Are You Sure We Don't Need Normally Distributed Data?

More about the misuses of probability theory

Donald J. Wheeler

Last year I discussed the problems of transforming data prior to analysis (see my August, September, and October 2009 Columns). There I demonstrated how nonlinear transformations distort the data, hide the signals, and change the results of the analysis. However, some are still not convinced. While I cannot say why they prefer to transform the data in spite of the disastrous results described above, I can address the fallacious reasoning contained in the current argument used to justify transforming the data. Since the question of transforming the data is closely associated with the idea that the data have to be normally distributed we shall discuss these two questions together.

AN ARGUMENT FOR TRANSFORMING THE DATA

This argument begins with the assumption that three-sigma limits are based on a normal distribution, and therefore the normal theory coverage of **P** = 0.9973 for three-sigma limits must be a characteristic of the process behavior chart. When three-sigma limits are used with other probability models we immediately find that we end up with different coverages **P**. So what do these different values of **P** mean? By analogy with other statistical procedures which are characterized by the risk of a false alarm, or alpha level, these different coverages **P** are converted into the sequential equivalent: the average run length between false alarms, or ARL_0. These average run length values are then interpreted as a characterization of how the three-sigma limits will work with different probability models.

When signals consist of points outside the three-sigma limits a well-known result establishes that the theoretical ARL_0 values will simply be the inverse of [1–**P**]. Thus, the normal theory value for **P** of 0.9973 corresponds to an ARL_0 of 370. See Table 1 for the results for some other probability models.

Table1: Average Run Lengths for Six Probability Models

Probability Model	Skewness Squared	Kurtosis	P	ARL_0
Uniform	0.00	1.80	1.000	infinity
Normal	0.00	3.00	0.9973	370
Chi-Square 32 d.f.	0.25	3.38	0.9946	185
Chi-Square 16 d.f.	0.50	3.75	0.9925	133
Chi-Square 4 d.f.	2.00	6.00	0.9859	70.9
Exponential	4.00	9.00	0.9817	54.6

The smaller ARL_0 values for the skewed models are cited as evidence that the process behavior chart does not work with skewed data. These values are interpreted to mean that the

chart will give too many false alarms. Therefore, skewed data need to be transformed to make them look "more normal." This is said to reduce the number of false alarms and make the chart work more like it does in the normal theory case. Thus, it is said that any analysis should begin with a test for normality. If there is a detectable lack of fit, then a normalizing transformation is needed before the data are placed on a process behavior chart.

AN EARLY CRITIQUE

Many years ago Francis Anscombe, a Fellow of the American Statistical Association, summarized the strength of the argument above when he said that, "Testing the data for normality prior to placing them on a control chart is like setting to sea in a rowboat to see if the Queen Mary can sail." Those of us who have published articles on lack-of-fit tests in refereed journals know about the weaknesses that are inherent in all such tests. However, there are more fundamental problems with the argument above than just the weakness of lack-of-fit tests. In order to fully explain these problems it is necessary to go into some detail about the origins of three-sigma limits and the inappropriateness of using the ARL_0 values in the manner they are used in the preceding argument.

SHEWHART'S ARGUMENT

On pages 275-277 of *Economic Control of Quality of Manufactured Product* Shewhart discussed the problem of "establishing an efficient method for detecting the presence" of assignable causes of exceptional variation. He began with a careful statement of the problem which can paraphrased as follows: If we know the probability model that characterizes the original measurements, X, when the process satisfies the differential equation of statistical control, then we can usually find a probability model, $f(y,n)$, for a statistic Y calculated from a sample of size n such that the integral:

$$\int_{\mathbf{A}}^{\mathbf{B}} f(y,n)\, dy \;=\; \mathbf{P}$$

will define the probability **P** that the statistic Y will have a value that falls in the interval defined by **A** and **B**. A good name for this approach would be the statistical approach.

The Statistical Approach:
Choose a fixed value for **P** that is close to 1.00,
then for a specific probability distribution $f(y,n)$,
find the critical values **A** and **B.**

When Y falls outside the interval **A** to **B** the observation may be said to be inconsistent with the conditions under which **A** and **B** were computed. This approach is used in all sorts of statistical inference techniques. And, as the argument says, once you fix the value for **P**, the values for **A** and **B** will depend upon the probability model $f(y,n)$.

But Shewhart did not stop here. He went on to observe that: "For the most part, however, we never know $f(y,n)$ in sufficient detail to set up such limits." Therefore, Shewhart effectively

reversed the whole argument above and proceeded as follows:

Shewhart's Approach:
Choose GENERIC values for **A** and **B** such that,
for ANY probability model *f(y,n)*,
the value of **P** will be reasonably close to 1.00.

Such generic limits will still allow a reasonable judgment that the process is unlikely to be predictable when Y is outside the interval **A** to **B**. Ultimately, we do not care about the exact value of **P**. As long as **P** is reasonably close to 1.00 we will end up making the correct decision virtually every time. After consideration of the economic issues of this decision process Shewhart summarized with the following:

> "For these reasons we usually choose a symmetrical range characterized by limits
>
> $$Average \pm t\ sigma$$
>
> symmetrically spaced in reference to [the Average}. Tchebycheff's theorem tells us that the probability **P** that an observed value of Y will lie within these limits so long as the quality standard is maintained satisfies the inequality
>
> $$\mathbf{P} > 1 - \frac{1}{t^2}$$
>
> "We are still faced with the choice of t. Experience indicates that $t = 3$ seems to be an acceptable economic value."

Thus, Shewhart's approach to the problem of detecting the presence of assignable causes is the complete opposite of the approach used with techniques for statistical inference. The very idea that the data have to be normally distributed before they can be placed on a process behavior chart is an expression of a misunderstanding of the relationship between the process behavior chart and the techniques of statistical inference. As Shewhart later noted, "*We are not concerned with the functional form of the universe, but merely with the assumption that a universe exists.*" The question of homogeneity is the fundamental question of data analysis.

If the data are homogeneous, then the process behavior chart will demonstrate that homogeneity and it will also provide reasonable estimates for both the process location and dispersion. At the same time the histogram will show how the process is performing relative to the specifications. Thus, virtually all of the interesting questions will be answered by the homogeneous data without having to make reference to a probability model. So while it might be plausible to fit a probability model to a set of homogeneous data, it is, in every practical sense, unnecessary.

If the data are not homogeneous, then the process behavior chart will demonstrate that the process is subject to the effects of assignable causes of exceptional variation. When this happens the histogram will be a mixture the process outcomes under two or more conditions. While the histogram will still describe the past, it looses any ability to characterize the future. Here it is not probability models or computations that are needed. Rather, you will need to take action to find the assignable causes. Fitting a probability model to nonhomogeneous data, or determining

which nonlinear transform to use on nonhomogeneous data, is simply a triumph of computation over common sense. It is also indicative of a fundamental lack of understanding about statistical analysis.

The software will give you lack-of-fit statistics and the statistics for skewness and kurtosis (all of which are numbers that you would never have computed by hand). However, just because they are provided by the software does not mean that these values are appropriate for your use. The computation of any lack-of-fit statistic, just like the computation of global skewness and kurtosis statistics, is predicated upon having a homogeneous data set. *When the data are nonhomogeneous all such computations are meaningless.* Any attempt to use these values to decide if the data are detectably nonnormal, or to decide which transformation to use, is complete nonsense.

Thus, fitting a probability model to your data is simply a red herring. When the data are homogeneous it will only be a waste of time. When the data are nonhomogeneous it will be meaningless. Moreover, whenever a lack-of-fit test leads you to use a nonlinear transformation on your data you will end up distorting your whole analysis.

THE COVERAGE OF THREE SIGMA LIMITS

In the passage above we saw how Shewhart used the Chebychev inequality to establish the *existence* of some number t that will result in values for **P** that are reasonably close to 1.00. Following this existence theorem, he then turned to empirical evidence to justify his choice of $t = 3$. While the Chebychev inequality will only guarantee that three-sigma limits will cover at least 89% of the area under the probability model, the reality is that three-sigma limits will give values for **P** that are much closer to 1.00 in practice.

To illustrate what happens in practice I decided to look at different probability models and to compute the area under the curve included within the "three-sigma interval" defined by the parameters:

$$MEAN(X) \pm 3\ SD(X)$$

I began with 100 binomial distributions. Next I looked at 112 Poisson distributions, 43 gamma distributions, 100 F-distributions, 398 beta distributions, 41 Weibull distributions, and 349 Burr distributions. These 1143 distributions contain all of the chi-square distributions, and they cover the regions occupied by lognormals, normals, and Student t distributions. They include U-shaped, J-shaped, and mound-shaped distributions. Since each of these probability models will have a skewness parameter and a kurtosis parameter, we can use the values of these two shape parameters to place each of these 1143 distributions in the shape characterization plane.

In Figure 1 each of the 1143 probability models is shown as a point in this shape characterization plane. For any given probability model the horizontal axis will show the value of the square of the skewness parameter, while the vertical axis will show the value for the kurtosis parameter. The lines shown divide the shape characterization plane into regions for mound-shaped probability models, J-shaped probability models, and U-shaped probability models.

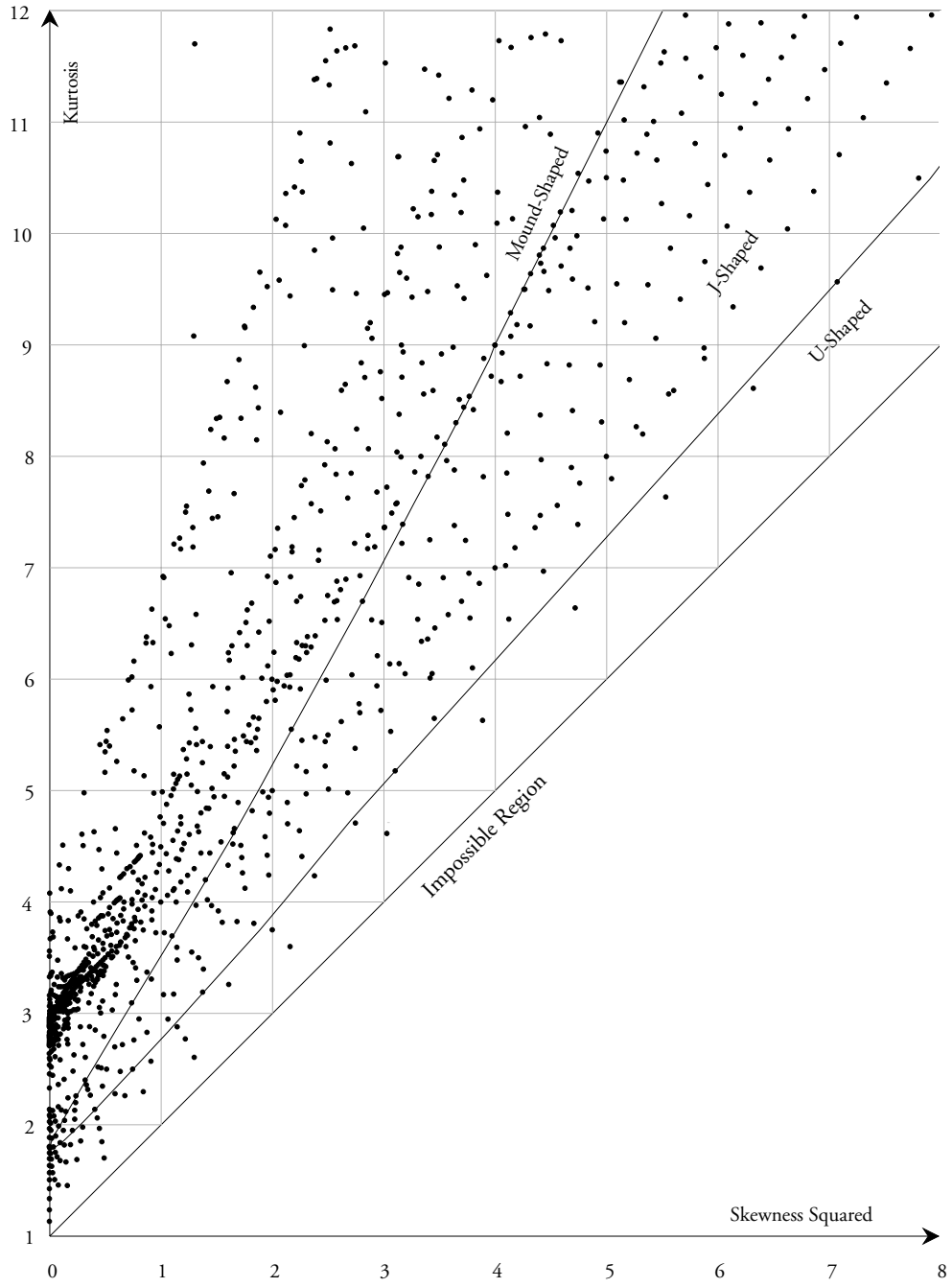

Figure 1: 1143 Probability Models in the Shape Characterization Plane

By considering the coverage of the three-sigma intervals for these 1143 models I was able to construct the contour map shown in Figure 2.

Starting on the lower left, the bottom region is the region where the three-sigma interval gives 100% coverage. The shaded slice above this region is the region where the three-sigma interval will give better than 99.5% coverage. Continuing in a clockwise direction, successive slices define regions where the three-sigma intervals will give better than 99% coverage, better

than 98.5% coverage, better than 98% coverage, etc. The six probability models of Table 1 are shown in Figure 2. This will allow you to compare the **P** values of Table 1 with the contours shown in Figure 2.

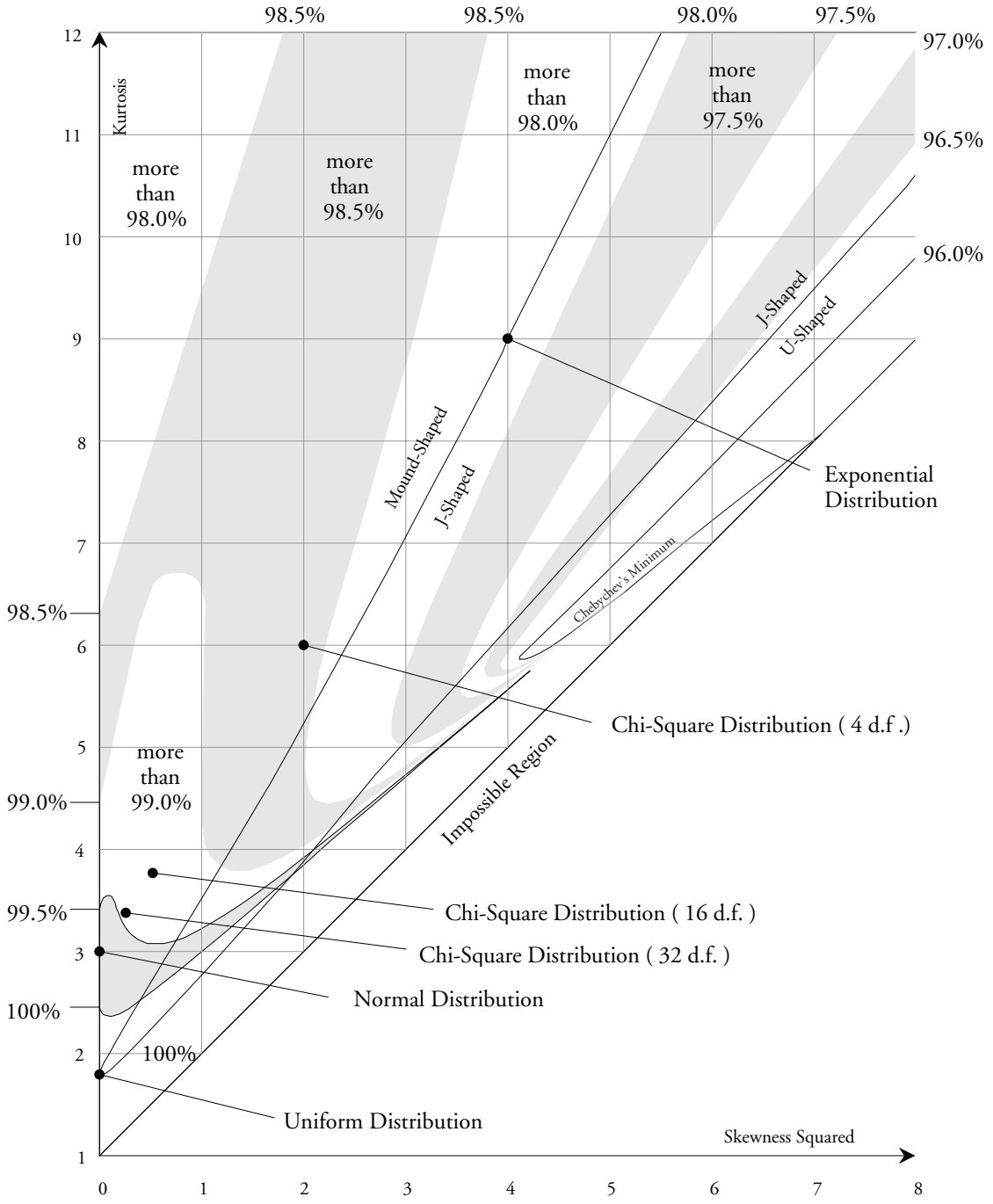

Figure 2: The Coverage of Three-Sigma Limits

Inspection of Figure 2 will reveal that three-sigma intervals will provide better than 98% coverage for *all* mound-shaped probability models. Since most histograms are mound-shaped,

this 98% or better coverage hints at why Shewhart found three-sigma limits to be satisfactory in practice.

Further inspection of Figure 2 will show that most of the J-shaped probability models will have better than 97.5% coverage with three-sigma intervals. Since these models tend to cover the remaining histograms we find in practice, we see that three-sigma limits actually do much better than the Chebychev inequality would lead us to believe.

In the U-shaped region most of the probability models will have 100% coverage. However, some very skewed and very heavy-tailed U-shaped probability models will fall into the Chebychev minimum. Offsetting this one area of weakness is the fact that when your histogram has two humps it is almost certain that you are looking at the mixture of two different processes, making the use of three-sigma limits unnecessary.

Thus, when the process is being operated predictably, Shewhart's generic three-sigma limits will result in values of **P** that are in the conservative zone of better than 97.5% coverage in virtually every case. When you find points outside these generic three-sigma limits you can be sure that either a rare event has occurred or else the process is not being operated predictably. (It is exceedingly important to note that this is *more conservative* than the traditional 5% risk of a false alarm that is used in virtually every other statistical procedure.)

Why were Figures 1 and 2 restricted to the values shown? Can not both skewness and kurtosis get bigger than the values in Figures 1 and 2? Yes, they can. The choice of the region used in Figures 1 and 2 was based on two considerations. First, whenever you end up with very large skewness or kurtosis statistics it will almost always be due to an unpredictable process. When this happens the skewness and kurtosis statistics are not a characteristic of the process, but are describing the mixture of two or more conditions.

Second, predictable processes are the result of routine variation. As I showed in my May column (Two Routes to Process Improvement) routine variation can be thought of as the result of a large number of common causes where no one cause has a dominant effect. In other words, a predictable process will generally correspond to some high-entropy condition, and the three distributions of maximum entropy are the normal distribution, the uniform distribution, and the exponential distribution. Figures 1 and 2 contain these three distributions and effectively bracket the whole region of high-entropy distributions that surround them. When considering probability models that are appropriate for predictable processes we do not need to go further afield. The region shown in Figures 1 and 2 is sufficient.

THE AVERAGE RUN LENGTH ARGUMENT

Any statistical technique that is used to detect signals can be characterized by a theoretical power function. With sequential procedures a characteristic number that is part of the power function is the average run length between false alarms (ARL_0). As we saw in Table 1 the ARL_0 value will be the inverse of [1 – **P**].

Of course, when we invert a small value we get a big value. This means that two values for **P** that are close to 1.00, and which are equivalent in practice, will be transformed into ARL_0 values that are substantially different. Interpreting the differences in the ARL_0 values in Table 1 to mean that we need to transform our data prior to placing them on a process behavior chart is nothing more or less than a bit of statistical slight-of-hand. There are three different problems with using

the ARL_0 values in Table 1 to argue that we need to have "normally distributed data."

The first problem with Table 1 is that this is an inappropriate use of the ARL_0 values. Average run length *curves* are used to compare different *techniques* using the *same* probability model. The ARL_0 values are merely the end points of these *ARL* curves. As the end points they are not the basis of the comparison, but only one part of a larger comparison. It is this comparison under the same conditions that makes the *ARL* curves useful. Using the same probability model with different techniques creates a mathematical environment where the techniques may be compared in a fair and comprehensive manner. Large differences in the *ARL* curves will translate into differences between the techniques in practice, while small differences in the *ARL* curves are unlikely to be seen in practice. But this is not what is happening in Table 1. There we are changing the probability model while holding the technique constant. As we will see in the next paragraph, regardless of the technique considered, changing the probability model will always result in dramatic changes in the ARL_0 values. *These changes are not a property of the way the technique works, but are merely a consequence of the differences between the probability models.* Thus, the first problem with Table 1 is that instead of holding the probability model constant and comparing techniques, it holds the technique constant and compares probability models (which we already know are different).

The second problem is that the differences in the ARL_0 values in Table 1 do not represent differences in practice. By definition, the ARL_0 values are the inverses of the areas under the extreme tails of the probability models. Since all histograms will have finite tails, there will always be discrepancies between our histograms and the extreme tails of our probability models. Given enough data, you will *always* find a lack of fit between your data and any probability model you might choose! Those who have studied lack-of-fit tests and are old enough to have computed the various lack-of-fit statistics by hand will understand this. Thus, the differences between the ARL_0 values in Table 1 tell us more about the differences in the extreme tails of the probability models than they tell us about how a process behavior chart will operate when used with a finite data set. Thus, the second problem is that the most artificial and unrealistic part of any *ARL* curve is the ARL_0 value.

The third problem with the ARL_0 values in Table 1 is the fact that they are computed under the assumption that you have an infinite number of degrees of freedom. They are expected values. In practice the three-sigma limits will vary. If we take this variation into account and look at how this variation translates into variation in the value for **P** for different probability models, then we can discover how the ARL_0 values vary. In Table 2 the limits are assumed to have been computed using 30 degrees of freedom. The ARL_0 values shown correspond to the **P** values that correspond to the middle 95% of the distribution of computed limits. As may be seen the differences in the Mean ARL_0 values disappear in the overwhelming uncertainties given in the last column. Moreover, there is no practical difference in the lower bounds on the ARL_0 values regardless of the probability model used. This is why transforming the data will not improve your ARL_0 to any appreciable degree. Thus, the third problem with comparing the Mean ARL_0 values is that the routine variation in practice will obliterate differences seen in Table 1.

Table 2 Average Run Lengths for Limits Having 30 Degrees of Freedom

Probability Model	Skewness Squared	Kurtosis	Mean ARL_0	95% Prediction Interval for ARL_0 Values
Normal	0.00	3.00	370	38 to 6,238
Chi-Square 32 d.f.	0.25	3.38	185	45 to 915
Chi-Square 16 d.f.	0.50	3.75	133	43 to 530
Chi-Square 4 d.f.	2.00	6.00	70.9	32 to 183
Exponential	4.00	9.00	54.6	28 to 118

SUMMARY

Therefore, while the argument presented at the start of this column has been used to justify transforming data to make them "more normal" it does not, in the end, justify this practice. The whole argument, from start to finish, is based on a strong presumption that the data are homogeneous and the underlying process is predictable. Of course, according to the argument, the purpose of the transformation is to make the data "suitable" for use in a technique that is intended to examine the data for homogeneity! Thus, this argument could be paraphrased as: "Three-sigma limits won't work properly unless you have good data."

Based on my experience, when you start using process behavior charts, you will find at least 9 out of 10 processes to be operated unpredictably. Both **P** and ARL_0 are concerned with false alarms. In the presence of real signals we do not need to worry about the ARL_0 value and trying to get a specific value for **P** is wasted effort. As long as **P** is reasonably close to 1.00, we will end up making the right decisions virtually every time. And, as we have seen, three-sigma limits will yield values of **P** that are greater than 0.975 with any probability distribution that provides a reasonable model for a predictable process.

Fortunately, the process behavior chart has been completely tested and proven in over 80 years of practice. No transformations were needed then. No transformations are needed now. Process behavior charts work, and they work well, *even when you use bad data*.

From the very beginning statisticians, who by nature and training are wedded to the statistical approach, have been nervous with Shewhart's approach. To them it sounds too simple. So they invariably want to jump into the gap and add things to fix what they perceive as problems. This started with E. S. Pearson's book in 1935, and it has continued to this day. Transforming your data prior to placing them on a process behavior chart is simply the latest manifestation of this nervousness. However, as perhaps you can begin to see from this column, the simplicity of Shewhart's process behavior chart is not the simplicity that comes from glossing over the statistical approach, but it is rather the simplicity that is found on the far side of the complexity of a careful and thoughtful analysis of the whole problem.

As Elton Trueblood said so many years ago, "There are people who are afraid of clarity because they fear that it may not seem profound." Beware those who try to make process behavior charts more complex than they need to be.