The Analysis of Observational Data

Shewhart's economic operation is the starting point for lean production

Donald J. Wheeler

Most of the World's data are obtained as by-products of operations These observational data track what happens over time and have a structure that requires a different approach to analysis than that used for experimental data. An understanding of this approach will reveal how Shewhart's generic, three-sigma limits are sufficient to define economic operation for all types of observational data.

Management requires prediction, yet all data are historical. To use historical data to make predictions we will have to use some sort of extrapolation. We might extrapolate from the product we have measured to the product not measured, or we might even extrapolate from the product measured to the product not yet made. Either way, the problem of prediction requires that we know when these extrapolations are reasonable and when they are not.

THE STRUCTURE OF OBSERVATIONAL DATA

Before we talk about prediction we need to consider the structure of observational data. For any one product characteristic we can usually list dozens, or even hundreds, of cause-and-effect relationships which affect that characteristic. Some of these causes will have larger effects than the others, so if we had perfect knowledge, we could arrange the causes in order according to their effects to obtain a Pareto like Figure 1.

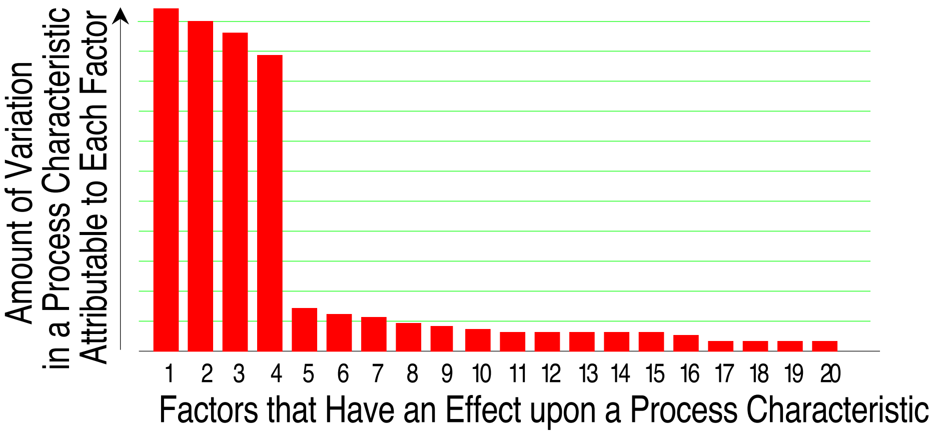


Figure 1: A Typical Cause-and-Effect Pareto

This gives us a model for the structure of the variation for a given product characteristic: A critical few causes have dominant effects, and the many other causes have trivial effects. In production it is never going to be economical to control all of the causes, so we seek to identify and control the critical few

ECONOMIC OPERATION

Given the model in Figure 1, economic production would require that we control the first four factors. Attempting to control more factors would involve diminishing returns, and controlling fewer would result in excess variation in production.

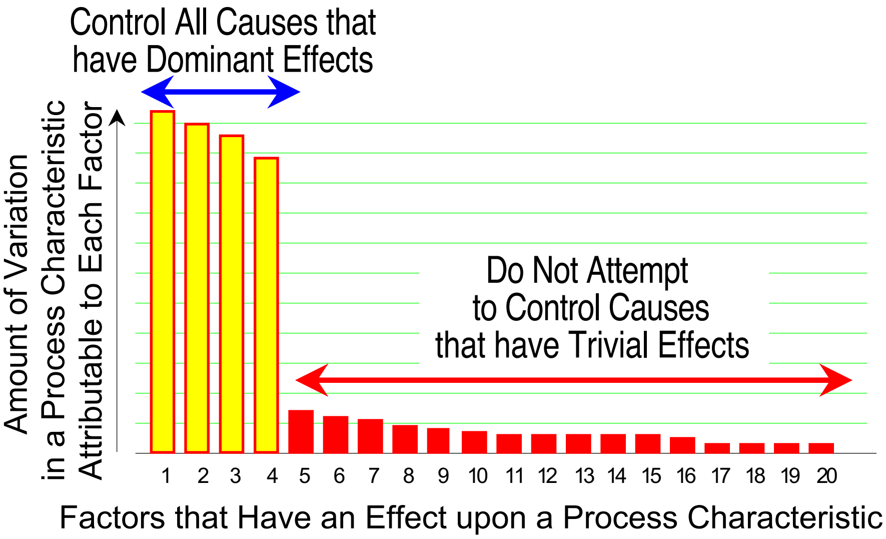
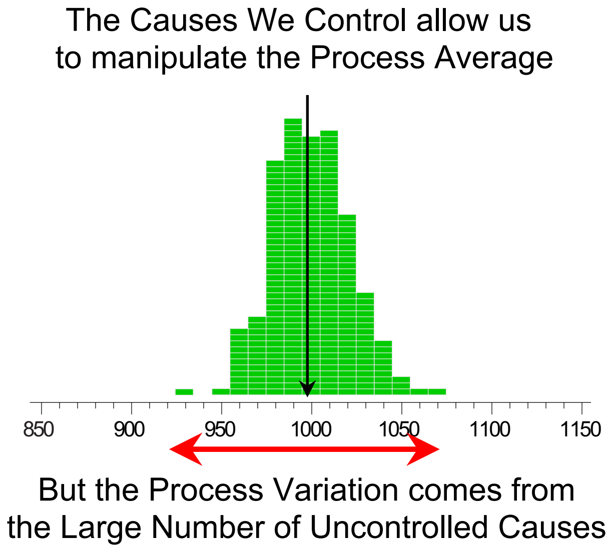


Figure 2: Economic Operation Requires a Balance

Figure 2 defines economic operation where all of the causes with large effects are controlled and the remaining causes with trivial effects are allowed to vary. The levels we choose for each of the controlled factors will determine the average value for the process outcomes. By holding these factors constant at the chosen levels we will also eliminate them as a source of variation.

As a result, virtually all of the variation in the product stream will come from the remaining, uncontrolled factors (causes 5 through 20 in Figure 2). Here the process variation will be the result of a large number of cause-and-effect relationships where no one cause has an effect that dominates the effects of the other causes. If these causes are independent of each other then their variation will combine in an additive manner, and the central limit theorem assures us that the histogram will be approximately normal.


The Causes We Control allow us to manipulate the Process Average

But the Process Variation comes from the Large Number of Uncontrolled Causes

Figure 3: Economic Operation Results in Approximate Normality

We have known about the central limit theorem for over two hundred years. Last month's column, "How Can the Sum of Skewed Variables Be Normally Distributed?" illustrated this fact of life. When we are operating economically the effects of the uncontrolled causes will sum up to produce a bell-shaped histogram.

But the effects of the trivial many do not have to combine as a sum. The normal distribution is the distribution of maximum entropy, and any operation that increases entropy will eventually have a histogram that approaches a normal. So regardless of how the effects of the trivial many combine, economic operation will result in a histogram that is approximately normal.

In his 1931 book "Economic Control of Manufactured Product" Shewhart described a "state of control" (a predictable process) as being one where the variation can be described as the result of a large number of uncontrolled causes where no one cause has a predominant effect (as in Figure 2).

Shewhart also defined a "state of maximum control" as one where the histogram of Figure 3 is approximately normal. Thus, Figures 2 and 3 provide useful models for what Shewhart defined as economic operation.

With this conceptual model of what economic operation should look like we are ready to begin to analyze observational data. We know what we are looking for, and we can begin to compare what we find with the conceptual model. However, we need to first consider what happens when a process is not operated economically.

WHAT USUALLY HAPPENS

We never have the perfect knowledge assumed by Figures 1, 2, and 3. While we do our best to identify and control all of the causes with dominant effects, it is difficult to find them all when the number of known causes may be in the hundreds. As a result, we will usually unknowingly end up with one or more dominant causes remaining in the uncontrolled set.

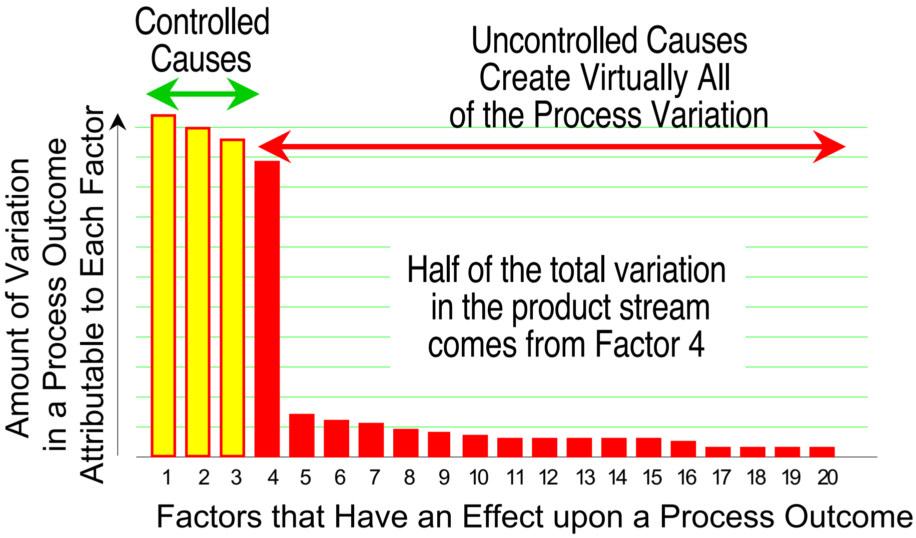


Figure 4: Cause-and effect Pareto for an Unpredictable Process

When the set of uncontrolled causes contains one or more causes with dominant effects everything changes. First, the set of controlled causes will only partially determine the process average. This happens because the large effect of Factor 4 means that it can take the whole process histogram on walkabout. As Factor 4 varies, you can count on having unexpected changes in the process average from time to time.

Second, as Factor 4 takes the process on walkabout, the operators may start adjusting Factors 1, 2, and 3 to keep the process on-target. This will create further variation.

Third, as Factor 4 and the operators take the process on walkabout, the variation from all the other uncontrolled causes will be taken along, making the histogram even fatter. The result for one such process is shown in Figure 5.

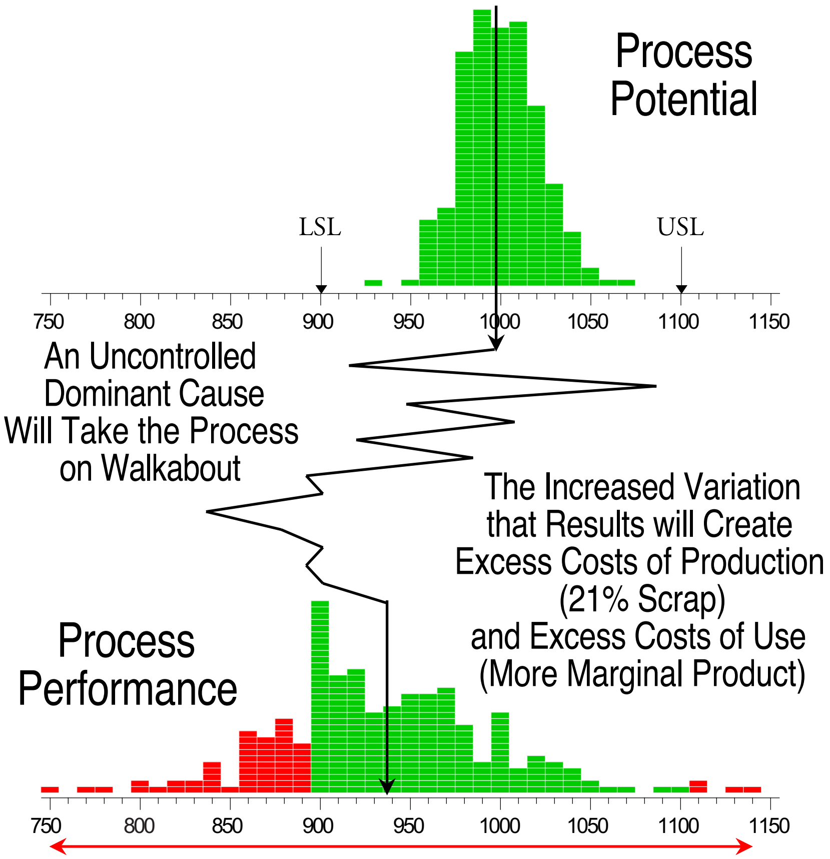


Figure 5: The Effect of an Uncontrolled Dominant Cause

Thus, the impact of having a dominant cause in the set of uncontrolled factors will inevitably be increased variation in the product stream, and this increased variation can result in substantial excess costs of production and use.

We can only reverse the situation shown in Figure 5 by identifying the assignable causes that have dominant effects (Factor 4 in Figure 4) and making them part of the set of controlled inputs for the process. Thus, the problem of analysis for observational data is the problem of detecting the dominant cause-and-effect relationships that remain in the set of uncontrolled factors. And this is precisely what Shewhart's process behavior charts allow us to do.

SHEWHART'S APPROACH

A process behavior chart makes no assumptions about your process. It makes no assumptions about the data that characterize the product stream. It simply compares the variation found in your observational data with the generic limits we expect when your process is operated in an economic equilibrium.

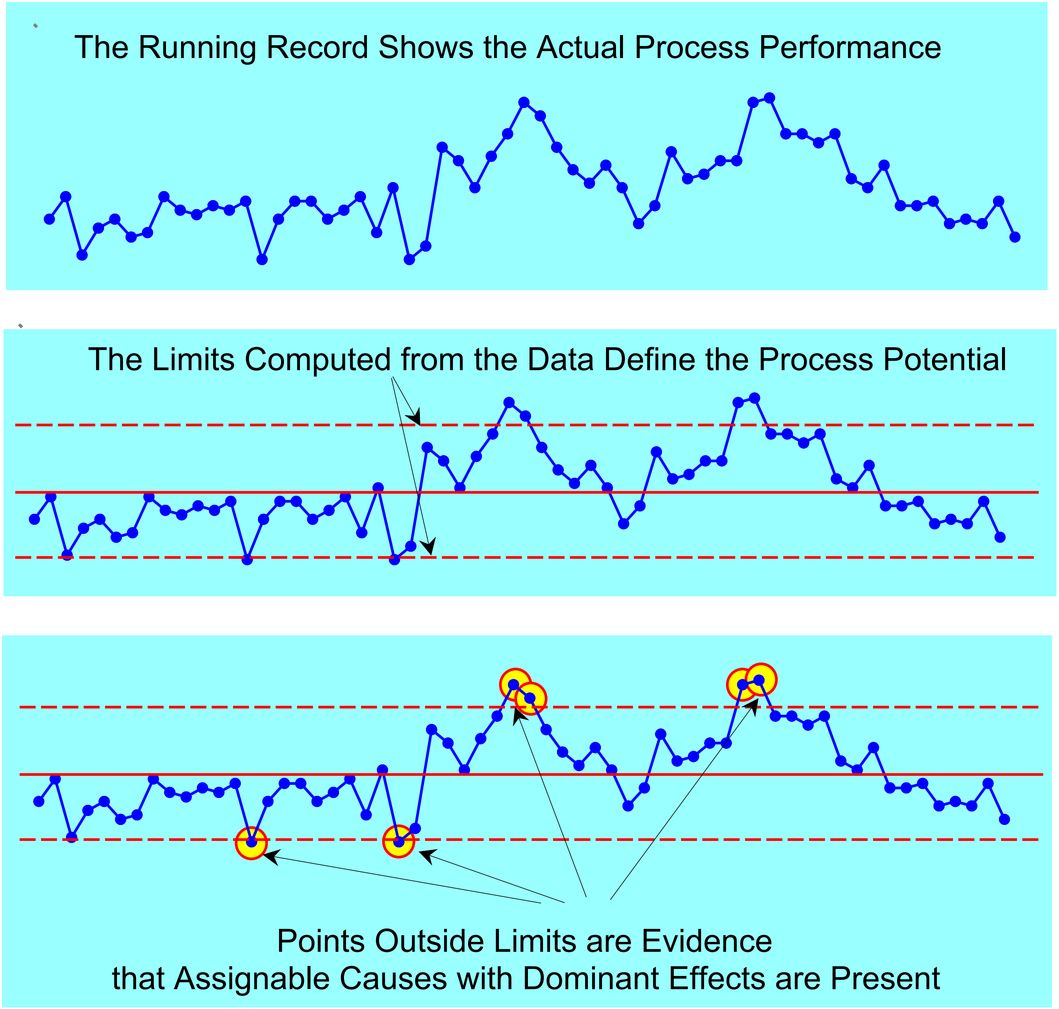


Figure 6: We Characterize a Process by Comparing Performance with Potential

To get limits that approximate the process potential we use an appropriate *within-subgroup* measure of dispersion to capture the routine variation inherent in the process. (In the case of individual values this will be either the average or the median of the set of two-point moving ranges.)

We then use this within-subgroup dispersion to compute generic, fixed-width, three-sigma limits centered on the average. Both theory and one hundred years of practice have shown that these limits approximate what the process can do when it is operated in a state of economic equilibrium.

If (as in Figure 2) the process variation is the result of a large number of causes where no one cause has a predominant effect, then the process performance should fall within the generic limits defining the process potential.

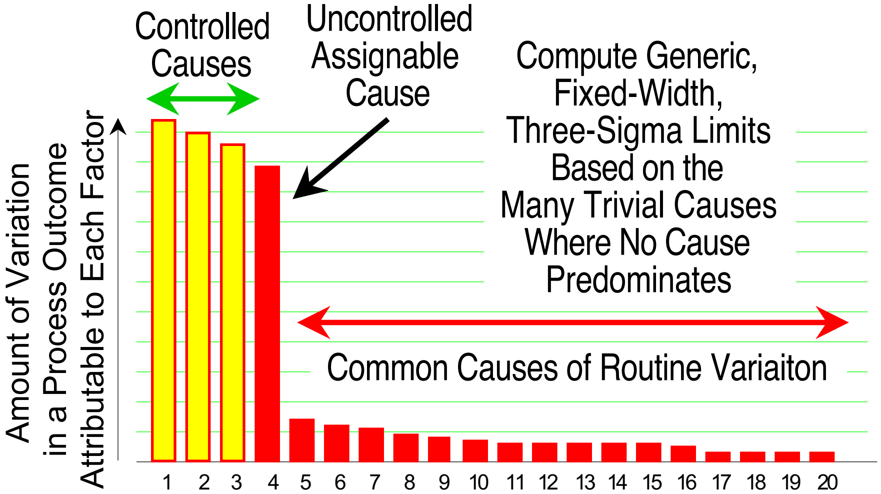


Figure 7: How Process Potential Works

However, as in Figure 7, if one or more assignable causes with dominant effects are present, the running record is likely to go outside the limits. When this happens we have strong evidence that an assignable cause is present.

PREDICTABLE OPERATION

When a process is operated predictably it is being operated up to its full potential. It will have the minimum variance that is consistent with economic operation. This is why seeking to control a common cause of routine variation is a low payback strategy.

When a process has been operated predictably in the past, then it is logical to expect that it will continue to be operated predictably in the future, and the past behavior forms a reasonable basis for predictions of what will be.

UNPREDICTABLE OPERATION

If the process shows evidence of assignable causes it will behave unpredictably. Such a process will be operating at less than its full potential, and it will almost always be economical to make the assignable causes part of the set of controlled factors.

This is a high payback strategy for two reasons. When we make an assignable cause part of the group of control factors we not only gain an additional process input to use in adjusting the process average, but we also remove a large chunk of variation from the product stream. In this way, even though the process may have been unpredictable in the past, we learn how to improve the process and come closer to operating it predictably and on-target in the future.

An unpredictable process is not going to spontaneously begin to be operated predictably in the future. Assignable causes will continue to take our process on walkabout, and no computation will let us predict what our unpredictable process will actually produce.

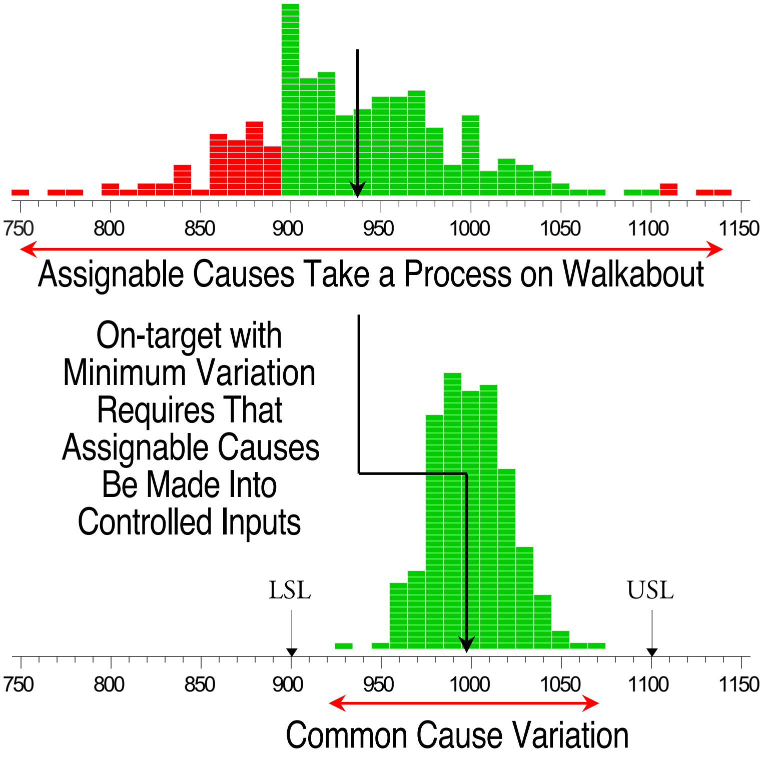


Figure 8: Assignable Causes Belong in the Set of Controlled Factors

SUMMARY

When analyzing observational data we have to focus on how the data vary. However, this focus has nothing to do with the shape of the histogram. We examine how the data vary over time in order to characterize the data as either predictable or unpredictable. These two broad categories suffice to tell us how to use the data. With a predictable process, prediction is feasible. With an unpredictable process, prediction is futile, but action may be taken to move the process closer to its full potential.

The generic, three-sigma limits of a process behavior chart define this broad model of predictable and economic behavior. When assignable causes are present they will disrupt the process and take the data outside these limits.

Because assignable causes can affect the histogram in various ways, it is always a mistake to start your analysis of observational data by looking at the histogram. Until you know if your process is being operated predictably or unpredictably you will not know how to interpret the histogram.

So, the first question in the analysis of observational data is the question of predictability. When the process is operated predictably, the data will be homogeneous, and the histogram will tend to be normally distributed due to the high-entropy condition known as routine variation.

When the process is operated unpredictably the data will not be homogeneous, and the histogram will be a smash-up of different conditions, resulting in all kinds of different shapes. So, with observational data, an irregular or skewed histogram is more likely to be the result of some assignable cause taking the process on walkabout than anything else.

This is why you do not need to "pre-qualify" your data. You do not need to fit a probability model to your data. Neither should you place your data on a normal probability plot. And you certainly do not need to transform your data to make them "more normal." All of these pre-qualification activities *assume* the process is already being operated predictably. Assignable causes completely undermine this assumption, making these activities nonsense.

So regardless of what your histogram may look like, put your data on a suitable process behavior chart, and characterize your process behavior as predictable or unpredictable. Use the data to make predictions for your predictable processes, and use the chart to look for the assignable causes that are taking your unpredictable processes on walkabout.

Learn how to operate your processes economically and you will improve quality, productivity, and competitive position. In their landmark book, "The Machine That Changed the World," Womack, Jones, and Roos explicitly, and repeatedly, state that lean production is predicated upon having predictable processes. Without the foundation of predictability all you have is wishful thinking.