Exact Answers to the Wrong Questions

Why Statisticians Still Do Not Understand Shewhart

Donald J. Wheeler

In a recent article that shall remain nameless a statistician carefully worked out the exact answer to the wrong question. Then, based on this exact answer, he made an erroneous recommendation regarding the use of a process behavior chart for individual values. In this column I will explain both why the question was wrong and how the recommendation is in error.

The question that the article claims to answer is "How does sample size impact Shewhart individuals control chart reliability?" To answer this question the article pulls out a standard tool of statistical inference, the "confidence" interval. These intervals are commonly used to express the uncertainty in an estimate of some parameter. As such, they complement the point estimate of a parameter by expressing the worst-case values that are consistent with the given data.

In this case the confidence intervals were computed for the distance between the upper and lower three-sigma limits for a chart for individual values. Without going into the details, the statistician correctly observes that the uncertainty in the width of these limits is a function of the inverse of the square root of the number of values used to compute the average moving range. If we have k values in our baseline period, the limits will be based on (k–1) moving ranges, thus:

$$\textit{Uncertainty in Width of Limits is proportional to } \frac{1}{\sqrt{k-1}}$$

This inverse square root function is shown in Figure 1. The values of (k–1) are shown on the horizontal axis while the values for the inverse square root of (k–1) are shown on the vertical axis.

This curve shows that you will cut the uncertainty in the limits in half whenever you increase the number of values in your baseline by a factor of four. The curve starts with the inverse square root of 2, which is 0.7071. By the time you get to eight values the curve has dropped to 0.3536. Increasing the amount of data from eight values to 32 values will cut the uncertainty in half again to 0.1768. At this point you will have obtained over 86% of the reductions in uncertainty that can reasonably be obtained by increasing the amount of data, and further increases in the amount of data will be of marginal benefit. The curve in Figure 1 underlies everything we do with statistics. It is the reason that, when we are concerned with the quality of an estimate, we will want to have at least 30 values in the computation.

Now how does the width of the computed limits affect the "reliability" of the *XmR* chart? In the article in question the statistician began with the worse case limits given by the confidence intervals. These limits were then converted into coverage probabilities, P, using a normal probability model. With this model, the expected coverage of the limits is P = 0.9973. Not surprisingly, the narrowest of the worst case limits had coverages that were substantially less than 0.9973, while the widest of the worst case limits had coverages that were essentially 1.000. These coverages were then plotted as a function of the number of values used to compute the limits.

Since [1 – P] will define the risks of a false alarm, the worst case coverages P were interpreted in the light of what was needed to obtain a false alarm near the theoretical value of 0.27%. Based on this analysis, the statistician recommended using at least 30 data when computing limits for an *XmR* chart. Every step of this argument appears to mathematically correct. The theory is sound. Unfortunately, the interpretation is not.

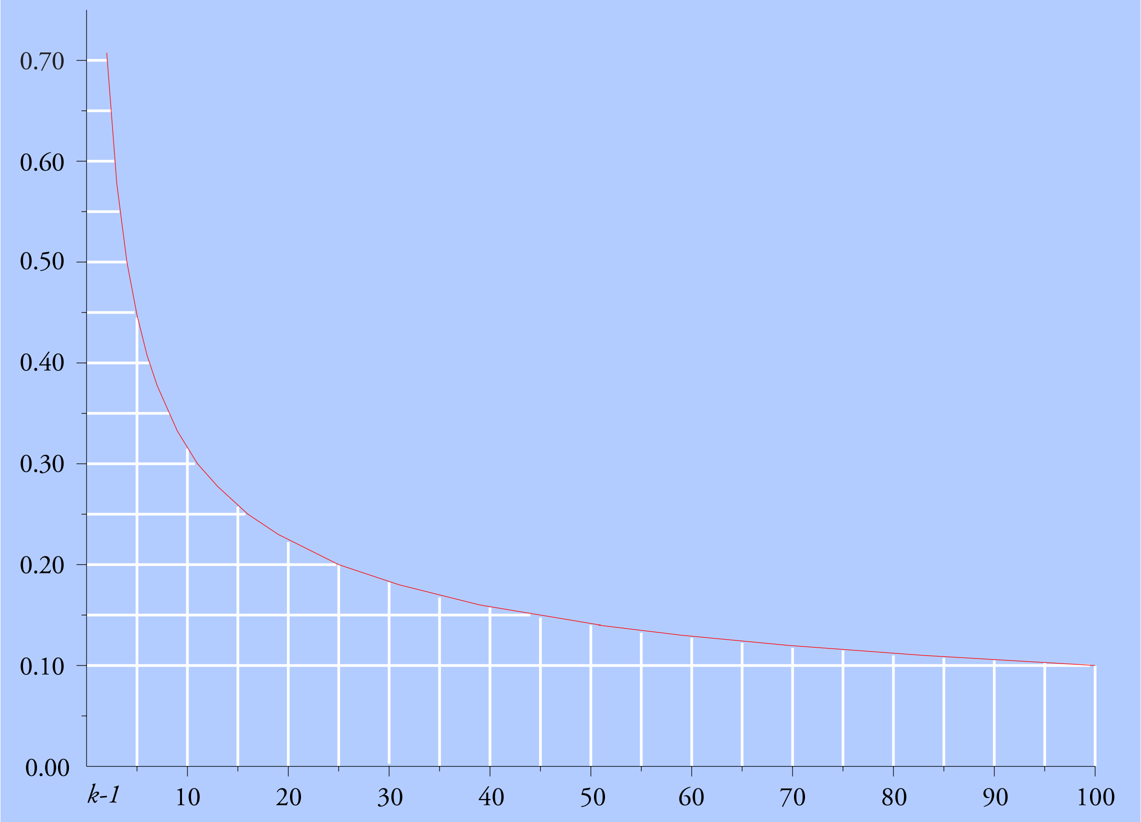

Figure 1 The Inverse Square Root Function

While we must use theory as a guide for practice, we have to also understand that theory only approximates what happens in practice. This is why Shewhart wrote that "the justification for a technique does not depend on a fine ancestry of highbrow statistical theorems, but rather upon empirical evidence that it works." While every part of the statistician's argument given above is mathematically correct and theoretically exact, it is disingenuous and incorrect on three levels.

The first flaw is the assumption that we need to somehow strive to achieve a specific risk of a false alarm near the theoretical value of 0.27%. This is a common mistake made whenever statisticians try to read Shewhart. By both training and practice, statisticians are used to techniques that follow the following recipe: Begin by fixing the value for the risk of a false alarm, then, for a given probability model, find and use the critical values that correspond to that risk. This approach works simply because the most commonly used statistical techniques are robust. This means that, regardless of how the original data behave, statistics such as Averages, Student's t-ratio, and the F-ratio for a test on means will all follow well-known probability models.

With this background, when statisticians try to assess the "reliability" of the limits for a process behavior chart, it is natural for them to look for what is needed to guarantee that the risk of a false alarm will be reasonably close to the theoretical 0.27%. However, this is completely opposite to what Shewhart did with the process behavior chart technique.

Shewhart made a crucial distinction that is missing from the statistician's argument. He observed that the process behavior chart is essentially applied to the original data, or to averages and ranges of small subgroups. After observing that we will never have sufficient data to fully specify a probability model for the original data, Shewhart completely reversed the statistical approach. Rather than holding the risk of a false alarm constant, he chose to hold the critical values constant and to let the risk of a false alarm vary. His argument was that as long as the risk of a false alarm is reasonably small, the decision procedure will work as desired. Since this distinction is at the very heart of Shewhart's approach, it undermines any attempt to evaluate the limits from the perspective of achieving a particular risk of a false alarm.

The second flaw in the mathematical argument is the inconsistent comparison made between using a small number of data to compute the limits and then looking at how such limits would work with an infinite number of data. While this contrast between a few values and an infinite number of values is inherent in any theoretical argument, it is disingenuous to base recommendations for practice solely on such a comparison.

When only a few data are available, only a few data will be placed on the chart. As more data become available, more data may be used to recompute the limits (when this is thought to be advisable and appropriate). This will generally preclude any great disparity between the number of values in the baseline and the number of data on the chart. So while the theoretical approach will describe the long-run performance of a set of limits based on a few values, these long-run numbers are not likely to materialize in practice.

Moreover, small data sets will tend to have small total ranges. This fact of life can be seen in Figure 2 which shows how the average distance between the maximum and minimum values for a histogram (in standard deviation units) will grow as the number of values in the histogram increases. (For those who wish to reproduce Figure 2, the values plotted on the vertical scale are the values for the bias correction factor known as d_2.)

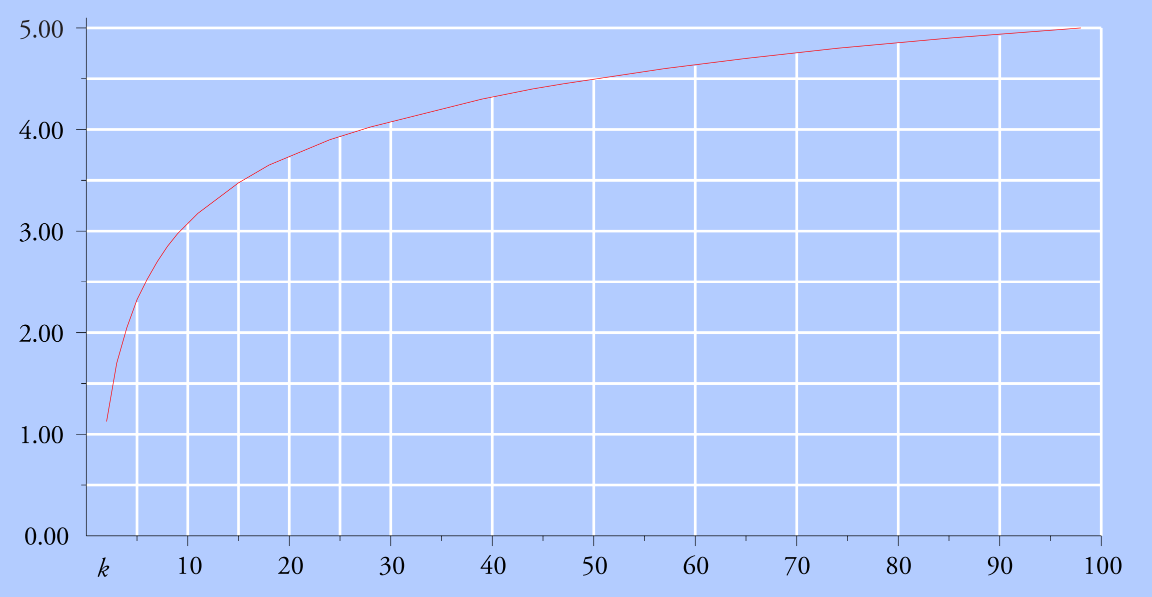

Figure 2: The Average Range for a Histogram of *k* Values

The curve in Figure 2 shows how the average distance from the minimum value to the maximum value will grow with increasing amounts of data. Data sets with less than 30 values will tend to spread out less than four standard deviations. While smaller data sets may have limits that are more uncertain, they are also less likely to have extreme values.

If we compare the curves in Figures 1 and 2 we see that they tend to cancel each other out. In fact the uncertainty in the limits drops faster than the histograms tend to grow. This may be seen by simply multiplying the two curves together and noting that the products get smaller with increasing k (when k exceeds 5). When k is small, the uncertainty in the limits is effectively canceled out by the fact that only a few values are being compared to those limits. This makes the question of the reliability of the limits one that cannot be successfully addressed from a purely theoretical perspective.

APPROXIMATE ANSWERS TO THE RIGHT QUESTIONS

When considering recommendations for practice it is important to look at how the technique is used. Since data occur in time, there will be situations where only a limited amount of data are available. How does the X chart work in these situations? There are two aspects to consider: what happens to the likelihood of a false alarm, and what happens to the ability of the chart to detect signals of process changes? We begin with a consideration of the initial likelihood of a false alarm within the baseline period.

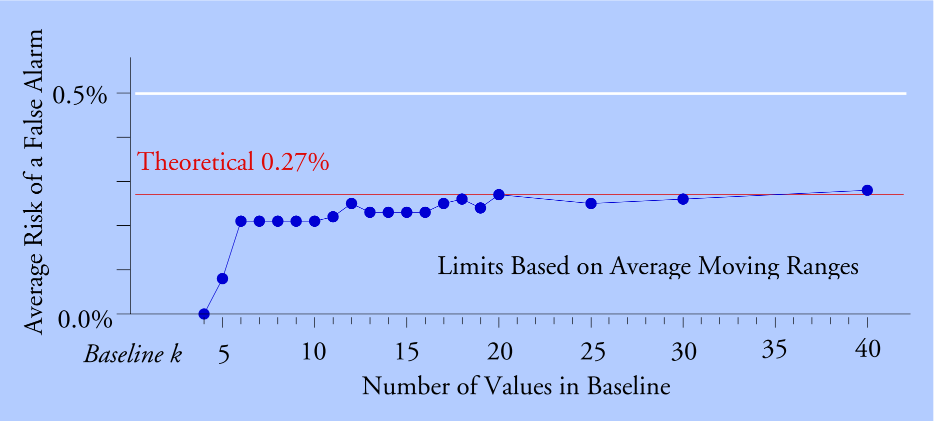

Figure 3: The Average Risk of a False Alarm Within the Baseline Period

Figure 3 comes from an extensive simulation study performed by Dr. Henry Neave, former Deming Professor of Management, Nottingham Trent University. There we see that regardless of how few data are used in computing the limits there is no real excess risk of a false alarm during the baseline period. For limits based on 6 or more values the average risk of a false alarm is indistinguishable from the theoretical value. So while the theoretical argument by the statistician is true, and while there will be some sets of limits with a greater risk of a false alarm, this is not a severe problem in practice. With average risks that are so small, and with a limit to how small the risks can go (zero), there simply cannot be many instances where the risk of a false alarm is very large. Thus, when we create an XmR chart using only a few values we will have a reasonably small chance of getting a false alarm within our baseline period. But what about our ability to detect a real signal within the baseline values?

When the baseline data for an XmR chart contain values that occur both before and after a change in the process location it is inevitable that the change will influence the limits. With short baseline periods any shift will have a pronounced impact upon the limits. As a consequence, as fewer data are used in the baseline period, it is inevitable that the X chart will become less sensitive to a shift that occurs within the baseline period.

To characterize this loss of sensitivity I chose a specific type of shift where the last point in the baseline period was shifted by some amount. Depending upon the number of points in the baseline period, the shift in this last point will inflate the limits by some amount and this inflation will make it less likely that the last point will fall outside the limits. By increasing the size of the shift and repeating the simulation study, it was possible to determine how large the shift would need to be in order to be detected at least half the time. The curve in Figure 4 shows the smallest size of shift that was detected in at least 5000 out of 10,000 simulations for baselines ranging from 6 to 40 data. The vertical scale begins at 3.0 simply because in theory, a three-sigma shift has a 50% chance of being detected on the first point following a shift. As expected, with increasing k, the curves in Figure 4 approach this theoretical value of 3.0.

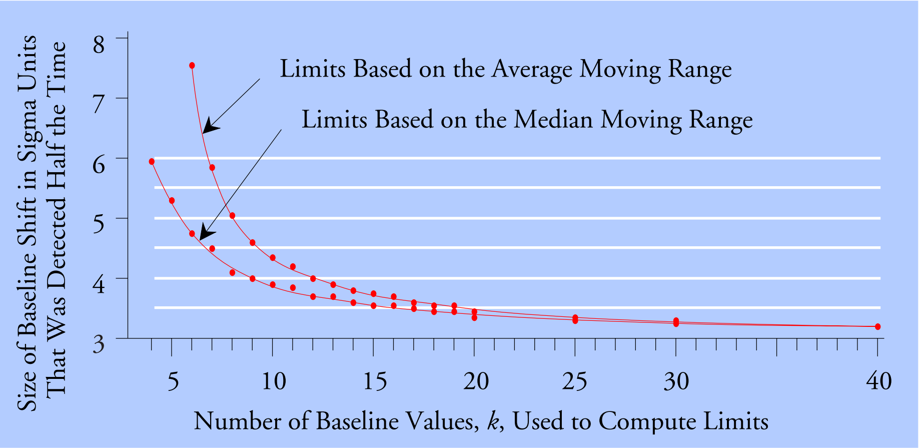

Figure 4: How Shifts Within the Baseline Period Affect the Sensitivity of the *X* Chart

The empirical values in Figure 4 show the expected decrease in sensitivity as *k* gets smaller. It also shows what happens when we use the median moving range as the basis for the limits rather than the average moving range.

For example, when using the average moving range to compute the limits when *k* is 12 we will detect the shifts at least half the time only when those shifts exceed 4.0 sigma. When using the median moving range to compute the limits when *k* is 9 we will detect the shifts at least half the time only when those shifts exceed 4.0 sigma.

The losses in sensitivity shown in Figure 4 suggest that we should exercise reasonable care when choosing a baseline period to make sure that it is free of known changes. While this is good advice for any value of *k*, it is imperative with short baseline periods.

Combining the results from Figures 3 and 4, we can see that we may construct an XmR chart using very few data indeed In the absence of actual signals, there will be a reasonably small chance of getting a false alarm. In the presence of actual signals, the limits will be inflated so that any signal we detect will be larger than it looks. All of this means that when you get a point outside the limits within the baseline period you can safely interpret that point as a signal regardless of how few data have been used in constructing the limits.

While we do run the risk of missing some smaller signals by constructing charts using few data, *not* computing the limits while waiting for more data will *guarantee* that you miss *all* of the signals. This is why there is no real penalty attached to using a small number of data to compute your initial limits for a process behavior chart. The objective is to take the right action, rather than to obtain the best possible estimates of the limits. The limits are a conservative filter, and as such they work even when our estimates are less than perfect.

But what about using the limits with future values? As argued above, you will always have the option of revising the limits as additional data become available. Figure 1 suggests that this can be worth considering until you have 20 to 30 data in your baseline. However, if your baseline period shows signals of exceptional variation, then you do not need to worry about revising the limits. Instead of worrying about the limits, you will need to find the assignable cause of the exceptional variation and take action to remove its effect from your process. Revising the limits is moot. Action is required. Once you have taken this action you will essentially have a new process and will therefore need to use a new baseline to compute new limits.

Therefore, if you have limits based on *k* values and your baseline period shows no signals of exceptional variation, then what happens if you choose to use your baseline limits with future values as they become available? This question has two parts: how well will you detect future signals, and what will happen to the risk of a false alarm with the future values?

To evaluate the question of sensitivity I used the fact that the theoretical probability of detecting a three-sigma shift in location on the first point following that shift is 0.5000. To see how the number of baseline values affected this probability I added a single point representing a three-sigma shift after the baseline period and looked at the number of times, out of 10,000, that the limits detected the shift. The results are shown in Figure 5.

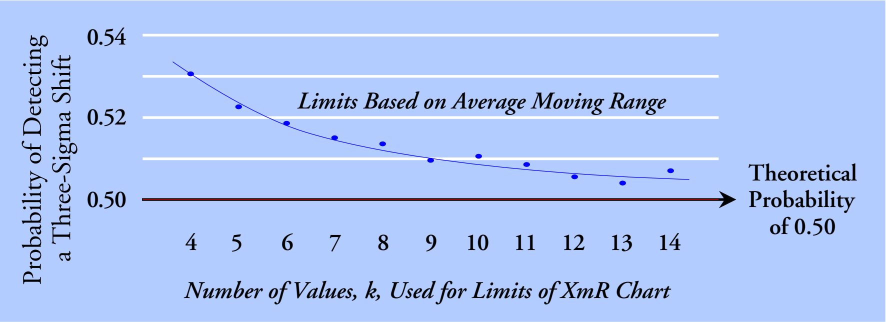

Figure 5: How the Baseline Period Affect the Sensitivity of the *X* Chart to Future Shifts

The observed frequencies shown in Figure 5, ranging from 53% down to 50.5%, are so similar to the 50% theoretical probability that the differences are too small to be of any practical interest. This means that the sensitivity of the *X* chart to future shifts is essentially unaffected by the number of points in the baseline period.

So that leaves the question of what is the risk of using limits from a short baseline with future values. Here we have to use theory to obtain an answer that is sufficiently general to be of practical use. Figure 6 shows the average risks of a false alarm when you use limits from a baseline of *k* values with future values.

From Dr. Neave’s simulation study, if your baseline contains at least ten original data, then the average risk of future false alarms will be a reasonably conservative 2% of less. If your baseline contains at least 15 data, then this average risk of a future false alarm will drop to less than 1.3%. With at least 20 baseline data this average risk drops to less than 1.0%, and with at least 30 baseline data this average risk of a future false alarm drops to less than 0.7%.

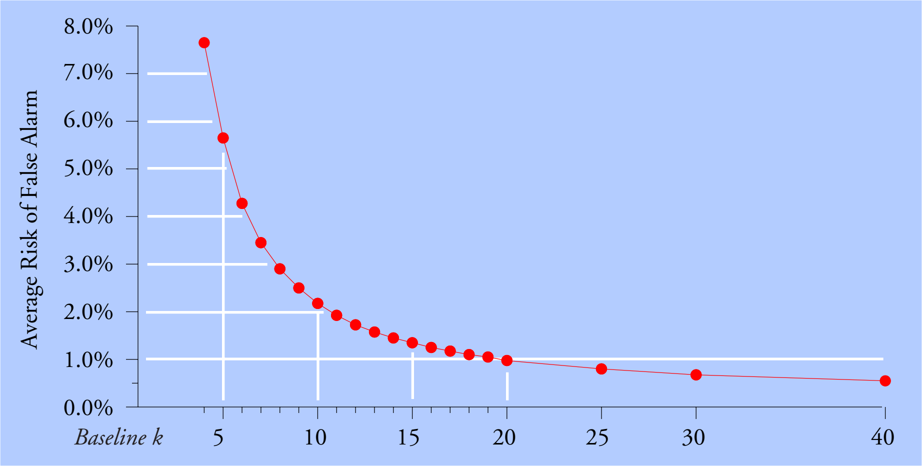

Figure 6: The Average Risks of Future False Alarms

However, when you have less than 10 values in your baseline, and you extend those limits without revising them with the new data, you will have an average risk of a false alarm in excess of 2.5%. Since 2.5% is the cut-off for a conservative risk, and since the operating principle behind the process behavior chart is to use a conservative analysis, this argues against extending the limits from a short baseline to future values. Since these are average risks, when you have less than 15 values in your baseline you will probably want to recompute the limits whenever additional values become available.

Combining the results from Figures 5 and 6, a short baseline will not undermine the sensitivity of the X chart to future changes, but it will increase the likelihood of future false alarms. Since the essence of the technique is to use a conservative analysis, this argues against extending limits from short baselines without revising them. However, by the time you have 15 to 20 values in your baseline period this recomputation does not purchase much additional insurance against false alarms.

But why not solve this problem by simply waiting until we have 30 or more data before computing our limits? If this is feasible, then there is little problem with this approach. However, in many situations, the data come along slowly and this simplistic answer to the problem is not realistic. Shewhart suggested using as few as two subgroups of size four to compute limits for an average and range chart. Here we looked as using as few as six values to compute limits for an XmR chart. While such limits will need to be updated as new values become available, any signals you find within your baseline period are likely to be real. And it is always better to detect the signal now than it is to detect it ten or twenty months from now.

In practice, the number of data used to compute the limits does not have an appreciable impact upon either the false alarm rate within the baseline period or the ability of the chart to detect future changes. As long as you exercise reasonable care to avoid known or obvious changes within your baseline period you can construct an XmR chart using as few as six original data. When additional data become available you will need to consider if you need to revise the limits to avoid an inflated risk of a false alarm with the future values.

Remember that the objective is not to compute the “right” limits, but to use the limits to characterize the behavior of the process. This characterization does not require high precision in our computations, but rather clear thinking about how we use and interpret the data.