
Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597292

# Using Graphical Diagnostics to Deal With Bad Data

To cite this Article: , 'Using Graphical Diagnostics to Deal With Bad Data', Quality Engineering, 19:2, 111 - 118
To link to this article: DOI: 10.1080/08982110701241434
URL: http://dx.doi.org/10.1080/08982110701241434

*Quality Engineering*, 19:111–118, 2007

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982110701241434

Taylor & Francis
Taylor & Francis Group



# Using Graphical Diagnostics to Deal With Bad Data

**Mark J. Anderson and Patrick J. Whitcomb**
Stat-Ease, Inc., Minneapolis, MN

**ABSTRACT** This article deals with thorny issues that confront every experimenter, i.e., 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 is 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: (1) deleting data that vary only due to common causes, thus introducing bias to the conclusions. (2) 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).

Two real life data sets are reviewed. Neither reveals its 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 engineers will be much more likely to draw the proper conclusions from experiments that produce bad (discrepant) data.

**KEYWORDS** Box-Cox plot, design of experiments (DOE), diagnostics, outliers, transformations

## INTRODUCTION

Personal computer software makes it very easy to fit models to experimental data via leastsquares 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, breakdowns in equipment, mistakes by operators, nonrepresentative samples, bad measurements and 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 that presence or absence of true outliers, that is, data produced by special causes.

Address correspondence to Mark J. Anderson, Stat-Ease, Inc., 2021 East Hennepin Avenue, Suite 480, Minneapolis, MN 55413. E-mail: mark@statease.com

**TABLE 1** Errors in Judging Whether or not Outliers are Present in Experimental Data

| | | What is presumed | |
|---|---|---|---|
| Outlier(s)? | | Yes (present) | No (absent) |
| The truth: | Yes | Correct | False negative |
| | No | False positive | Correct |

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 (Sparling).

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 desiring more 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) (Montgomery, 2005). They illustrate situations where an unwary experimenter 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).

## CASE STUDY ON IMPROVING BEARING LIFE

George Box reported a compelling success story for two-level factorial DOE that focused on improving the life of a deep groove rolling bearing (Box, 1990). 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). Although Box does not mention it, one would think that the many-fold increase in response to above 100 hours merits an immediate confirmation at the setting from which it was produced; the highest levels of the three test-factors:

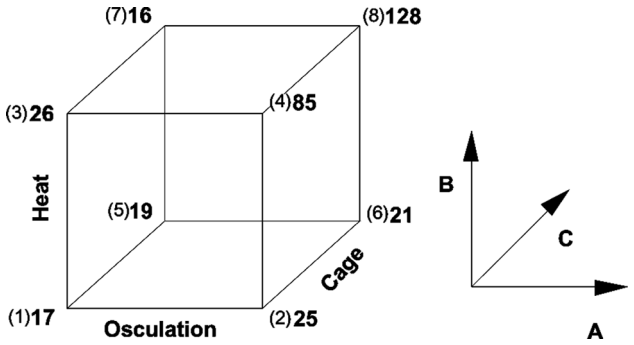


**FIGURE 1** Cube plot of bearing experiment.

(a) Osculation (the ratio of the radius of the cage to the radius of the bearing)
(b) Heat (level of treatment)
(c) Cage (one material versus another—neither revealed by the original experimenter)

Could these results be real and can they be explained?

We now delve into a statistical analysis of this data using techniques developed by Box and his predecessors. Figure 2 shows the half-normal plot of effects (Anderson and Whitcomb, 2000). 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, resulting in an abnormal pattern. Analysis of variance (ANOVA) for the modeled effects (A, B and AB)

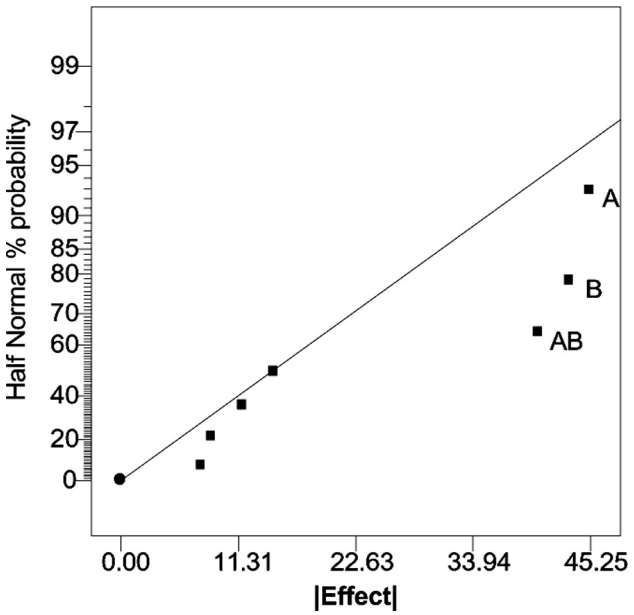


**FIGURE 2** Half-normal plot of effects for bearing experiment.

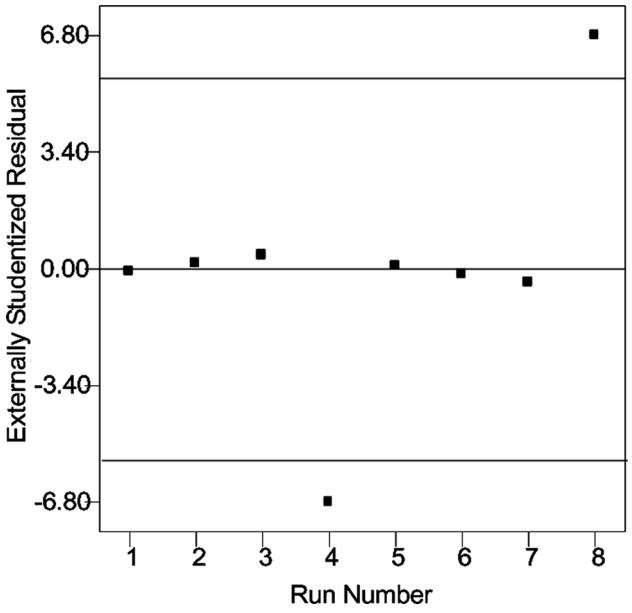


**FIGURE 3** **Externally studentized residual (outlier *t*) plot for bearing experiment.**

shows a high level of significance ($P < 0.05$), but, as shown in Figure 3, diagnosis of the externally studentized residuals (Weisberg, 1985), 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 generally 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), at about the 95% confidence level ($\alpha = 0.05$ risk) for the appropriate test of significance.

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 is modeled. This behavior becomes obvious upon inspection of two basic plots for diagnosing residuals: the normal plot (Figure 4a), which ideally shows a straight line, and a graph of residuals versus predicted values (Figure 4b) that normally exhibits a constant variation from left (low level of response) to right (highest predicted level).

Notice that in both plots the residuals have been studentized to account for potential variations in the leverage of the data points. This transformation rescales the residuals from actual units (in this case the life in hours) to units of standard deviation. We advise that one always use the studentized scale when assessing the relative magnitude of residuals. In this example, both plots exhibit nonnormality: an "S" shape on the normal plot and a megaphone (<) pattern on the residuals versus predicted plot.

These abnormal patterns are very typical of data that vary 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 requirement becomes evident in a plot of the scaled residual sum-of-squares (RSS) versus varying powers of response transformation, called "Box-Cox" after the originators. The Box-Cox plot (Figure 5) shows the current power (symbolized mathematically by the Greek letter lambda—$\lambda$) by the dotted line

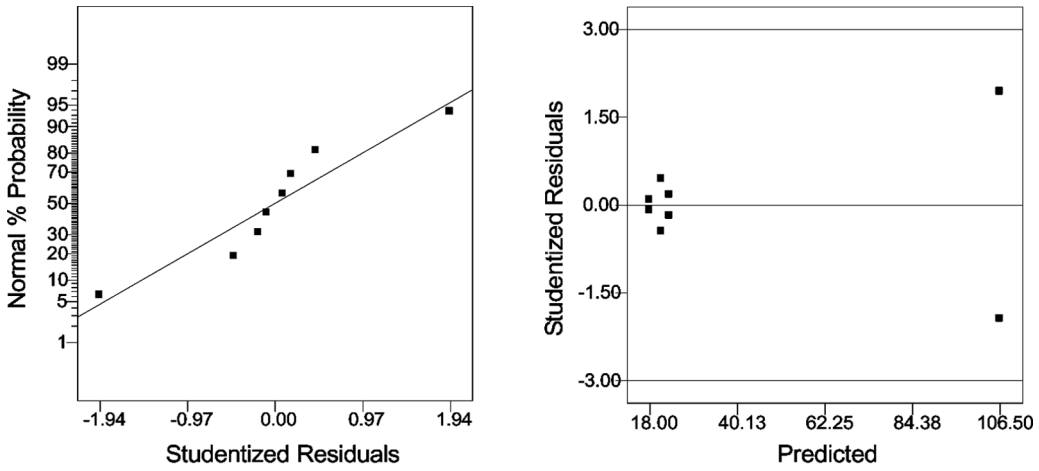


**FIGURE 4** **Normal plot of residuals (a) and residuals versus predicted (b) plot for bearing case.**

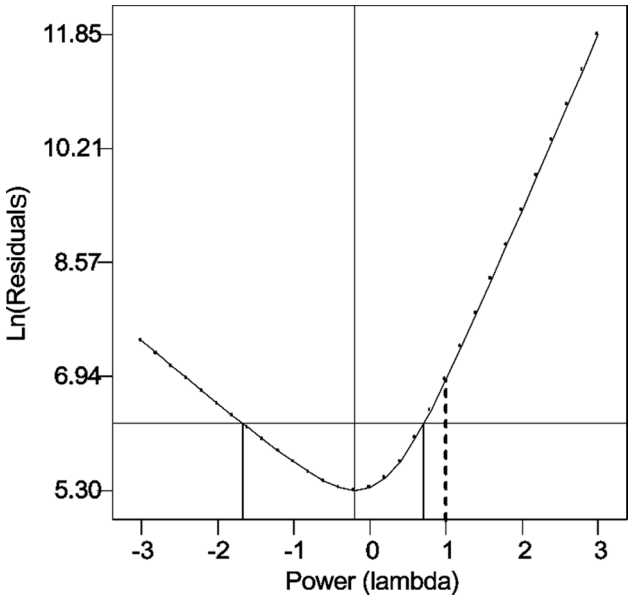


**FIGURE 5** **Box-Cox plot for bearing case.**

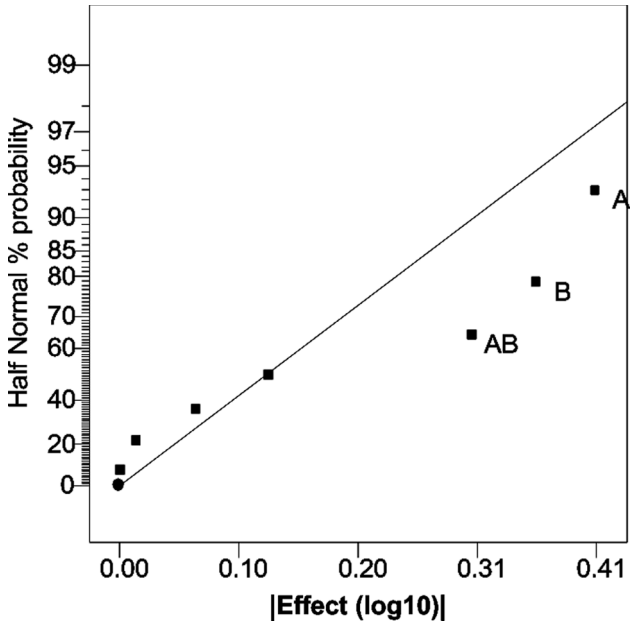


**FIGURE 6** **Half-normal plot of effects for bearing data transformed by log, base 10.**

at one on the x-axis. This location 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 scaled RSS generated. (Box and Cox recommended plotting against the natural logarithm (Ln) of the RSS, but this is not of critical importance.) The power ($\lambda$) that gives the minimum model RSS can then be found and a confidence interval calculated.

In this case notice that the current $\lambda$ (the dotted line) falls outside of the 95% confidence interval for the best $\lambda$. Therefore, applying a different power, one within the confidence interval at or near the minimum, will be advantageous. It is convenient in this case to select a power of zero, which represents the logarithmic transformation, either natural or base-10 it does not matter (Box and Draper, 1987). Let's try the base-10 log on the bearing data. Figure 6 shows the plot of effects in this new metric.

Notice that now the smaller effects (presumably insignificant) emanate from the origin—a normal pattern for two-level factorial design data. Now good patterns are seen on diagnostic plots of the residuals; straighter line on normal plot (Figure 7a) and more general scatter versus predicted level (Figure 7b).

Finally, what happened to the suspected outliers? As shown in Figure 8, they now fall into line with the other points. 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 influence becomes 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).

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 finding 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. The analyst now gains 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

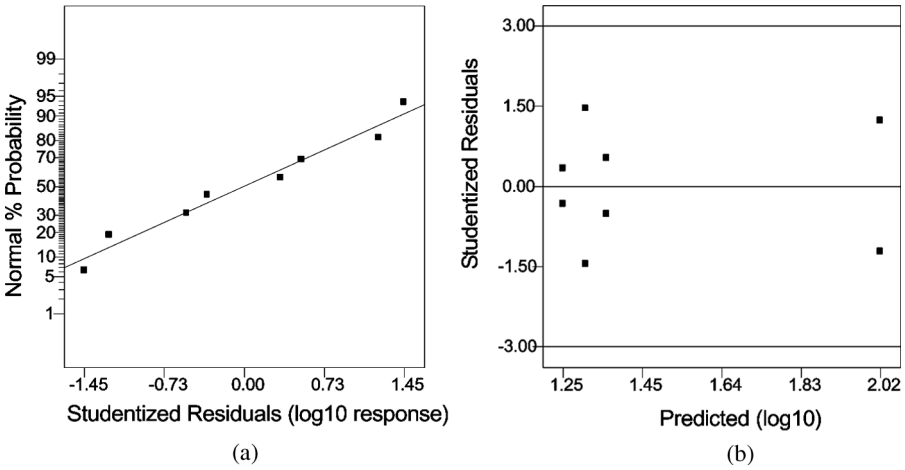


**FIGURE 7** **Normal plot of residuals (a) and residuals versus predicted (b) plot for transformed bearing data.**

otherwise be wrongly deleted as outliers—the false positive error defined in Table 1. The log transformation is just one member from a family of transformations, designated as "power law" by statisticians, which one should consider for "bad" response data. The Box-Cox plot is most helpful as a guide to use of the power law transformations. 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 are the square root (0.5 power), which works well for counts, such as the number of blemishes per unit area, and the inverse (−1 power), which often provides a better fit for rate data. For more detail on the inverse ("reciprocal") transformation see Box, Hunter and Hunter, Chapter 8 (Box et al., 2005). 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 scenario is discussed in the second case study that follows.

## CASE STUDY ON REDUCING DEFECTS IN DIE-CAST ALUMINUM PARTS

A manufacturer of die-cast aluminum parts wanted to reduce the defect rate on a diskdrive housing

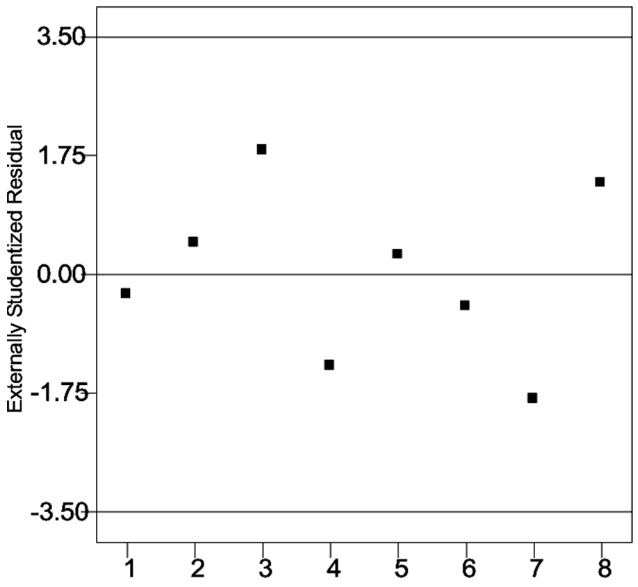


**FIGURE 8** **Externally studentized residual (outlier *t*) plot for bearing data in log-scale.**

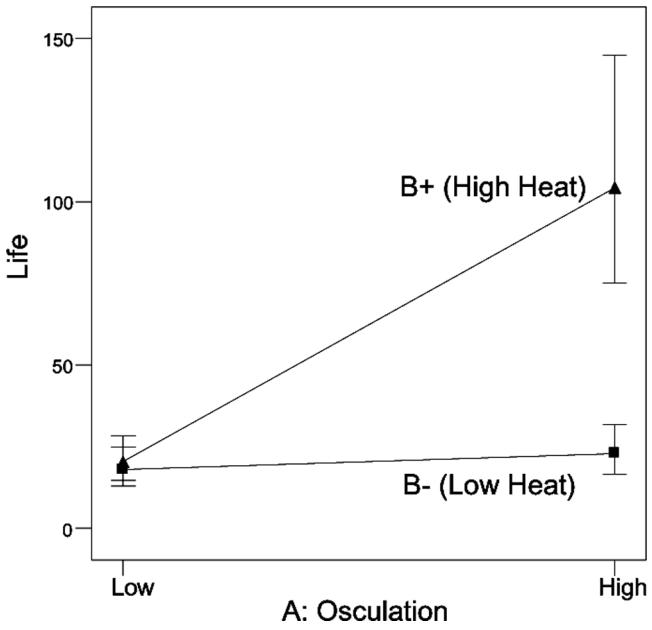


**FIGURE 9** **Interaction plot of AB from analysis of bearing data after transformation.**

(DeVowe, 1994). The process engineer 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 the engineer, who had been trained on the design and execution of statistically sound experiments. The results ranged from 0.06 (6% defective) to 1 (100% defective!). The defect rate had been running as high as 50% so the experiment looked promising.

However, to the experimenter's dismay, none of the effects stood out on the half-normal plot of effects (Figure 10). Having put his manufacturing staff through a great deal of effort and taken up a full week of production, the engineer could not accept the possibility of nothing being significant.

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 of the runs, thus creating potentially damaging statistical outlier(s). To check this, several of the biggest effects were chosen (Figure 11) to create a predictive model.

**TABLE 2** Data from die-casting experiment

| Std order | A: Hot oil temp (°F) | B: Trip (mm) | C: Metal temp (°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 |

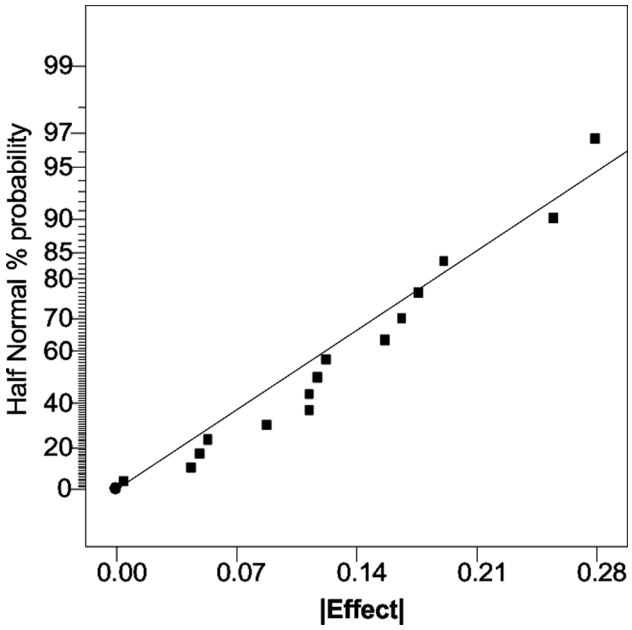


**FIGURE 10** Half-normal plot of effects from die-casting experiment.

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

Now the experimenter 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). The foreman, when confronted with this statistical evidence, broke down and confessed that the operators 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 the engineering staff came

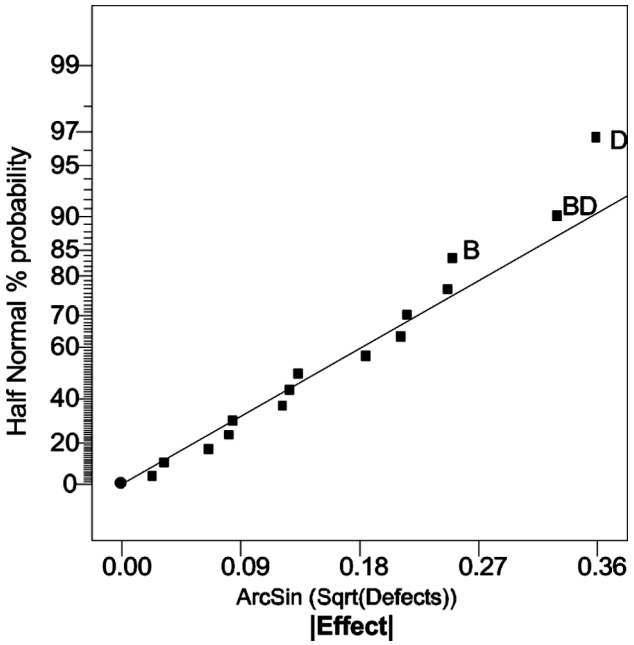


**FIGURE 11** **Half-normal plot of effects for die-casting data after doing a transformation.**

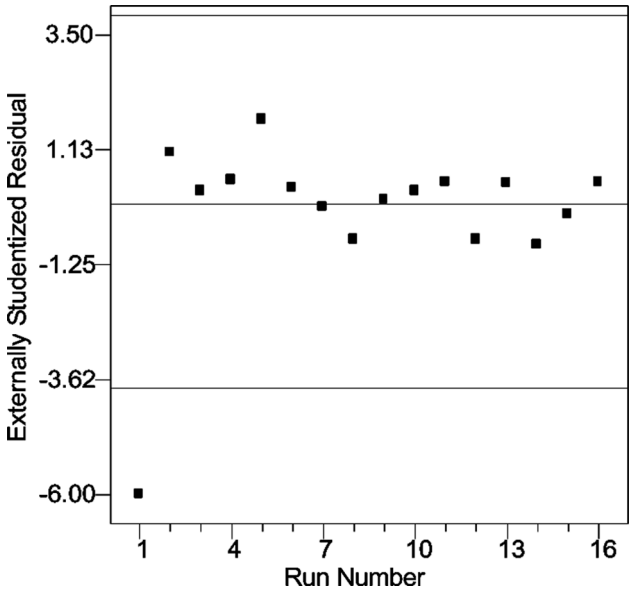


**FIGURE 13** **Externally studentized residual (outlier *t*) plot for die-casting data.**

in to work. Considering all that can happen during the start up of process like this that involves molten metal, it is fair to say that the statistical outlier occurred due to a special cause.

The next step is to try ignoring the discrepant run. The elimination of response data results in a loss of information on effects, not serious in this case, but something to be aware of. Box (1990–91) details a simple way, one that could be hand calculated, to plug in a "fitted value" for a missing or deleted value based on a method developed by Draper and Stoneman (1964). However, via more modern methods made possible by the computer (Larntz and Whitcomb, 1993), readily available statistical software makes it easy to recalculate effects after deleting outliers such as the one identified in the die-casting case. Figure 14 shows the half-normal

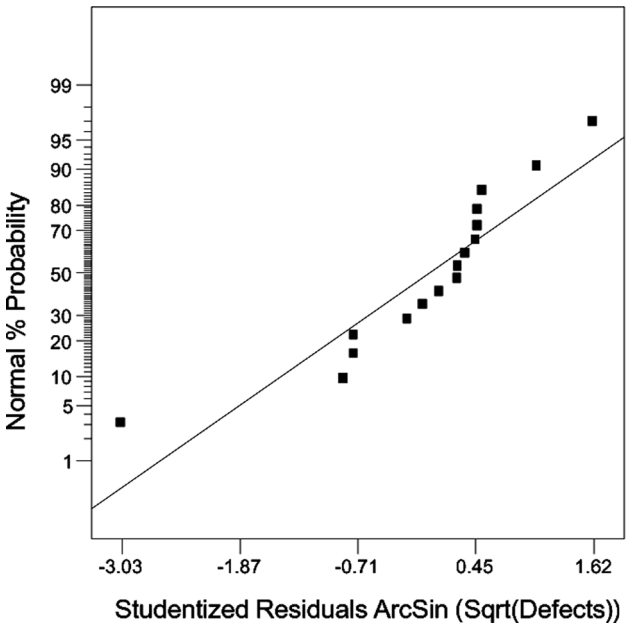


**FIGURE 12** **Normal plot of residuals from model for die-casting data.**

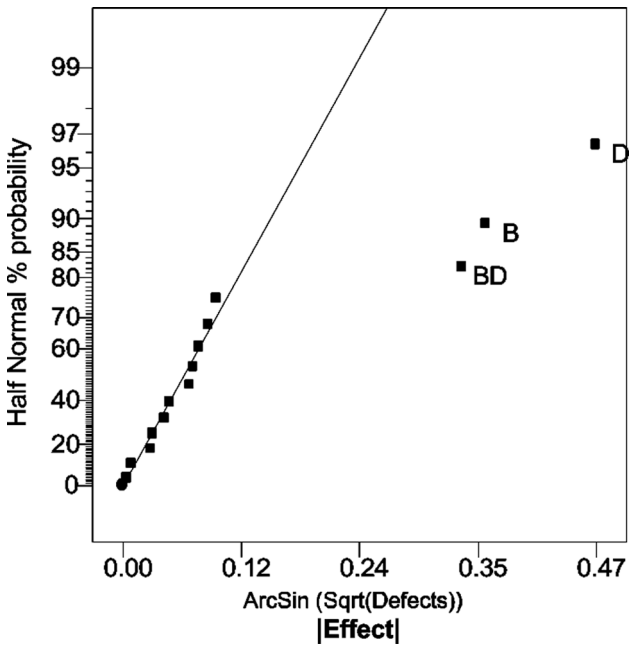


**FIGURE 14** **Half-normal plot of effects for die-casting data after ignoring the outlier.**

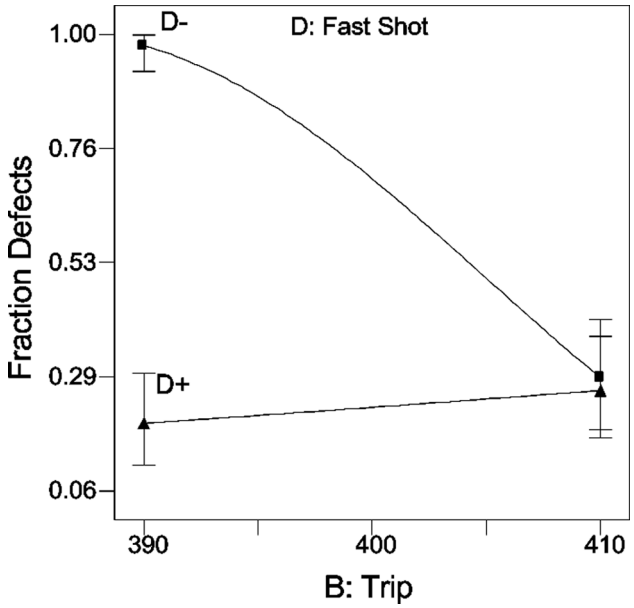


**FIGURE 15** **Interaction plot of BD from die-casting experiment.**

plot of effects with the outlier ignored and the effects recomputed using the methods of Larntz and Whitcomb.

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 from B, D and their interaction BD. We will leave the arcsin square root transformation in place because, not only is this standard practice for fraction defect data, it provides somewhat cleaner analysis. It is easy with some mathematics to reverse the transformation and put the response back into the original units before generating the effects plot. In this case the interaction, shown in Figure 15, proved to be the key.

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 drops way off. Inspired by these results, which at first remained obscured by the outlier, the engineer led his team through subsequent experimentation that reduced defects in their die-cast aluminum part to near zero percent. This case study illustrates the "false negative" error spelled out in Table 1– a real outlier that was mistakenly included in the initial analysis.

## CONCLUSION

An outlier is a response from an experiment that does not fit the proposed model. Before jumping to any conclusions, first consider that the model may be 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. In such cases one often sees a nonlinear (abnormal) pattern on the normal plot of studentized residuals and a significant deviation from the ideal transformation on the Box-Cox plot.

On the other hand, the result might really be an outlier due to an assignable cause. This proved to be the case in the study aimed at reducing defects in the die-cast aluminum part—the operators made a mistake. 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 (Feynman, 1974), "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, quality engineers can improve their odds of not being fooled into presenting findings that cannot be supported scientifically.

## REFERENCES

Anderson, M., Whitcomb, P. (2000). *DOE Simplified, Practical Tools for Experimentation*. Portland, OR: Productivity.

Box, G. E. P. (1990). George's column: Do interactions matter? *Quality Engineering*, 2(3):365–369.

Box, G. E. P. (1990–91). A simple way to deal with missing observations. *Quality Engineering*, 3(2):249–254.

Box, G. E. P., Draper, N. (1987). *Empirical Model-Building and Response Surfaces*. New York, NY: Wiley.

Box, G. E. P., Hunter, W., Hunter, S. (2005). *Statistics for Experimenters*, 2nd ed. New York, NY: Wiley.

DeVowe, D. (1994). Diecaster achieves zero-defect parts. *Quality in Manufacturing*, March/April 20.

Draper, N., Stoneman, D. (1964). Estimating missing values in unreplicated two-level factorial and fractional factorial designs. *Biometrics*, 20(3):443–458.

Feynman, R. (1974). ''Cargo Cult Science.'' Caltech commencement address. (Reprinted in *Surely You're Joking, Mr. Feynman*. Bantam Doubleday Dell, June 1999.)

Larntz, K., Whitcomb, P. (1993). *Analyzing Two-level Factorials Having Missing Data*. Rochester, NY: Fall Technical Conference of American Statistical Association (ASA) and ASQ Proceedings.

Montgomery, D. (2005). *Design and Analysis of Experiments*, 6th ed. New York, NY: Wiley.

Sparling, B. (2001). Ozone Depletion, History and Politics, NASA Advanced Supercomputing Division, Available at: http://www.nas.nasa.gov/About/Education/Ozone/history.html.

Weisberg, S. (1985). *Applied Linear Regression*, 2nd ed. New York, NY: Wiley.