http://www.qualitydigest.com/inside/quality-insider-column/individuals-control-chart-and-data-normality.html

Non-normal data: To Transform or Not to Transform

Sometimes you need to transform non-normal data.

Forrest Breyfogle III | 08/24/2009

"Do You Have Leptokurtophobia?" Don Wheeler stated, "'But the software suggests transforming the data!' Such advice is simply another piece of confusion. The fallacy of transforming the data is as follows:

"The first principle for understanding data is that no data have meaning apart from their context. Analysis begins with context, is driven by context, and ends with the results being interpreted in the context of the original data. This principle requires that there must always be a link between what you do with the data and the original context for the data. Any transformation of the data risks breaking this linkage. If a transformation makes sense both in terms of the original data and the objectives of the analysis, then it will be okay to use that transformation. Only you as the user can determine when a transformation will make sense in the context of the data. (The software cannot do this because it will never know the context.) Moreover, since these sensible transformations will tend to be fairly simple in nature, they do not tend to distort the data."

I agree with Wheeler in that data transformations that make no physical sense can lead to the wrong action or nonaction; however, his following statement concerns me. "Therefore, we do not have to pre-qualify our data before we place them on a process behavior chart. We do not need to check the data for normality, nor do we need to define a reference distribution prior to computing limits. Anyone who tells you anything to the contrary is simply trying to complicate your life unnecessarily."

I, too, do not want to complicate people's lives unnecessarily; however, it is important that someone's oversimplification does not cause inappropriate behavior.

The following illustrates, from a high level, or what I call a 30,000-foot-level, when and how to apply transformations and present results to others so that the data analysis leads to the most appropriate action or nonaction. Statistical software makes the application of transformations simple.

Why track a process?

There are three reasons for statistical tracking and reporting of transactional and manufacturing process outputs:

1. Is the process unstable or did something out of the ordinary occur, which requires action or no action?
2. Is the process stable and meeting internal and external customer needs? If so, no action is required.
3. Is the process stable but does not meet internal and external customer needs? If so, process improvement efforts are needed.

W. Edwards Deming, in his book, *Out of the Crisis* (Massachusetts Institute of Technology, 1982) stated, “We shall speak of faults of the system as common causes of trouble, and faults from fleeting events as special causes.... Confusion between common causes and special causes leads to frustration of everyone, and leads to greater variability and to higher costs, exactly contrary to what is needed. I should estimate that in my experience, most troubles and most possibilities for improvement add up to proportions something like this: 94 percent belong to the system (responsibility of management), 6 percent special.”

With this perspective, the second portion of item No. 1 could be considered a special-cause occurrence, while items No. 2 and 3 could be considered common-cause occurrence.

A tracking system is needed for determining which of the three above categories best describes a given situation.

Is the individuals control chart robust to non-normality?

The following will demonstrate how an individuals control chart is not robust to non-normally distributed data. The implication of this is that an erroneous decision could be made relative to the three listed reasons, if an appropriate transformation is not made.

To enhance the process of selecting the most appropriate action or nonaction from the three listed reasons, an alternate control charting approach will be presented, accompanied by a procedure to describe process capability/performance reporting in terms that are easy to understand and visualize.

Let’s consider a hypothetical application. A panel’s flatness, which historically had a 0.100 in. upper specification limit, was reduced by the customer to 0.035 in. Consider, for purpose of illustration, that the customer considered a manufacturing nonconformance rate above 1 percent to be unsatisfactory.

Physical limitations are that flatness measurements cannot go below zero, and experience has shown that common-cause variability for this type of situation often follows a log-normal distribution.

The person who was analyzing the data wanted to examine the process at a 30,000-foot-level view to determine how well the shipped parts met customers' needs. She thought that there might be differences between production machines, shifts of the day, material lot-to-lot thickness, and several other input variables. Because she wanted typical variability of these inputs as a source of common-cause variability relative to the overall dimensional requirement, she chose to use an individuals control chart that had a daily subgrouping interval. She chose to track the flatness of one randomly-selected, daily-shipped product during the last several years that the product had been produced.

She understood that a log-normal distribution might not be a perfect fit for a 30,000-foot-level assessment, since a multimodal distribution could be present if there were a significant difference between machines, etc. However, these issues could be checked out later since the log-normal distribution might be close enough for this customer-product-receipt point of view.

To model this situation, consider that 1,000 points were randomly generated from a log-normal distribution with a location parameter of two, a scale parameter of one, and a threshold of zero (i.e., log normal 2.0, 1.0, 0). The distribution from which these samples were drawn is shown in figure 1. A normal probability plot of the 1,000 sample data points is shown in figure 2.

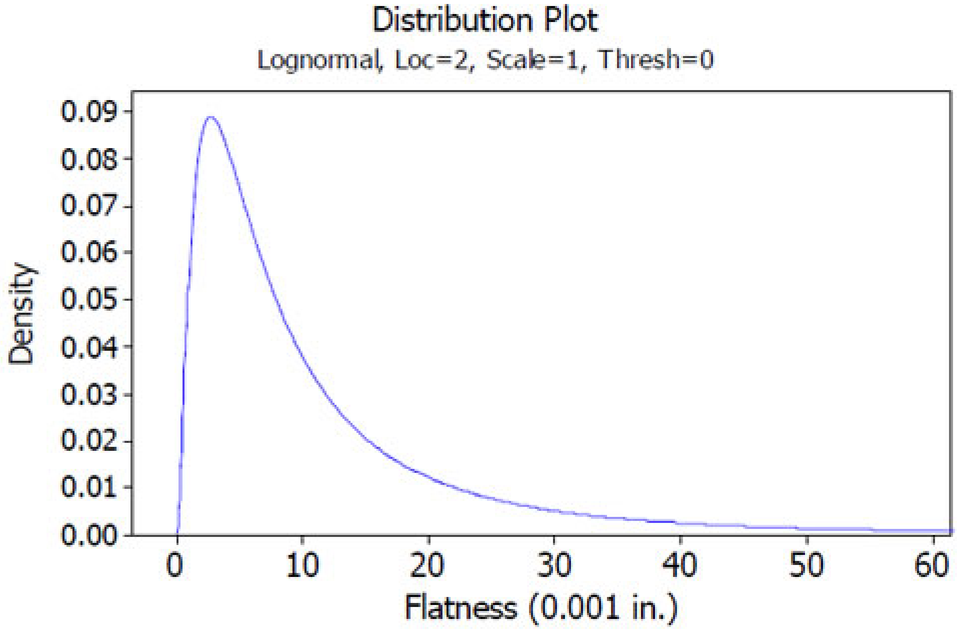


Figure 1: Distribution From Which Samples Were Selected

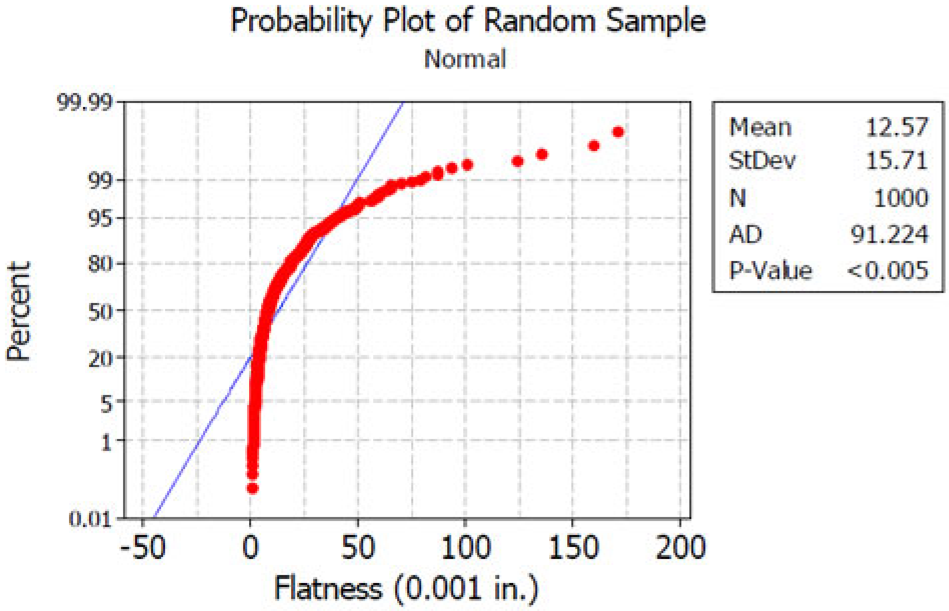


Figure 2: Normal Probability Plot of the Data

From figure 2, we statistically reject the null hypothesis of normality technically, because of the low p-value, and physically, since the normal probability plotted data does not follow a straight line. This is also logically consistent with the problem setting, where we do not expect a normal distribution for the output of such a process having a lower boundary of zero. A log-normal probability plot of the data is shown in figure 3.

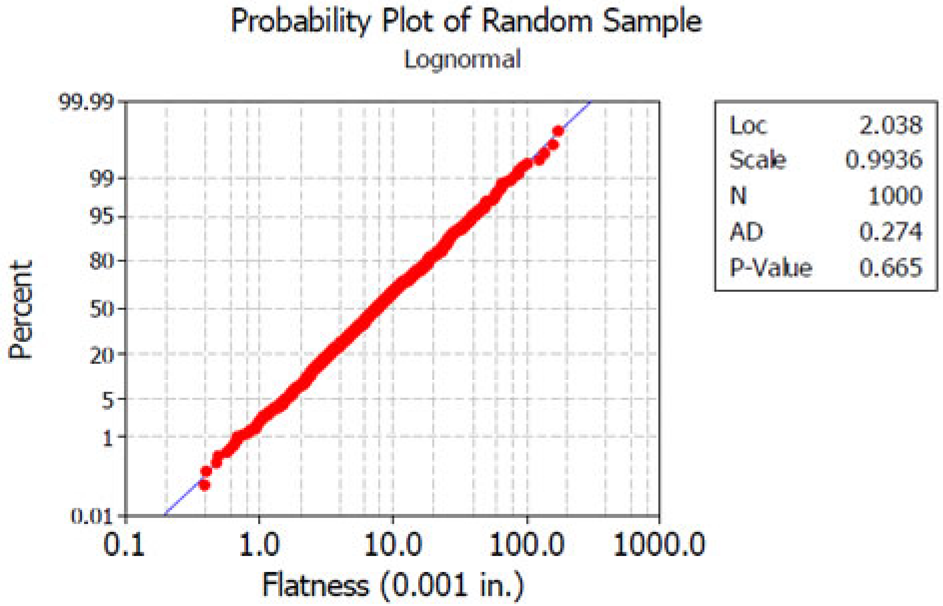


Figure 3: Log-Normal Probability Plot of the Data

From figure 3, we fail to statistically reject the null hypothesis of the data being from a log-normal distribution, since the p-value is not below our criteria of 0.05, and physically, since the log-normal probability plotted data tends to follow a straight line. Hence, it is reasonable to model the distribution of this variable as log normal.

If the individuals control chart is robust to data non-normality, an individuals control chart of the randomly generated log-normal data should be in statistical control. In the most basic sense, using the simplest run rule (a point is "out of control" when it is beyond the control limits), we would expect such data to give a false alarm on the average three or four times out of 1,000 points. Further, we would expect false alarms below the lower control limit to be equally likely to occur, as would false alarms above the upper control limit.

Figure 4 shows an individuals control chart of the randomly generated data.

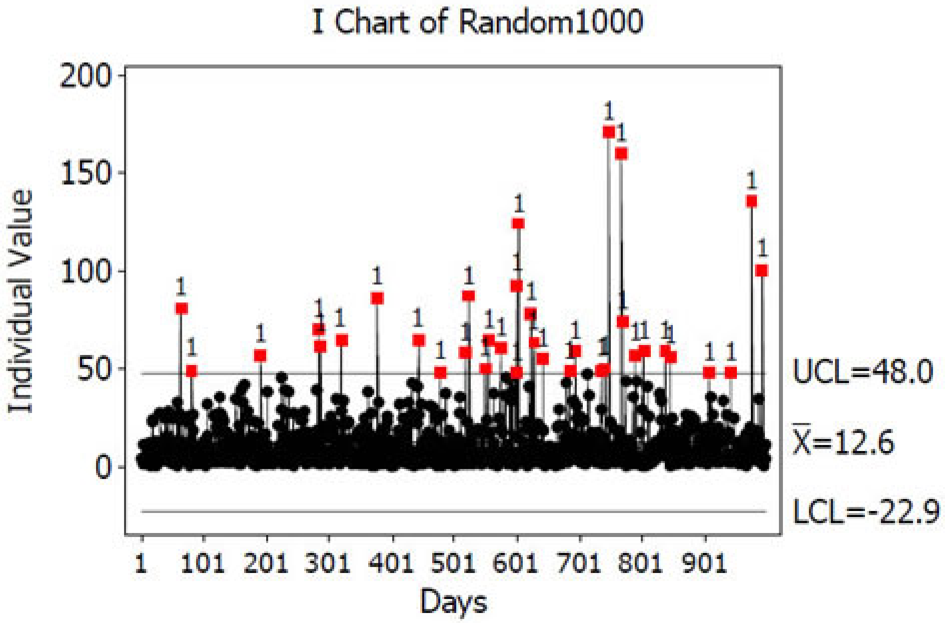


Figure 4: Individuals Control Chart of the Random Sample Data

The individuals control chart in figure 4 shows many out-of-control points beyond the upper control limit. In addition, the individuals control chart shows a physical lower boundary of zero for the data, which is well within the lower control limit of 22.9. If no transformation is needed when plotting non-normal data in a control chart, then we would expect to see a random scatter pattern within the control limits, which is not prevalent in the individuals control chart.

Figure 5 shows a control chart using a Box-Cox transformation with a lambda value of zero, the appropriate transformation for log-normally distributed data. This control chart is much better behaved than the control chart in figure 4. Almost all 1,000 points in this individuals control chart are in statistical control. The number of false alarms is consistent with the design and definition of the individuals control chart control limits.

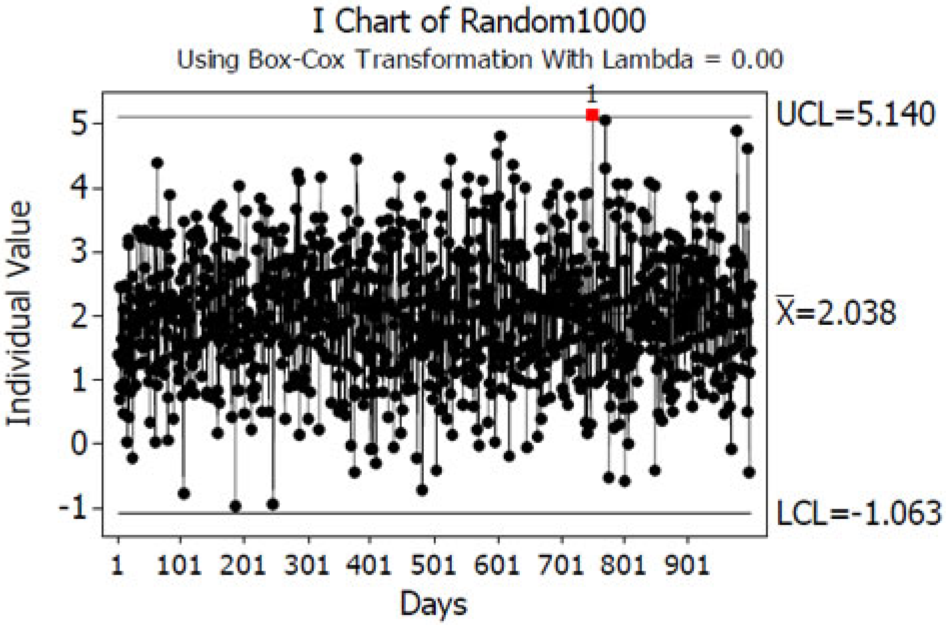


Figure 5: Individuals Control Chart With a Box-Cox Transformation Lambda Value of Zero

Determining actions to take

Previously three decision-making action options were described, where the first option was:
1. Is the process unstable or did something out of the ordinary occur, which requires action or no action?

For organizations that did not consider transforming data to address this question, as illustrated in figure 4, many investigations would need to be made where common-cause variability was being reacted to as though it were special cause. This can lead to much organizational firefighting and frustration, especially when considered on a plantwide or corporate basis with other control chart metrics. If data are not from a normal distribution, an individuals control chart can generate false signals, leading to unnecessary tampering with the process.

For organizations that did consider transforming data to address this question, as illustrated in figure 5, there is no over reaction to common-cause variability as though it were special cause.

For the transformed data analysis, let’s next address the other questions:
2. Is the process stable and meeting internal and external customer needs? If so, no action is

required.
3. Is the process stable but does not meet internal and external customer needs? If so, process improvement efforts are needed.

When a process has a recent region of stability, we can make a statement not only about how the process has performed in the stable region but also about the future, assuming nothing will change in the future either positively or negatively relative to the process inputs or the process itself. However, to do this, we need to have a distribution that adequately fits the data from which this estimate is to be made.

For the previous specification limit of 0.100 in., figure 6 shows a good distribution fit and best-estimate process capability/performance nonconformance estimate of 0.5 percent (100.0 - 99.5). For this situation, we would respond positively to item number two since the percent nonconformance is below 1 percent; i.e., we determined that the process is stable and meeting internal and external customer needs of a less than 1-percent nonconformance rate; hence, no action is required.

However, from figure 6 we also note that the nonconformance rate we expect to increase to about 6.3 percent (100−93.7) with the new specification limit of 0.35 in. Because of this, we would now respond positively to item number three, since the nonconformance percentage is above the 1-percent criterion. That is, we determined that the process is stable but does not meet internal and external customer needs; hence, process improvement efforts are needed. This metric improvement need would be “pulling” for the creation of an improvement project.

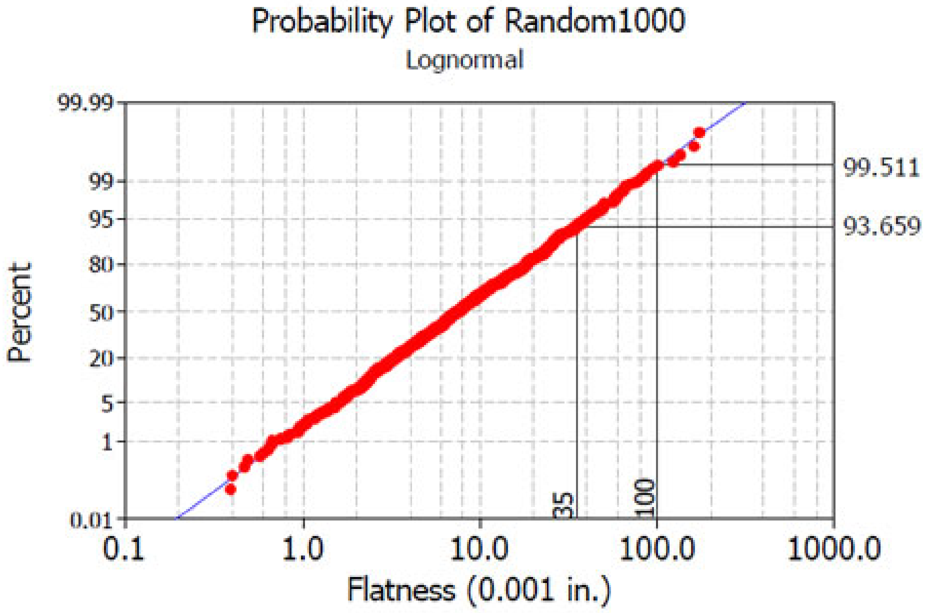
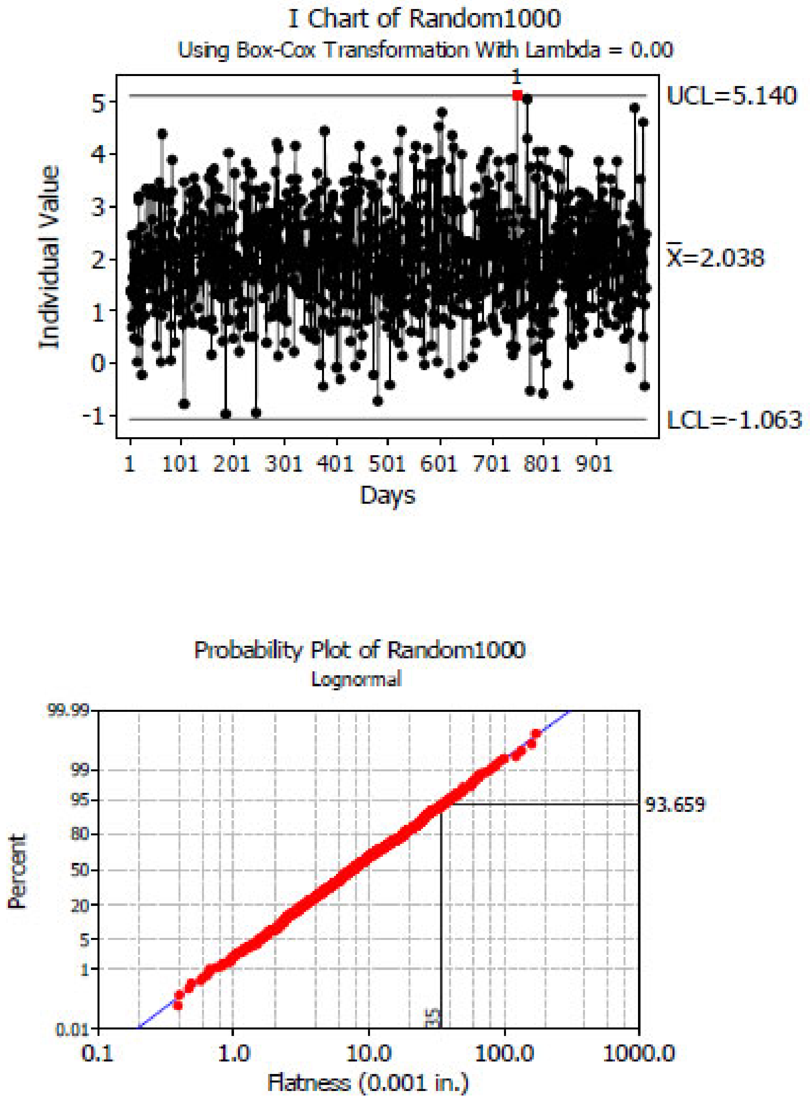


Figure 6: Log-Normal Plot of Data and Nonconformance Rate Determination for Specifications of 0.100 in. and 0.35 in.



Predictable Process with an estimated 6.3% Non-conformance Rate

Figure 7: Predictability Assessment Relative to a Specification of 0.35 in.

It is important to present the results from this analysis in a format that is easy to understand, such as described in figure 7. With this approach, we demonstrate process predictability using a control chart in the left corner of the report-out and then use, when appropriate, a probability plot to describe graphically the variability of the continuous-response process with its demonstrated predictability statement. With this form of reporting, I suggest including a box at the bottom of the plots that nets out how the process is performing; e.g., with the new specification requirement of 0.035, our process is predictable with an approximate nonconformance rate of 6.3 percent.

A lean Six Sigma improvement project that follows a project define-measure-analyze-improve-control execution roadmap could be used to determine what should be done differently in the process so that the new customer requirements are met. Within this project it might be determined in the analyze phase that there is a statistically significant difference in production machines that now needs to be addressed because of the tightened 0.035 tolerance. This statistical difference between machines was probably also prevalent before the new specification requirement; however, this difference was not of practical importance since the customer requirement of 0.100 was being met at the specified customer frequency level of a less than 1-percent nonconformance rate.

Upon satisfactory completion of an improvement project, the 30,000-foot-level control chart would need to shift to a new level of stability that had a process capability/performance metric that is satisfactory relative to a customer 1 percent maximum nonconformance criterion.

Generalized statistical assessment

The specific distribution used in the prior example, log normal (2.0, 1.0, 0), has an average run length (ARL) for false rule-one errors of 28 points. The single sample used showed 33 out-of-control points, close to the estimated value of 28. If we consider a less skewed log-normal distribution, log normal (4, 0.25, 0), the ARL for false rule-one errors drops to 101. Note that a normal distribution will have a false rule-one error ARL of around 250.

The log-normal (4, 0.25, 0) distribution passes a normality test over half the time with samples of 50 points. In one simulation, a majority, 75 percent, of the false rule-one errors occurred on the samples that tested as non-normal. This result reinforces the conclusion that normality or a near-normal distribution is required for a reasonable use of an individuals chart or a significantly higher false rule-one error rate will occur.

Conclusions

The output of a process is a function of its steps and inputs variables. Doesn’t it seem logical to expect some level of natural variability from input variables and the execution of process steps? If we agree to this presumption, shouldn’t we expect a large percentage of process output variability to have a natural state of fluctuation, that is, to be stable?

To me this statement is true for many transactional and manufacturing processes, with the exception of things like naturally auto-correlated data situations such as the stock market. However, with traditional control charting methods, it is often concluded that the process is not stable even when logic tells us that we should expect stability.

Why is there this disconnection between our belief and what traditional control charts tell us? The reason is that often underlying control-chart-creation assumptions are not valid in the real world. Figures 4 and 5 illustrate one of these points where an appropriate data transformation is not made.

The reason for tracking a process can be expressed as determining which actions or nonactions are most appropriate.
1. Is the process unstable or did something out of the ordinary occur, which requires action or no action?
2. Is the process stable and meeting internal and external customer needs? If so, no action is required.
3. Is the process stable but does not meet internal and external customer needs? If so, process improvement efforts are needed.

This article described why appropriate transformations from a physical point of view need to be a part of this decision-making process.

The box at the bottom of figure 7 describes the state of the examined process in terms that everyone can understand; i.e., the process is predictable with an estimate 6.7-percent nonconformance rate.

An organization gains much when this form of scorecard-value-chain reporting is used throughout its enterprise and is part of its decision-making process and improvement project selection.

ABOUT THE AUTHOR

CEO and president of Smarter Solutions Inc., Forrest W. Breyfogle III is the creator of the integrated enterprise excellence (IEE) management system, which takes lean Six Sigma and the balanced scorecard to the next level. A professional engineer, he's an ASQ fellow who serves on the board of advisors for the University of Texas Center for Performing Excellence. He received the 2004 Crosby Medal for his book, Implementing Six Sigma.

E-mail him at forrest@smartersolutions.com