All Outliers Are Evidence!

Removing the extreme values is always a serious mistake

Donald J. Wheeler

Many have been taught that they must remove outliers prior to analysis. This is because much of modern statistics is concerned with creating a mathematical model for the data. Because all these models are created using algorithms, they tend to be severely affected by any unusual or extreme values. Therefore, to use these mathematical techniques to obtain useful and appropriate models, it is often necessary to polish up the data by removing the outliers. However, the act of building a model implicitly assumes that the data are homogeneous enough to justify the use of a model.

For example, the histogram in Figure 1 has a bell-shaped curve superimposed. This curve is based on the average and standard deviation statistic for all 100 values in the histogram. It is neither wide enough nor tall enough to provide a good fit to the data. The histogram in Figure 2 contains the 93 values left after the seven extreme values (the 4 lowest and 3 highest) were deleted. Now the curve based on the average and the standard deviation statistic does a much better job of fitting the data. Thus, it is true that outliers can undermine our efforts to create a model for our data.

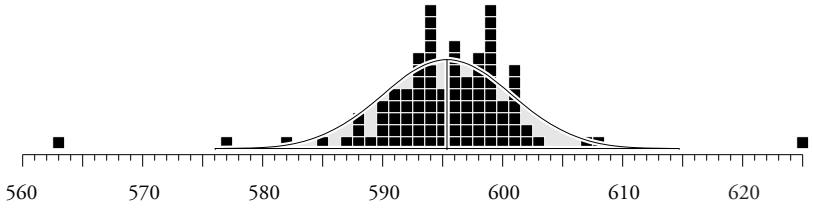

Figure 1. Histogram for 100 NB10 Values

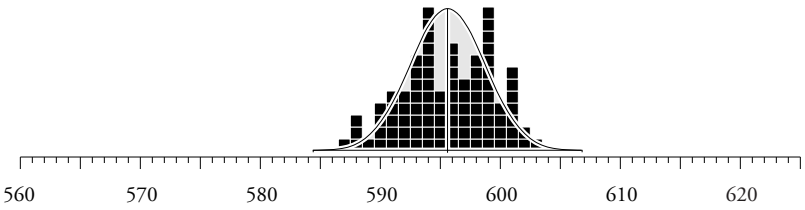

Figure 2. Histogram for 93 NB10 Values

But what about the seven values we simply deleted in order to obtain this better fit between our assumed model and our revised data set? What were these seven values trying to tell us about the underlying process that generated these data?

The whole operation of deleting outliers to obtain a better fit between our model and the data is based upon computations which implicitly assume that the data are homogeneous. When you have outliers, this assumption becomes questionable.

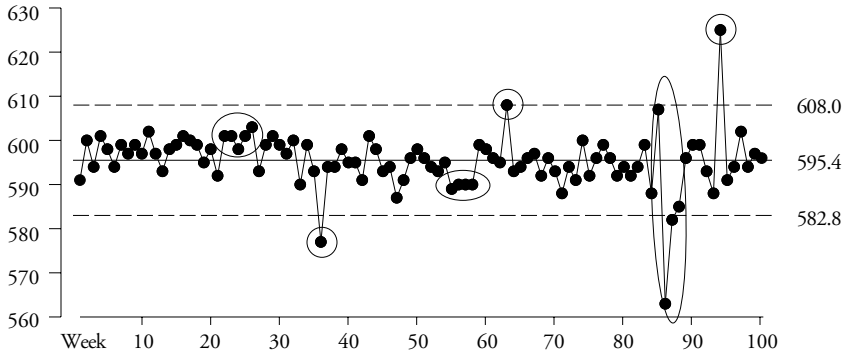

Figure 3. *X* Chart for NB10 Values

The *X* Chart in Figure 3 shows clear evidence of six upsets or changes in the underlying process. The seven "outliers" from Figure 1 are part of these signals. The outliers that we dismissed in Figure 2 are the signals that the values are not homogeneous, and that the models fitted in *both* Figures 1 and 2 are wrong. From the perspective of data analysis, the outliers are the most important values in the data set. We need to understand these values rather than dismissing them. Finding a model is premature. Here there is not one underlying process, but many.

"Are these data homogeneous?" must be the first question of any analysis. Process behavior charts provide the easiest way to address this question. Hence, any analysis that does not begin by organizing the data in some rational manner and placing those data on a process behavior chart is inherently flawed.