Transforming the Data Can Be Fatal to Your Analysis

Donald J. Wheeler
September 2009

Following my article on Leptokurtophobia it was almost inevitable that we should hear from a certified leptokurtophobe. We were fortunate to have someone as articulate as Forrest Breyfogle III to write the response. In reading his article I got the feeling that the concept of *not* transforming the data was so far outside of his thought pattern that he simply could not conceive of any other course of action. Hence, rather than offering a critique of the points raised in my original article, Mr. Breyfogle chose to ignore the arguments against transforming the data and to simply repeat his mantra of "transform, transform, transform." Thirty-five years ago I also thought that way, but now I know better, and out of respect for those who are interested in learning how to better analyze data, I feel the need to further explain why the transformation of data can be fatal to your analysis.

Starting on page 275 of *Economic Control of Quality of Manufactured Product* Dr. Shewhart described two completely different approaches to working with data. The first of these approaches we will call the statistical approach since it describes how we create a test statistic. This statistical approach consists of four steps: (1) Choose an appropriate probability model to use. (2) Choose some risk of a false alarm to use. (3) Find the exact critical values for the selected model that correspond to this risk of a false alarm, or else transform the selected model to match some known critical values. (4) Then use these critical values in your analysis. While this can make sense when working with *functions* of the data (i.e. statistics) it does not work when applied to the original data themselves. As Dr. Shewhart points out, we will *never* have enough data to uniquely identify a specific probability model. Probability models are limiting functions for infinite sequences, and therefore they can never be said to apply to any finite portion of that sequence. This is why any assumption of a probability model is just that—an unverifiable assumption. Dr. Shewhart goes on to say that even if we did assume a probability model, we still would not know the mean or the variance of that model.

So what are we to do when we try to analyze data? Dr. Shewhart suggests a different approach for the analysis of original data. Shewhart's approach also consists of four steps: (1) Choose some generic critical values for which (2) the risk of a false alarm *will be reasonably small* (3) *regardless* of what probability model we might choose, and (4) use these generic critical values in our analysis. This approach changes what is fixed and what is allowed to vary. With the statistical approach the alpha-level is fixed, and the critical values vary to match the specific probability model. With Shewhart's approach it is the critical values that are fixed and the alpha-level that is allowed to vary. This reversal of the statistical approach is what makes Shewhart's approach so hard for those with statistical training to understand.

Statistical Approach	**Shewhart's Approach**
1. Choose an appropriate probability model,	1. Choose some generic critical values for which
2. choose some risk of a false alarm,	2. the risk of a false alarm will be small
3. find exact critical values (or else transform data to match some known critical values),	3. *regardless* of what probability model we might choose, and
4. and use these critical values in the analysis.	4. use these generic critical values in the analysis.

Dr. Shewhart's generic solution to the problem of how to analyze a stream of data was to use three-sigma limits. These limits have been used around the world for over 70 years and they have been thoroughly proven. They strike a reasonable balance between the economic consequences of the twin errors of failing to detect signals and having false alarms. Three-sigma limits have been proven to work, and this is a fact of life, not a matter of opinion. To quote a note penciled into one of my books by Dr. Deming, "This means that even *wide* departures from normality will have virtually no effect upon the way the control chart functions."

However, when someone does not appreciate the difference between the statistical approach and Shewhart's approach it is almost inevitable that they will get lost trying to apply the statistical approach to original data.

In my earlier article (Quality Digest, August 5, 2009) I pointed out how three-sigma limits will filter out virtually all of the routine variation regardless of the shape of the histogram. I illustrated this with six specific examples ranging from the uniform to the exponential, and described a broader study encompassing over 1100 probability models where 97.3% of these models had better than 97.5% coverage at three-sigma limits. This kind of behavior is the very definition of robustness.

(Note that there has never been any claim that the area outside the three-sigma limits will remain constant regardless of the probability model, just that the alpha-level will remain reasonably small. With any statistical procedure we will change the risk of a false alarm whenever we change the original probability model. A *robust* procedure is one where a conservative alpha-level will remain conservative [generally taken as anything under 5%] and a traditional alpha-level will remain traditional [under 10%].)

But Mr. Breyfogle wants to prove that the *X* Chart is "not robust" to non-normal data. To this end he uses a lognormal probability model that has a skewness of 6.2 and a kurtosis of 113.9. (To understand just how extreme this model is it might be helpful to know that virtually all reasonable models for original data will have a skewness of less than 2.5 and a kurtosis of less than 10. If you have a histogram of real data that has skewness and kurtosis statistics that fall outside of the range above, it is a virtual certainty that the underlying process is out-of-control.)

So, in his quest to show that the *X* Chart is not robust to non-normal data Mr. Breyfogle selected a very, very extreme probability model. *This lognormal probability model has 98.19% of its area contained within the interval defined by the Mean plus or minus Three Standard Deviations.*

However, rather than looking at this theoretical value, Mr. Breyfogle generated 1000 observations from this model and placed them on an *X* Chart. He found 3.3% of the values in his sample outside the limits on this chart. He then complains that this is not the 3 out of a 1000 that we expect when using a normal distribution. But where is the surprise in this? The alpha-level is supposed to vary. When we go from a kurtosis of 3 to a kurtosis of 114, we should expect an increase in the area outside the three-sigma limits! The fact that it changes so little is the real surprise here. Thus, the very behavior that Mr. Breyfogle cites as evidence of a lack of robustness, is, in fact, a stunning demonstration of robustness—the *XmR* Chart still yields a conservative false alarm rate in spite of the extreme kurtosis!

Next Mr. Breyfogle complains that his observations on the *X* Chart are not symmetrically spread out around the central line in a "random scatter pattern." Once again, where is the surprise? These pseudo-data are lognormal! As I noted in my earlier article, whenever we have skewed data there will be a boundary value on one side that will fall inside the computed three-sigma limits. When this happens the boundary value takes precedence over the computed limit and we end up with a one-sided chart. Mr. Breyfogle's insistence on finding a set of values that are symmetrically spread out around the central line in a "random scatter pattern" is *an interpretative guideline for use with plots of residuals*. This is a completely different type of analysis than the analysis of the original data.

When we fit a regression model to a set of data we are trying to explain most of the variation in the response variable. When we do this successfully, the amount of variation that remains between the data and the fitted model are known as the residuals. When we look at these residuals, we would like to see a symmetric plot since any lack of symmetry would suggest that we have used the wrong model in our regression. Moreover, it is also appropriate to check these residuals for a detectable lack of normality since the residuals should be nothing but noise, and the classic model for noise is the bell shaped curve. Unfortunately, we cannot, like Mr. Breyfogle, simply take analyses and guidelines that are appropriate for the residuals from a regression model and apply them to the original data. To do so is to demonstrate a misunderstanding of the concepts behind the techniques of statistical analysis.

Finally Mr. Breyfogle complains that, from an operational perspective, the original data give too many false alarms, and that we need to transform the data to eliminate these false alarms. However, the situation he describes is one where these pseudo-data represent years of production and are being looked at in a retrospective manner. If you are interested in looking for assignable causes you will need to be using the process behavior chart (control chart) in real time. In a retrospective use of the chart you are unlikely to ever look for any assignable causes, so where is the problem?

In any sequence of 1000 real data I would expect to find signals, and the points outside the limits are the place to start looking for the assignable causes behind those signals. While some of the points outside the limits may be false alarms, with 1000 real data it is essentially inevitable that many of the points outside the limits will be signals. Occasional false alarms are a reasonable price to pay to avoid missing the signals that you need to know about. While a process behavior chart may be used for the one-time analysis of retrospective data, it is important to understand that the chart was created for use as a sequential analysis procedure, and that its retrospective use changes the way we interpret the chart. Here the emphasis is no longer upon using the individual points to identify assignable causes, but rather on the overall behavior displayed on the chart. (While we know that Mr. Breyfogle's observations are synthetic, and contain no signals, his original *X* Chart, if it was based upon real data, would justify the judgment that the underlying process is subject to assignable causes. This justification would come from the *extent* to which several of the points exceed the upper limit, rather than being based upon how many points fall outside the limits. Once again, this is because reasonable models for original data do not typically have kurtosis values greater than 10.)

When Mr. Breyfogle transforms his extremely skewed and leptokurtic data he gets a wonderful bell-shaped histogram, which should not be a surprise. He also changes the false alarm rate from the 3.3% back to a normal theory value of 1 in 1000. Again, no surprise, this is exactly what should happen. However, in practice, it is not the false-alarm rate that we are concerned with, but rather the ability to detect signals of process changes. And that is why I used real data in my article. There we saw that a non-linear transformation may make the histogram look more bell-shaped, but in addition to distorting the original data, *it also tend to hide all of the signals contained within those data*.

The important fact about nonlinear transformations is not that they reduce the false-alarm rate, but rather that they obscure the signals of process change. Since presumably, the purpose of analysis is discovery, this tendency of non-linear transformations to obliterate the signals contained within the data makes their use on the original data completely inappropriate. Nowhere in Mr. Breyfogle's article does he address this point.

The first step in data analysis has nothing to do with what probability model is appropriate. Data are never generated by a probability model. Rather they are generated by a process or system that can change without warning. This is why the primary question of data analysis is concerned with whether or not the data are homogeneous. If the data are homogeneous, then it might make sense to select some probability model to represent the data. But if the data are not homogeneous, then no single probability model will ever be appropriate since the process is changing. The process behavior chart was deliberately created and intended for use in this initial step of analysis. It examines the data for evidence of a lack of homogeneity that might indicate changes in the process where no changes ought to have occurred. If you transform the data to make them appear to be "more normal," you are likely to end up with a beautiful, but completely incorrect, analysis.

Remember Daniel Boorstin's caution: "The greatest obstacle to discovery is not ignorance, but rather the illusion of knowledge." Once you have been taught erroneous ideas, it is hard to change.