# How Can a Control Chart Work Without a Distribution?

## In spite of what everyone says to the contrary

Donald J. Wheeler

Students are told that they need to check their data for normality before doing virtually any data analysis. And today's software encourages this by automatically providing normal probability plots and lack-of-fit statistics as part of the output. So it is not surprising that many think that this is the first step in data analysis.

This practice of "checking for normality" has become so widespread that I have even found it listed as a prerequisite for using a distribution-free nonparametric technique! Yet there is little consensus about what to do if your data are found to "not be normally distributed." If you switch to some other analysis, you are likely to find it too is hidden behind the "check for normality" obstacle. So you are left needing to customize your analysis by fitting some probability model to your data before you can proceed. And this opens the door to all kinds of complexities.

The histogram in Figure 1 represents 20 days of production for one product. It is clearly not going to pass any test for normality. But is there some other probability model that might be used?

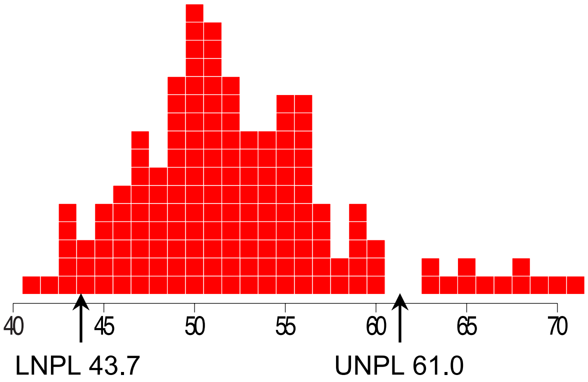

**Figure 1: Production Data for June**

If we attempt to fit some skewed probability model to these data we typically begin by estimating the mean and standard deviation. These 160 data have an average of 52.33, and an estimated within-subgroup standard deviation of 2.88. This results in natural process limits of 43.69 and 60.97. When we apply these limits to this histogram we find 19 of the 160 values outside this interval, which is 12% of the data.

So to fit this histogram we are looking for a skewed probability model having a mean of 52.33, a standard deviation of 2.88, with 12% outside the three-sigma limits. And this is where we come up against a mathematical fact of life. *No such probability model exists!*

No mound-shaped, unimodal probability model, regardless of how skewed that model might be, can ever have more than 2% outside the interval defined by the mean plus or minus three standard deviations. This limitation is imposed by the laws of rotational inertia and cannot be violated.

## SO WHAT CAN WE DO?

Rather than following the white rabbit down the hole of checking for normality, we need to think about a fundamental assumption that is an actual prerequisite for the use of most statistical techniques. This is the assumption that the data are logically homogeneous. This means that the data need to be "of a similar kind or nature, having no discordant elements, and of a uniform structure throughout." In practice this means that there are *no unknown changes* in the conditions under which the data are collected. And the primary technique for examining this assumption of homogeneity is the process behavior chart (also known as a control chart).

But how can a process behavior chart work without reference to a probability model? In answer to this question Walter Shewhart wrote the following on page 35 of "***Statistical Method from the Viewpoint of Quality Control***."

> "We next come to the requirement that the criterion [a process behavior chart] shall be as simple as possible and adaptable to a continuing and self-correcting operation. Experience shows that the process of detecting and eliminating assignable causes of variability so as to attain a state of statistical control is a long one. From time to time the chart limits must be revised as assignable causes are found and eliminated.
>
> "A simple procedure is used for establishing the limits without the use of probability tables because it does not seem that much is to be gained during the process of weeding out assignable causes by trying to set up exact probability limits upon the basis of assumptions that we know from experience do not hold until the state of statistical control has been reached. This is particularly true since such probabilities do not indicate the probability of detecting assignable causes but simply the probability of looking for such causes when they do not exist, which is of secondary importance until a state of statistical control has been reached. Then too, as already indicated, the design of an efficient criterion for the important job of indicating the presence of assignable causes

> depends more upon [rational sampling and rational subgrouping] than it does upon the use of any exact mathematical distribution."

As Shewhart observes, there are two mistakes we can make when we analyze data, but fortunately we cannot make them both at the same time. When our data contain signals we are open to the mistake of missing these signals. It is only when our data *do not contain signals* that we are open to the mistake of getting false alarms. Shewhart argues here that as long as your process is being operated unpredictably it is changing, and the only mistake you need concern yourself with is the mistake of missing a signal of these changes. You need not be concerned with the probability of false alarms until you have finally learned how to operate your process predictably.

But when we impose a probability model upon our data, and use that model to compute probability limits, we are only computing the probability of a false alarm. This is why checking your data for normality before placing them on a process behavior chart is to get things backward.

## PROCESS BEHAVIOR CHARTS

One of the few statistical techniques that is not built on any distributional assumption is the process behavior chart. So, while no probability model exists that will fit the mean, variance, and percentage outside three-sigma limits for Figure 1, we can still place the data from Figure 1 on a process behavior chart. When we do this we find ample evidence of a lack of homogeneity within these data.

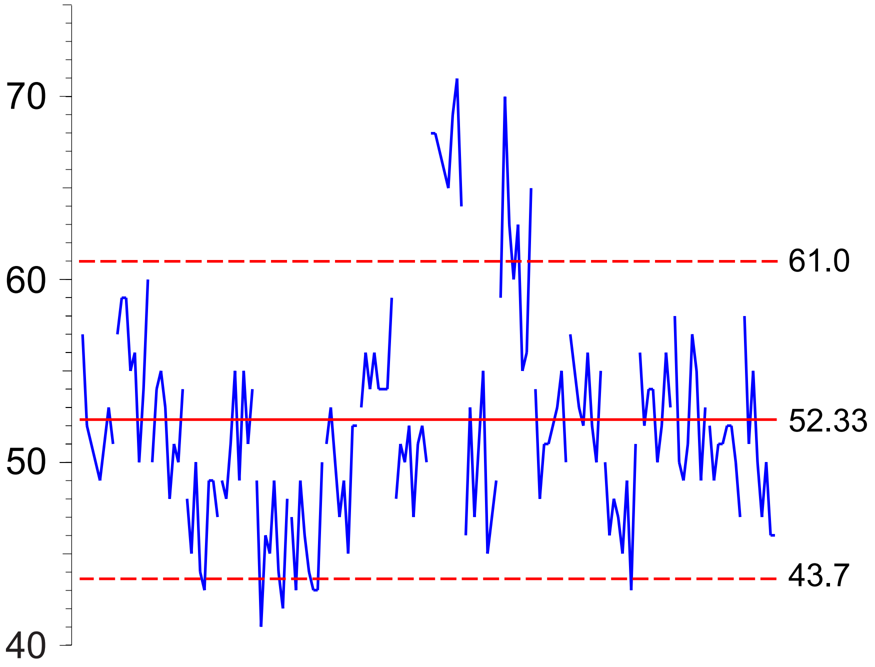

**Figure 2 *X*-chart for June Production Data**

So here is an example where the process behavior chart works *even when no probability model exists*.

The process behavior chart allows us to detect signals of process changes when they occur and as they occur. We need not be concerned with the probability of false alarms in Figure 2. Rather we need to identify the assignable causes of the unpredictable behavior from day to day.

The "skewness" of the histogram in Figure 1 is the result of the process going on walkabout. This process has different personalities on different days. Attempting to fit a probability model to such data is always going to be an exercise in futility.

The secret foundation of most statistical techniques is "Assume the data are homogeneous." But the secret of data analysis is "Your data are rarely homogeneous." And the premier technique for examining your data for homogeneity is the process behavior chart.

## SUMMARY

Shewhart clearly contradicts the claim that we have to fit a probability model to the data prior to computing the limits for a process behavior chart. If you and Shewhart see things differently, who do you think is right? Continuing in this vein, Shewhart concluded the epilogue of his 1939 book with the following:

> "Throughout this monograph care has been taken to keep in the foreground the distinction between the distribution theory of formal mathematical statistics and the use of such theory in statistical techniques designed to serve some practical end. Distribution theory rests upon a framework of mathematics, whereas the validity of statistical techniques can only be determined empirically. … The technique involved in the operation of statistical control [i.e. a process behavior chart] has been thoroughly tested and not found wanting, whereas the formal mathematical theory of distribution[s] constitutes a generating plant for new techniques to be tried."

Mathematical theory can only approximate what happens in practice. In order for any data analysis technique to be useful it will, of necessity, have to be robust to the assumptions of mathematical theory, otherwise it would not work in practice. When we turn the assumptions of mathematical theory into requirements that must be satisfied before we use some data analysis technique we do nothing but add unnecessary complexity to our analysis.

Another aspect of the absurdity of turning mathematical assumptions into preconditions for practice lies in the fact that the techniques for checking on the preconditions will generally be much less robust than the analysis techniques being qualified. As Francis Anscombe, Fellow of the American Statistical Association, said: "Checking your 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."