Don't the Outliers Distort the Limits?

When can we trust the limits on a process behavior chart?

Donald J. Wheeler

Last month we showed the *X* Chart in Figure 1. The four lowest values and the three highest values were seen to be "outliers" when we looked at the histogram. When we fitted a bell-shaped curve to the histogram the outliers corrupted the model and resulted in a poor fit. Yet we used all of the data to compute the limits seen in Figure 1. How can the outliers corrupt one computation but not corrupt another?

The answer to this question lies in how we compute limits for the *X* Chart. The central line is commonly taken to be the Average value. Now while it is true that the Average may be influenced by extreme values, this effect is generally smaller than you might expect. In this case, deleting the seven "outliers" would only change the Average from 595.4 to 595.6. The Average value is a very robust measure of location. However, in those cases where we think the Average may have been unduly influenced by extreme values, we may always resort to using the Median value instead. In this case the Median is 596. Thus, one way or another, we are going to have a reasonable estimate of location regardless of the outliers.

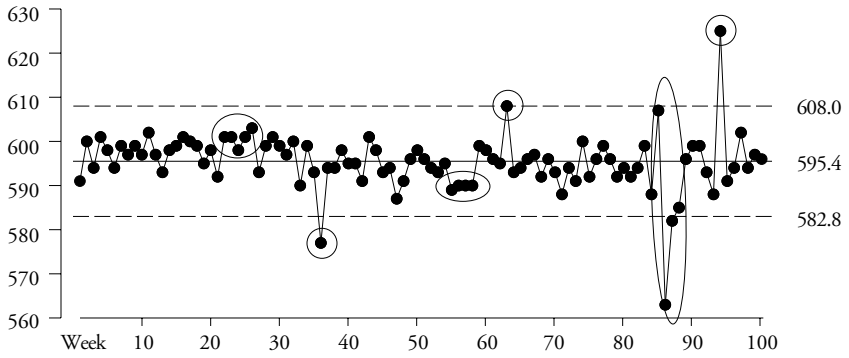

Figure 1. *X* Chart for NB10 Values

So what about dispersion? When working with process behavior charts we always use within-subgroup measures of dispersion. With an *X* Chart these within-subgroup measures are the two-point moving ranges (also known as the differences between successive values).

Once these values have been found they are usually summarized by the Average Moving Range. Since the operation of averaging provides a robust summary the Average Moving Range turns out to be a robust summary for dispersion. When the Average Moving Range is divided by the bias correction factor of 1.128 for the pseudo-subgroups of size two, it becomes a measure of dispersion known as *Sigma(X)*.

Whenever we think that the Average Moving Range may have been inflated by some very large moving ranges we can always shift over to use the Median Moving Range instead. When this summary is divided by its bias correction factor of 0.954 we again get a measure of dispersion known as *Sigma(X)*.

The limits on the *X* Chart are computed according to *Average* ± *3 Sigma(X)*. For the data of Figure 1 the Median Moving Range is 4, so *Sigma(X)* is 4.2, and the limits are 12.6 units on either side of the Average.

It is instructive to compare the value of *Sigma(X)* above with the standard deviation statistics computed in the usual way. For all 100 data the global standard deviation statistic is 6.47 units. Using only the central 93 values the global standard deviation statistic is 3.74 units. The *Sigma(X)* value of 4.2 compares favorably with the value *computed after the outliers had been removed.* This robustness which is built into the computations for the process behavior charts removes the need to polish the data prior to computing the limits. The computations work even in the presence of outliers and signals of exceptional variation. They allow you to get good limits from bad data.

Thus, the computations are robust. The key element in using an *X* Chart is to make sure that successive values are logically comparable. This requires some knowledge of the context for your data. But since the only reason to collect data is to take action, you should know the context well enough to make this judgment call.