

The Normality Myth

Be careful where you get your information

Donald J. Wheeler

The oldest myth about process behavior charts is the myth that they require “normally distributed data.” If you have ever heard this idea, or if you have ever taught this to others, then you need to read this article.

While this myth dates back to 1935, and while Walter Shewhart exposed this idea as a myth in 1938, it continually reappears in various forms even today. For example, a white paper put out by a software company recently called the process behavior chart a “normality control chart.” And a blurb for a workshop advertised the Western Electric zone tests as “depending upon a normal distribution.”

As I consider how these myths are perpetuated I do not detect any malicious intent, just unconscious confusion. Nevertheless, those that continue to spread these myths fall into three groups. One group spreads these myths because of their special interests, another group spreads these myths because of their unexamined assumptions, and the third group spreads these myths because they were taught that the myths are true.

SPECIAL INTERESTS

Software makes the difficult easy and the impossible possible. As part of their outreach software companies produce white papers and conduct seminars and tutorials. Inherent in these marketing efforts is an emphasis upon making the software seem indispensable to the user.

For example, if you are told that you have to qualify your data by checking to see if they are “normally distributed” then you will be interested in learning how the software might help you check for normality. As users come to depend upon the software to guide them through their analyses, all the bells and whistles begin to seem essential. Gradually all of these extra features, such as the checks for normality, the probability plots, and the histograms, turn into prerequisites for using a process behavior chart. Thus, as everyone tends to their own special interests, the myth that the data must be normally distributed prior to using a process behavior chart continues to get a new lease on life.

So, while specific training may be needed to fully benefit from a particular software package, and while it is reasonable and proper for software companies to provide this training, they are a poor resource for broader education in theory, technique, or practice. As a consequence of this software-based training, many users think of statistics as a grab-bag of techniques. Just enter your data, pick a procedure from the smorgasbord provided by the software, and read the output.

However, the first axiom of data analysis is that no data have any meaning apart from their context. When understanding does not guide the analysis, the extraneous can obscure the essence, and the analysis can go astray.

Figure 1: The Extraneous Can Obscure the Essence

A corollary to this axiom is that the software will never know the context for your data. Every meaningful analysis requires an understanding of the source of the data, an understanding of the analysis technique, and appropriate subject-matter knowledge.

UNEXAMINED ASSUMPTIONS

When someone does not recognize that SPC is *fundamentally different* from the traditional techniques of statistical inference they inevitably end up making unexamined assumptions that are incorrect. To understand why these assumptions are incorrect we will have to illustrate the difference between SPC and statistics by returning to first principles.

The purpose of every statistical analysis is to separate the *potential signals* from the *probable noise*. To this end we have to find a way to filter out the probable noise. Since we can never filter out all of the noise, we have to settle for filtering out most of the noise. So let us denote that proportion of the probable noise that we filter out by P .

The traditional statistical approach begins by picking some value for P that is reasonably close to 1.00. (Commonly used values for P are 95 percent, 98 percent, and 99 percent.) Next, we transform our data in some way to get a test statistic Y . (Y might be a t-statistic, a chi-square statistic, an F-statistic, a proportion, or some other statistic of interest.) Next we identify the appropriate probability model, $f(y)$, to use for our test statistic, Y . Finally, we use the equation for the area under a curve to find the critical values, A and B , which correspond to the chosen value for P .

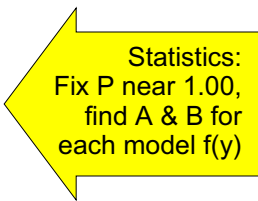
$$\int_A^B f(y) dy = P$$


Figure 2: The Fixed-Coverage Approach of Statistics

Once we know that the critical values A and B will filter out the proportion P of the probable noise, then we can compare our computed statistic with these critical values. If the statistic Y falls between A and B we conclude that our data contain no detectable signals. When our statistic Y falls outside the interval from A to B, we have evidence of a potential signal within our data. Once we have detected a signal, we can then estimate it and assess its practical importance.

This logical sequence for filtering out the noise is repeated over and over with different statistics and different analysis techniques. It describes the fundamental approach used by most statistical techniques. It has a proven track record when used with experimental studies, and it is so fundamental that it becomes automatic for those with statistical training to think in these terms: Fix the value for P, determine the probability model, and then find A and B. *But this is not the only approach to filtering out the noise.*

A DIFFERENT APPROACH

The statistical approach uses a fixed value for P which defines the coverage of the interval A to B. This coverage defines how much of the probable noise gets filtered out. Thus, the statistical approach can be said to use a *fixed-coverage filter*. Inherent in this approach is the necessity of defining a probability model, $f(y)$, to use in finding A and B. Hence, statisticians are by training and inclination prone to think in terms of “What is the probability model?”

While we can usually define a reasonable probability model for various well-known statistics, Y, we will never have enough data in practice to fully specify a probability model for the original data. To get around this problem Walter Shewhart chose to use a *fixed-width filter* rather than a fixed-coverage filter. Instead of beginning with a fixed value for P, he decided to fix the values for A and B instead. With appropriate values for A and B, regardless of what probability model might apply when the process is operated predictably, the value for P will always turn out to be reasonably close to 1.00.

In his search for how to define fixed values for A and B, Shewhart found symmetric, three-sigma limits to be sufficiently general to work as desired. As he observed, it is the potential signals that are of interest, not the noise, so in practice all that we need to know is that P is reasonably close to 1.00. *As long as P is close to 1.00 we will know that we are filtering out almost all of the noise, and the precise value of P will be moot.*

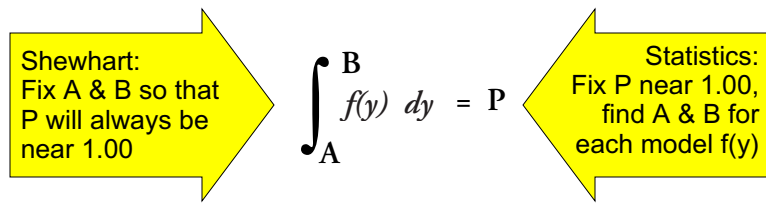


Figure 3: Shewhart's Fixed-Width Approach

Thus, Shewhart's approach using a fixed-width filter is exactly the opposite of the approach used by the fixed-coverage filter. The difference in these two approaches is absolutely fundamental. With Shewhart's fixed-width approach there never was a fixed value for P , and there never will be. This is why those who attempt to assign a fixed-coverage value to a process behavior chart are simply exposing their unexamined assumptions.

THE P VALUES STAY CLOSE TO 1.00

Shewhart's generic, fixed-width limits do not depend upon any specific probability model to work. There is no fixed P value for a point falling inside the limits. In fact, as I argued in last month's column, the notion of computing a P value only makes sense *when the process is being operated predictably*.

However, in order to illustrate the complete generality of Shewhart's approach we will consider the theoretical P values for each of the Western Electric zone tests with each of six different probability models.

Because of the inevitable gaps between theory and practice, theoretical probabilities are rarely meaningful beyond parts per thousand. (For more on this topic see "Invisible Probability Models," *QD* June 4, 2018.) However, in order to show the differences between the models, some of the following theoretical values are given to four decimal places.

DETECTION RULE ONE

A single point falling outside Shewhart's three-sigma limits is taken as evidence of a potential signal of a process change. If we assume that we have a predictable process that is characterized, in turn, by each of the following probability models, then the theoretical probabilities of a false alarm and the P values shown would apply. Figure 4 shows the probabilities of a point falling beyond the three-sigma limits for each tail of each probability model, and the complement of these tail probabilities will be the P value for Rule One for that model.

Theoretical False Alarm Probabilities and P-values
One Point beyond Three Sigma

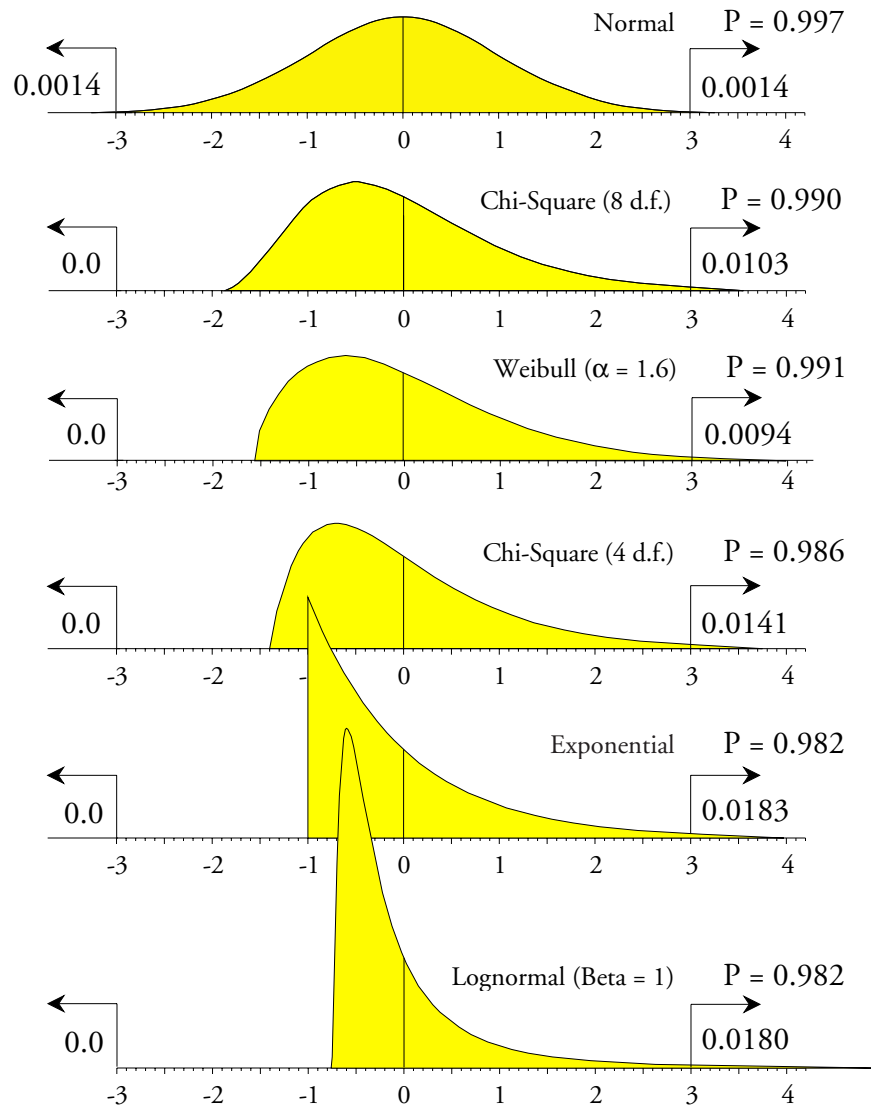


Figure 4: False Alarm Probabilities and Coverages P for Rule One

Statistical procedures with a fixed coverage P in excess of 0.975 are said to be conservative. Figure 4 shows that Shewhart's generic three-sigma, fixed-width limits will result in a conservative analysis regardless of what probability model we may use. Regardless of the shape of your histogram, Shewhart's three-sigma limits will filter out anywhere from 98 percent to 100 percent of the probable noise. Thus, P remains reasonably close to 1.00 and Rule One false alarms remain rare.

DETECTION RULE TWO

When at least two out of three successive values fall more than two sigma units above the

central line, or when at least two out of three successive values fall more than two sigma units below the central line, this *run beyond two-sigma* may be taken as evidence of a process change. In Figure 5 the probabilities of getting at least two out of three values beyond two-sigma are shown for each tail of each probability model. The complement of these tail probabilities will be equal to the P value for Rule Two for that model.

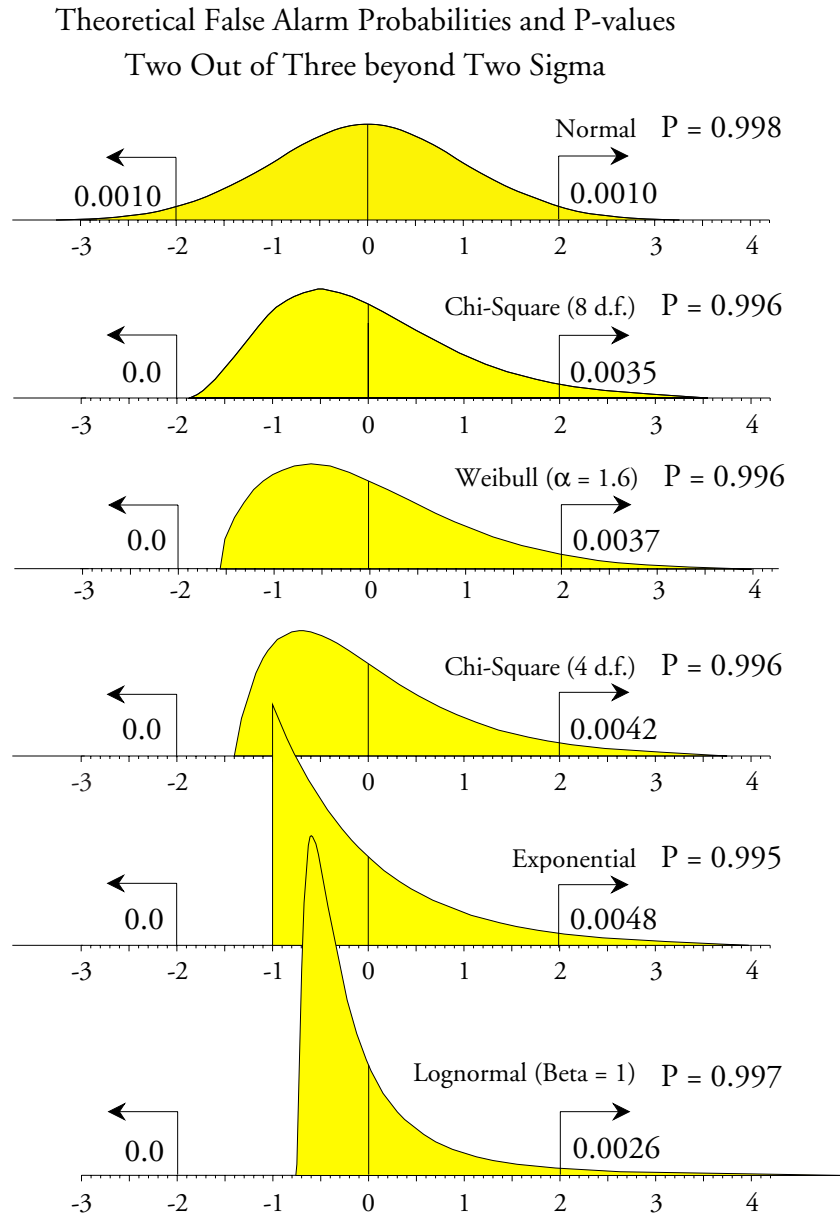


Figure 5: False Alarm Probabilities and Coverages P for Rule Two

Regardless of the shape of your histogram, when your process is operated predictably, the chance that a run beyond two-sigma will be a false alarm is less than one-half percent. Thus, Detection Rule Two is not dependent upon having a normal distribution. While the P values vary, they all remain very close to 1.00.

DETECTION RULE THREE

When at least four out of five successive values fall more than one sigma unit above the central line, or when at least four out of five successive values fall more than one sigma unit below the central line, this *run beyond one-sigma* may be taken as evidence of a process change. In Figure 6 the probabilities of getting at least four out of five values beyond one-sigma are shown for each tail of each probability model. The complement of these tail probabilities will be equal to the P value for Rule Three for that model.

Theoretical False Alarm Probabilities and P-values
Four Out of Five beyond One Sigma

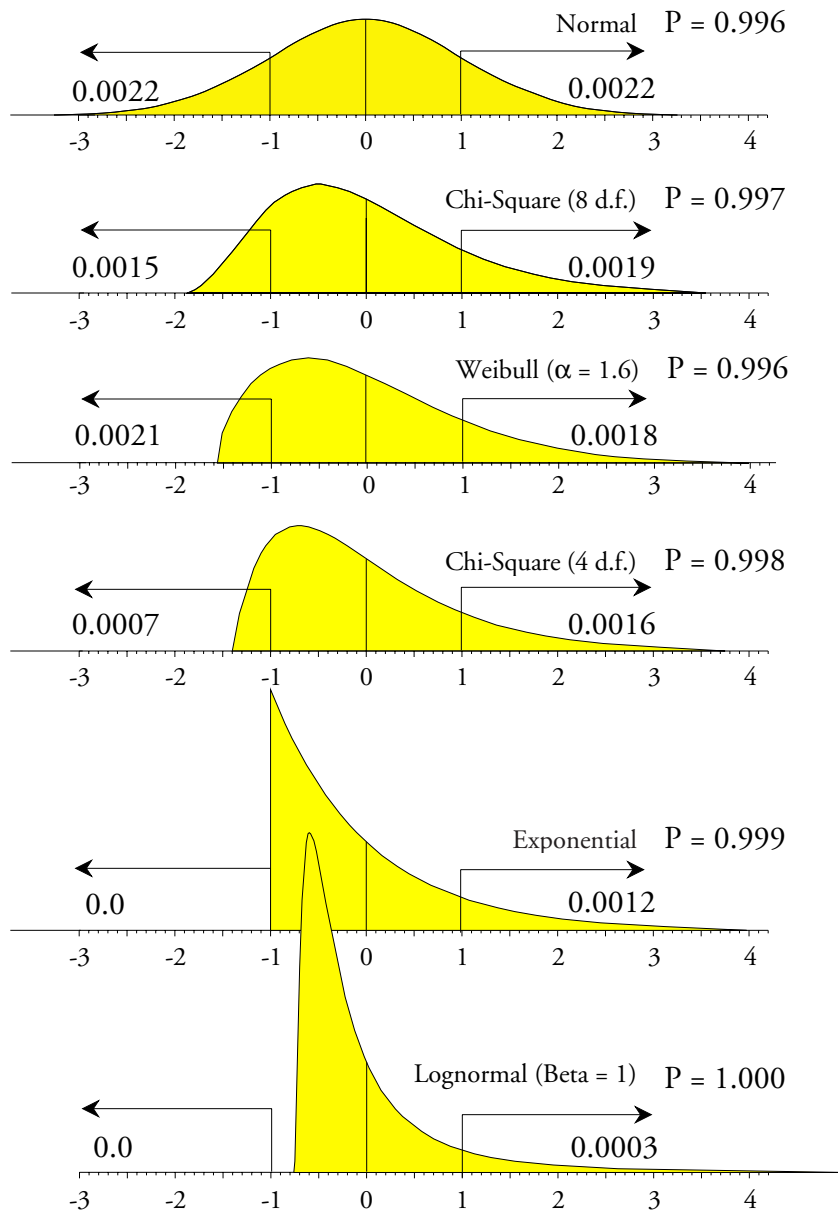


Figure 6: False Alarm Probabilities and Coverages P for Rule Three

Regardless of the shape of your histogram, when your process is operated predictably, the chance that a run beyond one-sigma will be a false alarm is less than one-half percent. Thus, Detection Rule Three is not dependent upon having a normal distribution. While the P values vary, they all remain reasonably close to 1.00.

DETECTION RULE FOUR

When eight successive values all fall on the same side of the central line this *run about the central line* may be taken as evidence of a process change.

Theoretical False Alarm Probabilities and P-values
Eight Successive Values on Same Side of Central Line

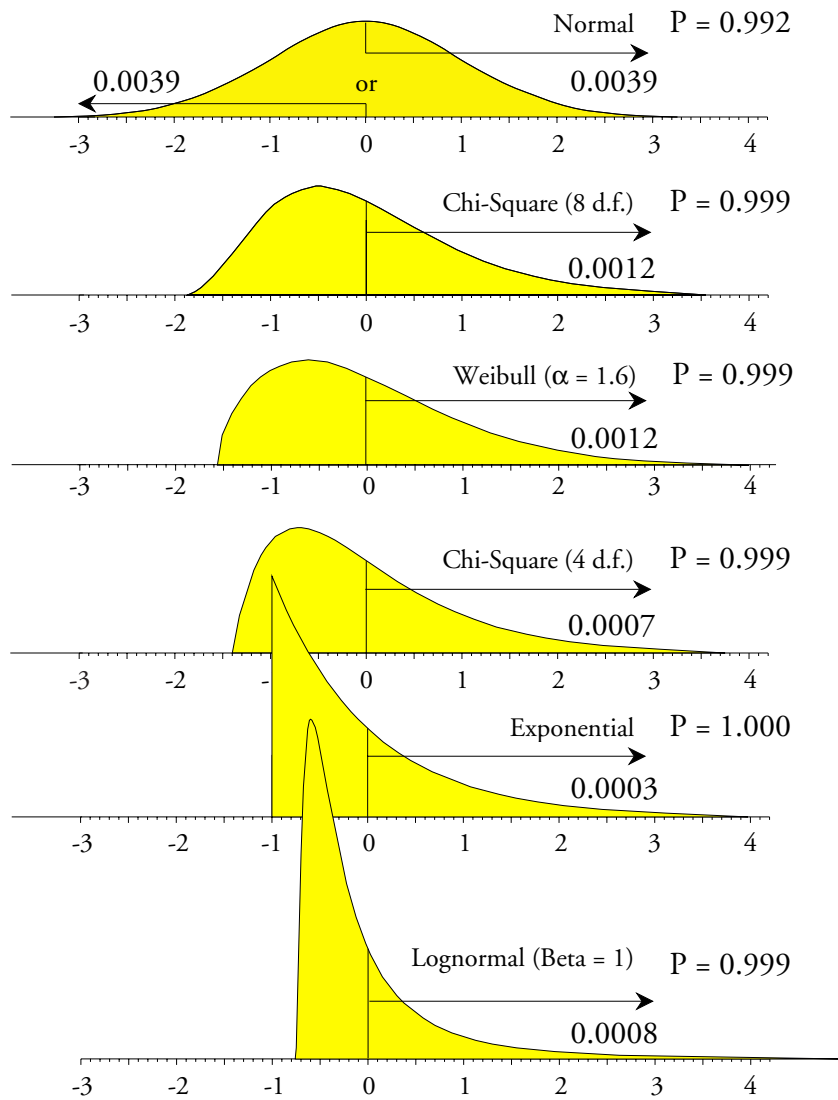


Figure 7: False Alarm Probabilities and Coverages P for Rule Four

Figure 7 shows the probabilities of getting eight successive values on either side of the mean

for the first model and above the mean for each of the last five models. The complement of these run-of-eight probabilities will be equal to the P value for Rule Four for that model.

When your predictable process has a histogram with one tail that is less than two sigma in extent you will almost certainly be operating near a boundary condition which limits your process on that side. In such a case detection rule four only makes sense when it is applied to the unbounded, long-tail side of the histogram. For this reason the bottom five probability models only give the false alarm probabilities for the upper tail. Once again, regardless of the shape of your histogram, when your process is operated predictably, the chance that a run about the central line on the unbounded side will be a false alarm is less than one percent. Thus, Detection Rule Four is not dependent upon having a normal distribution. While the P values vary, they all remain very close to 1.00.

SUMMARY

It is natural for those who think in terms of having a fixed value for P to be concerned about which probability model to use. But Shewhart's fixed-width filter is fundamentally different from the fixed-coverage filter used with traditional statistical techniques. Regardless of the shape of the histogram, Shewhart's generic, symmetric, three-sigma limits and the Western Electric zone tests will work reliably to filter out virtually all of the probable noise so you can detect any potential signals.

The purpose of a process behavior chart is simple: To characterize a process as being operated predictably or unpredictably. The technique of creating and using a process behavior chart is equally simple: Collect data, plot chart, compute limits, plot additional data on chart, and look for assignable causes associated with any signals shown on the chart. For years all of this was successfully done with pencil and paper. The complexity began when software came along and we started adding all the bells and whistles.

You do not have to have a "normal distribution" to use these techniques. Never have, never will. You do not have to achieve some magic value for P in order for the process behavior chart to work. Never have, never will. And you do not have to qualify your data before using a process behavior chart. Never have, never will. Anyone who says anything different is either promoting some special interest, or else has not taken the time, or has not had the opportunity, to learn how Shewhart's approach differs from the traditional statistical approach.

To paraphrase Shewhart, classical statistical techniques start with the assumption that a probability model exists, whereas a process behavior chart starts with the assumption that a probability model does *not* exist. Until you learn the difference, every time you open your mouth, you will be exposing your unexamined assumptions.