Myths About Data Analysis

Donald J. Wheeler

Data analysis is not generally thought of as being simple or easy, but it can be. The first step is to understand that the purpose of data analysis is to separate any signals that may be contained within the data from the noise in the data. Once you have filtered out the noise, anything left over will be your potential signals. The rest is just details.

Of course, the complexity comes when we try to figure out how to filter out the noise. Many different techniques have been developed for doing this. But before you get lost trying to choose between various techniques it is important to remember that you not only have to discover the signals, but you will also have to communicate them to others. For this reason, the best analysis will always be the simplest analysis that allows you to detect the interesting signals.

Whenever you use a conservative technique that filters out virtually all of the noise, then anything left over will be a potential signal. As long as you are finding the interesting signals with this conservative technique you will not need to be concerned about whether you are using the right technique. This is the idea behind the whole field of nonparametric statistics, and it can also be applied to what is, without a doubt, the simplest analysis technique ever invented—the process behavior chart.

The simplicity of the process behavior chart can be deceptive. This is because the simplicity of the charts is based on a completely different concept of data analysis than that which is used for the analysis of experimental data. When someone does not understand the conceptual basis for process behavior charts they are likely to view the simplicity of the charts as something that needs to be fixed. Out of these urges to fix the charts all kinds of myths have sprung up resulting in various levels of complexity and obstacles to the use of one of the most powerful analysis techniques ever invented. The purpose of this paper is to help you avoid this unnecessary complexity.

In order to provide a framework for our discussion of the myths we will begin with some simple distinctions which will organize our thoughts. The first of these is the four questions of data analysis, and the second is the two types of data analysis.

THE FOUR QUESTIONS OF DATA ANALYSIS

The four questions of data analysis are the questions of description, probability, inference, and homogeneity. Any data analyst needs to know how to organize and use these four questions in order to obtain meaningful and correct results.

THE DESCRIPTION QUESTION

Given a collection of numbers, are there arithmetic values
that will summarize the information contained in those numbers
in some meaningful way?

The objective is to capture those aspects of the data that are of interest. In order to be effective a descriptive statistic has to make sense—it has to distill some essential characteristic of the data into a value that is both appropriate and understandable. In every case, this distillation

takes on the form of some arithmetic operation:

$$Data + Arithmetic = Statistic$$

As soon as we have said this, it becomes apparent that the justification for computing any given statistic must come from the nature of the data themselves—it cannot come from the arithmetic, nor can it come from the statistic. If the data are a meaningless collection of values, then the summary statistics will also be meaningless—no arithmetic operation can magically create meaning out of nonsense. Therefore, the meaning of any statistic has to come from the context for the data, while the appropriateness of any statistic will depend upon the use we intend to make of that statistic.

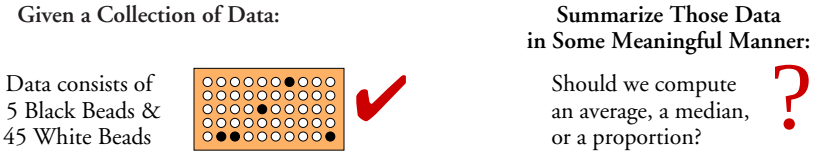

Figure 1: The Question of Description

THE PROBABILITY QUESTION

Given a known universe,
what can we say about samples drawn from this universe?

Here we enter the world of deductive logic, the enumeration of possible outcomes, and mathematical models. For simplicity we usually begin with a universe that consists of a bowl filled with known numbers of black and white beads. We then consider the likelihoods of various sample outcomes that might be drawn from this bowl. This is illustrated in Figure 2.

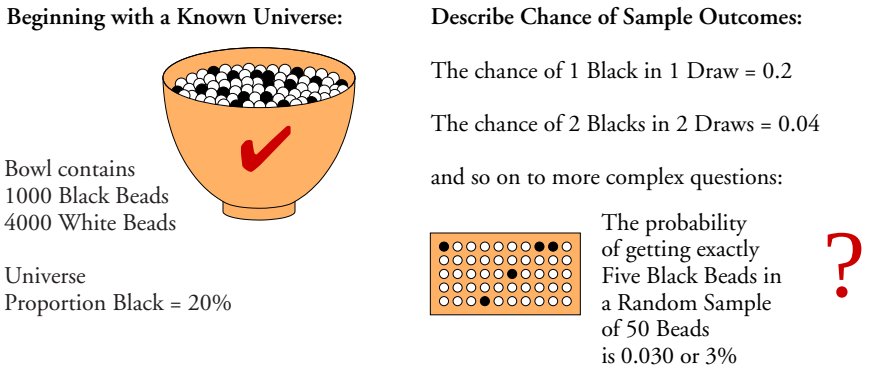

Figure 2: The Question of Probability

While the first two outcomes in Figure 2 have probabilities that are easy to find, the third outcome will require a mathematical model and some deductive logic. The mathematics involved in the calculation of probabilities provides the foundation for the development of analysis techniques. Without the probability models developed here we would be unable to

tackle the more complex problems encountered in practice. Fortunately, while the probability question has to be addressed in developing the theoretical foundations of data analysis, the mysteries of probability theory do not have to be mastered in order to analyze data effectively.

THE INFERENCE QUESTION

Given an unknown universe, and given a sample
that is known to have been drawn from that unknown universe,
and given that we know everything about the sample,
what can we say about the unknown universe?

This is usually thought of as the inverse of the problem addressed by the probability question. Here, it is the sample that is known and the universe that is unknown. Now the argument proceeds from the specific to the general, which makes it inductive in nature. Unfortunately, all inductive inference is fraught with uncertainty.

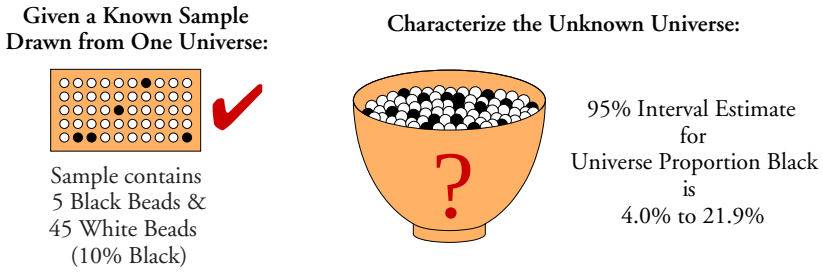

Figure 3: The Question of Inference

A sample result of 5 black beads and 45 white beads corresponds to a 95% Interval Estimate of 4.0% to 21.9% for the proportion of black beads in the bowl. Given this sample result, any percentage from 4.0% to 21.9% is plausible for the bowl.

The inference question is the realm of tests of hypotheses, confidence intervals, and regression. These techniques allow us to estimate and evaluate the parameters of the unknown universe—proportions, means, and standard deviations. Of course such estimates make sense only when our outcomes are all obtained from a single universe. This assumption of a single universe is equivalent to the assumption that the behavior of these outcomes is described by one probability model. Once we have made this assumption, it is possible to use the probability model in reverse—given this outcome, these are the parameter values that are most consistent with the outcome.

While the mathematics of using the probability model in reverse makes everything seem to be rigorous and scientific, you should note that the whole argument begins with an assumption and ends with an indefinite statement. The *assumption* is that all of the outcomes came from the same universe, and the *indefinite statement* is couched in terms of interval estimates. Again, with inductive inference there is not one right answer, but many plausible answers.

THE HOMOGENEITY QUESTION

Given a collection of observations,
is it reasonable to assume that they came from one universe,
or do they show evidence of having come from multiple universes?

To understand the fundamental nature of the homogeneity question, consider what happens if the collection of values does not come from one universe.

Descriptive statistics are built on the assumption that we can use a single value to characterize a single property for a single universe. If the data come from different sources, how can any single value be used to describe what is, in effect, not one property but many? In Figure 4 the sample has 10 percent black. But if the 50 beads are the result of three separate draws from the three bowls at the bottom of Figure 4, each of which has a different number of black beads, which bowl is characterized by the sample result?

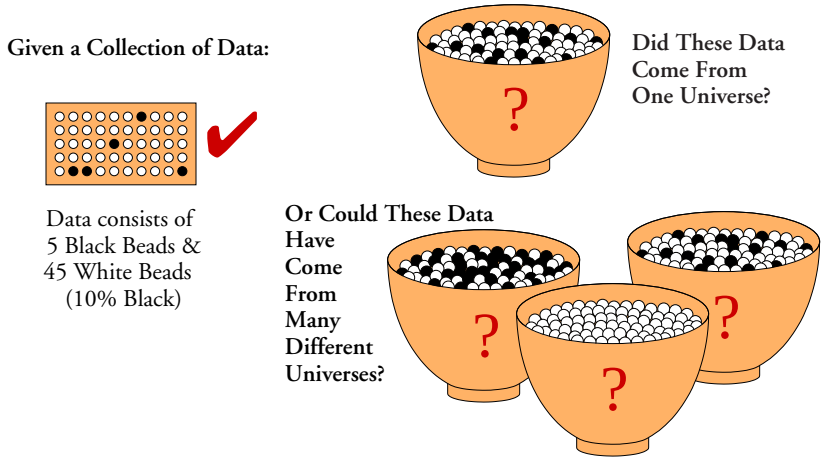

Figure 4: The Question of Homogeneity

Probability theory is focused on what happens to samples drawn from a known universe. If the data happen to come from different sources, then there are multiple universes with different probability models. If you cannot answer the homogeneity question, then you will not know if you have one probability model or many.

Statistical inference assumes that you have a sample that is known to have come from one universe. If the data come from different sources, what does your interval estimate represent? Which of the multiple universes does it characterize?

Therefore, before you can use the structure and techniques developed to answer the first three problems, you will need to examine your data for evidence of that homogeneity which is implicitly assumed by the use of descriptive statistics, by the concepts of probability theory, and by the techniques of statistical inference. While the implicit assumption of homogeneity is part of everything we do in traditional statistics classes, it becomes a real obstacle whenever we try to analyze data.

When we find evidence of a changing universe in a situation where there should be only one

universe we will be unable to learn anything from descriptive statistics. When the universe is changing we cannot gain from statistical inference, nor can we make predictions using probability theory. Any lack of homogeneity in our collection of values completely undermines the techniques developed to answer each of the first three questions. The lack of homogeneity is a signal that unknown things are happening, and until we discover what is happening and remove its causes, we will continue to suffer the consequences. Computations cannot remedy the problem of a lack of homogeneity; action is required.

How can we answer the homogeneity question? We can either *assume* that our data possess the appropriate homogeneity, or we can *examine* them for signs of nonhomogeneity. Since anomalous things happen in even the most carefully controlled experiments, prudence demands that we choose the second course. And the primary tool for examining a collection of values for homogeneity is the process behavior chart.

To examine our data for signs of nonhomogeneity we begin with the tentative assumption that the data are homogeneous and then look for evidence that is inconsistent with this assumption. When we reject the assumption of homogeneity we will have strong evidence which will justify taking action to remedy the situation. When we fail to reject the assumption of homogeneity we will know that any nonhomogeneity present is below the level of detection. While this is a weak result, we will at least have a reasonable basis for proceeding with estimation and prediction.

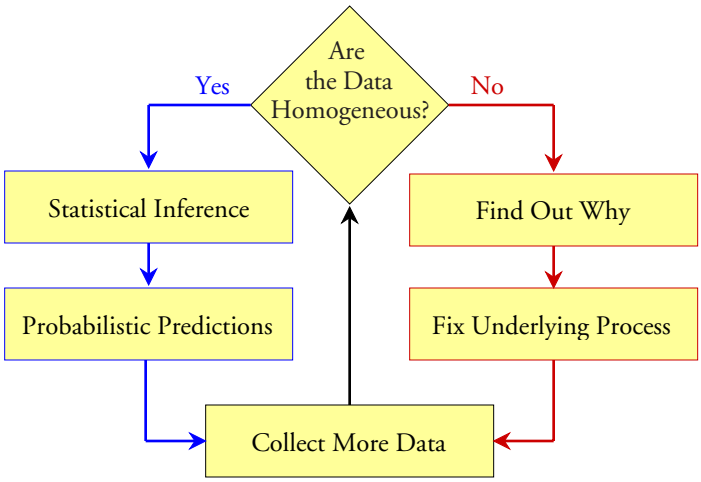

Figure 5: Homogeneity is the Primary Question of Data Analysis

Thus, for practitioners the first question must always be the question of homogeneity. Given a collection of data, did these data come from one universe? In fact, is it reasonable to assume that there is one universe? Only after this fundamental question has been addressed does the practitioner know how to proceed. If the assumption of a single universe is reasonable, then the techniques of statistical inference may be used to characterize that universe, and then, with reasonable estimates of the parameters, probability models may be used to make predictions. But if the assumption of a single universe is not justified, the practitioner needs to find out why.

This is not the way classes in statistics are taught, but it is the way you have to do data analysis. Look at your data on a process behavior chart. If there are surprises in your data, and

there often will be, then learn from these surprises. If there are no surprises, then you may proceed to analyze your data as if they came from a single universe. Any attempt to analyze data that does not begin by addressing the question of homogeneity is flawed.

At this point you might well ask why the question of homogeneity was not covered in your statistics class. The answer lies in the difference between the two types of data analysis.

TWO TYPES OF DATA ANALYSIS

In most textbooks and classes in statistics the emphasis is upon the analysis of experimental data. Since part of performing an experiment is taking the effort to assure homogeneity within each treatment condition, the assumption of homogeneity is simply taken as a given. However, since most industrial and business data are not experimental data, we risk making a serious mistake if we do not give the question of homogeneity some careful consideration.

Data that are obtained as a by-product of operations are known as observational data. Such data track what is happening during routine operations. Here we typically expect to find a stream of data that will show a process operating in a steady state. We do not expect to find signals of unknown changes. Therefore, if we find such a signal, we will want to be sure that the change is real before we sound the alarm that a change has occurred.

In contrast to observational data, experimental data are specifically collected under two or more conditions where the purpose of the experiment is to detect the differences between the conditions studied. Hence there will only be a finite amount of data available for analysis, and we will analyze these data all together at the end of the experiment. In this setting we do not want to fail to find the expected signals within the experimental data. To this end we are generally willing to accept a greater risk of getting a false alarm than we are willing to accept with observational data.

Observational Data	**Experimental Data**
Additional Data Available	Fixed Amount of Data
One Condition Present	Two or More Conditions Present
Should Be No Signals	Should Be Some Signals
Sequential Analysis Required	One-Time Analysis
Conservative Analysis Used	Traditional or Exploratory Analysis Used

Figure 6: Differences in the Analysis of Observational Data and Experimental Data

Therefore, the analysis of observational data differs from the analysis of experimental data in five major ways as summarized in Figure 6. With observational data we can always wait for more data rather than having to make a decision using a fixed amount of data. With observational data there will generally be one condition present rather than two of more conditions. With observational data there should be no signals of a change, while experimental data are collected in order to verify some difference. With observational data you will need a sequential procedure rather than a one-time analysis. And with observational data you will want each individual analysis to be performed in a conservative manner, while with an experiment you will use a traditional, or even an exploratory, analysis.

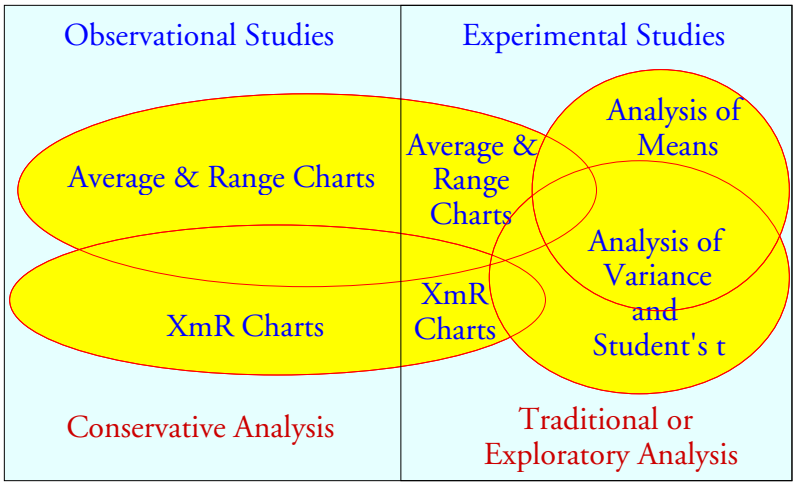

Figure 7: Techniques for the Analysis of Observational and Experimental Data

These differences in how we analyze observational data and experimental data result in differences in how we approach the analysis and require different techniques. In traditional classes in statistics you will learn the approaches and techniques that are suitable for the analysis of experimental data. However, if you try to use these approaches and techniques with observational data you are likely to make serious mistakes in your analysis. The first question of data analysis is not a question of descriptive statistics. It is not a question about what probability model to use. It is not a question of estimation or inference. And it is not a question about what critical values to use. The first question of data analysis is the question of homogeneity, and Shewhart showed us how to answer that question.

As we will see, most of the myths about data analysis and process behavior charts come from a failure to understand the relationships between the four questions of data analysis and the two different types of data analysis.

MYTH ONE

The standard deviation statistic is more efficient than the range and therefore we should use the standard deviation statistic when computing limits for a process behavior chart.

As is often the case with complex subjects, this myth begins with a statement that is partially true, and then it applies this statement in an incorrect manner. The global standard deviation statistic that everyone learns about in their introductory statistics class is a descriptive measure of dispersion. To understand just what this statistics does we have to understand how descriptive statistics work.

The purpose of descriptive statistics is to summarize the data. The average is commonly used to characterize the location of the data, while the standard deviation statistic is commonly used to characterize the dispersion of the data. The way that these two descriptive statistics work can be summarized by a set of nested intervals centered on the average with radii equal to multiples of the standard deviation statistic. Each of these intervals will bracket certain

proportions of the data. In the following figures we will show these coverages for six data sets.

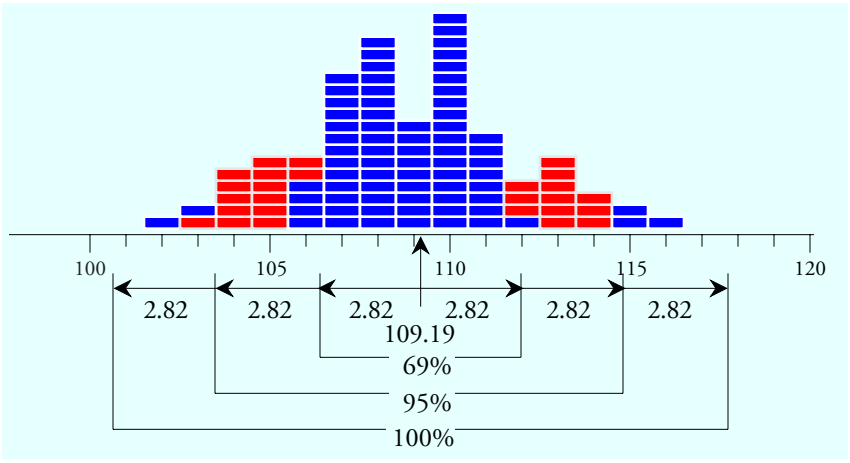

Figure 8: The Wire Length Data

In Figure 8 we see the 100 observations of the Wire Length Data. The first interval brackets 69% of these data. The second interval brackets 95% of these data. The third interval brackets all of these data.

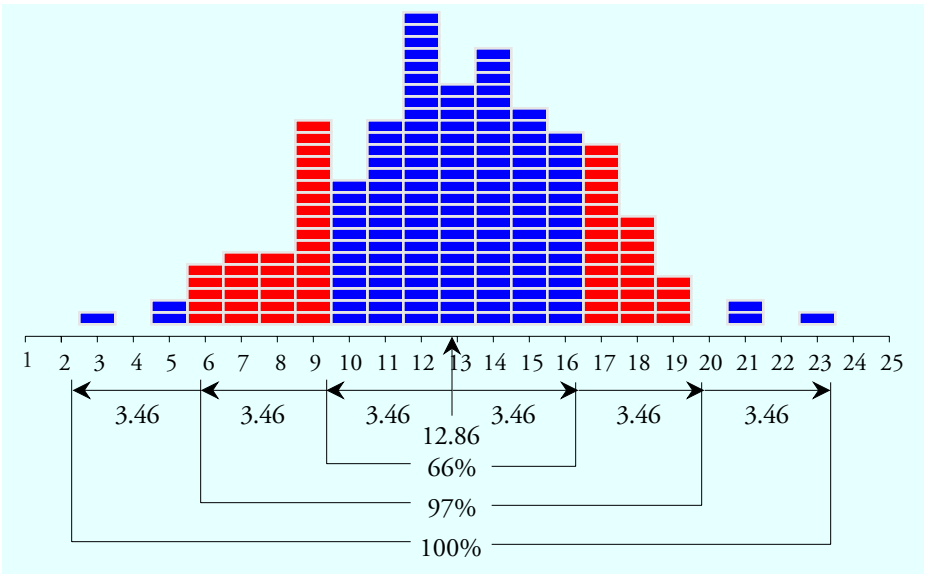

Figure 9: The Bead Board Data

In Figure 9 we see the 200 observations of the Bead Board Data. The first interval brackets 66% of these data. The second interval brackets 97% of these data. The third interval brackets all of these data.

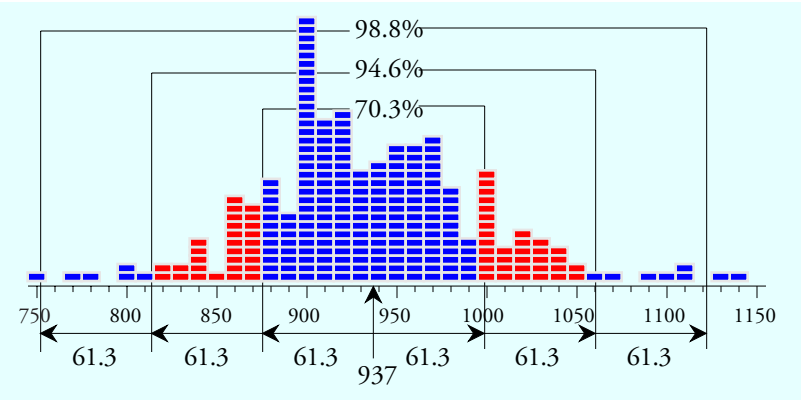

Figure 10: The Batch Weight Data

In Figure 10 we see the 259 observations of the Batch Weight Data. The first interval brackets 70% of these data. The second interval brackets 95% of these data. The third interval brackets 99% of these data.

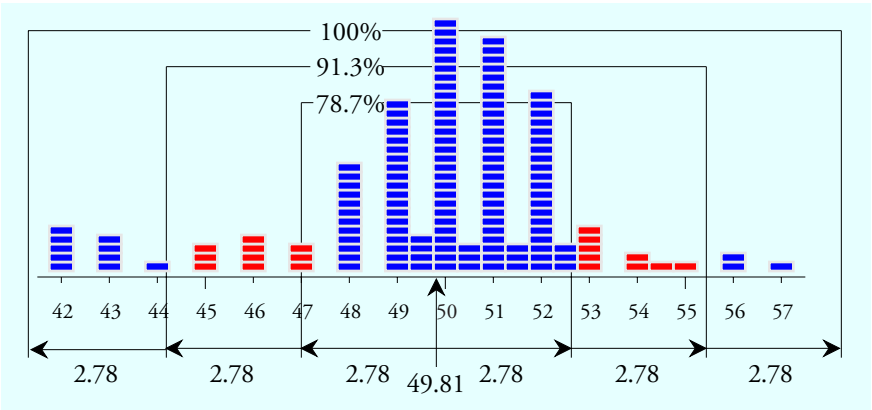

Figure 11: The Camshaft Bearing Diameter Data

In Figure 11 we see the 150 observations of the Camshaft Bearing Diameter Data. The first interval brackets 79% of these data. The second interval brackets 91% of these data. The third interval brackets all of these data.

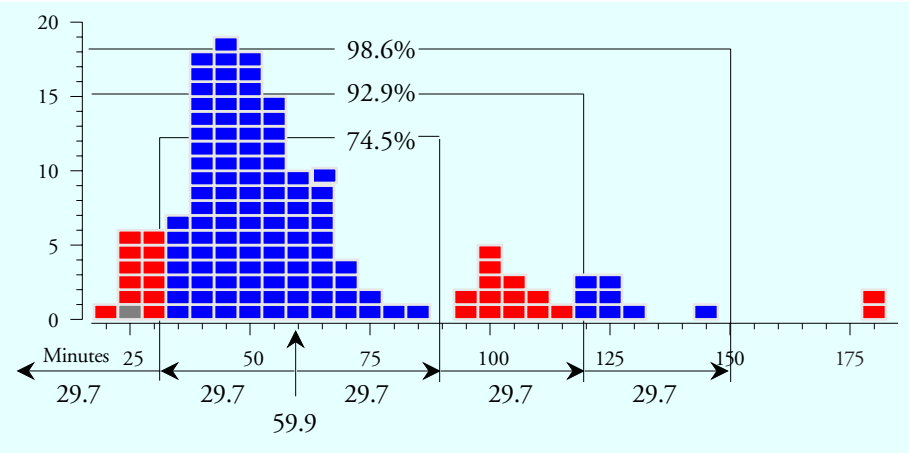

Figure 12: The Hot Metal Delivery Time Data

In Figure 12 we see the 141 observations of the Hot Metal Delivery Time Data. The first interval brackets 75% of these data. The second interval brackets 93% of these data. The third interval brackets 99% of these data.

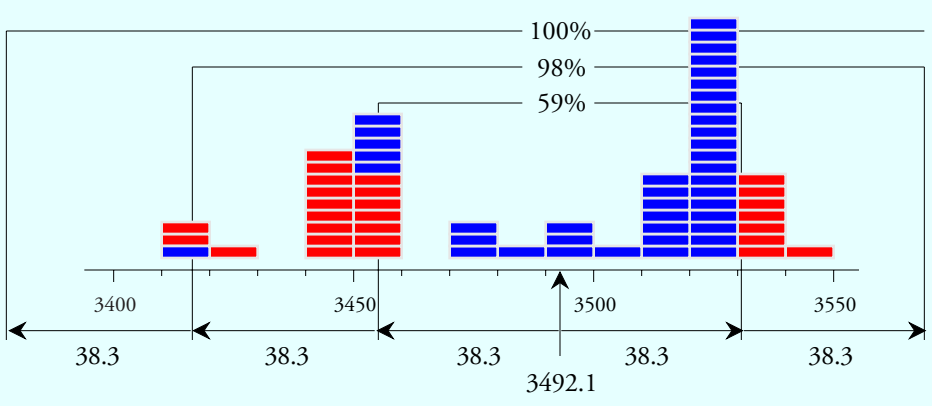

Figure 13: The Creel Yield Data

In Figure 13 we see the 68 observations of the Creel Yield Data. The first interval brackets 59% of these data. The second interval brackets 98% of these data. The third interval brackets all of these data.

Table 1 summarizes the coverages. There we see that the interval defined by the average plus or minus the standard deviation statistic will bracket the bulk of the data. Here this interval covers from 59% to 79% of the data. The interval defined by the average plus or minus twice the standard deviation statistic will bracket at least 90% of the data. And the interval defined by the average plus or minus three times the standard deviation statistic will bracket virtually all of the data, regardless of how ragged the histogram may be.

Table 1: What Descriptive Statistics Do

	Coverage *Avg. ± s*	Coverage *Avg ± 2s*	Coverage *Avg. ± 3s*
Wire Length Data	69%	95%	100%
Bead Board Data	66%	97%	100%
Batch Weight Data	70.3%	94.6%	98.8%
Bearing Diameter Data	78.7%	91.3%	100%
Hot Metal Delivery Time Data	74.5%	92.9%	98.6%
Creel Yield Data	59%	98%	100%

The intervals in Table 1 show how descriptive statistics summarize the data. This is what they are supposed to do. These descriptive statistics are focused solely upon the data. They do not attempt to characterize the underlying process that generated the data. In effect, the computations implicitly assume that the underlying process is operating steady-state and that, as a consequence, the data are completely homogeneous. This assumption of homogeneity is implicit in every descriptive statistic and is a consequence of these statistics being computed globally. As a result, descriptive statistics do not provide a way to check on the assumption of homogeneity.

So how can we use data to learn about the underlying process that generates those data? This is going to require more than descriptive statistics. In this case we are going to have to use some localized measure of dispersion. When the process is changing the data will be nonhomogeneous. When the data are not homogeneous, the global standard deviation statistic will be inflated by this lack of homogeneity. However, localized measures of dispersion will not be inflated in the same way, or to the same extent, as the global standard deviation statistic. Thus, to examine the data for possible evidence of a lack of homogeneity, and to thereby characterize the underlying process, we have to use some localized measures of dispersion rather than a global measure of dispersion such as the global standard deviation statistic.

When the data possess a time order we can arrange the data in this time order and find the differences between successive values. These differences summarize the local, short-term variation within the data and are commonly called moving ranges. These moving ranges are then summarized by computing either the average moving range or a median moving range. When this summary is divided by the appropriate bias correction factor we have a local measure of dispersion which shall be designated by the notation *Sigma(X)*:

$$Sigma(X) = \frac{Average\ Moving\ Range}{1.128}$$

$$\text{or} \quad Sigma(X) = \frac{Median\ Moving\ Range}{0.954}$$

Alternately, when the data have been organized into rational subgroups of size n, we compute the range for each subgroup (or possibly the standard deviation statistic for each subgroup). Then we average these within-subgroup measures of dispersion and divide by the appropriate bias correction factor to obtain a local measure of dispersion:

$$Sigma(X) = \frac{Average\ Range}{d_2}$$

$$\text{or} \quad Sigma(X) = \frac{Average\ Standard\ Deviation\ Statistic}{c_4}$$

Figure 14 uses a collection of nine values to illustrate the difference between the global and local approaches to characterizing the dispersion of a data set. The global approach effectively collects all the data together and computes a standard deviation statistic. The local approach organizes the data in time order, computes the differences between the successive values, and uses the average moving range to characterize the dispersion.

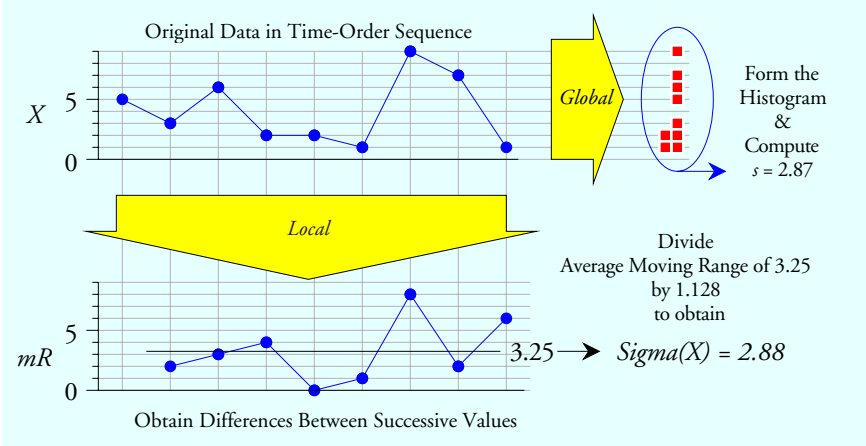

Figure 14: Global and Local Measures of Dispersion

When the data are homogeneous the global measure of dispersion and the local measure of dispersion will be quite similar. However, when the data are not homogeneous the global measure of dispersion will be inflated relative to the local measure of dispersion. In practice we do not actually compare the local and global measures of dispersion. Instead we use the local measure of dispersion to compute *three-sigma limits* and then we look for points outside these limits.

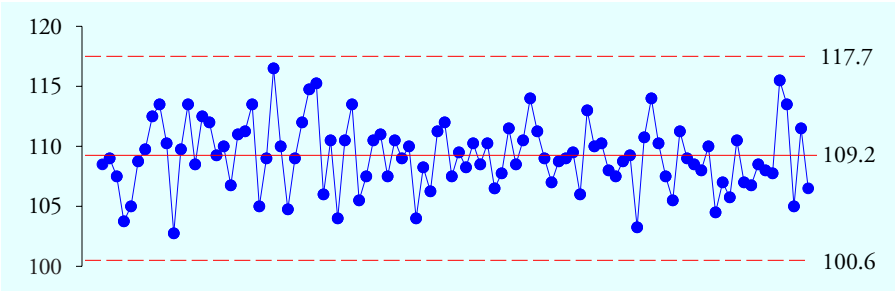

Figure 15: *X* Chart for the Wire Length Data

For the Wire Length Data the global standard deviation statistic was $s = 2.82$ and the three-standard-deviation interval bracketed 100% of the data. Here the value for *Sigma(X)* is 2.85, and the *X* chart has 100% of the data within the three-sigma limits. Thus, the data appear to be homogeneous, and therefore it is reasonable to assume that the underlying process was being operated predictably.

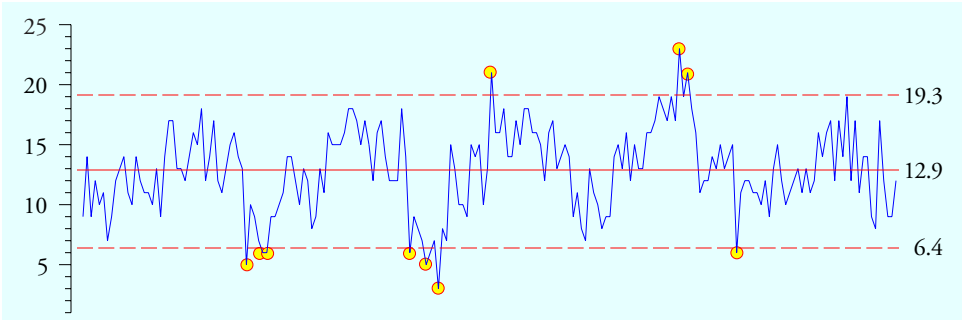

Figure 16: *X* Chart for the Bead Board Data

For the Bead Board Data the global standard deviation statistic was s = 3.46 and the three-standard-deviation interval bracketed 100% of the data. Here the value for *Sigma(X)* is 2.14, and the *X* chart has 11 out of 200 values outside the three-sigma limits. Thus, these data appear to be nonhomogeneous, and we must conclude that the underlying process was changing.

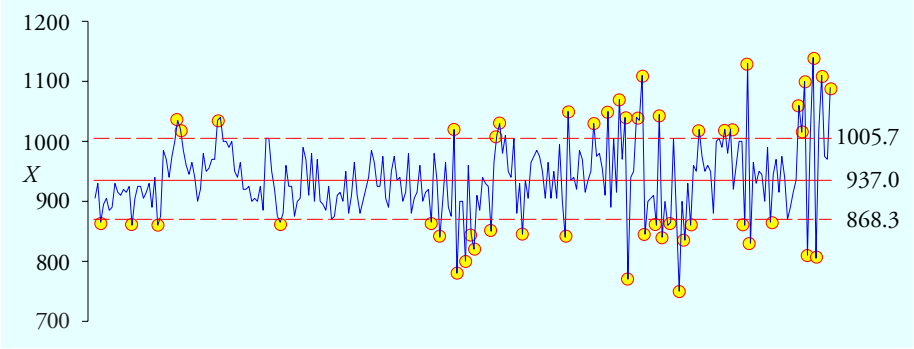

Figure 17: *X* Chart for the Batch Weight Data

For the Batch Weight Data the global standard deviation statistic was s = 61.3 and the three-standard-deviation interval bracketed 99% of the data. Here the value for *Sigma(X)* is 22.9, and the *X* chart has 56 out of 259 values outside the three-sigma limits. Thus, these data appear to be nonhomogeneous, and we must conclude that the underlying process was changing.

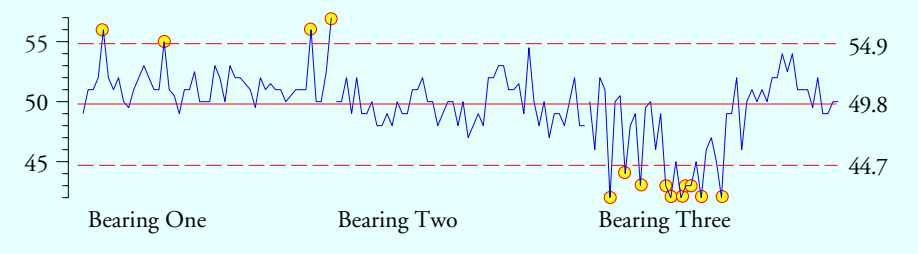

Figure 18: *X* Chart for the Camshaft Bearing Diameter Data

For the Camshaft Bearing Diameter Data the global standard deviation statistic was s = 2.78 and the three-standard-deviation interval bracketed 100% of the data. Here the value for

Sigma(X) is 1.70, and the *X* chart has 14 out of 150 values outside the three-sigma limits. Thus, these data appear to be nonhomogeneous, and we must conclude that the underlying process was changing.

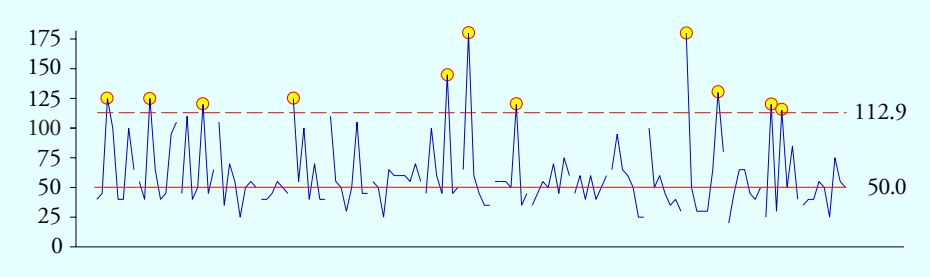

Figure 19: *X* Chart for the Hot Metal Delivery Time Data

For the Hot Metal Delivery Time Data the global standard deviation statistic was $s = 29.7$ and the three-standard-deviation interval bracketed 99% of the data. Here the value for *Sigma(X)* is 20.96, and the *X* chart has 11 out of 141 values outside the three-sigma limits. Thus, these data appear to be nonhomogeneous, and we must conclude that the underlying process was changing.

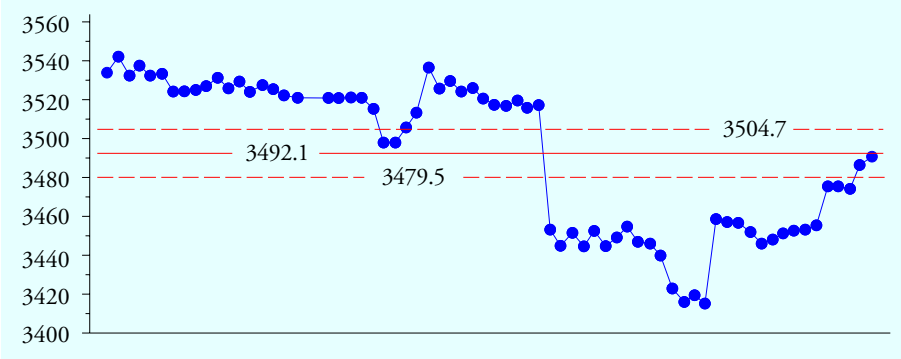

Figure 20: *X* Chart for the Creel Yield Data

For the Creel Yield Data the global standard deviation statistic was $s = 38.3$ and the three-standard-deviation interval bracketed 100% of the data. Here the value for *Sigma(X)* is 4.2, and the *X* chart has 64 out of 68 values outside the three-sigma limits. Thus, these data appear to be nonhomogeneous, and we must conclude that the underlying process was changing.

Table 2 summarizes the performance of three-standard-deviation limits and three-sigma limits. Only one of our six data sets was homogeneous. The three-standard-deviation limits computed using the descriptive statistics bracketed all of the data for three of the five sets of nonhomogeneous data. In contrast to this, the three-sigma limits detected all five of the nonhomogeneous data sets. This is why the only reliable way to check for homogeneity is to use localized measures of dispersion. As we have seen, the question of homogeneity is the fundamental question of data analysis. And any attempt to answer this question will require the use of local measures of dispersion.

Table 2: Points Outside Global and Local Limits

	Global		Local		Homogeneous?
	Avg. ± 3 s		*Avg ± 3 Sigma(X)*		
Wire Length Data	0 of 100	0%	0 of 100	0%	Yes
Bead Board Data	0 of 200	0%	11 of 200	5.5%	No
Batch Weight Data	3 of 259	1.2%	56 of 259	21.6%	No
Bearing Diameter Data	0 of 150	0%	14 of 150	9.3%	No
Hot Metal Delivery Time Data	2 of 141	1.4%	11 of 141	7.8%	No
Creel Yield Data	0 of 68	0%	64 of 68	94.1%	No

So the truth about Myth One is this: *How* you compute the measure of dispersion is more important than *which* measure of dispersion is used. Correctly computed limits will always use a *within-subgroup measure of dispersion*. It is never correct to compute three-standard-deviation limits using some global measure of dispersion. Thus, there are correct and incorrect ways of computing limits for a process behavior chart. Unfortunately, every piece of software available today will allow you to compute the limits incorrectly. This is why, until you know the difference between the correct and incorrect ways of computing limits, you are not competent to use your software.

MYTH TWO

It has been said that the data must be normally distributed before they can be placed on a process behavior chart.

In discussing this myth some historical background may be helpful. Walter Shewhart published his "*Economic Control of Quality of Manufactured Product*" in 1931. When the British statistician E. S. Pearson read Shewhart's book he immediately felt that there were gaps in Shewhart's approach, and so he set out to fill in these perceived gaps. The result was Pearson's 1935 book entitled "*The Application of Statistical Methods to Industrial Standardization and Quality Control.*" In this book Pearson wrote on page 34: "Statistical methods and tables are available to test whether the assumption is justified that the variation in a certain measured characteristic may be represented by the Normal curve."

After reading Pearson's book, Shewhart gave a series of lectures that W. Edwards Deming edited into Shewhart's 1939 book, "*Statistical Method from the Viewpoint of Quality Control.*" In choosing this title Shewhart effectively reversed Pearson's title to emphasize that his approach solved a real problem rather than being a collection of techniques looking for an application. On page 54 of this second book Shewhart wrote: "*we are not concerned with the functional form of the universe, but merely with the assumption that a universe exists.* [Italics in the original]." Here Shewhart went to the heart of the matter. While Pearson essentially assumed that the use of a probability model would always be justified, Shewhart created a technique to examine this assumption. The question addressed by a process behavior chart is more basic than "What is the shape of the histogram?" or "What is the probability model?" It has to do with whether we can meaningfully use *any* probability model with our data.

Shewhart then went on to note that having a symmetric, bell-shaped histogram is neither a prerequisite for the use of a process behavior chart, nor is it a consequence of having a predictable process. Figure 21 shows Shewhart's Figure 9 from the 1931 book. He characterized these data as

"at least approximately [in] a state of control." This skewed histogram is certainly not one that anyone would claim to be "normally distributed." So, while Shewhart had thoroughly examined this topic in his 1931 book, his approach was so different from traditional statistical thinking that Pearson and countless others (including this author on his first reading) completely missed this crucial point.

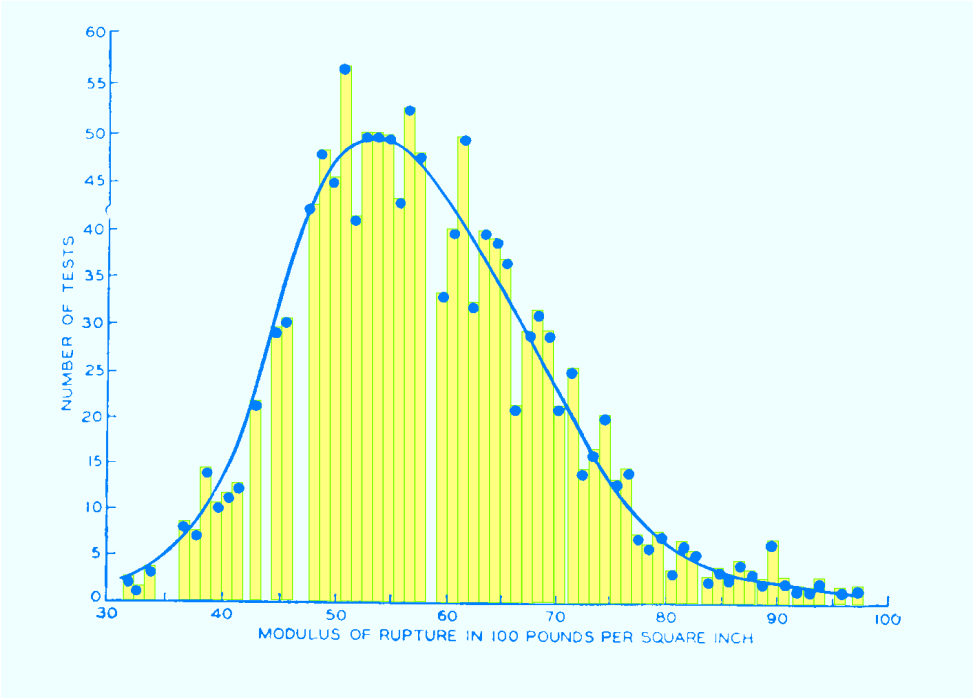

Figure 21: Shewhart's Figure 9: Variability in Modulus of Rupture of Clear Specimens of Green Sitka Spruce Typical of the Statistical Nature of Physical Properties

To begin to understand how a process behavior chart can be used with all sorts of data we need to begin with a simple equation from page 275 of Shewhart's 1931 book:

$$\int_{A}^{B} f(x)\ dx \quad = P$$

Shewhart described two completely different approaches this equation. The first of these approaches I call the statistical approach since it describes how we approach statistical inference:

1. Choose an appropriate probability model $f(x)$ to use;
2. Choose some small risk of a false alarm ($1 - P$) to use;
3. Find the exact critical values A and B for the selected model $f(x)$ that correspond to the selected risk of a false alarm ($1 - P$);
4. Then use these critical values in your analysis.

While this approach makes sense when working with *functions* of the data (i.e. statistics) for which we know the appropriate probability model, it encounters a huge problem when it is applied to the original data. As Shewhart pointed out, we will *never* have enough data to uniquely identify a specific probability model for the original data. In the mathematical sense all

probability models are limiting functions for infinite sequences of random variables. This means that they can never be said to apply to any finite portion of that sequence. This is why any assumption of a probability model for the original data is just that—an assumption that cannot be verified in practice. (While lack-of-fit tests will sometimes allow us to reject this assumption, they can never verify an assumed probability model.)

So what are we to do when we try to analyze data? Shewhart suggested a different approach for the analysis of original data. Shewhart's approach to the equation above was:

1. Choose some *generic* critical values A and B for which
2. the risk of a false alarm ($1 - P$) *will be reasonably small*
3. *regardless* of what probability model $f(x)$ we might choose,
4. then use these generic critical values in your analysis.

This approach changes what is fixed and what is allowed to vary. With the statistical approach the alpha-level ($1 - P$) is fixed, and the critical values vary to match the specific probability model. With Shewhart's approach it is the critical values that are fixed (three-sigma limits) and the alpha-level that is allowed to vary. This complete reversal of the statistical approach is what makes Shewhart's approach so hard for those with statistical training to understand.

Since Shewhart's approach is so radical it might help to look at how it works with different probability models. First we begin with a selection of six probability models ranging from the uniform to the exponential. Figure 22 shows the values for P that correspond to three-sigma limits for each of these six distributions. Traditionally, an analysis is said to be conservative whenever P exceeds 97.5%. Figure 22 shows that Shewhart's three-sigma limits will provide a conservative analysis over a wide range of probability models.

Next we look at more than 1100 probability models. Since it is impractical to plot over 1100 probability distributions, we use the shape characterization plane. Here each probability model is represented as a point by using the skewness and kurtosis parameters for that distribution. Thus, the six probability models from Figure 22 become six points in Figure 23. When we compute the value for P for each of the 1143 distributions shown and plot these P values on the shape characterization plane we end up with the contour map shown in Figure 23.

Probability models to the left of the upper red line will have a mound shaped distribution. There we see that *any mound shaped probability model will have a P value in excess of* 98%.

Probability models in between the red lines will have a J-shaped distribution. Figure 23 shows that all but the most extreme of the J-shaped probability models will have a P value in excess of 97.5%. Since virtually all data sets have a mound shaped or J-shaped histogram, Figure 23 shows that Shewhart's three-sigma limits will provide a conservative analysis in virtually every case. Once you have detected potential signals using a conservative analysis, you no longer need to worry about the appropriateness of your analysis, but rather need to get busy finding the assignable causes of the process changes.

Given the difference between the statistics approach and Shewhart's approach you can begin to see why Pearson and others have been concerned with the probability model $f(x)$, why they have sought to maintain a fixed alpha level $(1 - P)$, and why they have been obsessed with the computation of exact values for A and B. And more recently, you can see how others have become obsessed with transforming the data prior to placing them on a process behavior chart.

Their presuppositions prevent them from understanding how Shewhart's choice of three-sigma limits is completely independent of the choice of a probability model. In fact, to their way of thinking, you cannot even get started without a probability model. Hence, as people keep misunderstanding the basis for process behavior charts, they continue to recreate Myth Two.

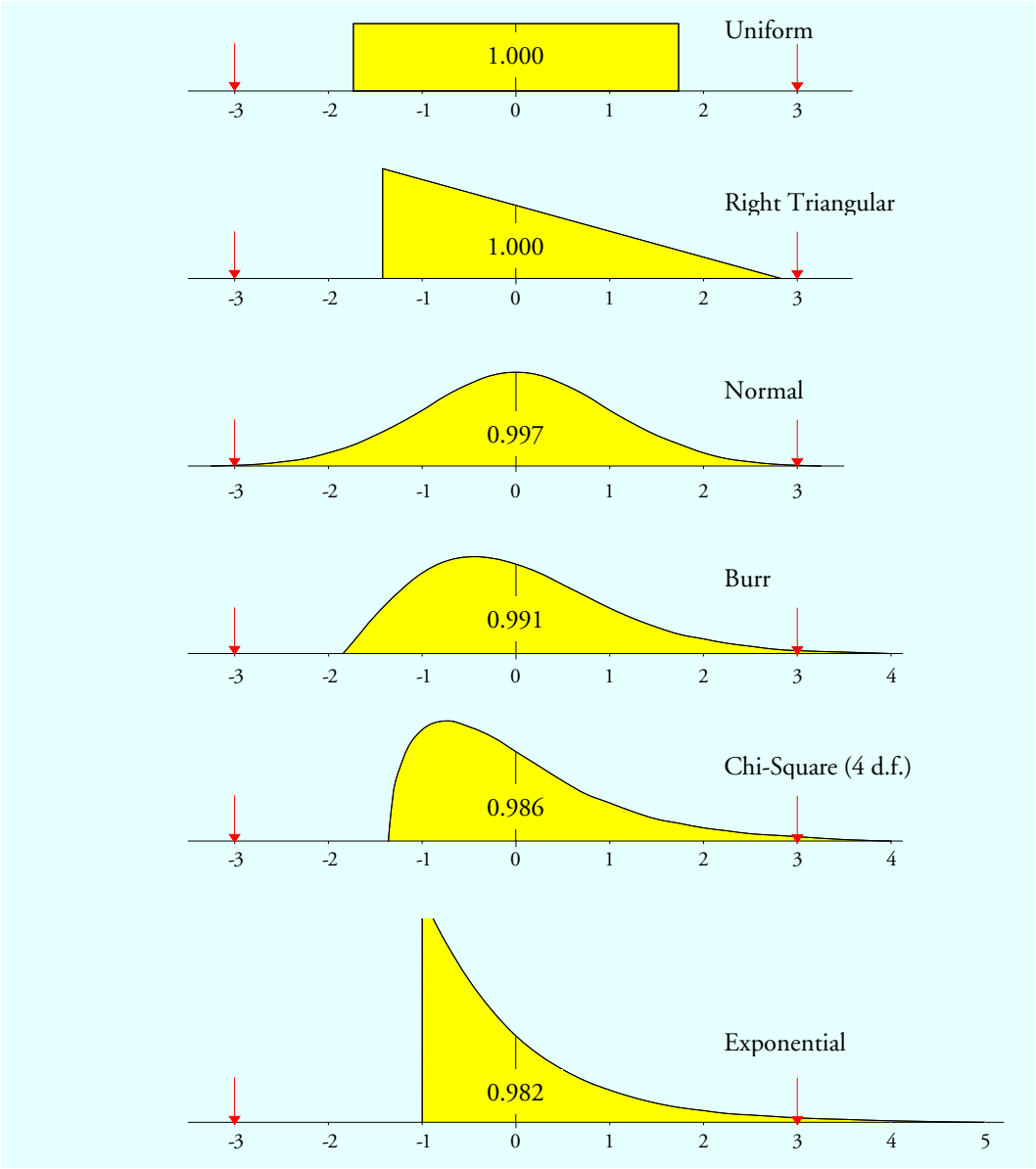

Figure 22: The Coverages *P* for Three Sigma Limits for Six Probability Models

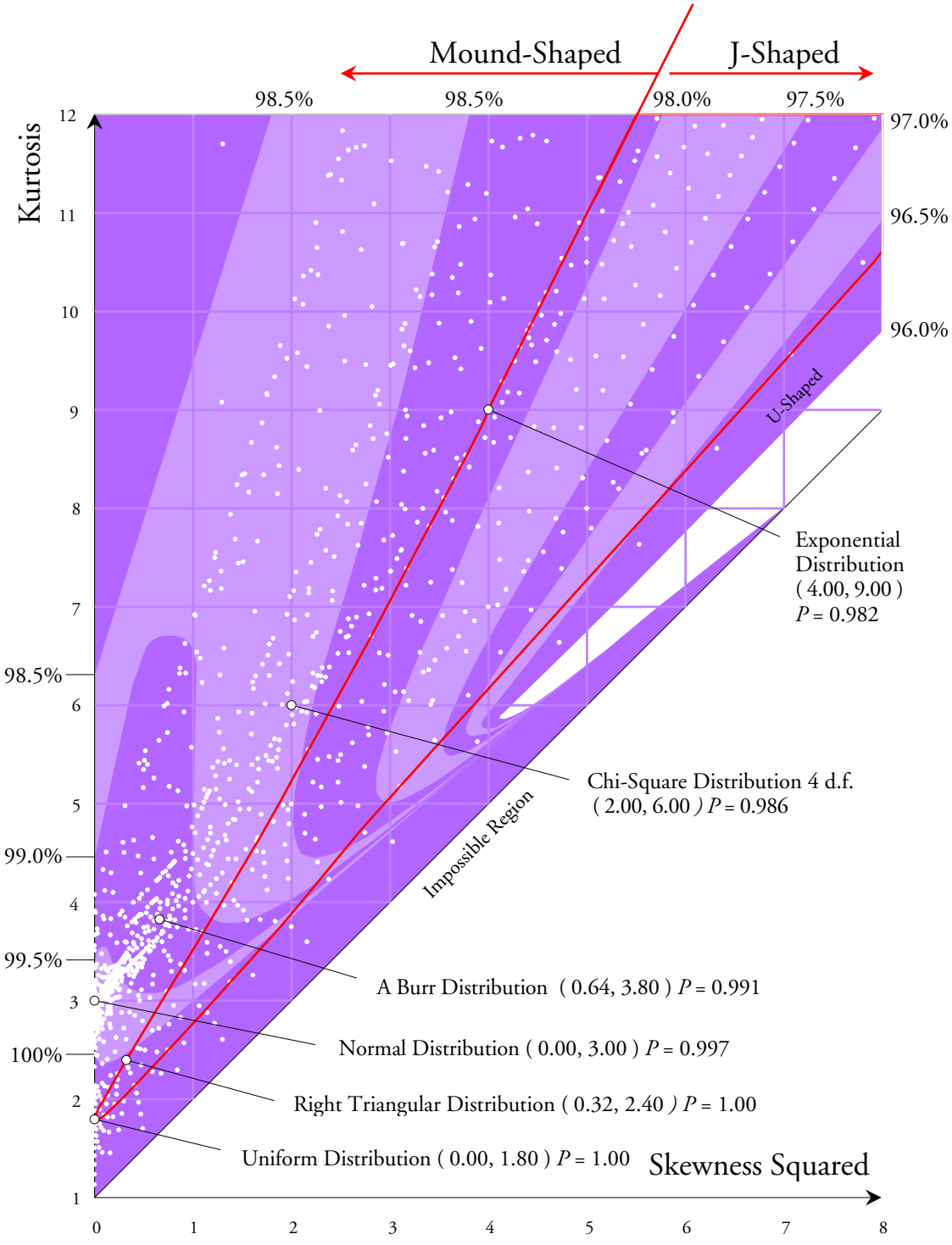

Figure 23: The Coverage Contours for Three Sigma Limits in the Shape Characterization Plane

MYTH THREE

It has been said that process behavior charts work because of the central limit theorem.

The central limit theorem was published by Laplace in 1810. This fundamental theorem shows how, regardless of the shape of the histogram of the original data, the histograms of subgroup averages will tend to have a "normal" shape as the subgroup size gets larger. This is illustrated in Figure 24 where the histograms for 1000 subgroup averages are shown for each of three different subgroup sizes for data obtained from two completely different sets of original data. There we see that even though the histograms for the individual values differ, the histograms for the subgroup averages tend to look more alike and become more bell-shaped as the subgroup size increases.

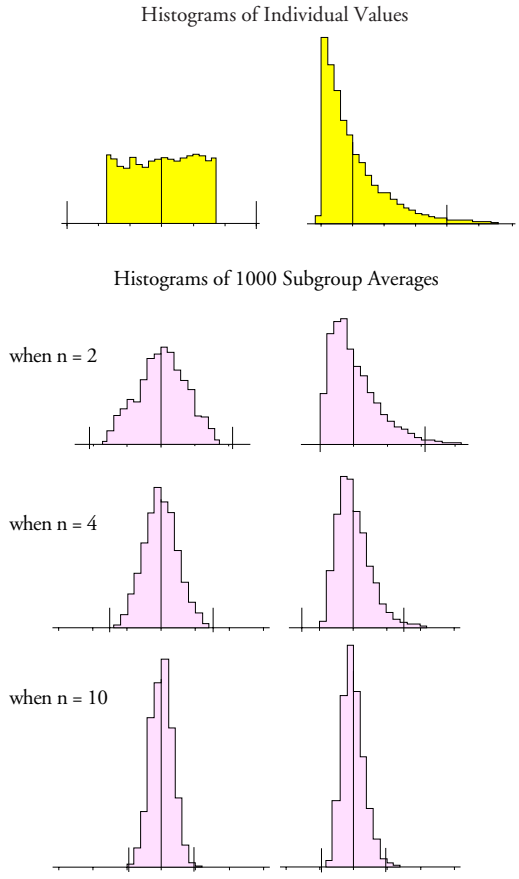

Figure 24: The Central Limit Theorem for Subgroup Averages

Many statistical techniques that are based on averages utilize the central limit theorem. While we may not know what the histogram for the original data looks like, we can be reasonably sure that the histogram of the subgroup averages may be approximated by a normal distribution. From this point we can then use the statistical approach outlined in the preceding section to carry out our analysis using the subgroup averages.

However, while we have a central limit theorem for subgroup averages, there is no central limit theorem for subgroup ranges. This is illustrated in Figure 25 where we see the histograms

of the subgroup ranges obtained from the two different sets of original data. Each histogram shows the ranges of 1000 subgroups, for each of three subgroup sizes, obtained from each of the two data sets shown. As the subgroup size increases the histograms for the subgroup ranges become more dissimilar and do not even begin to look bell-shaped.

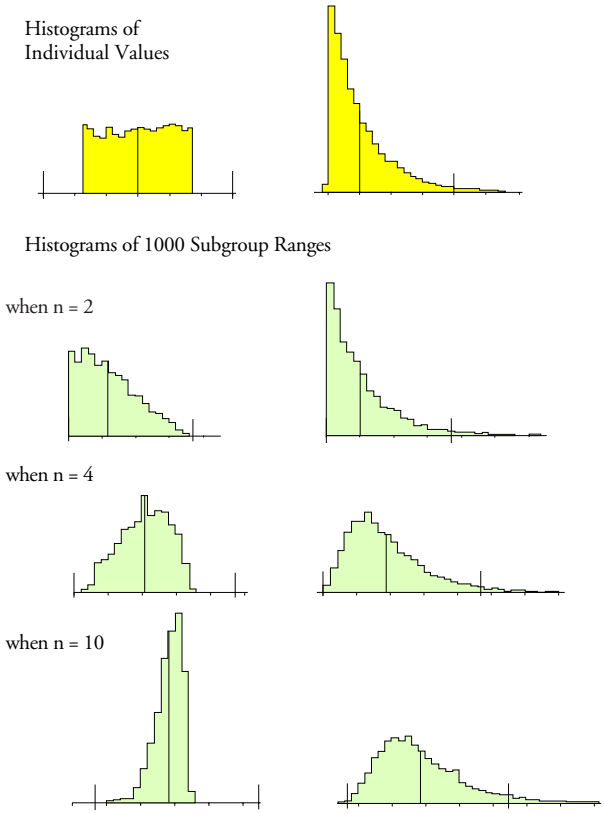

Figure 25: The Lack of a Central Limit Theorem for Subgroup Ranges

Therefore, Myth Three has no basis in reality. If the central limit theorem was the foundation for process behavior charts, then the range chart would not work.

Rather, as we saw in the preceding section, Shewhart chose three-sigma limits to use with the process behavior chart simply because, when the data are homogeneous, these limits will bracket virtually all of the histogram regardless of the shape of that histogram. Three-sigma limits are shown on each of the 16 histograms in Figures 24 and 25. There they bracket better than 98 percent of each histogram, leaving less than a 2 percent chance of a false alarm in each case. In practice, as long as (1–*P*) is known to be small, you do not need to know the exact risk of a false alarm. This means that when you find a point outside the limits of a process behavior chart the odds are very good that the underlying process has changed and you will be justified in taking action. Three-sigma limits provide you with a suitably conservative analysis without requiring a lot of preliminary work. It is this conservative nature of three-sigma limits that eliminates the need to appeal to the central limit theorem to justify the process behavior chart.

Undoubtedly, Myth Three has been one of the greatest barriers to the use of process behavior charts with management data and process-industry data. Whenever data are obtained one-value-

per-time-period it will be logical to use subgroups of size one. However, if you believe Myth Three you will feel compelled to average something in order to invoke the blessing of the central limit theorem, and the rationality of your data analysis will be sacrificed to superstition. The conservative nature of three-sigma limits allows you to use the chart for individual values with all sorts of original data without reference to the shape of the histogram.

MYTH FOUR:

It has been said that the observations must be independent—
data with autocorrelation are inappropriate for process behavior charts.

Again we have an artificial barrier to the use of a process behavior chart which ignores both the nature of real data and the robustness of the process behavior chart technique. Virtually all data coming from a production process will display some amount of autocorrelation. Autocorrelation is simply a measure of the correlation between a time series and itself. A positive autocorrelation (lag one) simply means that the data display two characteristics: (1) successive values are generally quite similar while (2) values that are far apart can be quite dissimilar. These two properties mean that when the data have a large positive autocorrelation the underlying process will be changing.

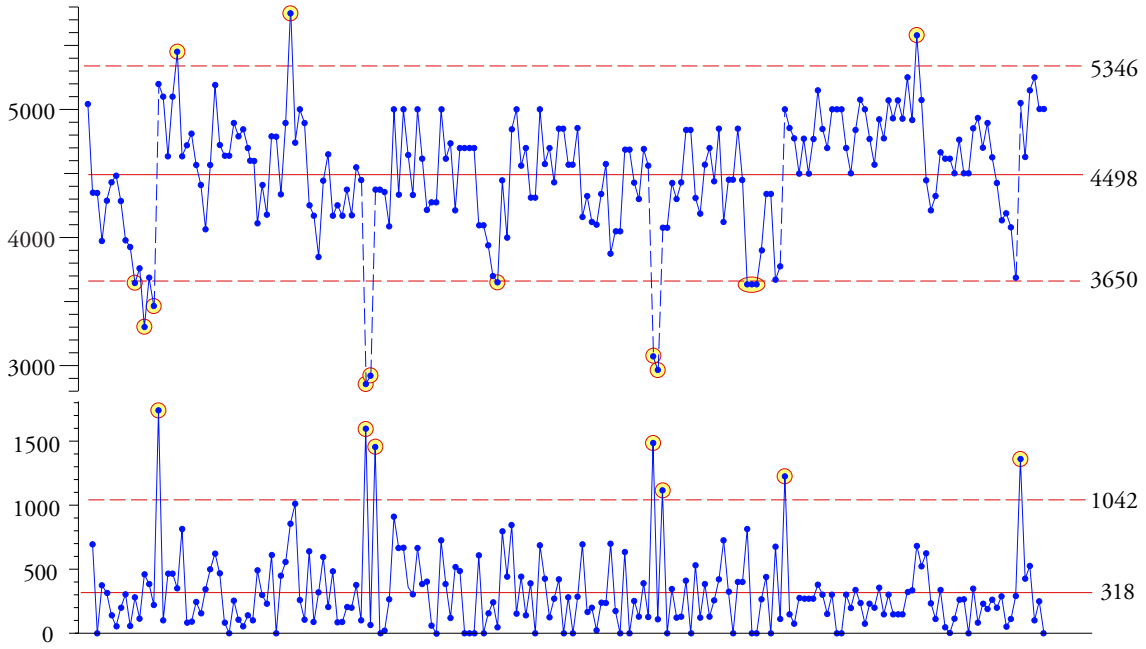

Figure 26: *XmR* Chart for 204 Resistances from Shewhart (1931) Page 20

To illustrate this property I will use data from Shewhart's 1931 book. These data are the measured resistances of insulation material. These data have an autocorrelation of 0.549, which is detectably different from zero (also known as significantly different from zero). While Shewhart organized these data into 51 subgroups of size four and placed them on an average chart, it could be argued that this subgrouping obscures the effects of the autocorrelation upon the chart. To avoid this problem I have placed these 204 data on an *XmR* chart in Figure 26.

Shewhart found 8 averages outside his limits. We find 14 individual values and 7 moving ranges outside our limits. So both Shewhart's average chart and our *XmR* chart tell the same story. This process was not being operated predictably.

As they found the assignable causes and took steps to remove their effects from this process they collected some new data. These data, shown in Figure 27, show no evidence of unpredictable behavior. Notice that the new limits are only 60% as wide as the original limits. By removing the assignable causes of exceptional variation they not only got rid of the process upsets and the extreme values, but they also removed a substantial amount of process variation. The autocorrelation for the data in Figure 27 is 0.091, which is not detectably different from zero.

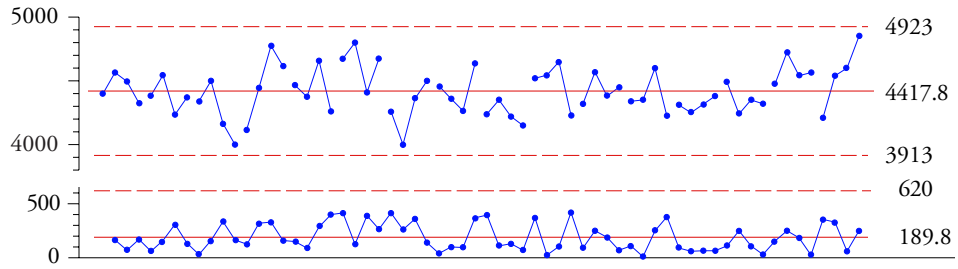

Figure 27: *XmR* Chart for 64 Additional Resistances from Shewhart (1931) Page 20

This example illustrates an important point. Whenever the data have a substantial autocorrelation the underlying process will be moving around. Autocorrelation is simply one way that the data have of revealing that the underlying process is changing. On the other hand, when the process is operated predictably, the data are unlikely to possess a substantial autocorrelation. (Be careful, these sentences cannot be reversed. While the Bead Board Data of Figure 16 have an autocorrelation of 0.60, and the Creel Yield Data of Figure 20 have an autocorrelation of 0.95, the Batch Weight Data of Figure 17 only have an autocorrelation of 0.09.)

Remember that the purpose of analysis is insight rather than numbers. The process behavior chart is not concerned with creating a model for the data, or whether the data fit a specific model, but rather with using data for making decisions in the real world. To insist that the data be independent is to add something to Shewhart's work that Shewhart was careful to avoid. This example from Shewhart's first book illustrates that process behavior charts have worked with autocorrelated data from the very beginning. Do not let those who do not understand this point keep you from placing your data on a chart because the values might not be independent.

While a complete treatment of the effects of autocorrelation is beyond the scope of this article, the following observation is in order. While it is true that when the autocorrelation gets close to +1.00 or –1.00 the autocorrelation can have an impact upon the computation of the limits, such autocorrelations will also simultaneously create running records that are easy to interpret at face value. This increased interpretability of the running record will usually provide the insight needed for process improvement and further computations become unnecessary. (Did you need the limits to know that the process was changing in Figures 16, 18, or 20?)

Finally, because of the link between having a significant autocorrelation and having a process that is moving around, Myth Four essentially insists that you only use a process behavior chart when your process is being operated predictably. This same catch-22 also shows up in Myth Five.

MYTH FIVE

It has been said that the process must be operating in control before you can place the data on a process behavior chart.

I first encountered this myth when I was refereeing a paper written by a professor of statistics at a land-grant university in the South, which goes to prove my point that even an extensive knowledge of statistics does not guarantee that you will understand Shewhart.

A second form of Myth Five is:

A control chart is a tool for maintaining the status-quo— it was created to monitor a process after that process has been brought to a satisfactory level of operation.

I suspect that the origin of Myth Five is a failure to appreciate the point of Myth One that there are correct and incorrect ways of computing the limits for a process behavior chart. The fallacy of using three-standard-deviation limits was shown on page 302 of Shewhart's 1931 book, yet it is found in virtually every piece of software available today. While three-standard-deviation limits will mimic three-sigma limits whenever the process is operated predictably, three-standard-deviation limits will be severely inflated when the process is being operated unpredictably. Thus, when someone is using the incorrect way of computing the limits, they might come to believe Myth Five.

Of course, as soon as you believe Myth Five you will begin to look for a way to remedy this perceived defect in the technique. Among the absurdities which have been perpetrated in the name of Myth Five are the censoring of the data prior to placing them on the chart (removing the outliers) and the use of two-standard-deviation limits. (As Dr. Henry Neave observed in a letter to the Royal Statistical Society, calculating the limits incorrectly and then using the wrong multiplier is an example of how two wrongs still do not make one right.) Needless to say that these, and all other associated manipulations are unnecessary.

The express purpose of the process behavior chart is to detect when a process is changing, and to do this we have to be able to get good limits from bad data. As illustrated in Figures 16, 17, 18, 19, 20, and 26, *process behavior charts give you limits that allow you to detect the potential signals even when the limits are based on the data produced by a changing process*. We do not have to wait until the process is "well-behaved" before we compute our limits. The correct computations are robust. And this is why Myth Five is patent nonsense.

HOW THE PROCESS BEHAVIOR CHART WORKS

In his second book Shewhart showed how a process behavior chart provides an operational definition of how to get the most out of any process. The three elements of this operational definition are:

1. A process behavior chart defines the *Ideal* of what a process can do when operated up to its full potential.
2. A process behavior chart provides a *Method* for moving a process towards the *Ideal*.
3. A process behavior chart allows you to make a *Judgment* about how close to the *Ideal* your process is operating.

The Ideal: The three-sigma limits of a process behavior chart characterize the process potential. When the process is being operated predictably, these limits define the actual capability of the process. When a process is being operated unpredictably these limits will approximate what the process can be made to do. In either case, these limits approximate the ideal of what your process can achieve when it is operated with maximum consistency.

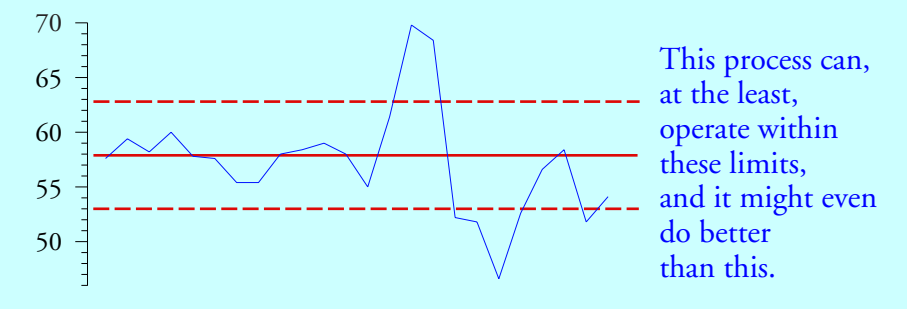

Figure 28: What Do You Want to Accomplish?

The Method: The running record displays the actual process performance. By highlighting the exceptional values, the chart gives you points to investigate, and thereby gives you a methodology you can use to improve your process. The idea that you can discover what affects a process by studying those points where the process changes is an idea that dates back to Aristotle. It is an idea that has stood the test of time.

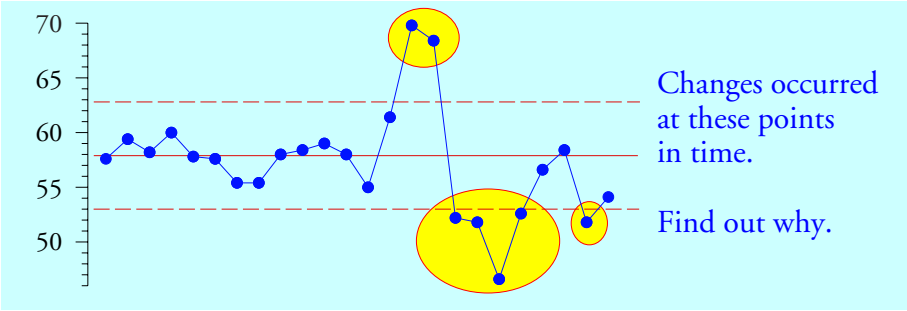

Figure 29: By What Method?

The Judgment: By combining both the process performance and the process potential on a single graph, the process behavior chart allows you to make a judgment about how close to the ideal your process may be operating at a given point in time.

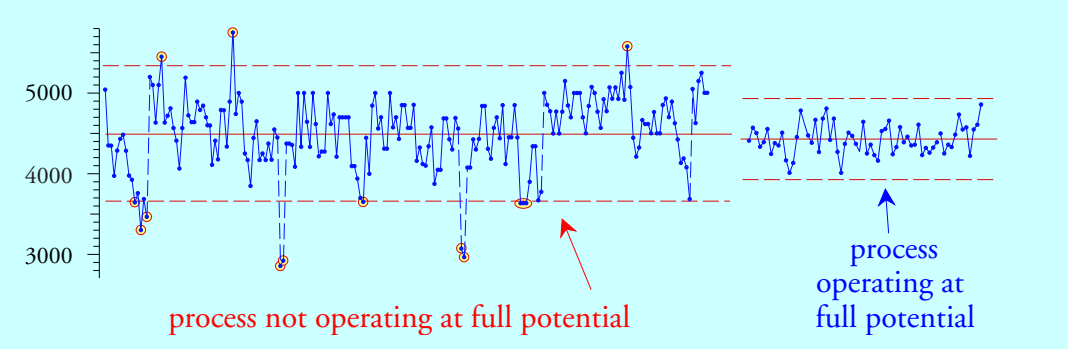

Figure 30: How Will You Know?

While we are good at defining what we want to accomplish, the trick is how to answer the second and third questions. Until you know the answers to "By what method?" and "How will you know?" all of your hopes, all of your plans, all of your goals will be nothing more than wishful thinking.

SO, WHAT IS REQUIRED TO USE A PROCESS BEHAVIOR CHART ?

In the course of discussing the preceding myths we have covered the first two foundations of the process behavior chart technique:

1. The robustness of the control chart technique is founded upon the universality of three-sigma limits.
2. To get good limits from bad data we have to use a within-subgroup measure of dispersion.

The third foundation is rational sampling and rational subgrouping. You have to organize the data so that the within-subgroup variation will properly capture and represent the routine variation present in the production process. Here you will have to use process knowledge to make a judgment about how to organize the data in a rational manner.

When you organize your data into subgroups you have to think about the context for each of the values. Two or more values may be placed in the same subgroup when they may be judged to have been obtained under essentially the same conditions. Values that might have been obtained under different conditions belong in different subgroups. Mixing apples, oranges, and bananas together may make a great fruit salad, but it makes for lousy subgroups and ineffective charts.

When working with individual values this requirement of rational subgrouping means that successive values must be logically comparable. In addition, the moving ranges will need to capture the routine variation present in the process. Once again, this means that a chart for individual values will only work properly when you are comparing apples to apples. If you have apples and oranges, put them on separate charts.

SUMMARY

Shewhart's approach to the analysis of data is profoundly different from the statistical approach. This is why people end up with such confusion when they try to "update" Shewhart by attaching bits and pieces from the statistical approach to what Shewhart has already done. The bits and pieces were not left out because the computations were too hard iback in the 1930s, they were left out because they were not needed. Shewhart provided us with an operational definition of how to get the most out of any process. Nothing extra is needed to make process behavior charts work.

We do not need to check for normality.

We do not need to transform the data to make them "more normal."

We do not have to use subgrouped data in order to receive the blessing of the central limit theorem before the chart will work.

We do not need to examine our data for autocorrelation.

We do not need to wait until our process is "well-behaved" before computing limits.

And we do not need to "update the computations" by using different dispersion statistics, or by using limits with different widths.

All such "extras" will just mess you up, get in your way, leave you confused, and keep you from using one of the most powerful data analysis techniques ever invented.