# More to Beware About Tukey Control Charts

## Answers to questions asked

Donald J. Wheeler

Last month's column "Beware the Tukey Control Chart" generated several questions of a fundamental nature that deserve expanded answers. These questions and their answers will be considered here.

### INTERQUARTILE RANGES

The Tukey control chart uses the interquartile range or IQR to characterize dispersion. One statistician wrote that he had found the IQR to be useful, especially in the presence of outliers. While he is right in saying that the IQR is less sensitive to outlying data points than is the global standard deviation statistic, this does not make it an appropriate measure of dispersion for a process behavior chart. The problem with the IQR is that it is a *global* measure of dispersion, i.e., it uses *all* of the data in the computation. As a global measure of dispersion it makes an implicit assumption that the data are homogeneous. This assumption is a distinct problem when you are examining the data for evidence of a lack of homogeneity which is the purpose of a process behavior chart.

One of the major breakthroughs of the Twentieth Century in statistical analysis was the use of within-subgroup measures of dispersion to filter out the noise. The use of the IQR to compute limits for a control chart completely ignores this fundamental principle of modern statistical analysis. Thus, while the IQR is a useful descriptive statistic, it completely inappropriate for use in the analysis of data where we are trying to separate the potential signals from the probable noise.

### TWO SIGMA LIMITS

The Tukey control chart essentially uses two sigma limits. The problems of using two-sigma limits are explained in my recent column "Contra Two-Sigma" ***QDD***, May 1, 2013. For those who have not yet read that article, I would suggest starting there. However, even after using the arguments in that article, one correspondent wrote that he still had people who wanted to use two sigma limits in spite of the large number of false alarms that they generate. He asked for another explanation of why the process behavior chart uses three-sigma limits. The following explanation will use the fundamentals of the decision problem to show how any sequential procedure will require a more conservative decision rule than that used for a one-time analysis.

The purpose of collecting data is to take action. If you "cry wolf" too often you will loose your creditability. One of my colleagues noted that USA Today had "predicted 15 of the last three downturns in the stock market." This is why we wish to avoid false alarms. At the same time, we want to detect those signals that are large enough to be of practical interest so that we

can take appropriate action. So there are two mistakes we can make when interpreting data: we can miss a signal, or we can sound a false alarm. The trick is to strike a balance between these two mistakes. But to understand how to strike this balance we have to understand how the nature of the decision problem changes with different situations.

When we are performing an experiment we are deliberately trying to create signals. Here we only get to analyze the data one time and we are going to make a single decision. We have paid good money to try to create signals within our data and we do not want to miss those signals, so we accept a larger risk of a false alarm in order to minimize the risk of missing a signal. A traditional risk of a false alarm is 5 percent which will roughly correspond to the use of two-sigma limits in virtually every one-time analysis.

When we are observing a process over time and placing our data on a process behavior chart we are essentially performing a sequential analysis on a process where we hope there are no signals. In the absence of any signals, the risk of missing a signal is not our primary concern. However, every time we plot a point on the chart we perform an act of analysis and risk a false alarm. Since we do not have to detect a signal immediately the emphasis here is to avoid false alarms . Given the sequential nature of the process behavior chart we will detect any signals that are large enough to be of economic interest reasonably quickly, and we are free to concentrate on avoiding the mistake of having too many false alarms. We do this by using three-sigma limits.

Thus there are different situations with different types of analysis and these different situations require different approaches to the decision problem. This is why your software gives you the option of doing things in different ways. It is up to you to understand the different situations. To illustrate how the different situations require different approaches I will use a graphical technique that presents a unified approach to the decision problem. This technique is known as an ROC curve.

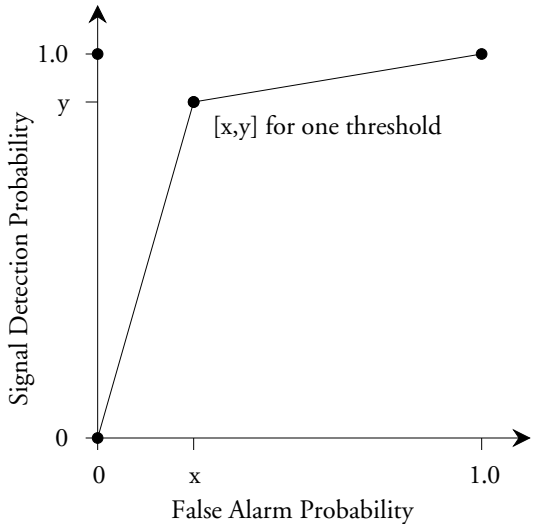


Figure 1: A Simple OC Curve for a Given Threshold Value

In World War II the problem of setting a signal threshold for a radar system was the stimulus for creating the Receiver Operating Characteristic graph or ROC curve. It provided a way to

determine the appropriate cut-off point between potential signals and probable noise. The horizontal axis was used to show the probability of a false alarm. This scale would run from zero to one. The vertical axis would show the probability of detecting a particular signal. This scale would also run from zero to one.

For a given cut-off threshold value there would be a single false alarm probability, x, and a signal detection probability, y. The operating characteristic (OC) for this single threshold value would start at [0,0], connect with the point [x,y] and then connect with the point [1,1] as shown in figure 1.

Since a perfect decision rule would have x = 0 and y = 1, a given decision rule can be evaluated by how close the point [x,y] is to [0,1]. For a given radar receiver different threshold values would be considered. Each threshold value would have its own value for [x,y]. To determine the best threshold value for a given receiver these different [x,y] points would all be combined to create a composite OC curve. Since this composite OC curve was specific to a particular receiver, it came to be called a Receiver Operating Characteristic curve. And the best threshold for a given receiver would be the threshold value that corresponds to the point closest to [0,1]. In Figure 2 this best threshold point is circled.

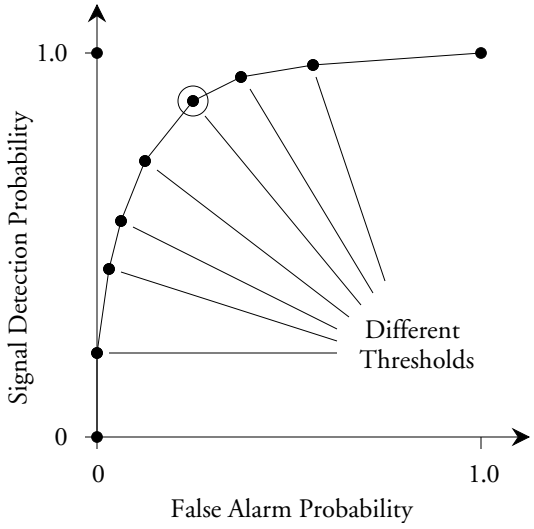


Figure 2: An ROC curve

In statistical jargon, a receiver is a test procedure, and a threshold is the decision rule for deciding what constitutes a signal. If we consider the decision problem for a one-time analysis and want to detect a four-sigma shift whenever it occurs, we can construct a theoretical ROC curve. We shall compare the ROC curve for a one-time analysis with that for a sequential analysis. Without going into the calculus of probabilities, Figure 3 lists the probabilities of a false alarm and the signal detection probabilities for detection limits of the form [average ± k sigma] for a one-time analysis.

| k | False Alarm | Signal Detection |
|---|---|---|
| 1 | 0.317 | 0.999 |
| 1.5 | 0.134 | 0.994 |
| 2 | 0.046 | 0.977 |
| 2.5 | 0.012 | 0.933 |
| 3 | 0.003 | 0.841 |
| 3.5 | 0.000 | 0.691 |

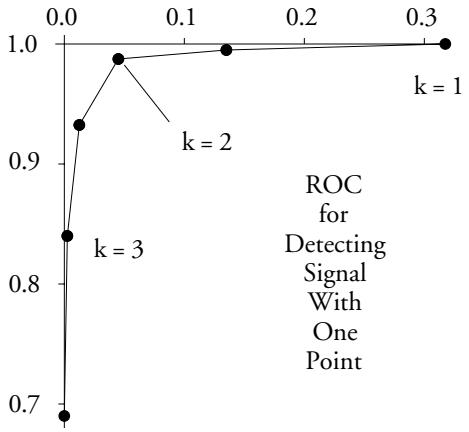


Figure 3: ROC Curve for Detecting a Four Sigma Shift with a Single Analysis

Figure 3 shows that two-sigma limits will be reasonably optimal when performing a single analysis. Three-sigma limits would be overly conservative here. This is why the traditional alpha level for a one--time test is five percent. Since classes in statistics emphasize the one-time analysis of experimental data, this is why students are taught to use two-sigma decision points.

However, with a sequential test we do not have to make a decision after the first point. We have the wait-and-see option. If we allow ourselves up to three points before detecting a four-sigma shift,, the ROC curve changes from that in Figure 3 to the curve shown in Figure 4.

| k | False Alarm | Signal Detection |
|---|---|---|
| 2 | 0.130 | 1.000 |
| 2.5 | 0.037 | 1.000 |
| 3 | 0.008 | 0.996 |
| 3.5 | 0.001 | 0.971 |
| 4 | 0.000 | 0.875 |

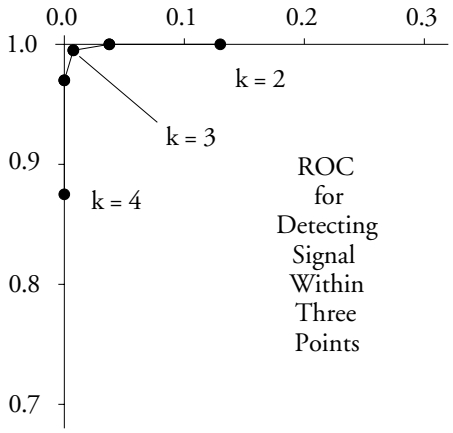


Figure 4: ROC Curve for Detecting a Four Sigma Shift Using Three Sequential Analyses

Figure 4 shows that with a sequential procedure we need to shift to using three-sigma limits. Using two-sigma limits with a sequential procedure is simply not appropriate. While two-sigma limits do increase the sensitivity by 4 parts per thousand they also increase the risk of a false alarm by over 1600 percent! This is a trade-off that simply cannot be justified.

We should also notice in Figure 4 that the ROC curve with three-sigma limits is very nearly perfect. Both Figure 3 and Figure 4 are drawn to the same scale, and you can see the increased power of the sequential procedure in Figure 4.

ROC curves are not for everyone. Simpler explanations of the problem of two-sigma limits

are given in my column entitled "Contra Two Sigma" cited earlier. However, whether you use the rigorous approach given here, or the simpler approach of the earlier article, it is clear that there is simply no way to justify the use of two-sigma limits with a sequential procedure such as a process behavior chart.

## YOU MUST KNOW THE PROBABILITY DISTRIBUTION

Other questions and comments implicitly assumed that you cannot use a process behavior chart of any type until you know what probability model to use. This idea is incorrect. It can be traced back to 1935 when E.S. Pearson erroneously wrote that the process behavior chart required normally distributed data. Walter Shewhart corrected Peason's error in his 1938 book when he wrote "*we are not concerned with the functional form of the universe* [i. e. probability model] *but merely with the assumption that a universe exists.*" The whole purpose of using a process behavior chart is to determine if the data display that degree of homogeneity that would allow us to assume that the data might have come from a predictable process which might be characterized by a single probability model.

If the data are homogeneous, then it is logical to assume that the process is being operated predictably and that the histogram might be approximated by some probability model. But if the data are not homogeneous, no probability model will ever be right. For more about this see my column entitled "Why We Keep Having 100-Year Floods" ***QDD***, June 3, 2013.

Since E. S. Pearson's error in 1935 there have been many who have made the same mistake of thinking you have to have a probability model before you can compute appropriate limits for a process behavior chart. Fortunately, this idea is completely wrong. Three-sigma limits are sufficiently conservative to work with all sorts of data. Three-sigma limits will filter out 99% to 100% of the routine variation regardless of the shape of the histogram. Therefore, any points that fall outside the three-sigma limits may be regarded as potential signals and investigated as such. Symmetric, three-sigma limits work with skewed data simply because they manage to cover the bulk of the elongated tail, regardless of how skewed the data may be. Finally, three-sigma limits are conservative enough to be robust to the uncertainties of estimation even when based on few data. For more on these points I suggest my article entitled "Do You Have Leptokurtophobia?" ***QDD***, August 5, 2009.

## SUMMARY

From the beginning,, those who have had a incomplete grasp of the purpose and scope of Shewhart's technique have taken his simple computations and labored to make them more complex, more obscure, and harder to use. Why is this? As one who made these mistakes and who was fortunate enough to have David S. Chambers and W. Edwards Deming as mentors, I can only tell of my journey. As a statistician I was trained to think one way. To understand Shewhart I had to learn how to look at the decision problem from a different perspective. As I have explained in "Myths About Process Behavior Charts" ***QDD***, Sept. 8, 2011, this perspective is 180 degrees opposite to the way I thought as a statistician.

So you have a choice. You can learn to use Shewhart's simple and profound process behavior charts as he intended, or you can follow the modern day Pied-Pipers into the complexity of their hybridized techniques that do not work as well as the simple original.