How to Use Graphs to Diagnose and Deal with Bad Experimental Data

By Mark J. Anderson and Patrick J. Whitcomb, Stat-Ease, Inc. *www.StatEase.com*

SUMMARY

This article deals with thorny issues that confront every experimenter – how to handle individual results that do not appear to fit with the rest of the data. It provides graphical tools that make it easy to diagnose what's really wrong with response data – damaging outliers and/or a need for transformation. The trick is to maintain a reasonable balance between two types of errors:

- Deleting data that vary only due to common causes, thus introducing bias to the conclusions.
- Not detecting true outliers that occur due to special causes. Such outliers can obscure real effects or lead to false conclusions. Furthermore, an opportunity may be lost to learn about preventable causes for failure or reproducible conditions leading to breakthrough improvements (making discoveries more or less by accident).

You will see two real-life data sets that don't reveal their secrets at first glance. However, with the aid of various diagnostic plots (readily available in off-the-shelf statistical software), it becomes much clearer what needs to be done. Armed with this knowledge, quality professionals will be much more likely to draw the proper conclusions from experiments that produce bad (discrepant) data.

INTRODUCTION

Personal computer software makes it very easy to fit models to experimental data via least-squares regression. However, these models often prove susceptible to outliers created by special causes. Such outliers occur with alarming frequency due to:

- Errors in data entry. It's very easy to miss a decimal point or accidentally press the wrong key. (Suggestion: If you type data from top to bottom into a response column on the computer, proof-read them from bottom to top.)
- Breakdowns in equipment.
- Mistakes on the part of the people operating the process.
- Non-representative samples.
- Bad measurements.
- Unknown lurking variables that appear only intermittently.

On the other hand, all experimenters must be careful not to bias their results by deleting data that does not meet their preconceived notions. In many cases the data deviates from the standard assumptions that variations are normally distributed with zero mean and a fixed variance. In such cases, outliers may be falsely reported when the real problem is that the response needs to be transformed by the log or some other function.

Table 1 shows how an experimenter can be correct or in error about the presence or absence of true outliers, that is, data produced by special causes.

Outlier(s)?		**What you say:**	
		Yes (present)	**No (absent)**
The truth:	**Yes**	Correct	False Negative
	No	False Positive	Correct

Table 1: Errors in judging whether or not outliers are present in experimental data

Correctly identified outliers should not just be thrown away. They might reveal something of great value. For example, despite the presence of a satellite that collected the necessary data, it took many years before scientists realized the presence of a hole in the ozone layer over the Antarctic. Unfortunately the data acquisition system automatically deleted outliers caused by the intermittent hole so it never got reported.[1]

Statisticians have developed very powerful graphical methods for diagnosing abnormalities in data, detecting potential outliers, and suggesting possibly beneficial transformations. Many of these diagnostics will be shown in this article, with references provided for those who want to dig up the details. As will be demonstrated via case study, it would be a serious mistake not to take advantage of these methods before drawing conclusions about the outcome of an experiment.

Two case studies follow, both of which detail results from design of experiments (DOE). They illustrate situations where an unwary experiment might either overlook real outliers that obscure the true effects (false negative) or throw out data that can be explained via an appropriate response transformation (false positive). See how early you can guess which error is illustrated in each case.

THE SECRET TO LONG LIFE REVEALED!

George Box reported a great success story for two-level factorial DOE that focused on improving the life of a deep-groove rolling bearing.[2] Figure 1 shows the factors and the astonishing results (hours of bearing life) in the form of a cube plot (the numbers in parentheses show the standard design order).

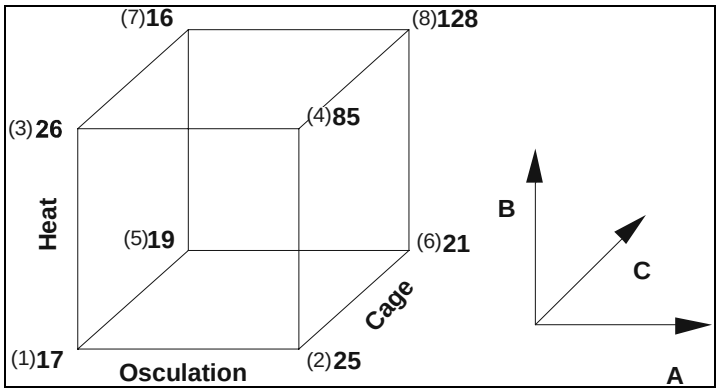


Figure 1: Cube plot of bearing experiment

Let's do a statistical analysis of this data using techniques developed by Box and his predecessors. Figure 2 shows the half-normal plot of effects.[3] Factors A, B and their interaction AB stand out on the absolute scale of effect on bearing life. However, notice that the smaller effects (points not labeled) do not line up with the origin of the half-normal plot. This is an abnormal pattern.

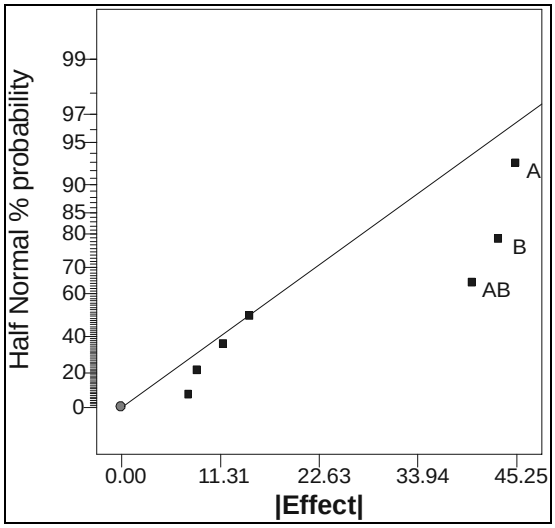


Figure 2: Half-normal plot of effects for bearing experiment

Analysis of variance (ANOVA) for the modeled effects (A, B and AB) shows a high level of significance ($p < 0.05$), but, as shown in Figure 3, diagnosis of the externally studentized residuals,[4] a common method for detecting discrepant data that some software labels "outlier t," reveals two potential outliers in the data – points 4 and 8. (Note: the x-axis on this plot displays "Run" number, presumably randomized, but it's shown in standard order to be consistent with Figure 1.) These two discrepant points fall more than six standard deviations from their expected value (the zero line on the plot), well above the 99% confidence level (alpha = 0.01 risk) for the appropriate test of significance.

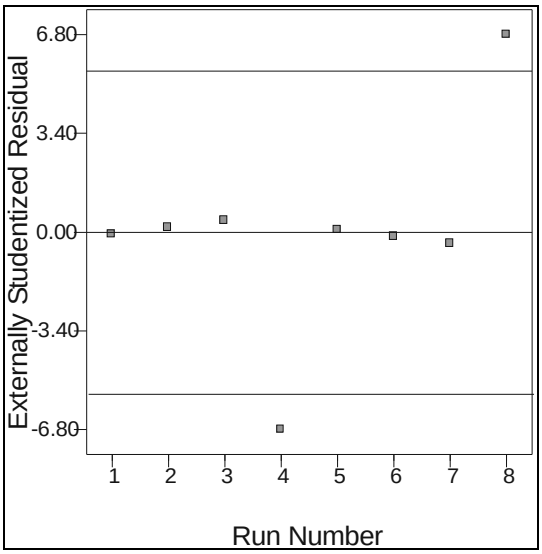
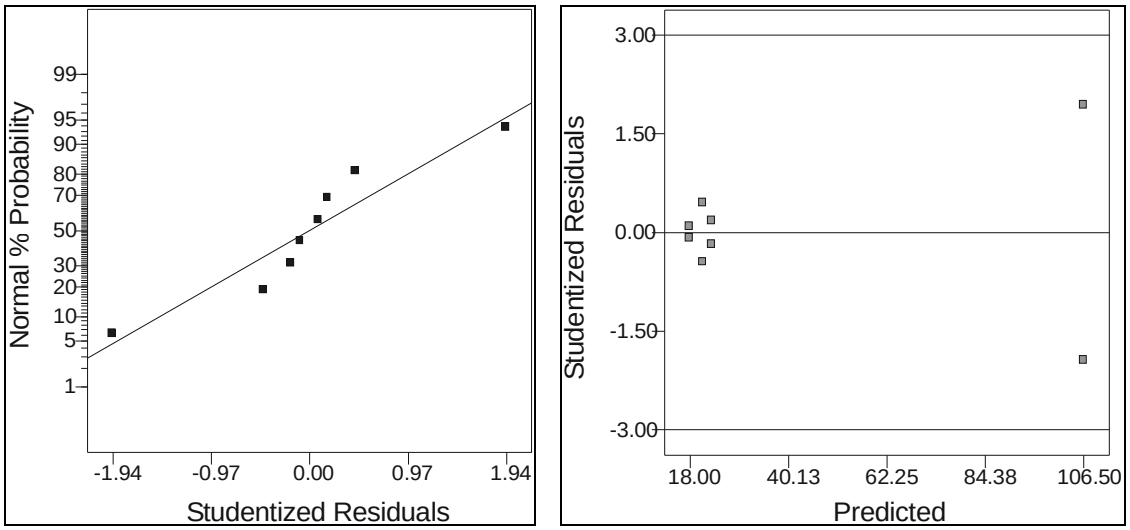


Figure 3: Externally studentized residual (outlier t) plot for bearing experiment

It would be very easy at this stage to delete the two discrepant values, but this would be a big mistake, because as shown in Figure 1, points 4 and 8 represent the breakthrough improvement in bearing life. Perhaps the problem lies not in the data, but in how it's modeled. This becomes obvious upon inspection of two basic plots for diagnosing residuals:

- Normal plot (Figure 4a), which ideally shows a straight line, and
- Residuals versus predicted values (Figure 4b) that ideally exhibits a constant variation from left (low level of response) to right (highest predicted level).



Figures 4a,b: Normal plot of residuals and residuals versus predicted plot for bearing case

Notice that in both plots the residuals have been studentized to account for potential variations in the leverage of the data points. This re-scales the residuals from actual units (in this case the life in hours) to units of standard deviation. We advise that you always use the studentized scale when assessing the relative magnitude of residuals. In this case, the patterns on both plots exhibit non-normality:

- An "S" shape on the normal plot
- A "<" (megaphone) shape on the residuals versus predicted plot.

These patterns are very typical of data that varies over such a broad range (eight-fold in this case) that it needs to be transformed via a logarithm to get a decent fit with a factorial model. This becomes evident in a plot, called "Box-Cox" after the originators, of the residuals versus varying powers of response transformation.[5] The plot (Figure 5) shows the current power (symbolized mathematically by the Greek letter lambda) by the dotted line at 1 on the x-axis. This represents no transformation of the response data. Alternatively, the response is transformed by a range of powers from -3 (inverse cubed) to +3 (cubed). The transformed data is then refitted with the proposed model (in this case A, B, AB) and the residual sum of squares (SS) generated. (Box and Cox recommended plotting against the natural logarithm (Ln) of the residual SS, but this is not of critical importance.) The minimum model residual can then be found and the confidence interval calculated.

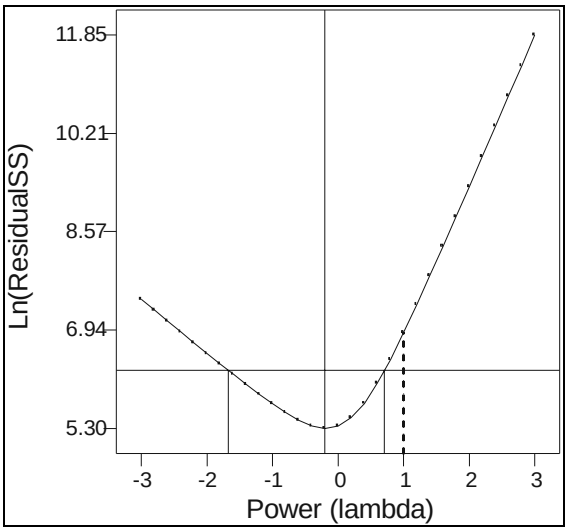


Figure 5: Box-Cox plot for bearing case

In this case notice that that the current point (the dotted line) falls outside of the 95% confidence interval. Therefore applying a different power, one within the confidence interval at or near the minimum, will be advantageous. It's convenient in this case to select a power of 0, which represents the logarithmic transformation, either natural or base 10 – it does not matter (Box and Draper, page 289). Let's try the base-ten log on the bearing data. Figure 6 shows the plot of effects in this new metric.

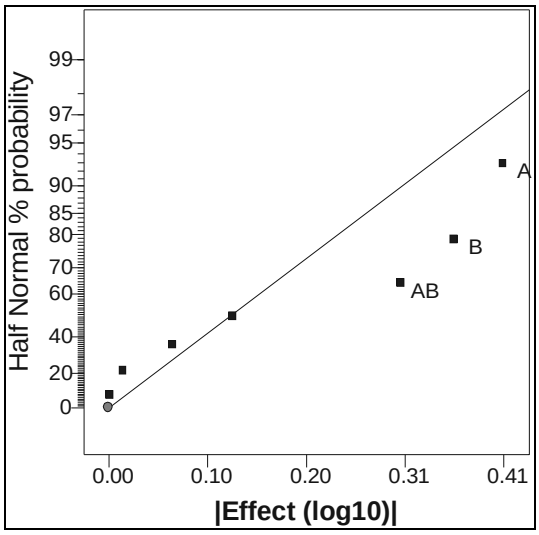
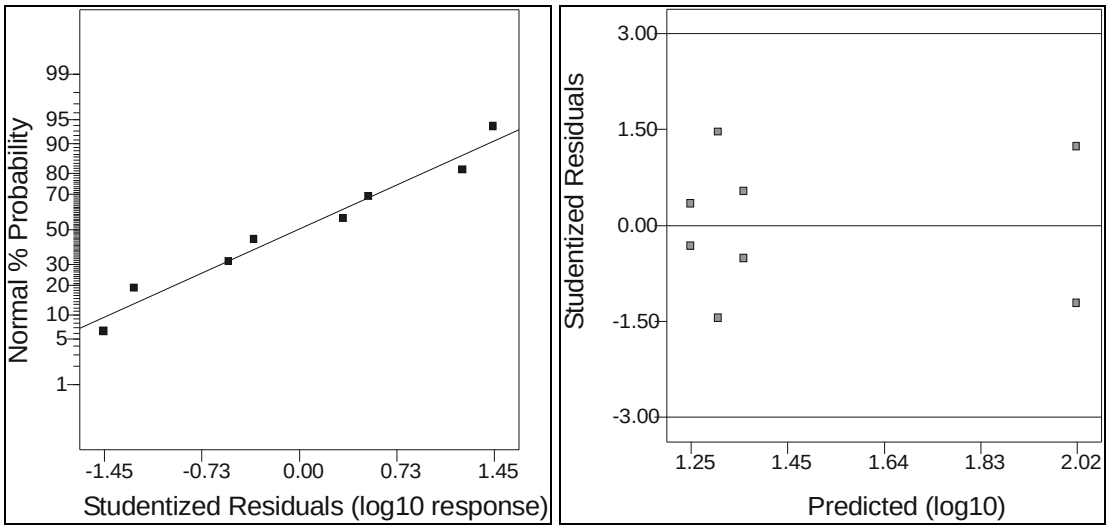


Figure 6: Half-normal plot of effects for bearing data transformed by log, base 10

Notice that now the smaller effects (presumably insignificant) emanate from the origin – a normal pattern for two-level factorial design data. That's good! More good patterns can be seen on diagnostic plots of the residuals: a straighter line on the normal plot (Figure 7a) and more general scatter versus predicted level (Figure 7b).



Figures 7a,b: Normal plot of residuals and residuals versus predicted plot for transformed bearing data

Finally, let's see what happened to the suspected outliers. As shown in Figure 8, they now fall into line with the other points.

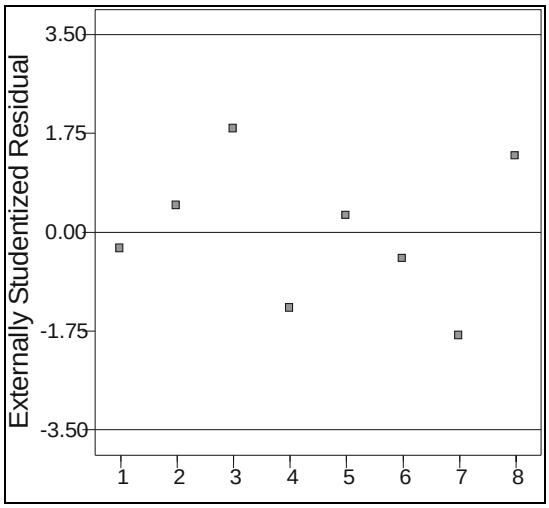


Figure 8: Externally studentized residual (outlier t) plot for bearing data in log-scale

Now we can focus on what George Box wanted to show with the bearing case – how proper DOE revealed a powerful interaction that could not be seen by simple one-factor-at-a-time (OFAT) methods. This becomes very obvious in the interaction graph of AB (Figure 9) constructed from the analysis of data in log scale, but with the response untransformed to the original units of measure (life in hours).

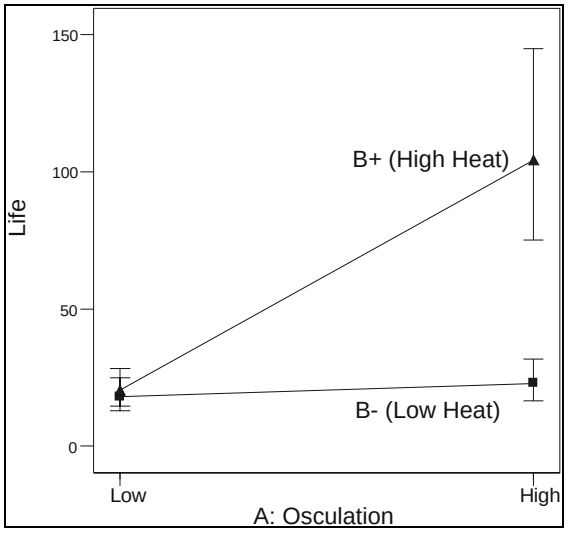


Figure 9: Interaction plot of AB from analysis of bearing data after transformation

Notice how wide the interval, representing the least significant difference (LSD) for 95% confidence, becomes at the increased level of life with both A (osculation) and B (heat) at their high levels. This is the reason for doing the analysis in the log scale, which counteracts the direct dependence of variation on predicted level observed in Figure 4b. We now gain a subtle benefit from applying the response transformation: What looks like a large difference in life, 85 versus 128 hours (Figure 1), obviously must be due only to chance based on the length of the

LSD interval. Thus it becomes more apparent why factor C (cage design) did not emerge as a significant factor. According to Box, the engineers who conducted the bearing experiment did not expect this outcome. It saved their company a lot of money that would otherwise have been spent on reconfiguring their production for a new design.

This case study illustrates the application of a log transformation to better fit good data that might otherwise be wrongly deleted as outliers – the false positive error defined in Table 1. This is just one member from a family of transformations, designated as "power law" by statisticians, you should consider for "bad" response data. Use the Box-Cox plot as a guide to which of the power law transformations, if any, will help you the most. Remember that the log transformation represents a special case where the power will be labeled "0" on the X-axis of the Box-Cox plot. Other transformations that might be revealed by it are:

- Square root (0.5 power), which works well for counts, such as the number of blemishes per unit area
- Inverse (-1 power), which often provides a better fit for rate data.

George Box and his colleagues offer these general comments on transformations, in particular the inverse: "The possibility of transformation should always be kept in mind. Often there is nothing in particular to recommend the original metric in which the measurements happen to be taken. A research worker studying athletics may measure the time *t* in seconds that a subject takes to run 1000 meters, but he could equally well have considered 1000/*t*, which is the athlete's speed in meters per second."

Other transformations, not part of the power law family, may be better for certain types of data, such as pass/fail from quality control records. This is discussed in the second case study that follows.

A CASE TO TEST YOUR METAL

A manufacturer of die-cast aluminum parts wanted to reduce the defect rate on a disk-drive housing.[6] The process engineer, Dave DeVowe designed a 16-run, two-level fractional factorial experiment to screen the following five factors:

A. Hot oil temperature
B. Trip in mm
C. Molten aluminum temperature
D. Fast shot velocity
E. Dwell time

The operators measured fraction defective out of 50 parts made at each set of conditions. The results can be seen in Table 2. It lists them in standard order, but they were actually performed in random fashion at the insistence of Dave, who had just completed of a workshop on design of experiments taught by the authors. The results ranged from 0.06 (6% defective) to 1 (100% defective!). The defect rate had been running as high as 50% so it looked promising!

Std Order	A: Hot Oil Temp Deg F	B: Trip mm	C: Metal Temp Deg F	D: Fast Shot mm	E: Dwell Time Sec	Defects Fraction
1	350	390	1260	1.60	5.50	0.14
2	450	390	1260	1.60	3.50	0.98
3	350	410	1260	1.60	3.50	0.36
4	450	410	1260	1.60	5.50	0.42
5	350	390	1300	1.60	3.50	1.00
6	450	390	1300	1.60	5.50	0.90
7	350	410	1300	1.60	5.50	0.28
8	450	410	1300	1.60	3.50	0.14
9	350	390	1260	2.20	3.50	0.22
10	450	390	1260	2.20	5.50	0.26
11	350	410	1260	2.20	5.50	0.38
12	450	410	1260	2.20	3.50	0.12
13	350	390	1300	2.20	5.50	0.30
14	450	390	1300	2.20	3.50	0.06
15	350	410	1300	2.20	3.50	0.22
16	450	410	1300	2.20	5.50	0.38

Table 2: Data from die-casting experiment

However, to Dave's dismay, none of the effects stood out on the half-normal plot of effects (Figure 10).

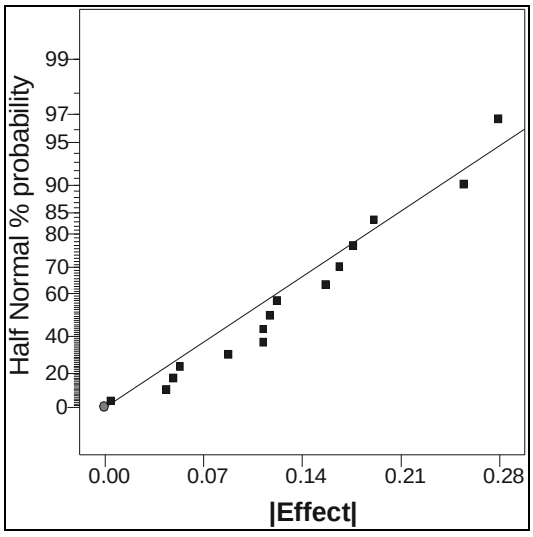


Figure 10: Half-normal plot of effects from die-casting experiment

Having put his manufacturing staff through a great deal of effort and taken up a full week of production, Dave could not accept the possibility of nothing being significant. He appealed for help from his teachers (us!).

The first thing that came to mind was the possibility that the response data needed to be transformed. The standard transformation for binomial data such as fraction defect (pass/fail) is the arcsin square root. However, this made very little difference in the pattern of effects – again, none stood out. But one possibility remained: Something may have gone wrong with one or more the runs, thus creating a damaging statistical outlier. To check this, several of the biggest effects were chosen (Figure 11) to create a predictive model.

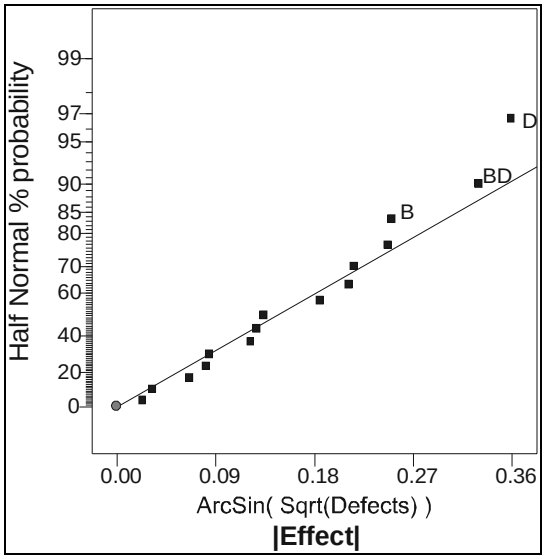


Figure 11: Half-normal plot of effects for die-casting data after doing a transformation

Not surprisingly, the ANOVA for this model does not show much significance. The real surprise comes when you look at the normal plot of residuals from the model (see Figure 12).

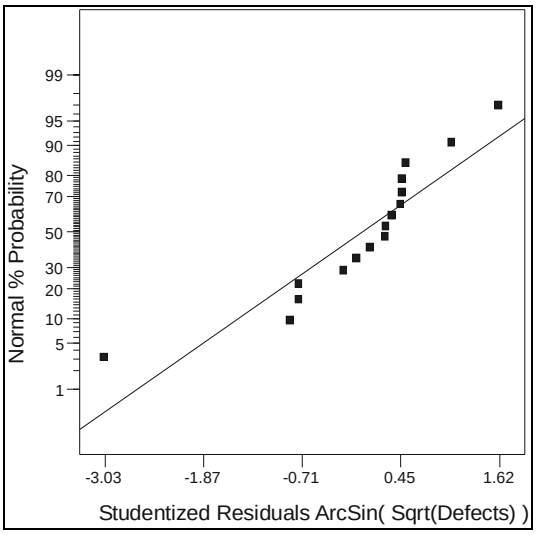


Figure 12: Normal plot of residuals from model for die-casting data

Obviously, one of the experimental runs stands out from the rest. This becomes even more apparent in the plot of externally studentized residuals (Figure 13), which, as noted earlier, helps to detect outliers.

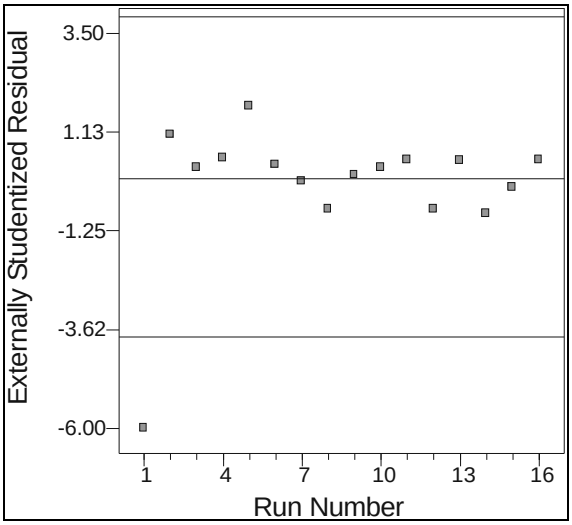


Figure 13: Externally studentized residual (outlier t) plot for die-casting data

Now Dave could identify the culprit: run number 1 (actually done in randomized order, but reported here in standard order to match the layout of Table 1). His foreman, when confronted with this statistical evidence, broke down and confessed that his crew overlooked this particular combination of factors. They then tried to make up for it by coming in early the following week, after shutting down the foundry over the weekend, to sneak the missing run in before Dave came in to work. Considering all that can happen during the start up of process like this that involves molten metal, it's fair to say that the statistical outlier occurred due to a special cause. The next step is to try ignoring the discrepant run. (Note: the elimination of response data results in a loss of information on effects, not serious in this case, but something to be aware of. See Larntz and Whitcomb[7] for advice on how to deal with missing data in two-level factorial designs.)

Figure 14 shows the resulting half-normal plot of effects. At this stage it makes little difference whether the response is transformed or left in the original units of measure, the effects become amazingly clear: B, D and their interaction BD. We will leave the arcsin square root transformation in place because it does provide a somewhat cleaner analysis and it's recommended by statisticians for fraction defect data. It's easy with the aid of software to reverse the transformation and put the response back into the original units before generating the effects plot.

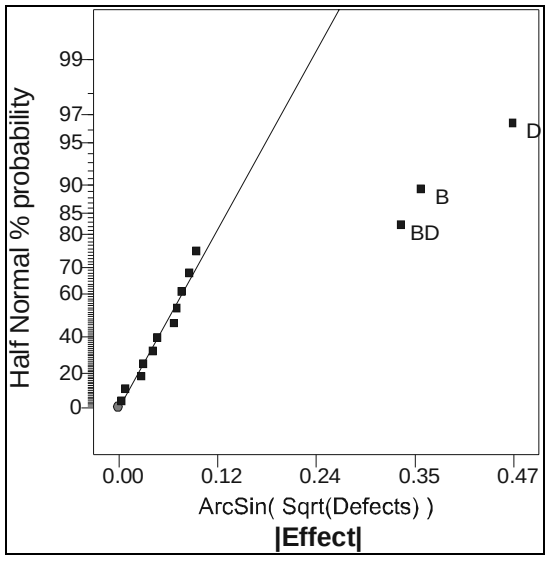


Figure 14: Half-normal plot of effects for die-casting data after ignoring the outlier

In this case the interaction, shown in Figure 15, proved to be the key.

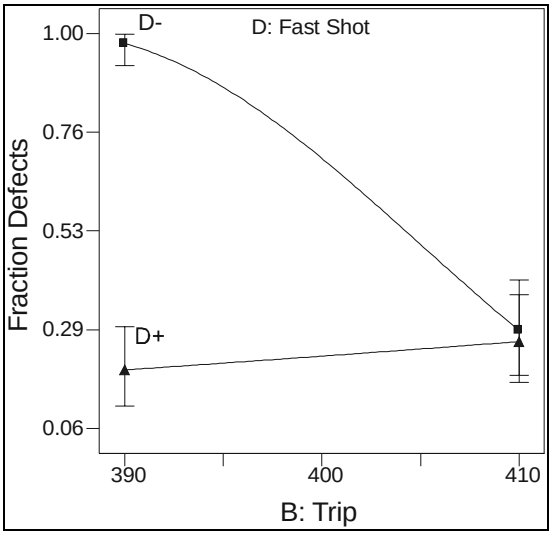


Figure 15: Interaction plot of BD from die-casting experiment

The combination of low B (trip) and low D (fast shot) causes the process to fail. By simply increasing the level of factor B and/or D, the fraction defects drop way off. Inspired by these results, which at first remained obscured by the outlier, Dave led his team through subsequent experimentation that reduced defects in their die-cast aluminum part to near zero percent (DeVowe).

CONCLUSION

An outlier is a response from an experiment that does not fit the proposed model. Before jumping to any conclusions, consider these possibilities:

1. The model is faulty, not the data. The bearing case showed an example of this – the best results cropped up as outliers, which naturally provokes a search for alternatives to deleting data. Here's what should be done in such cases:
    a. Examine the following residual plots (all studentized) to diagnose non-normality (always do this!):
        - Normal plot
        - Residuals versus predicted (check on assumption of constant variance)
        - Externally studentized (outlier t) versus run number
        - Box-Cox plot (to look for power transformations)
    b. Consider a response transformation as a remedy, such as:
        - The logarithm (base ten or natural, it does not matter)
        - Another one from the power-law family such as square root (for counts) or inverse (for rates)
        - Arcsin square root (for fraction defects) and other functions not from the power-law family
2. The result really is an outlier. This proved to be the case in the study aimed at reducing defects in the die-cast aluminum part. Look for possible errors in data entry, or in response measurement, or in the conduct of that particular experimental run. True outliers should not be dismissed - the response may actually be different at that particular combination of the design factors. Further study may lead to an important discovery!

As the famous physicist Richard Feynman said[8] "The first principle is that you must not fool yourself--and you are the easiest person to fool." By using the appropriate graphs to diagnose and deal with potentially bad experimental data, you improve the odds of not being fooled into presenting findings that cannot be supported scientifically.

REFERENCES

1. Sparling, B. Ozone Depletion, History and politics, NASA Advanced Supercomputing Division website: http://www.nas.nasa.gov/About/Education/Ozone/history.html .
2. Box, G. 1990. George's Column: Do Interactions Matter? *Quality Engineering.* Vol. 2, No. 3, p365.
3. Anderson, M. and P. Whitcomb. 2000. *DOE Simplified, Practical Tools for Experimentation*. Portland, Oregon: Productivity, Inc.
4. Weisberg, S. 1985. *Applied Linear Regression*, Second Edition. New York, New York: John Wiley and Sons.
5. Box, G., Hunter W. and S. Hunter. 1978. *Statistics for Experimenters*. New York, New York: John Wiley and Sons.
6. DeVowe, D. 1994. Diecaster achieves zero-defect parts. *Quality in Manufacturing*. March/April.

7. Larntz, K. and P. Whitcomb. 1993. Analyzing Two-level Factorials Having Missing Data. *Proceedings from Fall Technical Conference of American Statistical Association (ASA) and ASQ*. Rochester, NY.
8. Feynman, R. 1974. Cargo Cult Science. Caltech commencement address. (*Surely You're Joking, Mr. Feynman*. Bantam Doubleday Dell Pub. June 1999.)