Center for Quality and Productivity Improvement
UNIVERSITY OF WISCONSIN
610 Walnut Street
Madison, Wisconsin 53705

(608) 263–2520
(608) 263–1425 FAX

Report No. 119

Analysis of Factorial Experiments with Defects or Defectives as the Response

Søren Bisgaard and Howard Fuller

June 1994

Analysis of Factorial Experiments with Defects or Defectives as the Response

Søren Bisgaard Howard Fuller

Center for Quality and Productivity Improvement

University of Wisconsin
Madison, Wisconsin

ABSTRACT

The performance of a production process is often characterized by the number of defects in its products or the number of defective products. Typically, reduction of the number of defects or defectives is paramount to improving the quality of such a process. A powerful tool used for identifying variables that influence the process level of defects or defectives is experimental design. However, when using counts of defects or defectives as the experimental response, the assumption of constant variance made with almost all standard analyses is violated. A common method for dealing with this problem is to transform the data before the analysis so that the assumption of constant variance is more likely. In this paper, we present various transformations that can be used to approximately stabilize the variance of counts of defects and the variance of proportion of defectives. We also re-analyze examples of each case where transformation of the experimental data followed by a simple analysis of the data led to significantly different conclusions.

KEYWORDS: Defects; Defectives; Variance stabilizing transformations.

This work was supported by a grant from the Alfred P. Sloan Foundation and by National Science Foundation Grant ECD8721545.

CQPI Report No. 119, June 1994

Analysis of Factorial Experiments with Defects or Defectives as the Response

Søren Bisgaard and Howard Fuller

The performance of a production process is often characterized by the number of defects in its products or the number of defective products. Typically, reduction of the number of defects or defectives is paramount to improving the quality of such a process. A powerful tool used for identifying variables that influence the process level of defects or defectives is experimental design. However, when using counts of defects or defectives as the experimental response, the assumption of constant variance made with almost all standard analyses is violated. A common method for dealing with this problem is to transform the data before the analysis so that the assumption of constant variance is more likely. In this paper, we present various transformations that can be used to approximately stabilize the variance of counts of defects and the variance of proportion of defectives. We also re-analyze examples of each case where transformation of the experimental data followed by a simple analysis of the data led to significantly different conclusions.

Although the modern concept of quality improvement by no means is limited to defect reduction, such efforts nevertheless remain important. Control chart studies are often useful in the initial phases of projects with that objective. Later, however, when the obvious assignable causes have been identified and removed, more potent tools are necessary to achieve further improvements. When that is the case it is worth recalling George Box's often quoted statement (Box, 1966) that "to find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)." Thus instead of just observing a process we need to "kick" it so that it reveals how we can improve it. Two-level factorial and fractional factorial experiments provide simple and economical recipes for systematically "kicking" complex systems from a number of directions.

In several previous columns (Box, 1992, 1993a, 1993b; Bisgaard, 1993a, 1993b) we have discussed how to analyze two-level factorial and fractional factorial designs. Typically in the examples we have used, the response has been measured on a continuous scale. Frequently, however, the only economical measure of quality is a simple count of the number of defects or defectives. Almost all the standard statistical methods, in particular the methods used for the analysis of factorial experiments, are based on the assumptions that the response is measured on a continuous scale and has constant variance. When using two-level factorials and fractional factorials the more important of these assumptions is that of constant variance. Unfortunately, when dealing with counts of defects and defectives the constant variance assumption is violated.

Of course, we may simply ignore the problem and use the standard tools anyway and fortunately factorial and fractional factorials are so powerful that such an analysis will often not mislead. However, a simple remedy that provides a more sensitive analysis and adds only a few more minutes to the analysis is to "bend the data into shape" by a *variance stabilizing transformation**. The beauty of using transformations is that we can get an efficient analysis by using the standard techniques on the transformed data. They therefore significantly enlarge the applicability of our existing "toolbox."

DEFECTS AND DEFECTIVES

You may have noticed that in the title of this column we have carefully distinguished between *defects* and *defectives*. To illustrate the difference suppose we are going to conduct a factorial experiment on cleaning

* Experience shows that frequently when we have achieved constant variance other assumptions are also more closely met.

optical lenses. For some reason, presently unknown to us, the lenses develop scratches in the cleaning process. Thus we would like to see which of a number of factors – such as temperature of the water, type of detergent, etc. – may reduce or eliminate the problem. Under each set of conditions determined by the factorial design we might manufacture n lenses and look for scratches. We can then classify each as either "bad" or "good" according to whether they have scratches or not and use the count, X, of bad lenses in the sample as the response. Counts of this kind are called *defectives*. It is easy to see that defective counts will possibly take on integer values 0, 1, 2, up to n. Notice, however, that it will be impossible for X to assume values larger than n.

Alternatively we might use only one lens for each trial but look at them individually and count the *number* of scratches on the surface of each lens. Such a count is called the number of *defects*. It is easy to see that, at least theoretically, there may be any number of scratches on a lens; therefore, X may assume any of the integer values 0, 1, 2, ... with no upper limit. If you are in doubt whether you are dealing with either a defect or defective count, a simple test is to ask if there is a natural upper limit. If the answer is yes, then you are dealing with defectives. Otherwise it is defects.

TRANSFORMING PROPORTIONS DEFECTIVE

Suppose that we are counting defective lenses out of a sample of n, and within a particular trial the individual lenses have an equal chance p of being "bad" and $1-p$ of being "good." Then for each trial the count will be distributed according to the *Binomial* distribution with parameters n and p. For a more detailed discussion see BH^2 pages 135-137 (Box, Hunter and Hunter, 1978)

The Binomial distribution characterizes how the observed proportion of defectives $\hat{p} = X/n$ varies from one sample of n lenses to the next when we do not change the factors in the experiment. The observed proportion $\hat{p}$ has mean $\mu = p$ and variance $\sigma^2 = p(1-p)/n$. Thus the variance of our responses *depends* on the value of the mean p. But that is unfortunate because the objective of a factorial screening experiment is to try to change $\hat{p}$ so that we can learn about factors that may help reduce the number of defectives. Therefore, if we are successful in our overall mission we unfortunately end up violating the assumption of constant variance needed in the analysis.

For illustration suppose we conduct an experiment with a sample size of $n = 20$ lenses for each trial, and that for three different factor combinations the true proportion of defectives p is 0.1, 0.3, and 0.5, respectively. We see from Figure 1 that this will change the variance and indeed the whole shape of the distribution of $\hat{p} = X/n$. In Figure 1a the probability of an individual defective is $p = 0.1$, and we see that the distribution of $\hat{p}$ is rather narrow with a standard deviation of

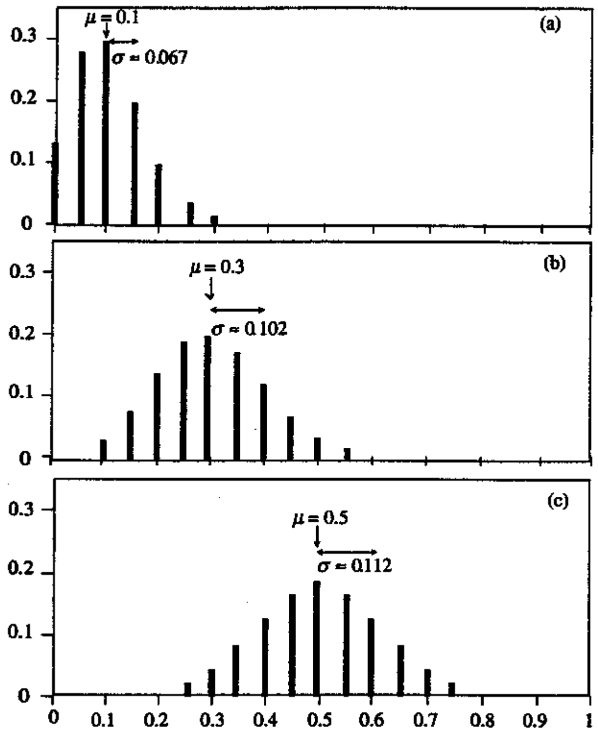

Figure 1. *The distributions of proportions for n=20 and (a) p = 0.1, (b) p = 0.3, and (c) p = 0.5.*

$\sigma = \sqrt{0.1(1-0.1)/20} \approx 0.067$. However, in Figure 1b, $p = 0.3$ and now the distribution is wider with $\sigma = \sqrt{0.3(1-0.3)/20} \approx 0.102$. Finally, in Figure 1c, $p = 0.5$, and we see that the distribution of $\hat{p}$ is again slightly wider with a standard deviation of $\sigma = \sqrt{0.5(1-0.5)/20} \approx 0.112$. Figures 2a and 2d show how $\sigma^2 = p(1-p)/n$ the variance of $\hat{p}$ depends on p for samples of $n = 20$ and $n = 50$. Over the range of possible values of p we see that the variance is far from constant and changes like a parabola. It is that problem we are concerned about.

Fortunately we can overcome the problem by using a variance stabilizing transformation that makes the variance of the transformed $\hat{p}$ approximately constant. Table 1 shows two types of transformations useful for dealing with defectives and defects. Specifically for defective type data that follow the Binomial distribution the table shows that the appropriate transformation is the *arcsin square root* function. Thus, if instead of using $\hat{p}$ directly as our response we use $arcsin\sqrt{\hat{p}}$ then the variance will be approximately constant. Note that you may use either radians or degrees for the arcsin function. In our example below we have used radians.

Figure 2b shows for $n = 20$ how the variance changes with p if instead of $\hat{p}$ we use the arcsin square root of $\hat{p}$ as the response. From about $p = 0.2$ to $p = 0.8$ the variance curve is now quite constant but outside this interval the curve begins to swing up and down. This happens because the smaller p is, the greater the probability of getting samples containing zero defectives, which causes the variance to become unstable. The same thing happens at the other end of the scale as p gets close to one.

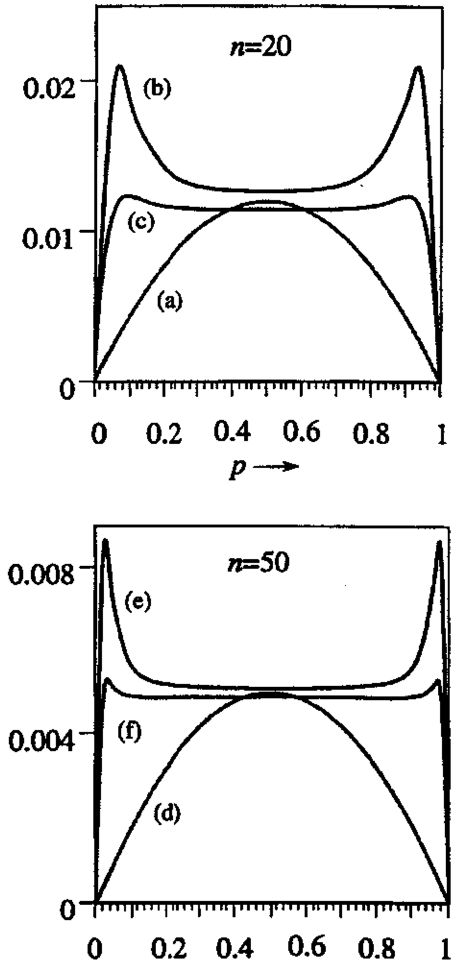

Figure 2. The variance functions for sample sizes of n=20 and n=50 for Binomial distributed proportions when using no transformation ((a) and (d)), the arcsin square root of proportions ((b) and (e)), and Freeman and Tukey's modification ((c) and (f)).

To deal with this problem Bartlett (1937) and later Freeman and Tukey (1950) suggested minor modifications to the arcsin square root transformation. We prefer Freeman and Tukey's modification, also provided in Table 1, because it works slightly better and is easier to implement on a computer. Figure 2c shows the variance function after Freeman and Tukey's modification. Notice that it is now quite stable from about $p = 0.05$ to $p = 0.95$; hence, their modification extends the region of constant variance.

By comparing Figures 2c and 2f we see that the effect of increasing the sample size to $n = 50$ is to further increase the interval of near constant variance from about $0.05 \le p \le 0.95$ to about $0.02 \le p \le 0.98$. It is true that even Freeman and Tukey's modification does not work well at the extremes of the range of p especially if $n = 20$. However, in practice, if we were planning a factorial experiment expecting to estimate with reasonable confidence defective probabilities as low as $0.02 < p < 0.05$ then both formal calculations and intuition show that a sample size of $n = 20$ would be inadequate. Instead we should use sample sizes of say 50 or larger and when doing so, as we see from the figure, the range of stable variance would automatically be extended.

TRANSFORMING COUNTS OF DEFECTS

A similar scenario applies when dealing with a count of defects which we usually denote by the symbol c or $\hat{c}$ if estimated from a sample. If the probability of getting a defect in any small sub-unit, say a square millimeter of a lens, is constant from one sub-unit to the next and the chance of a defect in any sub-unit is independent of getting a defect in any other sub-unit then the distribution of the number of defects per unit tends to be distributed as the Poisson distribution. The defect rate or the expected number of defects per unit is then denoted by the parameter λ. A Poisson distributed count c then has mean $\mu = \lambda$ and variance $\sigma^2 = \lambda$. Hence as in the previous case the variance depends on the mean, and if we are successful in finding factors that change the defect

Table 1.
The standard transformations and Freeman and Tukey's (F & T) modifications when using proportion of defectives or count of defects as the response

TYPE OF DATA	*TYPE OF DISTRIBUTION*	*TRANSFORMATION*	*F & T'S MODIFICATION*
Proportions $(\hat{p})$ (Defective units in a sample of n units)	Binomial	$\arcsin\sqrt{\hat{p}}$	$\dfrac{\left(\arcsin\sqrt{\dfrac{n\hat{p}}{n+1}}+\arcsin\sqrt{\dfrac{n\hat{p}+1}{n+1}}\right)}{2}$
Counts $(\hat{c})$ (Defects on a unit)	Poisson	$\sqrt{\hat{c}}$	$\left(\sqrt{\hat{c}}+\sqrt{\hat{c}+1}\right)/2$

rate we end up violating the assumption of constant variance. However, theoretical derivations show that to get approximately constant variance for counts of defects we need to take the square root of the observed count.

Again, with count data we may get zeros in our samples causing similar problems to those discussed for defectives. It has been suggested, therefore, that we avoid zero's by first adding 1 to all the counts. However, as a better compromise, Freeman and Tukey (1950) suggested that we instead take the average of the two transformations, that is use $(\sqrt{\hat{c}}+\sqrt{\hat{c}+1})/2$, and surprisingly it works really well. To see how well, let us consider the graphs in Figure 3. Figure 3a shows the steadily increasing variance for the count itself*. Figure 3b shows that by using the square root of the count, the variance after a large initial swing quickly stabilizes. As in the Binomial case, the reason for this initial instability is the large chance of getting zero's in the samples for small λ's. By using Freeman and Tukey's modified transformation we get an even more stable variance as shown in Figure 3c.

As in the Binomial case the early unstable region for the Poisson case using the modified transformation is of only minor practical relevance. If we operate in the region where the variance is not stable, then we are dealing with so rare defects that the chosen unit size will contain little information about the defect rate. Hence, as before, it is the proposed experiment and not the transformation that is to blame. A practical remedy, therefore, is to redefine the basic experimental unit size on which we are counting defects. In our lens experiment instead of counting the number of defects per lens we might count the number of defects on, say, ten lenses. Thus, if the mean defect rate per lens was 2 then the defect rate per ten lenses will be 20 and we would move into the region where the variance is stable.

As already indicated above, these transformations are based on several assumptions regarding the exact distributional form. In practice, these idealizing assumptions are likely to be only approximately true. The transformations are nevertheless useful as a starting point for an iterative data analysis (Bartlett, 1947).

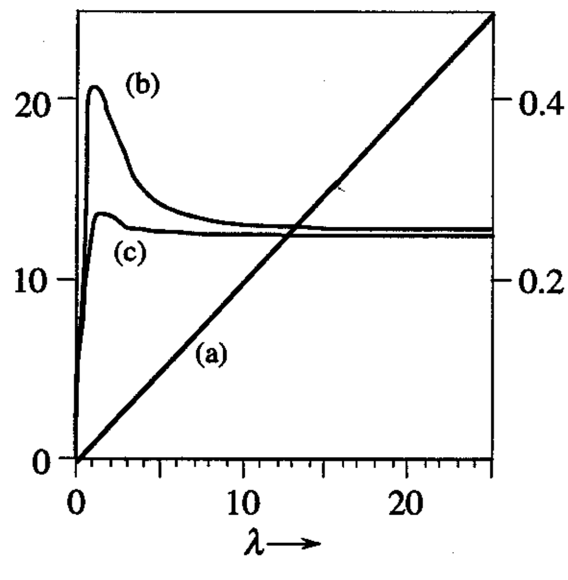

Figure 3. *The variance functions for Poisson distributed counts when using (a) no transformation, (b) the square root of counts, and (c) Freeman and Tukey's modification. The y-axis scale on the left applies to (a) and the scale on the right applies to (b) and (c)*

* Notice that for plotting purposes the y-axis scale is different for Figure 3a and the two other graphs in Figure 3.

TWO ILLUSTRATIVE EXAMPLES

Let us now consider two examples from the Fourth Symposium on Taguchi Methods (Becknell, 1986; Hsieh and Goodwin, 1986) held in 1986. In the first example the object of the experiment was to reduce the proportion of defective castings at the Essex Aluminum Plant of Ford Motor Company. The second example aimed at reducing the number of defects in the paint finish of a car grille panel was conducted by Chrysler Motors Engineering Department. In the paper presented at the symposium the engineers involved in the first experiment used the much discredited (Box and Jones, 1992; Nair, 1990) Accumulation Analysis Method suggested by Dr. Taguchi as a way of dealing with count data. As a supplementary analysis they also used analysis of variance (ANOVA) directly on the observed proportions. In the second example the team again analyzed the count data directly with ANOVA. In both cases the assumptions required for ANOVA were violated. In our re-analysis of these examples we will use the transformations discussed above and as a simple alternative to ANOVA employ the much simpler and more informative Normal Plots.

If you should compare our discussion below with the original write-up of these two examples notice that we have converted the one-two notation for the factor levels preferred by Taguchi with the more standard minus and plus notation. The reason for this conversion is that the structure of the design and, in particular, the confounding becomes much more transparent and easy to understand with the minus and plus notation. Another minor alteration is that we have changed the order of the columns and the definition of high and low level of some of the factors. These changes are intended to make it easier for you to re-analyze these examples using the spreadsheets presented in a previous column (Bisgaard, 1993a).

EXAMPLE 1: ANALYSIS OF DEFECTIVES.

In Table 2 are shown the results from a 16-run two-level fractional factorial experiment on sand-castings of engine manifolds conducted by the engineers at the Essex Aluminum Plant of the Ford Motor Company. The objective of the experiment was to determine which of 10 factors, Q4 Program *(A)*, Q6 Washers *(B)*, Q3 Washers *(C)*, Q2 Washers *(D)*, Q4 Washers *(E)*, Flask Thickness *(F)*, Flask Width *(G)*, Q3 Program *(H)*, Q2 Program *(J)*, and Q6 Program *(K)*, had an effect on the proportion of defective castings of a 3.0 liter intake manifold. The sample size per trial is not provided but the author said that it was "very large". However, that is not a serious problem for our analysis. The data, $\hat{p}$, is the proportion of non-defective castings. The generators for this 2^{10-6} resolution *III* design are $E = CD$, $F = BD$, $G = BC$, $H = AC$, $J = AB$ and $K = ABC$. From these we can derive the confounding pattern shown in Table 2.

For the purpose of demonstration let us first analyze the proportion non-defectives without any transformation ignoring the problem of the non-constant variance. In Figure 4a we show a Normal Plot of the effects. We see that factors F and K seem to have a significant effect on the number of defective castings. If we now proceeded to fit a simple linear model with these two factors and plot the residuals against the predicted values then we can check if they seem to have constant variance. The residual plot in Figure 4b shows that the closer to one the predicted values are the smaller the residual scatter appears. Although one should always interpret residual patterns with caution, this is likely because the variance as indicated in Figure 2a depends on p and decreases as p gets closer to 1. Since the factors F and K are so obviously significant, it is most likely of little practical consequence in this case for our overall purpose of screening out the most important factors.

Let us proceed to apply the arcsin square root transformation. From Figure 4c we see that factors F and K stand out even more clearly after this transformation as would be expected by the increased sensitivity of the test imparted by the transformation. The straight line part of the effects in the center of the Normal Plot may also appear slightly more straight and the two significant factors slightly more off the line. Turning to the residual plot shown in Figure 4d we see that they look more as if they have constant variance because those in the right-hand side of the plot spread out more. In general, however, as we already indicated earlier two-level factorials are so powerful that effects that are truly significant often are so obvious that it does not matter which transformation we use. This data set is a good example of that. In general, however, that will not necessarily be the case.

To apply Freeman and Tukey's modification we would need to know the sample size. For illustration we have here assumed that $n = 1000$. Since the data only contained two observations with $\hat{p} = 1.0$ and with a large sample size the modification would not be expected to show much difference in either the Normal Plot of effects or the plot of residuals which

Table 2.

The design matrix, data, and confounding pattern for the sand-casting experiment. Note that the arcsin square root and Freeman and Tukey's (F & T) modification to the arcsin square root were computed using radians.

RUN	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*J*	*K*	$\hat{p}$	$\arcsin\sqrt{\hat{p}}$	*F & T'S MODIFICATION*
1	−	−	−	−	+	+	+	+	+	−	0.958	1.364	1.363
2	+	−	−	−	+	+	+	−	−	+	1.000	1.571	1.555
3	−	+	−	−	+	−	−	+	−	+	0.977	1.419	1.417
4	+	+	−	−	+	−	−	−	+	−	0.775	1.077	1.076
5	−	−	+	−	−	+	−	−	+	+	0.958	1.364	1.363
6	+	−	+	−	−	+	−	+	−	−	0.958	1.364	1.363
7	−	+	+	−	−	−	+	−	−	−	0.813	1.124	1.123
8	+	+	+	−	−	−	+	+	+	+	0.906	1.259	1.259
9	−	−	−	+	−	−	+	+	+	−	0.679	0.969	0.968
10	+	−	−	+	−	−	+	−	−	+	0.781	1.081	1.083
11	−	+	−	+	−	+	−	+	−	+	1.000	1.571	1.556
12	+	+	−	+	−	+	−	−	+	−	0.896	1.241	1.242
13	−	−	+	+	+	−	−	−	+	+	0.958	1.364	1.363
14	+	−	+	+	+	−	−	+	−	−	0.818	1.130	1.130
15	−	+	+	+	+	+	+	−	−	−	0.841	1.161	1.160
16	+	+	+	+	+	+	+	+	+	+	0.955	1.357	1.356

$\ell_1 = A + BJ + CH + GK$

$\ell_2 = B + AJ + CG + DF + HK$

$\ell_3 = J + AB + CK + GH$

$\ell_4 = C + AH + BG + DE + JK$

$\ell_5 = H + AC + BK + GJ$

$\ell_6 = G + BC + AK + EF + HJ$

$\ell_7 = K + AG + BH + CJ$

$\ell_8 = D + BF + CE$

$\ell_9 = AD + EH + FJ$

$\ell_{10} = F + BD + EG$

$\ell_{11} = E + CD + FG$

$\ell_{12} = JD + AF + EK$

$\ell_{13} = HD + AE + FK$

$\ell_{14} = GD + BE + CF$

$\ell_{15} = KD + EJ + FH$

are shown in Figures 4e and 4f, respectively. Again we stress that for other data sets the improvements achieved by using the modified transformation are likely to be more obvious.

These results differ from those reported in the original analysis based on Taguchi's Accumulation Analysis and the analysis of variance of the untransformed proportions which indicated that seven main effects and interactions were significant. The most significant were as in our analysis, factors *F* and *K*. However, in addition the Taguchi style analysis found factors *D, E, G,* and the interactions *DG* and *FK*, to be significant. In the Normal Plots shown in Figure 4 we have indicated by solid dots the contrasts that the experimenters called out as significant and with open dots those they pooled for estimating error. From this we clearly see what went wrong in their analysis. They used a standard error estimate that grossly underestimated the real standard error and hence with that smaller yardstick several more effects seemed significant. For a further discussion of this see Box (1988).

You may ask what is wrong with calling out factors as significant when they are not. The experimenters certainly did not miss the real significant factors. If we are just going to adjust the factor levels to a combination that produces the highest number of non-defectives then there seems to be little harm in this. We agree partly with this argument. The experimenters did not fail to find the truly significant effects; the experiment certainly was a success and is by far superior to the alternatives of doing nothing or experimenting one-factor-at-a-time. However, we have several concerns.

Most importantly, as you can see above, a *less complicated* analysis got *more* out of the experiment and did not mislead. Moreover, experimentation should not necessarily lead only to a quick fix to a problem. It should allow the investigators to *learn* more about the process. The Taguchi approach used

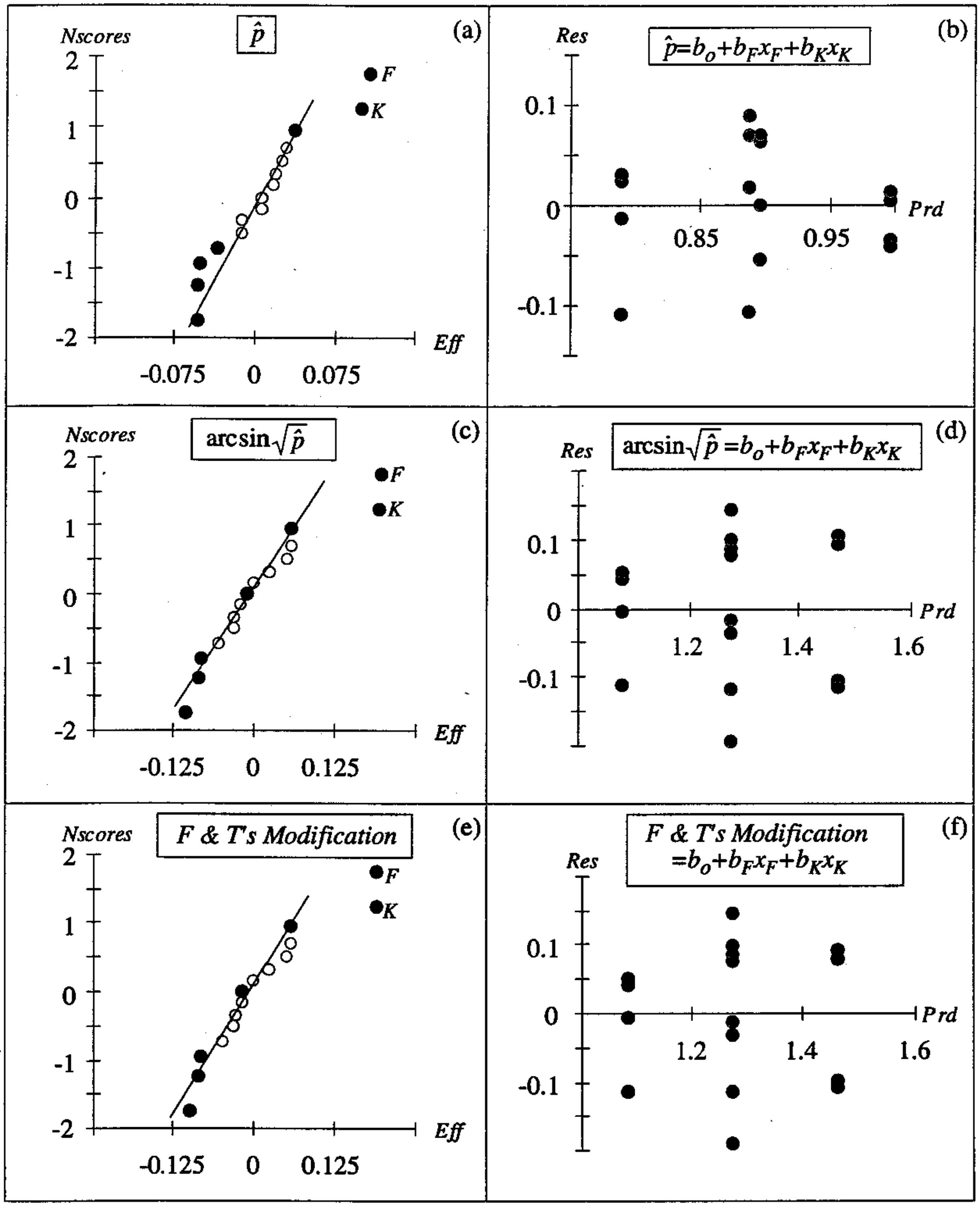

Figure 4. *The Normal Plots of effects and residuals vs. predicted values plots for the sand-castings using* $\hat{p}$, $\arcsin\sqrt{\hat{p}}$, *and Freeman and Tukey's (F & T) modification* $\left(\arcsin\sqrt{n\hat{p}/(n+1)}+\arcsin\sqrt{(n\hat{p}+1)/(n+1)}\right)/2$.

by the experimenters indicates that certain factors or interactions are active when they are not and that leads to confusion. Specifically, it will lead to a earnest effort by the engineers to explain effects that are not there and will thus bedevil attempts at further improvement. Thus the valuable catalytic effects of *scientific feedback* (Box and Draper, 1969) will be lost.

A further concern is in regard to the Taguchi strategy of experimentation described by the experimenters. This involves running a single experimental design, assessing the significant factors, determining the factor combination that provides the best predicted response, and performing a confirmatory trial. The obvious disadvantages of such a one-shot approach as opposed to the sequential strategy to experimentation have already been discussed in a previous column by George Box(1993b) and therefore need not be repeated here. However, there is also the disadvantage that if we include factors that are not significant because of misleading analysis and adjust, as the experimenters in our example did, those to the levels that supposedly yield the best predicted response, then we might end up prescribing expensive treatments that are not necessary. As a contrived but illustrative example suppose one of the factors were gold plating versus no gold plating then it would be unfortunate if our analysis misled us to believe that gold plating was beneficial when in fact it had no effect.

Our final concern is in regard to robustness. Robustness simply means that a system is insensitive to certain factors or their interactions. Hence we are looking for effects that are not significant. The irony therefore is that Taguchi who is credited for having emphasized the idea of robustness also seems to recommend methods of analysis that may pronounce effects significant when they are not and hence defeat this very purpose.

EXAMPLE 2: ANALYSIS OF DEFECTS

In this second example we re-examine an experiment conducted by Chrysler Motors Engineering in which the goal was to reduce the number of defects in the finish of sheet molded grille opening panels. The experiment was a 16-run, two-level fractional factorial in nine factors: Mold Cycle *(A)*, Viscosity *(B)*, Mold Temperature *(C)*, Mold Pressure *(D)*, Weight *(E)*, Priming *(F)*, Thickening Process *(G)*, Glass Type *(H)*, and Cutting Pattern *(J)*. The generators for this 2^{9-5} resolution *III* design were

Table 3.
The design matrix, data, and confounding pattern for the car grille opening panel experiment.

RUN	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*J*	$\hat{c}$	$\sqrt{\hat{c}}$	*F & T'S MODIFICATION*
1	−	−	−	−	+	−	+	−	+	56	7.48	7.52
2	+	−	−	−	+	−	−	+	−	17	4.12	4.18
3	−	+	−	−	−	+	+	−	−	2	1.41	1.57
4	+	+	−	−	−	+	−	+	+	4	2.00	2.12
5	−	−	+	−	+	+	−	+	+	3	1.73	1.87
6	+	−	+	−	+	+	+	−	−	4	2.00	2.12
7	−	+	+	−	−	−	−	+	−	50	7.07	7.12
8	+	+	+	−	−	−	+	−	+	2	1.41	1.57
9	−	−	−	+	−	+	+	+	+	1	1.00	1.21
10	+	−	−	+	−	+	−	−	−	0	0.00	0.50
11	−	+	−	+	+	−	+	+	−	3	1.73	1.87
12	+	+	−	+	+	−	−	−	+	12	3.46	3.54
13	−	−	+	+	−	−	−	−	+	3	1.73	1.87
14	+	−	+	+	−	−	+	+	−	4	2.00	2.12
15	−	+	+	+	+	+	−	−	−	0	0.00	0.50
16	+	+	+	+	+	+	+	+	+	0	0.00	0.50

$\ell_1 = A + BJ + CG$

$\ell_2 = B + AJ + DE$

$\ell_3 = J + AB + FH$

$\ell_4 = C + AG + EF$

$\ell_5 = G + AC + DH$

$\ell_6 = BC + DF + GJ$

$\ell_7 = BG + CJ + EH$

$\ell_8 = D + BE + GH$

$\ell_9 = AD + CH + EJ$

$\ell_{10} = E + BD + CF$

$\ell_{11} = CD + AH + BF$

$\ell_{12} = JD + AE + FG$

$\ell_{13} = H + DG + FJ$

$\ell_{14} = F + CE + HJ$

$\ell_{15} = AF + BH + EG$

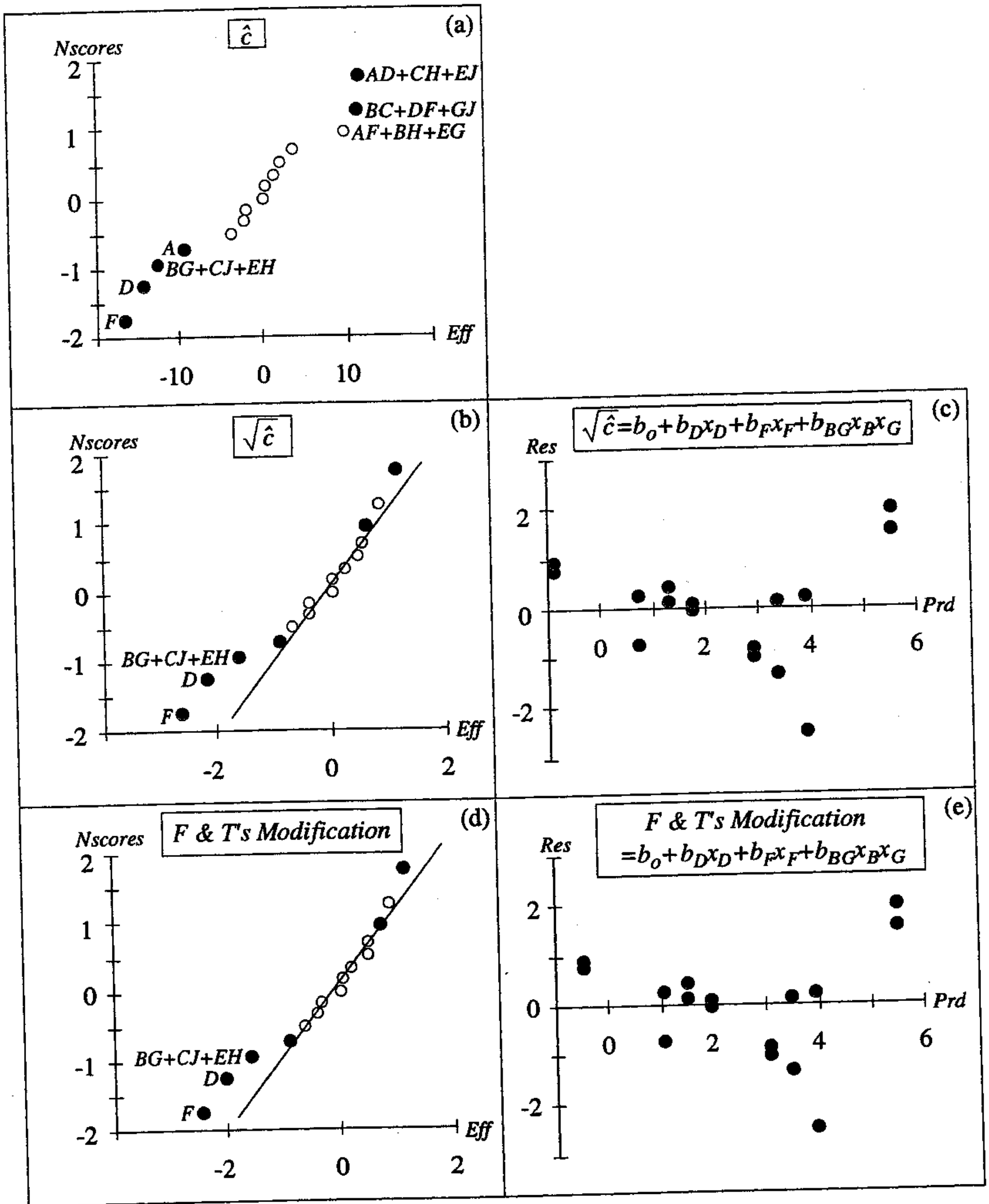

Figure 5. *The Normal Plots of effects and residuals vs. predicted values plots for the car grille opening panels using $\hat{c}$, $\sqrt{\hat{c}}$, and Freeman and Tukey's (F & T) modification* $\left(\sqrt{\hat{c}}+\sqrt{\hat{c}+1}\right)/2$.

$E = BD$, $F = BCD$, $G = AC$, $H = ACD$ and $J = AB$. The experimental design with the notional modifications described above, the confounding pattern, and the responses, $\hat{c}$, are displayed in Table 3. The experimental responses in this case are counts of defects per grille. The original analysis presented by the engineers from Chrysler was performed directly on the counts using analysis of variance. However, we will instead use the square root of the defects followed by a Normal Plot of the effects.

As in Example 1 we will again first analyze the data without transformation. Figure 5a shows aNormal Plot of the effects calculated directly on the raw data $(\hat{c})$. We see that seven effects and interactions are falling off the upward sloping line formed by the dots in the center of the plot. Hence it would appear that those effects are significant. Indeed except for the *AF* interaction and its aliases these factors were the effects that the experimenters reported as significant in their paper. On the Normal Plots in Figure 5, as in example 1, we have indicated with solid dots those effects that in the original report were pronounced significant and with open dots those that were used for error in the analysis of variance. (We have not shown the residual plot in the first case because when fitting seven effects to sixteen data points the residuals begin to lose their meaning)

By taking the square root of the counts we see in Figure 5b that now only the main effects *F* and *D* and the interaction *BG* (and its aliases) appear as significant. In this example, therefore, the additional interaction effects found by the engineers were likely due to curvature of the response surface induced by failing to use an appropriate transformation.

The response for several runs was zero and in Figures 5d and 5e we analyzed that data with the modified transformation $(\sqrt{c}+\sqrt{c+1})/2$. This modification does not in this case make a difference. In general, however, we do recommend the use of the modified transformation for the reasons illustrated in Figure 3.

Notice, incidentally that the residual plots in Figures 5c and 5e may indicate that the residual variance increases as a function of the predicted value. This in turn may indicate the need for further transformation of the already transformed counts. To do so the empirical approach described in a previous column (Bisgaard, 1993b) would be applicable but that will not be pursued here. However, remember that for real data there is no guarantee that defect counts follow exactly a Poisson distribution so further transformation is legitimate (Bartlett, 1947).

As a conclusion, it is interesting to note that the experimenters reported with enthusiasm that "this experiment is considered unique and successful [within Chrysler and among its suppliers]." We certainly agree and would like to encourage further applications of factorial experiments. Furthermore, the confirmatory experiment that they ran at the conditions that looked most favorable based on seven significant effects and interactions produced an increase in "first-time-through capability" from 77% to 96%. This we think shows the enormous power of two-level factorial experiments. As before, however, the analysis reported by the experimenters would be somewhat misleading for further development of the process. A less misleading analysis, which would most likely have pointed to the same most favorable factor combination and is a lot easier to use, is an appropriate transformation and Normal Plot rather than Accumulation Analysis and ANOVA. By doing so we are better able to use statistics to catalyze our creativity and to use engineering knowledge. Continuous improvement also applies to the tools and approaches we use!

ACKNOWLEDGMENT

The work on this manuscript was supported by a grant from the Alfred P. Sloan Foundation and by the National Science Foundation (NSF) under grant number ECD-8721545. The authors would like to thank George Box, Spencer Graves, Haim Shore and the CQPI Reports Committee for their helpful comments.

REFERENCES

Bartlett, M.S. (1937), "Some Examples of Statistical Methods of Research in Agriculture and Applied Biology", *Journal of the Royal Statistical Society*, Supplement Vol. 4, No. 2. p. 137-170.

Bartlett, M.S. (1947), "The Use of Transformations", *Biometrics*, Supplement Vol. 4, No. 2. p. 137-170.

Becknell, D. (1986), "Evaporative Cast Process 3.0 Liter Intake Manifold Poor Sandfill Study", *Fourth Symposium on Taguchi Methods*. American Supplier Institute; Dearborn, MI. p. 120-130.

Bisgaard, S. (1993a), "Spreadsheets for Analysis of Two-Level Factorials", *Quality Engineering*, Vol. 6, No. 1. p. 149-157.

Bisgaard, S.(1993b), "Iterative Analysis of Data from Two-level Factorials," *Quality Engineering*, Vol. 6, No. 2. p. 319-330.

Box, G.E.P. (1966), "Use and Abuse of Regression", *Technometrics*, Vol. 8, No. 4. p. 625-629.

Box, G.E.P. (1988), "Signal-to-Noise Ratios, Performance Criteria, and Transformations", *Technometrics*, Vol. 30. p. 1-17, with discussion p. 18-40.

Box, G.E.P. (1992), "What Can You Find Out From Eight Experimental Runs?", *Quality Engineering*, Vol. 4, No. 4. p. 619-625.

Box, G.E.P. (1993a), "What Can You Find Out From Sixteen Experimental Runs?", *Quality Engineering*, Vol. 5, No. 1. p. 164-178.

Box, G.E.P. (1993b). "Sequential Experimentation," *Quality Engineering*, Vol. 5, No. 2. p. 321-330.

Box, G.E.P. and N.R. Draper (1969), *Evolutionary Operation*. John Wiley & Sons, Inc.; NY.

Box, G.E.P., W.G. Hunter and J.S. Hunter (a.k.a. BH^2) (1978), "*Statistics for Experimenters*. John Wiley & Sons, Inc.; NY.

Box, G.E.P. and S. Jones (1992), "An Investigation of the Method of Accumulation Analysis", *Total Quality Management*, Vol. 1, No. 1. p. 101-113.

Freeman, M.F. and J.W. Tukey (1950), "Transformations Related to the Angular and the Square Root", *Annals of Mathematical Statistics*, Vol. 21, No. 4. p. 607-611.

Hsieh, P.I. and D.E. Goodwin (1986), "Sheet Molded Compound Process Improvement", *Fourth Symposium on Taguchi Methods*. American Supplier Institute; Dearborn, MI. p. 13-21.

Nair, V.J. (1990), "A Critical Look at Accumulation Analysis and Related Methods", *Technometrics*, Vol. 32. p. 151-152.