Don't We Need to Remove the Outliers?

Characterization and estimation are different.

Donald J. Wheeler

Much of modern statistics is concerned with creating models which contain parameters that need to be estimated. In many cases these estimates can be severely affected by unusual or extreme values in the data. For this reason students are often taught to polish up the data by removing the outliers. Last month we looked at a popular test for outliers. In this column we shall look at the difference between estimating parameters and characterizing process behavior.

ESTIMATION

To illustrate how polishing the data can improve our estimates we will use the data of Figure 1. These values are 100 determinations of the weight of a ten-gram chrome steel standard known as NB10. These values were obtained once each week at the Bureau of Standards, by one of two individuals, using the same instrument each time. The weights were recorded to the nearest microgram. Since each value has the form of 9,999,xxx micrograms, the four nines at the start of each value are not shown in the table—only the last three values in the xxx positions are recorded. The values are in time order by column.

Weeks	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
	591	602	592	597	595	596	596	588	592	599
	600	597	601	600	591	594	595	594	594	593
	594	593	601	590	601	593	608	591	599	588
	601	598	598	599	598	595	593	600	588	625
	598	599	601	593	593	589	594	592	607	591
	594	601	603	577	594	590	596	596	563	594
	599	600	593	594	587	590	597	599	582	602
	597	599	599	594	591	590	592	596	585	594
	599	595	601	598	596	599	596	592	596	597
	597	598	599	595	598	598	593	594	599	596

Figure 1: NB10 Values for Weeks 1 to 100

If we compute the usual descriptive statistics we find that the average of the tabled values is 595.4 micrograms and their standard deviation statistic is 6.47 micrograms. Using these two values to define a normal distribution we would end up with the curve shown superimposed upon the histogram in Figure 2. Both the area under the curve and the area of the histogram are the same. Yet the curve does not really match up with the histogram. It is too heavy in the regions around 585 and 605, and not high enough near 595.

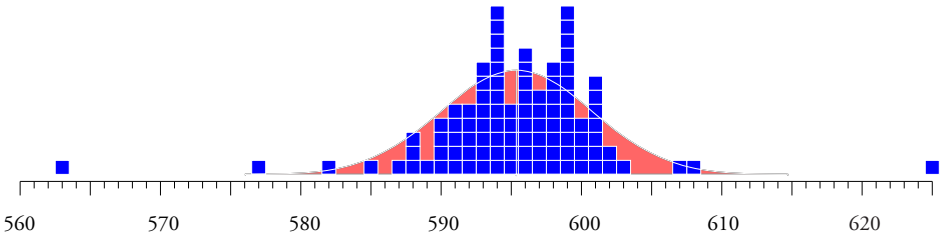

Figure 2: Histogram and Normal Curve for NB10 Values

The outliers in the histogram create the mismatch between the fitted model and the data. Seven values look like outliers in Figure 2. If we delete the four values below 586 and the three values above 606, and recompute our descriptive statistics we find the revised histogram has an average of 595.6 micrograms and a standard deviation statistic of 3.74 micrograms. Using these two values to define a normal distribution we end up with the curve shown in Figure 3. Now we have a much better fit between our model and the histogram.

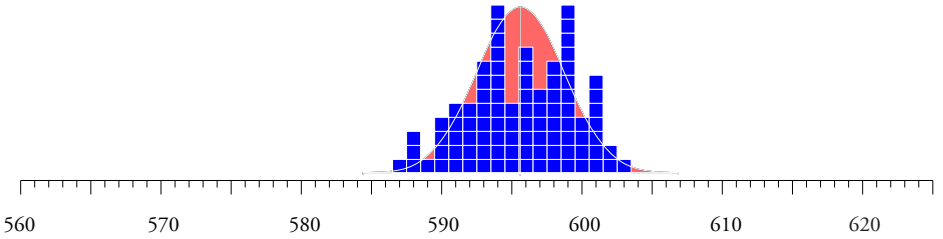

Figure 3: Histogram and Normal Curve for Revised NB10 Values

The whole operation of deleting outliers to obtain a better fit between the model and the data is based upon computations which implicitly assume that the data are homogeneous. However, when you have outliers, this assumption becomes questionable. If the data are homogeneous, where did the outliers come from? Thus, whether the data are homogeneous or not must be the primary question for any type of analysis. While this is the one question we do not address in our statistics classes, it is percisely the question considered by the process behavior chart.

THE CHARACTERIZATION OF PROCESS BEHAVIOR

What about the seven values we simply deleted in order to obtain the better fit between our *assumed* model and our *revised* data set? What were these values trying to tell us about this process? Here the question is not one of estimation, but rather one of using the data to *characterize* the underlying process represented by the data.

Figure 4 contains the *XmR* Chart for the 100 values of Figure 1. The limits are based upon the Median Moving Range of 4.0 micrograms. Here we have clear evidence of at least three upsets or changes in the process of weighing NB10. Five of the seven outliers that we deleted in order to fit the model in Figure 3 are signals that reveal that this set of values is not homogeneous. This lack of homogeneity undermines the model of Figure 3 and makes it inappropriate. If you want to use your data to gain insight into the underlying process that creates the data, then the outliers are

the most important values in the data set! Yet students are routinely taught to delete those pesky outliers. After all, when you are looking for iron and tin, you should not let silver and gold get in the way.

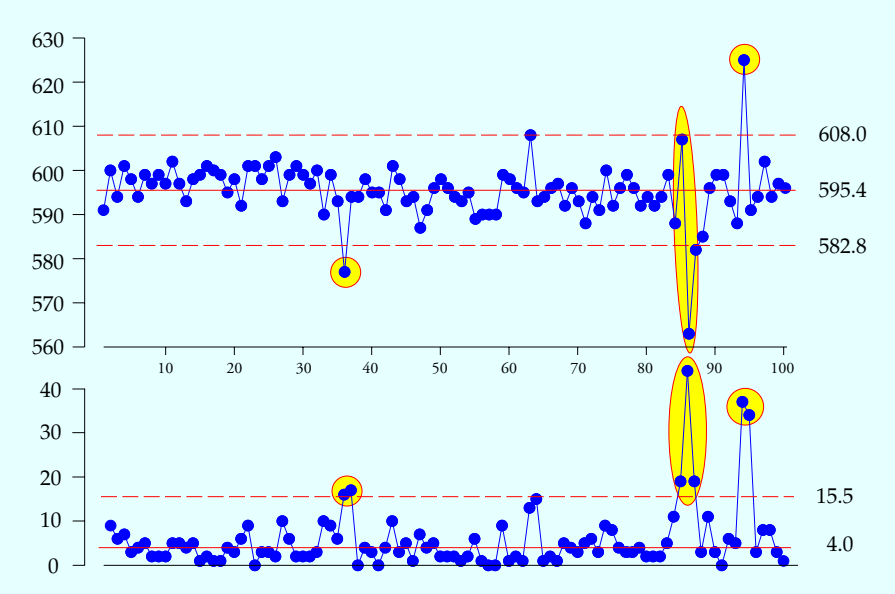

Figure 4: *XmR* Chart for 100 Weighings of NB10

DON'T THE OUTLIERS DISTORT THE LIMITS ?

But don't we need to remove the outliers to compute good estimates of location and dispersion? No, we don't. To see why this is so it is helpful to consider the impact of outliers upon the limits of a process behavior chart.

We commonly base our limits on the average and an average range. The average may be affected by some very extreme values, but this effect is usually much smaller than people think it will be. In Figure 1 some values are out of line with the bulk of the data by as much as 30 micrograms. However, the average value of approximately 595 micrograms was found by dividing 59,500 by 100. If the total of 59,500 is adjusted up or down by 30, 60, or even 90 units, it will have a very small effect upon the average. In this example deleting the outliers changed the average from 595.4 to 595.6. Thus, the average is a very robust measure of location, which is why we use it as our main statistic for location. Of course, whenever we have reason to think that the average may have been affected by the presence of several extreme values that are all on the same side, we can always use the median instead. Hence, while our primary measure of location is robust, we have an alternative for those cases where one is needed.

Likewise, when we compute an average range, we are once again diluting the impact of any extreme values that are present in the set of ranges. In general, a few large ranges will not have an undue impact upon the average range. However, if they do appear to have inflated the average range, we can resort to using the median range. In Figure 4 the limits are based upon the Median Moving Range of 4.0 micrograms. This results in an estimated dispersion for the

individual values of:

$$Sigma(X) = \frac{Median\ Range}{0.954} = 4.2 \text{ micrograms}$$

It is instructive to compare this with the two values for the standard deviation statistic computed from these data. Using all 100 values from Figure 1 we found s = 6.47 micrograms. Using only the 93 values shown in Figure 3 we found s = 3.74 micrograms. Thus, the Median Moving Range (based on all 100 values) gives an estimate for dispersion that is quite similar to the descriptive statistic coumputed 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.

DON'T WE NEED A PREDICTABLE RANGE CHART ?

The fact that the computations work even in the presence of outliers is important in the light of the advice given in some SPC texts. These texts warn the student to check the Range Chart before computing limits for the *X* Chart or the Average Chart. If the Range Chart is found to display evidence of unpredictable behavior, then the student is advised to avoid computing limits for the Average Chart or the *X* Chart. The idea being that signals on the Range Chart will corrupt the average range and hence corrupt the limits on the other chart. This advice is motivated by a desire to avoid using anything less than the best estimates possible. However, the objective of a process behavior chart is not *to estimate*, but rather *to characterize* the process as being either predictable or unpredictable.

Given the conservative nature of three-sigma limits we do not need high precision in our computations. Three sigma limits are so conservative that any uncertainty in where our computed limits fall will not greatly affect the coverage of the limits. To characterize this effect Figure 5 shows the coverages associated with limits ranging from 2.8 sigma to 3.2 sigma on either side of the mean. There we see that regardless of the shape of the distribution, and regardless of the fact that there is uncertainty in our computations, our three-sigma limits are going to filter out virtually all of the routine variation. This is what allows us to compute limits and characterize process behavior without having to first delete the outliers. The computations are robust, and as a consequence, the technique is sensitive.

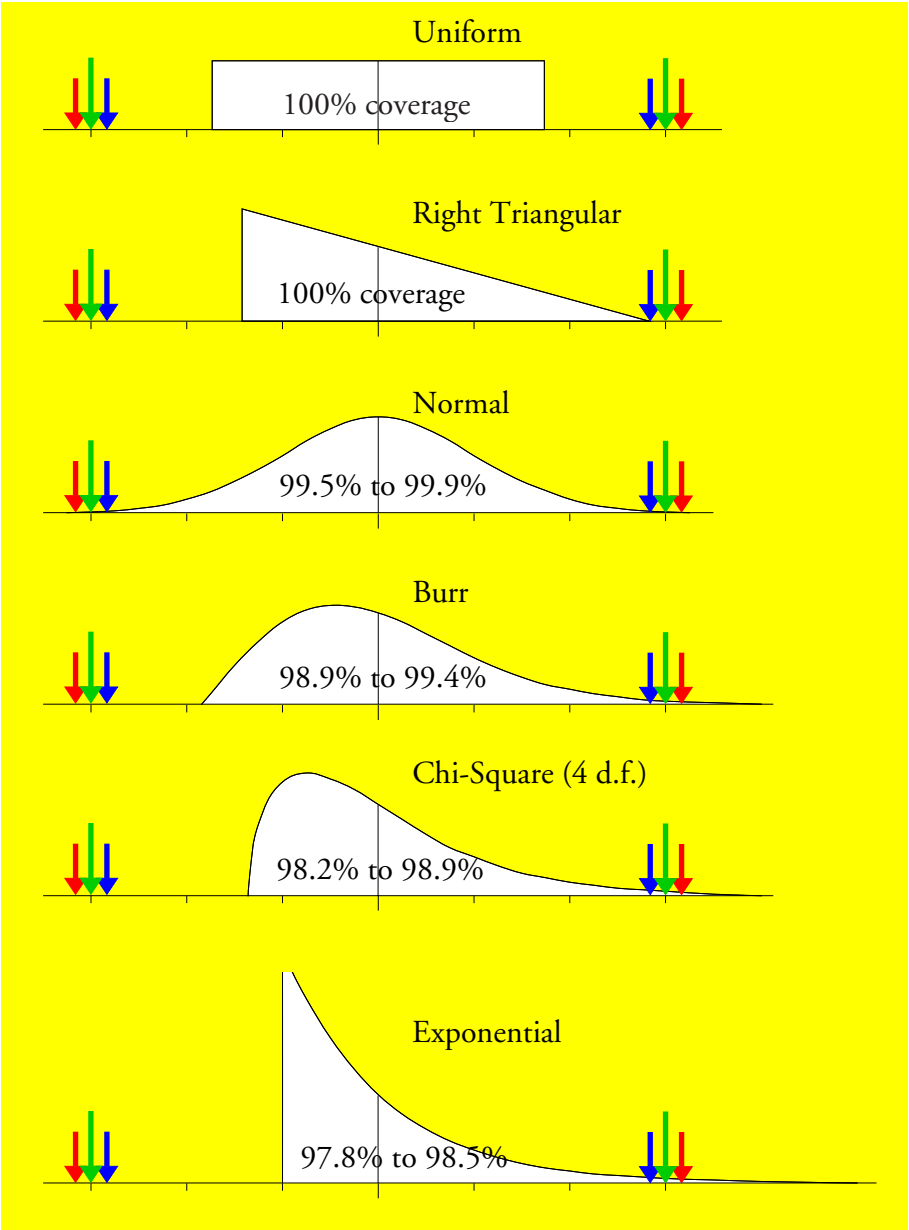

Figure 5: Three-Sigma Limits Filter Out Virtually All of the Routine Variation Regardless of the Shape of the Histogram and Regardless of the Uncertainty in Our Estimates of the Limits

Thus, the advice to make sure the Range Chart is predictable prior to computing limits for the Average Chart or the *X* Chart is just another version of the delete the outliers argument. These arguments are built on a misunderstanding of the objective of process behavior charts and a failure to appreciate that the computations are already robust.

SUMMARY

So, should you delete outliers before you place your data on a process behavior chart? Only if you want to throw away the most interesting and valuable part of your data!

If you fail to identify the signals of exceptional variation as such, if you assume that a collection of data is homogeneous when it is not, then you are likely to have both your analysis and your conclusions undermined. The outliers are the interesting part of your data. In the words of George Box, "The key to discovery is to get alignment between an interesting event and an interested observer." Outliers tell you where to look in order to learn something about the underlying process that is generating your data. When you delete your outliers you are losing an opportunity for discovery.