Many Versions Output
- Many Versions Outliers

Cast Selected Columns into Roles

- Y, Response: Version Outliers
- X, Factor: Researcher
- Block: optional
- Weight: optional numeric
- Freq: optional numeric
- By: Version

A second is to add a Researcher*Version interaction term to the Fit Model analysis of the nominal variation in the Version Outliers column:


Fit Model

Model Specification

Select Columns

13 Columns
- Week
- Experiment
- Researcher
- True Signal
- Random Noise
- Measured Output
- Outliers
- Version
- Version Output
- Version Outliers
- Many Versions
- Many Versions Output
- Many Versions Outliers

Pick Role Variables

- Y: Version Outliers
- Weight: optional numeric
- Freq: optional numeric
- Validation: optional
- By: optional

Personality: Nominal Logistic

Target Level: no

□ Keep dialog open

Construct Model Effects

- Version
- Researcher
- Researcher*Version

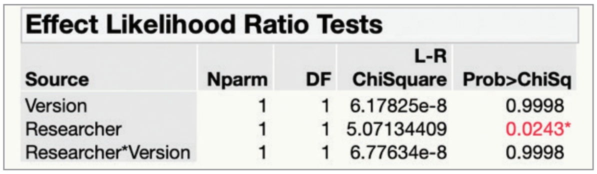
Degree 2

□ No Intercept

The disadvantage of using the "By" variable for testing the two versions is that we lose statistical power when we perform two independent analyses on each of two smaller data tables. Unsurprisingly, all of the various p-values calculated this way are >0.05 (not shown). By using Fit Model instead, we can use the full data table to test multiple hypotheses simultaneously and with maximum statistical power:

1) Assuming researcher doesn't matter, is there evidence that different versions of our positive control generate outlying results with different frequencies?
2) Assuming version doesn't matter, is there evidence that Joe and Jane generate outliers with different frequencies?
3) Is there evidence that a particular combination of version and researcher is especially prone to outlying behavior?

In this case we see that there is only evidence ($p < 0.05$) that the second hypothesis is true:



Effect Likelihood Ratio Tests

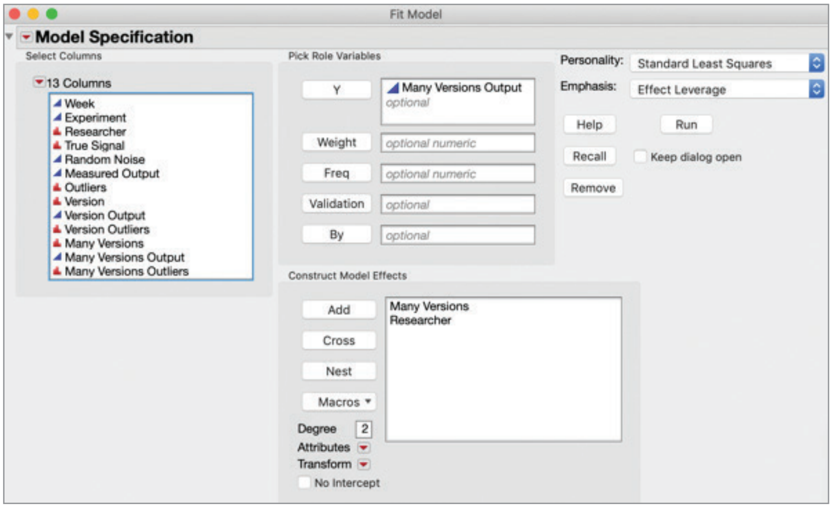
			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Version	1	1	6.17825e-8	0.9998
Researcher	1	1	5.07134409	0.0243*
Researcher*Version	1	1	6.77634e-8	0.9998

Although it's possible the other hypotheses are also true, we just don't have enough data yet to resolve those effects if they exist.

Part 3: How to use JMP to test for differences between researchers running rapidly evolving positive controls

For many of us in R&D, the previous examples may have seemed foreign – even fictitious. Who among us ever runs the same positive control 50 consecutive times? Or even 25 times? Maybe the analytical chemists on our team have the luxury of measuring the same standard samples over and over and over again,[2] but for many of us, each run of our experiment consumes a lot of time and money. If we're going to replicate anything at all, we'd rather direct our restricted replication budget toward our latest, greatest results.

The columns in our sample data table called "Many Versions" and "Many Versions Output" simulate a scenario in which we change what we run as a positive control every five experiments or so. Even though the 10 versions of our positive control were run as few as three times each – and some were never run by more than one of the researchers – we can still use Fit Model to test for a version-independent difference in means between researchers:


Fit Model

Model Specification

Select Columns

13 Columns
- Week
- Experiment
- Researcher
- True Signal
- Random Noise
- Measured Output
- Outliers
- Version
- Version Output
- Version Outliers
- Many Versions
- Many Versions Output
- Many Versions Outliers

Pick Role Variables

Y: Many Versions Output

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

Personality: Standard Least Squares

Emphasis: Effect Leverage

□ Keep dialog open

Construct Model Effects

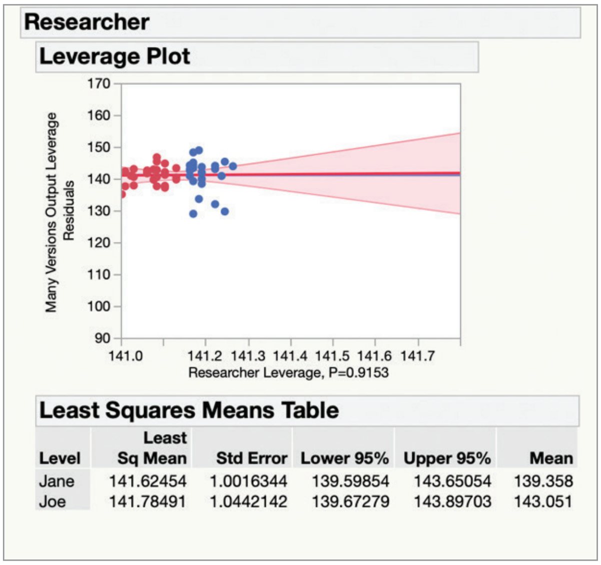
Many Versions
Researcher

Degree: 2

□ No Intercept

[2] In this author's experience, even relatively simple and routine analytical methods are more likely than not to exhibit nonrandom variation until a deliberate effort has been made to stabilize them. As much as we might wish our team's experimental methods to be inherently stable (or its eyes to be especially observant or its hands to be especially reliable or its hearts to be especially caring), we should neither overestimate the cost of doing the work described in this manuscript nor underestimate the costs of misinterpreting nonrandom variation as the signals we had intended to study.

Our first analysis using all 50 measurements made with all 10 versions suggests there is no difference between researchers:



Researcher

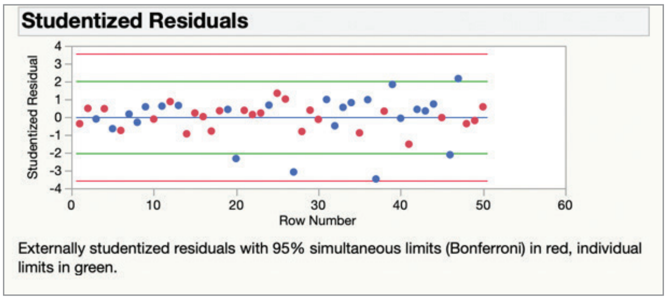
Leverage Plot

Many Versions Output Leverage Residuals
170
160
150
140
130
120
110
100
90
141.0
141.2 141.3 141.4 141.5 141.6 141.7
Researcher Leverage, P=0.9153

Least Squares Means Table

Level	Least Sq Mean	Std Error	Lower 95%	Upper 95%	Mean
Jane	141.62454	1.0016344	139.59854	143.65054	139.358
Joe	141.78491	1.0442142	139.67279	143.89703	143.051

However, we should always look for evidence of nonrandom variation in our data set before accepting the conclusions of our first analysis. When we add the Studentized Residuals chart to our report, we see that none of our data appears outside the red limits for outlier detection:

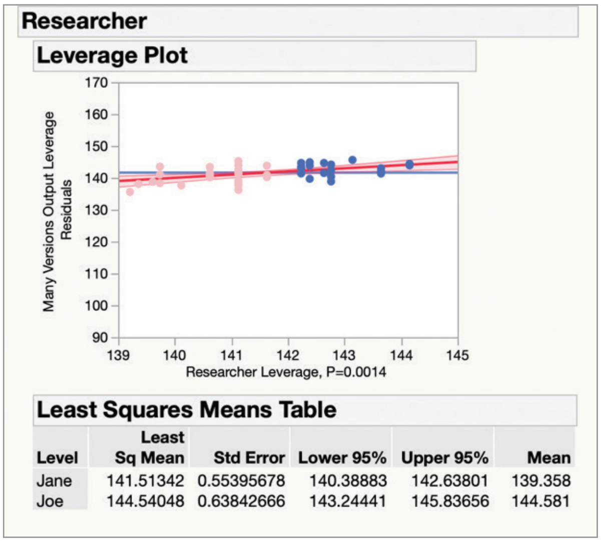


Studentized Residuals

Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

However, as discussed above, these outlier tests can have substantial false negative rates, as each additional outlier in the data set inflates the estimate of random normal variation used to test for outlying behavior. If excluding any one of the suspected outliers would place another outside the red limits, it's reasonable to tag both of those points as outliers. In this case, as soon as we tag and filter just the one residual furthest from 0, our updated model immediately identifies a second outlier (not shown). Filtering this second outlier reveals a third. Filtering this third outlier then reveals a fourth.[3] When we filter out all four outliers, we can now detect a statistically significant Researcher effect (p = 0.0014):

[3] Because we can now count some numbers of version replicates on one hand, some of the studentized residuals will appear strongly anti-correlated with each other. For example, if one of Joe's two replicates running Version I has a negative residual in our model, the other will likely have a positive residual. If only one of those replicates is a true outlier, there is some risk that both of them will appear outside the red limits. In order to mitigate this false positive risk, consider tagging and filtering only one additional outlier at a time – generally the one furthest from zero – until no more of the data appears outside the red limits. If both points are outlying and equally distant from zero – especially if we have no prior reason to expect either positive or negative outliers – filter out both of the outlying points to reduce the risk that our estimate of the researcher effect would be biased in either direction.



Researcher

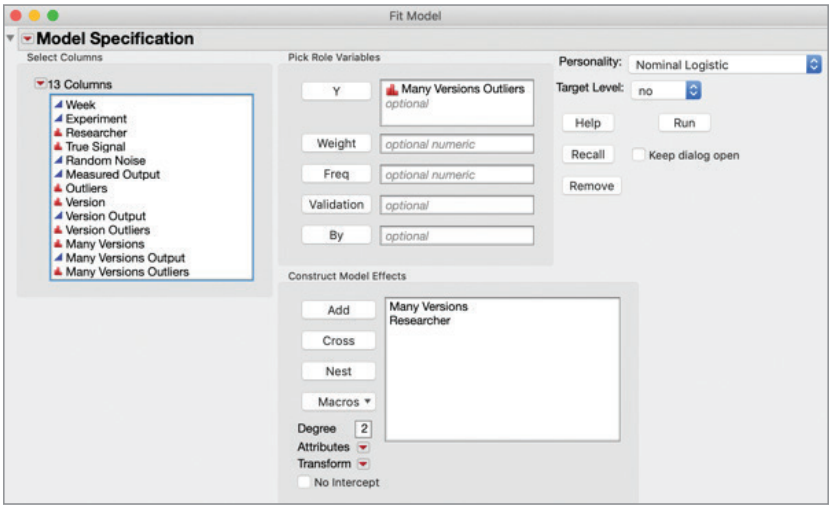
Leverage Plot

Least Squares Means Table

Level	Least Sq Mean	Std Error	Lower 95%	Upper 95%	Mean
Jane	141.51342	0.55395678	140.38883	142.63801	139.358
Joe	144.54048	0.63842666	143.24441	145.83656	144.581

If we want to relax the implicit assumption that the researcher effect has the same magnitude across all versions, we are going to lose an awful lot of statistical power. Each additional version in our data set requires an additional interaction term in our model, which reduces the degrees of freedom we use to measure its error. In other words, even if there were a real difference between researchers for just Version F, it will be harder to detect when Joe has run Version F only three times and Jane has run Version F only once; and, of course, it will be impossible to detect that difference for Version J, which Joe has never run at all.

Having saved the outlier information to its own column, we can again test for a significant difference in outlier frequency between researchers using either the Fit Y by X or Fit Model platforms (only the latter is shown here):



Personality: Nominal Logistic

Target Level: no

Y: Many Versions Outliers

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

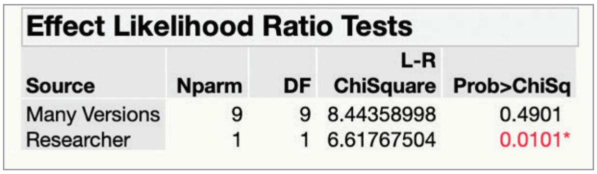
□ Keep dialog open

Model Effects:
Many Versions
Researcher

Degree 2

□ No Intercept

In both cases, we observe evidence ($p < 0.05$) that there is a real difference between Joe's and Jane's outlier frequencies (under the implicit assumption that those rates are constant across all versions):



Effect Likelihood Ratio Tests

			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Many Versions	9	9	8.44358998	0.4901
Researcher	1	1	6.61767504	0.0101*

It is similarly straightforward to test for unequal variances among the residuals of our fit to the Many Versions output data. That said, in Part 2 this test already failed to detect a difference in standard deviations when Joe's real drift was instead detected as a statistically significant interaction between the researcher and the two versions of his positive control. In Part 3 we have been forced by our highly stratified data to drop the interaction terms from our model, but the fact that our model is constructed by minimizing a single sum of squared residuals across both researchers and all 10 versions means that Joe's real drift is now hopelessly confounded with – and biasing – our estimates of the many version means. We suspect this multiparameter minimization will end up balancing the unexplained variation in our model between Joe's and Jane's residuals, and a test for unequal variances confirms that suspicion by exhibiting p-values > 0.05 (not shown).

Additional Commentary

What can happen when we change our positive controls too frequently

If we imagine the number of positive control versions approaching the total number of rows in our data table, the real shifts in Joe's outcomes become even harder to detect. For one, we expect the p-value for the Researcher effect on means to climb as the number of parameters in our multivariate linear model approaches the number of data points we feed it. For another, even Joe's outlying results may go undetected if they are no longer outlying enough to stick out from the real variation our experiments were intended to study.

Perhaps the best-case scenario is that these undetected signals merely inflate the Root Mean Square Errors (RMSEs) of our multivariate linear models. Of course, as our RMSE increases, so does the false negative risk that we measure $p > 0.05$ for many of the truly significant factors we are testing with our experiments. To be clear, the 95% confidence intervals of our parameter estimates do not necessarily become less accurate – i.e., on average, 95% of them should still contain their true values – but they can become so large that many of them now contain zero as well.

A more likely scenario when we rarely repeat the same controls is that these undetected signals become so confounded with the factors tested by our experiments that well over 5% of our 95% confidence intervals no longer contain their true values. Some parameters may be observed to be statistically significant when they are truly zero, while other parameters could be estimated as positive when they are in fact negative.

Although Parts 1, 2 and 3 of this article illustrated how important it is to tag and filter the outliers in any data set, the tools that make it so easy to identify real outliers in repeated data can make it too easy to tag the wrong data when no conditions are repeated. The false positive risk for mistakenly tagging the wrong data as outliers increases as (1) the degrees of freedom in our model decrease and (2) the real variation we intend to study gets closer in size (either larger or smaller) to the nonrandom variation we wish were zero. Although it's tempting to remove that one leverage point from our experimental design that singlehandedly accounts for most of our model's RMSE, we should be aware of this risk we incur as we do so.

Of course, model-building should never be the last thing we do on our projects; it is always prudent to validate our models with at least one final experiment. We can even take care to focus those validation runs on whichever regions of the design space would be most sensitive to the different assumptions we might make when fitting our data (e.g., which predictions are most sensitive to whether or not we train our model using the rows suspected of being outliers?).

What can happen when we normalize each experiment's results to its positive control
Some researchers on our R&D teams may be thinking, "But that's why we always run positive controls. It doesn't matter if Joe's results are shifting or drifting or occasionally outlying. As long as he includes some kind of positive control or reference condition in each of his experiments, he can still measure the signals his experiments were designed to measure by differencing or normalizing each experiment's results to its reference condition."

Sometimes this is true, but this claim makes two implicit assumptions that are too often untrue if we have not done the work to validate them explicitly. When we difference or normalize our measurements against a control, we implicitly assume:

1) The nonrandom drivers of variation within each experiment influence each of its measurements identically.

2) The magnitude of random variation in each measurement is small relative to the signals we are trying to measure.

Imagine that we've investigated the drift in Joe's data revealed by the positive controls of Part 1 (or, equivalently, the shift in Joe's data revealed by the positive controls of Part 2), and we've determined that what's driving Joe's variation is that his thermocouple has been drifting more quickly than expected since its last scheduled calibration. Going forward, we'll check our thermocouple calibrations more frequently, and we'll replace Joe's misbehaving thermocouple as well. Looking backward, what do we make of the last three months of Joe's experiments?

If it turns out the thermocouple drift has been fast relative to the time between runs in each of Joe's experiments, then each run has its own unique bias that cannot be completely removed by simply differencing or normalizing to a contemporaneous reference condition. We could, in this case, develop a more sophisticated way to process our data that accounts for the measured rate of thermocouple drift (and assumes it has been constant). But if we had never measured that rate by compiling our positive controls in the first place, the undetected thermocouple drift would confound our results, leading to higher rates of false positives and false negatives than we would predict from the random component of variation in Joe's and Jane's data. Those preventable false negatives would slow our R&D team's progress. Those preventable false positives would keep consuming precious time and money until the day we stop complaining about our lab's "reproducibility problem" and start dedicating more resources to solving it.

If we can demonstrate that the rate of materially significant thermocouple drift has been slow relative to the time between the runs in each of his experiments, we can have some confidence that a simple differencing or normalization of the measurements within each experiment is correcting those data for the nonrandom and correlated variation attributable to this nuisance factor. However, since there is always some amount of random and uncorrelated variation present in every measurement, we can never completely isolate the shift, drift or outlying variation in our positive controls from the signals in the rest of our experiment. Even when the normalization of data within experiments genuinely increases the comparability of data between experiments, data that are unbiased to begin with will always generate higher signal-to-noise ratios than differenced or normalized data.

It doesn't matter if the random fluctuations within each experiment are small relative to its nonrandom bias. What matters is that these random fluctuations are small relative to the signals that, if detected, would move our R&D program substantially closer to its goals. When we don't explicitly demonstrate that this is true, we can waste considerable resources running insufficiently powered experiments or chasing the random noise that is introduced specifically by normalizing our data against our positive controls. When the random variation in a single positive control happens to be positive, we may misinterpret all our experiment's test conditions as disappointments. When the random variation in a single reference condition happens to be negative, we may conclude that whoever designed this week's experiment is somehow smarter than whoever designed last week's experiment. When we then fail to reproduce any of those promising results, those undeserved reputational differences may linger and bias important resourcing decisions that can't wait for our statistical software to stop making bad predictions with the ineffectively normalized data we've been feeding it.

Conclusion

Every detail of this article would remain relevant if we substituted Instrument, Day of the Week, Site or any number of other nominal factors everywhere that it currently reads Researcher. Regardless of which nuisance factor turns out to explain a surprising amount of the variation in our data tables – whether it's a who, a what, a when or a where – too often we assume the noise in our data is normally distributed when it is not. In other words, too often we assume our research processes are sufficiently stable before we have invested the effort to make them so.

If we are already running one or more positive controls in every experiment we do, let's make sure we are learning as much as possible from this significant commitment of our time and money. Here are the five keys to getting this right:

1) Compile positive control data for all of our related research processes into a single data table that spans numerous experiments.
2) At minimum, add columns to this data table that document who ran these positive controls, when they were run and where they were run (e.g., which reactor or instrument was used?). Record any other nuisance factors that could explain some of today's observed variation, even if our ultimate goal is to develop processes that are independent of those factors.
3) Use statistical software such as JMP to analyze our data as a function of both the factors we intended to study and those we didn't.
4) Resist the urge to invent a new positive control with every experiment. Although we will still need to change our positive controls from time to time, we should take deliberate care to avoid changing our positive controls so often that we can no longer measure any materially significant effects of who, what, when or where on our experiments.
5) Whenever an unexpected signal in our data is flagged as statistically significant, prioritize the investigation to determine its root cause. If necessary, negotiate an extension on the deadline for the next experiment. Sometimes these underappreciated root causes can be identified in less than an hour! Even when it takes days to identify the root cause, this investment can save us weeks, if not months, of drawing the wrong conclusions from the work we do every day. Unless it's already December, these investigations will almost always improve our chances of hitting our year-end targets for technology development, regardless of whatever delay they impart to next week's previously scheduled experiment.

There's just no substitute for detecting and addressing those root causes of the unexplained variation in our data tables! The first step is to stop thinking about our positive controls as something we do to normalize the results of each experiment and to start planning to use them to measure the day-to-day, week-to-week and month-to-month stability of our various research processes.

About SAS and JMP

JMP® is a software solution from SAS that was first launched in 1989. John Sall, SAS co-founder and Executive Vice President, is the chief architect of JMP. SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 83,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.

To contact your local JMP office, please visit: jmp.com/offices