Contra Two Sigma

The consequences of using the wrong limits

Donald J. Wheeler

You may occasionally encounter charts with two-sigma limits. The origins of this practice are not clear, and no real justification of this practice has been given in the literature. In this article I will consider the theoretical and practical consequences of using two-sigma limits on a process behavior chart.

THE ORIGINS OF TWO – SIGMA LIMITS

One possible origin for the idea of using two-sigma limits is an extrapolation from our introductory classes in statistics where we were taught to use a 95 percent confidence interval and a 5 percent alpha-level. Since two-sigma limits cover approximately 95 percent of the outcomes we might think we could also use such limits on a process behavior chart. However, this extrapolation fails to consider the difference between a one-time analysis such as a confidence interval or a t-test and a sequential procedure such as a process behavior chart. The alpha level we use with a one-time procedure is much too large to use with any sequential procedure. In a typical t-test you are working with experimental data and are trying to find a signal that you have attempted to create. Here the traditional alpha-level of 5 percent accepts a larger risk of a false alarm to reduce the chance of a missed signal. With a process behavior chart you are examining data where there should be no signals. Every time you add a point to the chart you perform an act of analysis, and you run the risk of having a false alarm. Since this is a recurring risk, the collective risk will be considerably larger than the individual risks. Here, as with all sequential procedures, we have to tighten up on the individual risks of a false alarm in order to be sure that we are looking at a signal when a point goes outside the limits. With a process behavior chart we do this by using three-sigma limits.

Another possible origin for two-sigma limits is the use of the incorrect method of computing a measure of dispersion. When people erroneously use a global standard deviation statistic to construct "three-standard-deviation" limits they will often end up with very wide limits. As a remedy they then shift to using "two-standard-deviation limits." I have explained why the global standard deviation is the wrong way to compute limits for a process behavior chart in "The Right and Wrong Ways of Computing Limits," ***Quality Digest Daily*** Jan. 7. 2010. Trying to adjust for this mistake by using two-standard-deviation limits is just compounding one mistake by another. Two wrongs still do not make one right.

A third possible origin for the use of two-sigma limits is the idea of increasing the sensitivity of a process behavior chart. Those who are new to the use of process behavior charts will often feel that three-sigma limits are too conservative. This thought is often expressed in the form of "We'll never find a point outside those limits" or "We can't wait that long to react." Once they learn how to use the charts these comments often turn into "We have too many signals to get around to them all." Three-sigma limits have been thoroughly proven in years of use in all kinds of applications and in all kinds of industries. Most of the time we do not need to increase the

sensitivity.

INCREASING THE SENSITIVITY

While it is true that two-sigma limits will result in increased sensitivity, they are not the best means to this end. Since 1956 the recognized and accepted way to increase the sensitivity of a process behavior chart has been to use the Western Electric Run-Tests in addition to the primary detection rule of a point falling outside the three-sigma limits. For the sake of clarity these detection rules are:

Detection Rule One: A point outside the three-sigma limits is likely to signal a large process change.

Detection Rule Two: Two out of three successive values that are beyond one of the two-sigma lines (both on the same side of the average) are likely to signal a moderate process change.

Detection Rule Three: Four out of five successive values that are beyond one of the one-sigma lines (all four on the same side of the average) are likely to signal a moderate, sustained shift in the process

Detection Rule Four: Eight successive values on the same side of the average are likely to signal a small, sustained shift in the process.

Detection Rules Two, Three, and Four are the Western Electric run-tests. Collectively, all four rules are often referred to as the Western Electric Zone Tests. Since these run-tests look for smaller signals they increase the sensitivity of a process behavior chart when they are used with Rule One. In the sections that follow I shall compare the use of these detection rules with the use of two-sigma limits as a means of increasing the sensitivity of a process behavior chart. This will be done in three ways. First we shall look at the power functions. Next we shall look at the average run length curves. Finally we shall look at the probabilities of a false alarm.

THE POWER FUNCTIONS

The power function for a statistical technique describes the probability of detecting a signal. Of course this probability will depend upon the size of the signal, the number of data available, and the technique itself. When the signal is large, useful techniques will have a 100% probability of detecting that signal. However, as the size of the signal gets smaller the probability of detection will generally drop. Finally, in the limiting case where there is no signal present, desirable techniques will have a small probability of a false alarm. If we plot the probability of detecting a signal on the vertical axis, and plot the size of the signal on the horizontal axis, then we would like to see a curve that starts near zero on the left and climbs rapidly up to 1.00 on the right. I published the formulas for the power function for a process behavior chart 30 years ago. They may be found in my text *Advanced Topics in Statistical Process Control*. These formulas are for detecting a shift in location using either an *X*-chart or an Average chart. To remove the effects of subgroup size the shifts are expressed in standard error units. The curves shown are the power functions for exactly $k = 10$ subgroups. Ten subgroups were used because Rule Four cannot be used with less than 8 subgroups.

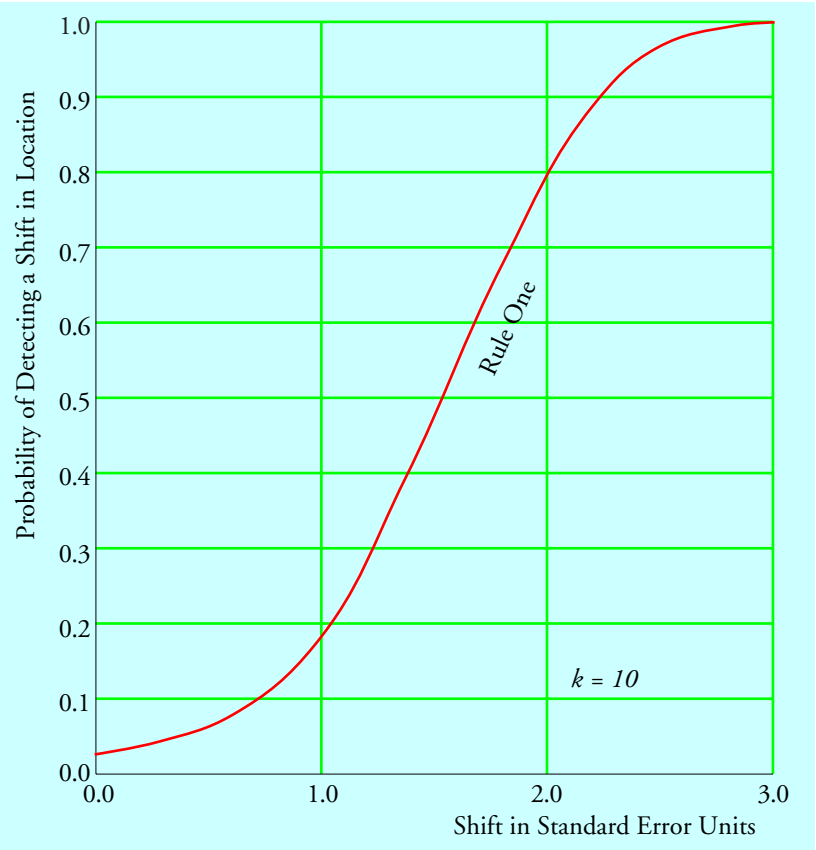

Figure 1: Power Function for Detection Rule One Alone

Figure 1 shows the power function for using Detection Rule One alone. The left end-point of the power function defines the risk of a false alarm. Here we find that there is a 2.7% chance of a false alarm for an average chart using ten subgroups. When a shift occurs the probability of detecting that shift within ten subgroups of when it actually occurs climbs as the size of the shift increases. An average chart using Detection Rule One alone will have a 100% chance of detecting a 3.0 standard error shift in location within ten subgroups of when that shift occurs. Since the objective is to find those shifts that are large enough to justify the expense of fixing the problem, this curve shows why Rule One is usually sufficient.

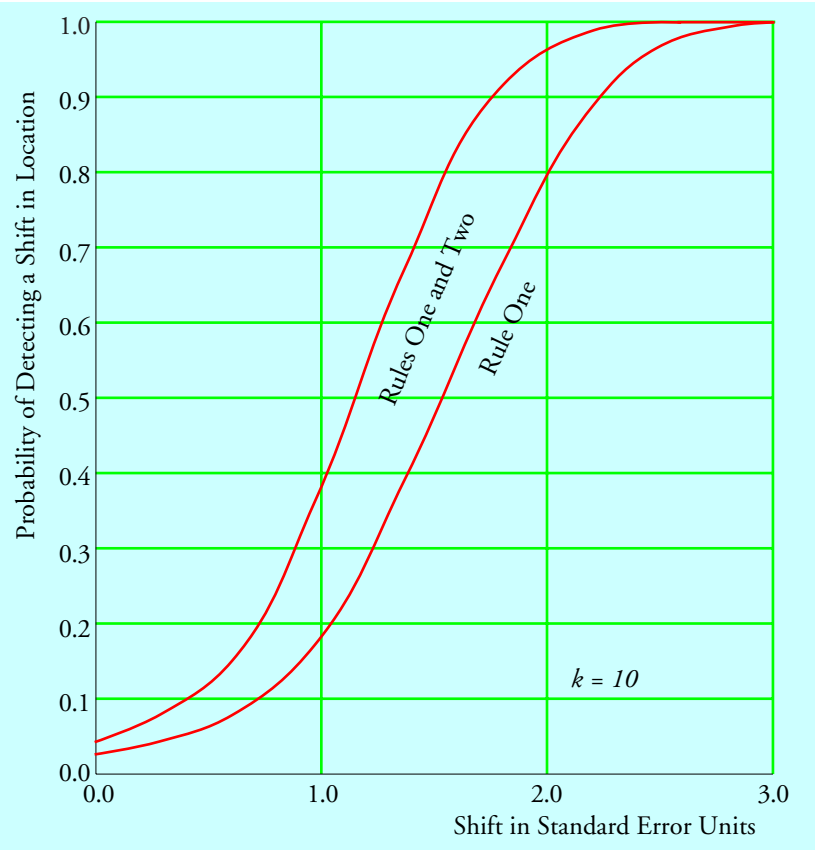

Figure 2: Power Function for Detection Rules One and Two

Figure 2 shows the power function for using Detection Rules One and Two. The increased sensitivity can be seen in the steeper power function curve. With Rules One and Two you have a 100% chance of detecting a 2.5 standard error shift within ten subgroups of when that shift occurred. However, as is always the case, the use of additional detection rules results in an increased risk of a false alarm. Here it is 4.3% for an average chart with ten subgroups.

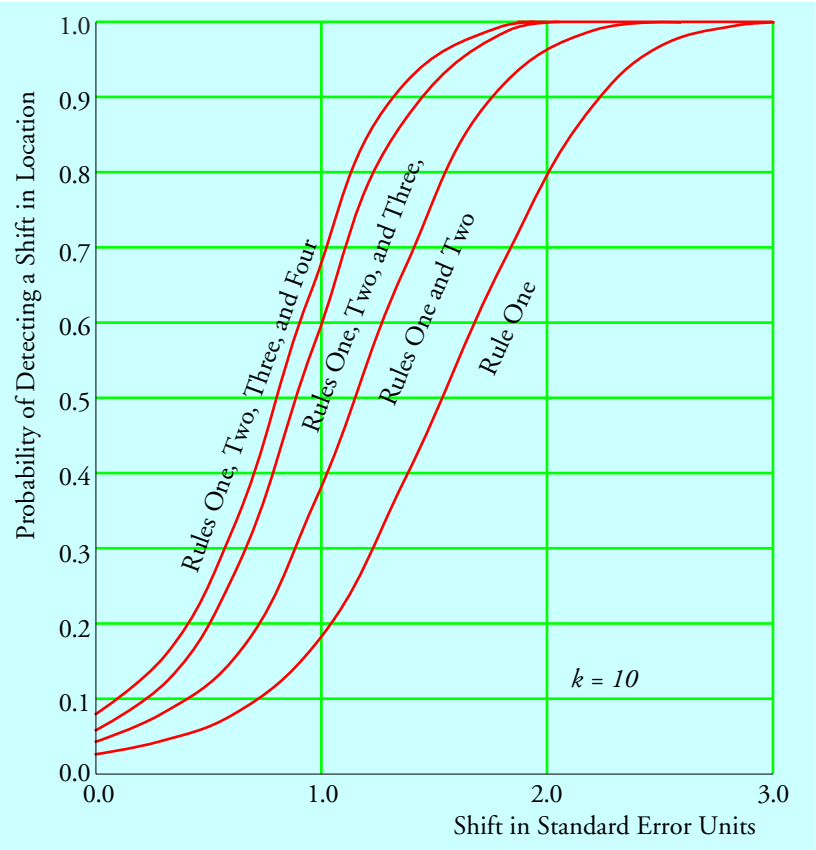

Figure 3: Power Functions for Western Electric Zone Tests

Figure 3 shows the power function for using Rules One, Two, and Three and the power function for using all of the Western Electric zone tests. These curves are slightly steeper than the curve for Rules One and Two combined. They both show a 100% probability of detecting a 2.0 standard error shift in location within ten subgroups of when it occurred. The false alarm risks for these two curves for an average chart with ten subgroups are approximately 6% and 8%. In fact, these last two power function curves show probabilities that differ by less than 0.10 for shifts smaller than 2.0 standard errors. Such small differences in power are hard to detect in practice. Using Rules One, Two, and Three will work about as well as using all of the detection rules. Rule Four will only add some sensitivity to small and sustained shifts.

The curves in Figure 3 are getting squeezed together because there is a limit to steepness of the power function and these curves are approaching that limit. Once you hit this limit, the only way to raise the power function curve is by raising the left-hand end-point of the curve. We essentially see this beginning to happen in the last two curves of Figure 3.

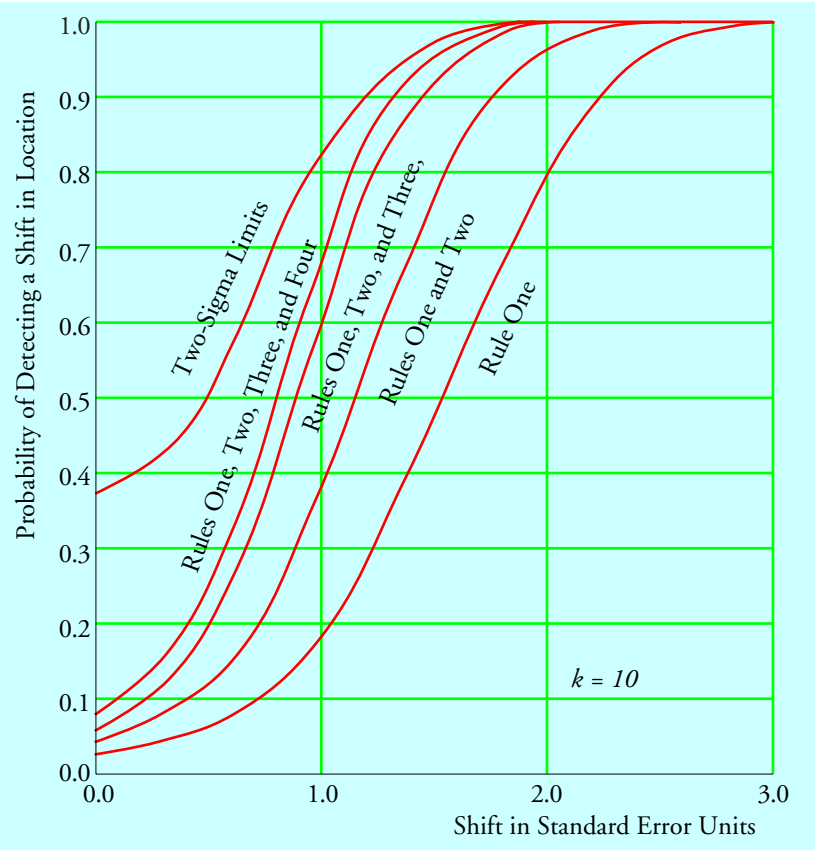

Figure 4: Power Function for Two Sigma Limits

Figure 4 shows the power function curve for using two-sigma limits. It has a 100% chance of detecting a 2.0 standard error shift in location within ten subgroups of when that shift occurs. This is the same as for the third and fourth curves in Figure 3. In fact, for shifts larger than 1.5 the top three curves of Figure 4 are so close together than there will no discernible difference in practice. Yet with two-sigma limits you will incur about a 38% risk of a false alarm on an average chart having ten subgroups. Thus. for the same ultimate sensitivity, the use of two-sigma limits has a false alarm risk that is more than four times larger than that of the Western Electric zone tests.

THE AVERAGE RUN LENGTH CURVES

A different perspective is provided by the Average Run Length (*ARL*) curves. The average run length is the average number of subgroups between the occurrence of a signal and the detection of that signal. When these *ARL* values are plotted against the size of the signal we end up with the curves in Figure 5.

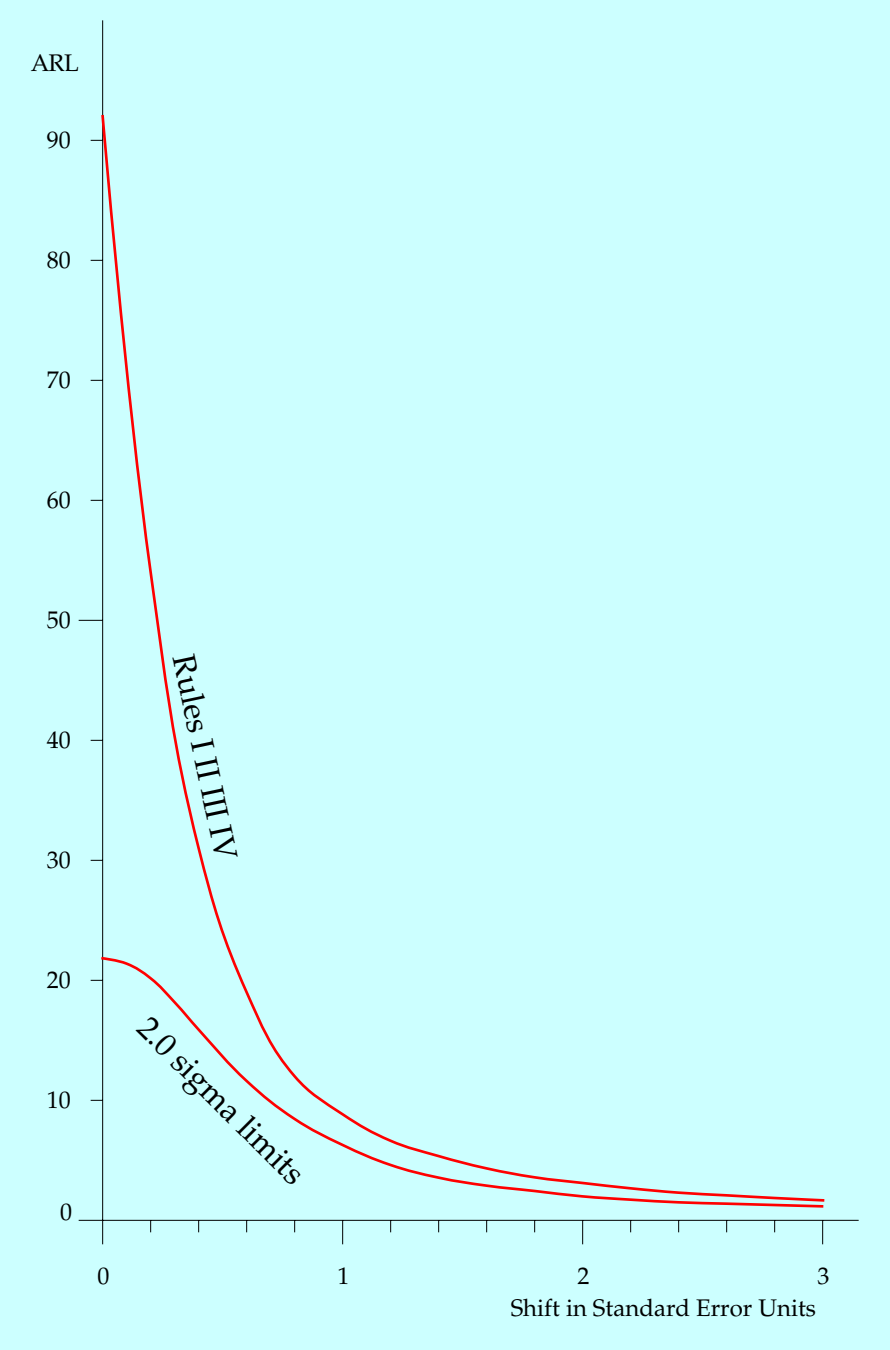

Figure 5: ARL Curves for Two-Sigma Limits and Western Electric Zone Tests

For shifts in excess of 2.0 two sigma limits will have an average run length that is less than one subgroup smaller than that of all four detection rules. However, when there are no signals, the use of two-sigma limits will result in one false alarm every 22 subgroups on the average. In contrast to this, the use of all four detection rules will result in one false alarm every 91 subgroups on the average. So, by using two-sigma limits you are increasing your false alarm rate four-fold in return for a very slight advantage in detecting a signal that is large enough to be of any practical consequence.

THE EFFECT OF THE NUMBER OF SUBGROUPS

Figure 4 showed a comparison while holding the number of subgroups constant. Figure 5 showed the average number of subgroups between the signal and the detection of that signal. As noted earlier, the risk of a false alarm on each step of a sequential procedure is not the same as the overall risk of a false alarm across several steps. Here we shall look at how the false alarm probability increases as the number of subgroups increases.

Figure 6 shows the probability of a false alarm on the vertical axis and the number of subgroups considered on the horizontal axis. Here we compare the probabilities of false alarms for the traditional process behavior charts and a chart using two-sigma limits.

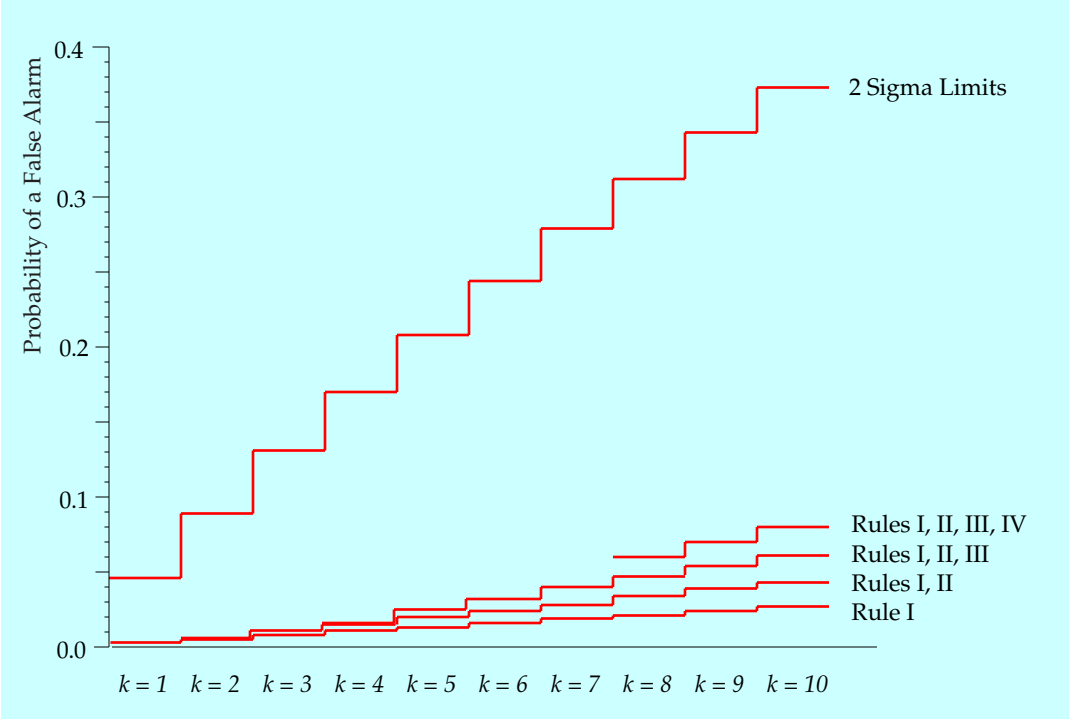

Figure 6: False Alarms for Two-Sigma Limits and Western Electric Zone Tests

The use of two-sigma limits will result in a dramatic increase in the number of false alarms compared to the other charts. This dramatic increase begins immediately, and just gets bigger as more data are collected. Since, in order to use the charts effectively, you must investigate each and every out-of-limits point in search of an assignable cause, this excessive number of false alarms will inevitably undermine the credibility of both the chart and the person using the chart. In short, an excessive number of false alarms will kill the use of the charts.

In practice, most people who use process behavior charts effectively find that they have plenty of signals using Detection Rule One. In fact, the problem is usually one of needing a procedure that is *less* sensitive, rather than more sensitive. However, for those situations where an increased sensitivity is desired, the addition of Detection Rules Two, Three and Four will suffice.

IMPLICATIONS

The only reason to collect data is to use it to take action. In order to use data to take action you have to have a properly balanced decision rule for interpreting your data. If your decision rule is not properly balanced you will either err in the direction of missing signals, or else you will err in the direction of taking action based on noise.

The purpose of a process behavior chart is to tell you when to take action. The action is to look for an assignable cause of exceptional variation. The idea behind the chart is as old as Aristotle, who taught us that the time to identify a cause is at that point where a change occurs. This means that you either look for an assignable cause, or you do not. If you do not, then nothing will happen, and you will continue to suffer to consequences of having a process that operates at less than its full potential. This will mean increased scrap, increased rework, and increased problems in making the conforming product work in the subsequent process steps.

Trying to tighten up the limits on a process behavior chart is not like tightening up the specifications. It does not make things better. You cannot squeeze the voice of the process. Tightening up the limits on a process behavior chart will only increase the false alarm rate. When you look for non-existent assignable causes, you will be wasting time and effort while undermining the usefulness of the process behavior chart. Three-sigma limits strike a balance between the economic consequences of the dual mistakes of missing signals and getting false alarms.

For a discussion of the role of specifications and their use see "Right and Wrong Ways to Use Specifications," ***Quality Digest Daily*** March 4, 2013. However, if you have a say in setting specifications, you should avoid setting them at the average plus and minus two sigma. One of my clients had been setting their product specifications using these so-called 95 percent specification limits. One of their products had six specified characteristics. When I pointed out to them that 0.95 raised to the 6th power would only be 73.5 percent, they suddenly understood why they were rejecting almost 30% of their batches of this product! Specifications are for sorting good product from bad product after the fact. Process behavior charts are for learning how to improve process performance. When you confuse the voice of the customer with the voice of the process, chaos is inevitable.

Finally, in this article I have used the power functions, ARL curves, and false alarm probabilities computed in the usual way simply because that is the only way to obtain valid and global comparisons between different techniques. These usual assumptions are that the shift in location can be modeled by a step function, that the measurements are continuous, that the measurements are independent of each other, and that the measurements are normally distributed. These assumptions are necessary to carry out the mathematics. In practice none of these assumptions are realistic. This is why theory only provides a starting place for practice. So, while theory suggests that three-sigma limits should work, over 80 years of practice has proven beyond any doubt that they do work as expected. Make no changes and accept no substitutes.