Probability Limits

A long standing controversy

Donald J. Wheeler

Shewhart explored many ways of detecting process changes. Along the way he considered the analysis of variance, the use of bivariate plots of location and dispersion, and the idea of probability limits. In the end he settled on the use of a generic approach using symmetric three-sigma limits based on a within-subgroup measure of dispersion. This article updates and expands Shewhart's arguments regarding probability l;imits

CAVEAT LECTOR

Now the idea of using probability limits has been around since 1935 and it still has many proponents and adherents today. Here I explain why, after a lifetime of studying this topic, and with the advantages of having been mentored by both Dr. Deming and David S. Chambers, I do not teach people how to use probability limits.

It is my intention to persuade you to avoid using probability limits. If you are not open to be persuaded, this article is not for you. I do not wish to engage anyone in a debate, nor do I wish to raise anyone's blood pressure. This paper explains why I teach what I teach. As always, it is offered with the intent of helping my readers to see a better way to understand their data.

PROBABILITY LIMITS

The problem to be addressed is the practical one of how to define limits for an observable variable, such as individual values, subgroup averages, or subgroup ranges, so that we will know when the underlying process that produced our data has changed. To do this we will need to separate any potential signals of a process change from the probable noise of routine process variation. Thus we shall need to be able to filter out the routine variation found in the data stream created by our process.

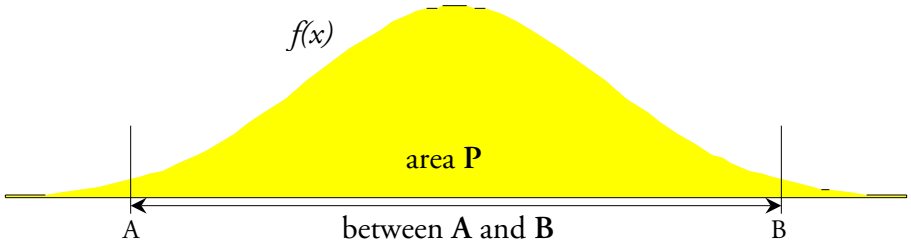

Figure 1: How Critical Values A and B Define a Central Area P for a Probability Model *f(x)*

If we know the appropriate probability model, $f(x)$, then we can use a straightforward

approach to computing limits. We would begin by choosing some proportion, **P**, of the routine variation that we wish to filter out as probable noise. Our limits would then be those points **A** and **B** that define the central proportion **P** under the probability model, *f(x)*, as shown in Figure 1. Typically we would choose **P** to be close to 1.000 so that we would filter out virtually all of the routine variation. Of course, the usual way of expressing the relationships shown in Figure 1 is by means of the integral equation:

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

Thus, using the elements of this integral equation we can summarize the probability approach to process behavior charts in the following manner: For a *given* probability model *f(x)*, and for a *given* proportion **P**, find the *specific* critical values **A** and **B**. Since such limits depend upon the probability **P**, values of **A** and **B** found in this way are known as probability limits.

Shewhart identified this approach on page 275 of *Economic Control of Quality of Manufactured Product* This approach is consistent with what is typically done in statistical inference, and is well understood by statisticians. Having thus defined probability limits, Shewhart continued: "For the most part, however, we never know [the probability model] in sufficient detail to set up such limits."

At this point Shewhart leaves the probability approach behind and presents a different way of handling the integral equation above. Rather than fixing the value for **P** in advance, he uses the Chebychev inequality as an existence theorem to argue that we can use *fixed*, *generic* values **A** and **B** that will automatically result in a value for **P** that is *close to 1.00 regardless* of what probability model *f(x)* is used.

Notice that Shewhart's approach to the integral equation is the exact opposite of the probability approach. The probability approach fixes the value for **P** and finds critical values for a *specific* probability model. Shewhart's approach *fixes* the critical values **A** and **B** and lets **P** vary so that the limits will work for *any* probability model. So, while the probability approach *requires* that you start out with a specific probability model, Shewhart's approach does not.

Shewhart's argument is that as long as **P** remains reasonably close to 1.00, we will end up making the right decision essentially every time. In general, whenever a procedure has a value for **P** that is larger than 0.98 that procedure is considered to be conservative. And as we shall see, Shewhart's symmetric three-sigma limits are sufficiently conservative to be completely general and non-specific.

"Well, that was in 1931. Today we have computer software that will identify a probability model for us."

Okay, let's consider how that works.

FROM DATA TO PROBABILITY MODEL

Since we have to identify a specific probability model in order to compute probability limits, we need to look at how software can "fit" a model to a data set. The software cannot look at the histogram and make a judgment, so it will have to use values computed from the data, i.e. statistics.

The average statistic will tell us everything there is to know about the location of the data set,

so we use the average to characterize location. The standard deviation statistic will tell us everything there is to know about the dispersion of the data set, so we use the standard deviation statistic to characterize dispersion. But characterizing the location and dispersion is not enough to specify a particular probability model. Figure 2 shows six probability models that all have the same means and standard deviations, and yet they have radically different shapes.

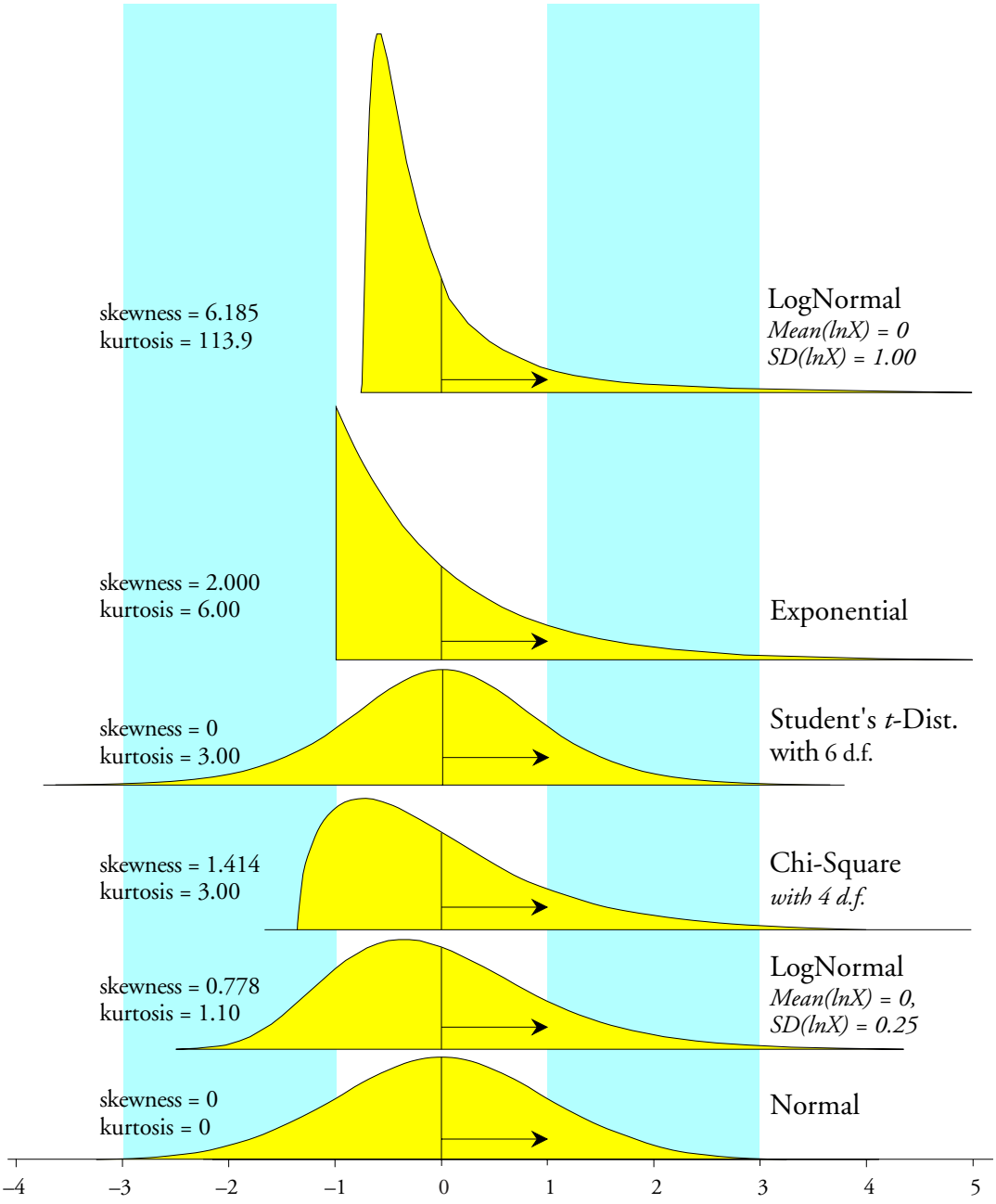

Figure 2: Six Probability Models with the Same Mean and Standard Deviation

So, the first lesson of Figure 2 is that knowing the location and dispersion is not sufficient to determine the shape of the probability model. The second lesson is that generic, three-sigma

limits will cover virtually all of a probability model *regardless of its shape*.

In order for any distribution to get even a tiny bit of area out beyond three-sigma it has to compensate for the increased rotational inertia by concentrating a much larger amount of area close to the mean. This can be seen in Figure 2 by starting at the bottom. Figure 3 gives the areas beyond three sigma and the areas within one sigma for the six distributions of Figure 2.

Distribution	Area Beyond Three Sigma	Increase vs. Normal	Area Within One Sigma	Increase vs. Normal
Normal	0.003	—	0.683	—
Lognormal (1, 0.25)	0.008	0.005	0.701	0.017
Chi-square with 4 d.f.	0.009	0.006	0.726	0.043
Student's-t with 6 d.f.	0.010	0.007	0.733	0.050
Exponential	0.018	0.015	0.865	0.182
Lognormal (1, 1)	0.018	0.015	0.910	0.227

Figure 3: Areas Beyond Three Sigma and Within One Sigma

- For the lognormal (1, 0.25) to get an extra 5 parts per thousand (ppt) outside three sigma (beyond what the normal has) it has to compensate by increasing the area within one sigma by 17 parts per thousand.
- For the chi-square with 4 degrees of freedom to get an extra 6 ppt outside three sigma it has to compensate by increasing the area within one sigma by 43 ppt.
- For the Student's-t with 6 degrees of freedom to get an extra 7 ppt outside three sigma it has to compensate by increasing the area within one sigma by 50 ppt.
- For the exponential to get an extra 15 ppt outside three sigma it has to compensate by increasing the area within one sigma by 182 ppt.
- Finally, the lognormal (1, 1) only has 15 parts per thousand more area outside three sigma than the normal, but it has 227 ppt more area within one sigma of the mean.

Thus, compared to the normal distribution, any increase in the infinitesimal areas out beyond three sigma will require a much larger compensating increase in the area within one sigma of the mean. This is an unavoidable consequence of using rotational inertia to characterize dispersion. There is much more to a skewed distribution than merely having an elongated tail. No matter how much you may stretch that tail, you are going to stretch sigma at essentially the same rate. ***In consequence, no mound-shaped distribution can ever have more than 1.9 percent outside the mean ± three sigma***.

So, as may be seen from Figure 3, the use of any mound-shaped or J-shaped model with greater kurtosis than the normal will impose a *requirement* that more of the observations fall within one standard deviation of the mean. *Thus, if all you have are measures of location and dispersion, then the absolute worst case probability model that you can fit to your data is a normal distribution.* The normal distribution is the distribution of maximum entropy. It spreads the middle 90 percent of the probability out to the maximum extent possible, so that the outer ten percent of a normal distribution is as far, or further, away from the mean than the outer ten percent of any other probability model.

Read the above paragraph again. It is completely contrary to what many students of statistics think, yet with the computing power we have today it is easy to verify. For more information see my ***QDD*** articles "What They Forgot to Tell You About the Normal Distribution," and "The

Heavy-Tailed Normal," from September and October 2012.

FINDING A SHAPE FOR YOUR PROBABILITY MODEL

So how can your software determine a shape for your probability model? As we saw in Figure 2, estimates of location and dispersion will not suffice. Therefore, absolutely the only way your software can fit a non-normal model to your data is to use the shape statistics of skewness and kurtosis. Whether you are aware of this or not, your software has no alternative.

In most cases your software will ask you to choose some family of probability models, and then, based on your data, the software will pick an appropriate member from that family of distributions. Three commonly used families of distributions are the Lognormals, Gammas, and Weibulls. Figure 4 shows each of these three families of distributions on the shape characterization plane in the broader context of all mound and J-shaped distributions.

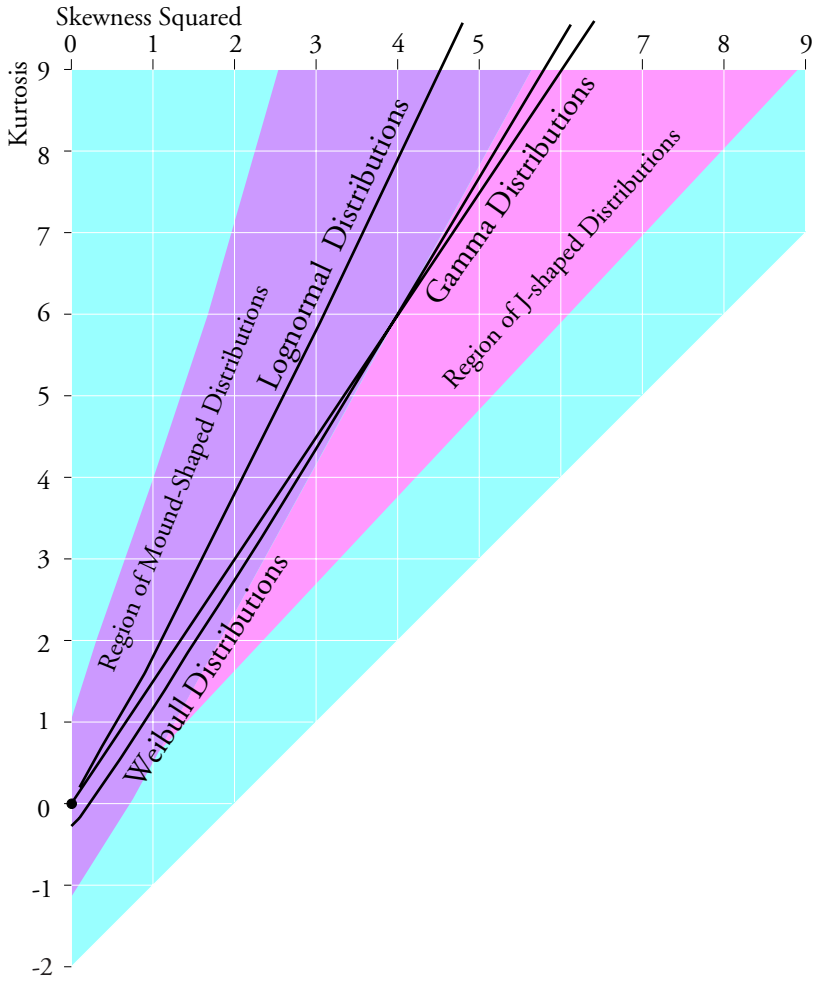

Figure 4: The Lognormal, Weibull, and Gamma Families of Probability Models

The Lognormal, Gamma, and Weibull families are shown as lines in Figure 4 because they each have only a single shape parameter. To fit these models your software will use some algorithm to estimate the single shape parameter for the model using the skewness and kurtosis

statistics of your data. It may not tell you that it is doing this, but it is. It may use some fancy name for the algorithm such as "maximum likelihood," or "least squares," or "minimum variance unbiased," but in the end it absolutely, positively has to make use of the shape statistics to choose between the various models. It cannot do otherwise.

In the following examples I will use Burr and Beta probability models to fit my data because these families each have two shape parameters. This will allow models from anywhere in the mound-shaped or J-shaped regions of Figure 4 to be used. By fitting both skewness and kurtosis separately we can obtain very close fits between the data and the probability models without imposing any presuppositions as to which family of probability models is appropriate. We begin with a simple set of 25 values. These values were all generated from the same probability model using a random number generator in Excel.

-0.30	0.61	-0.47	-1.00	-0.65	-0.69	-0.88	0.93	0.36	0.22
-0.31	1.07	0.27	-0.84	-0.86	0.48	0.21	1.40	0.14	-0.53
			-0.03	-0.70	3.81	0.22	0.09		

Figure 5: Twenty-five Observations from a Specific Probability Model

The histogram for these 25 values is shown in Figure 6. This histogram has an average of 0.10, and standard deviation statistic of 1.01, a skewness statistic (Excel formula) of 2.14, and a kurtosis statistic (Excel formula) of 6.78. (To see the formulas for these shape statistics see my article "Problems with Skewness and Kurtosis, Part Two," ***QDD*** August 2, 2011.)

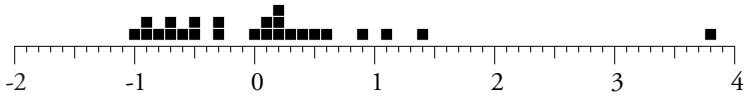

Figure 6: Histogram of the Twenty-five Observations of Figure 5

Both skewness and kurtosis statistics are highly dependent upon the extreme values of the data set. This can be seen without having to get involved in the formulas: simply move the most extreme value and watch how the skewness and kurtosis change. Since we have a very large extreme value here, I will move it closer to the mean. As long as the value you are moving is the most extreme value, each change will have a pronounced effect upon both the skewness and the kurtosis. As soon as the value you are changing is no longer the most extreme value the skewness and kurtosis statistics will stabilize. Figure 7 shows several such modifications of the data from Figure 5, along with the first four descriptive statistics for each set.

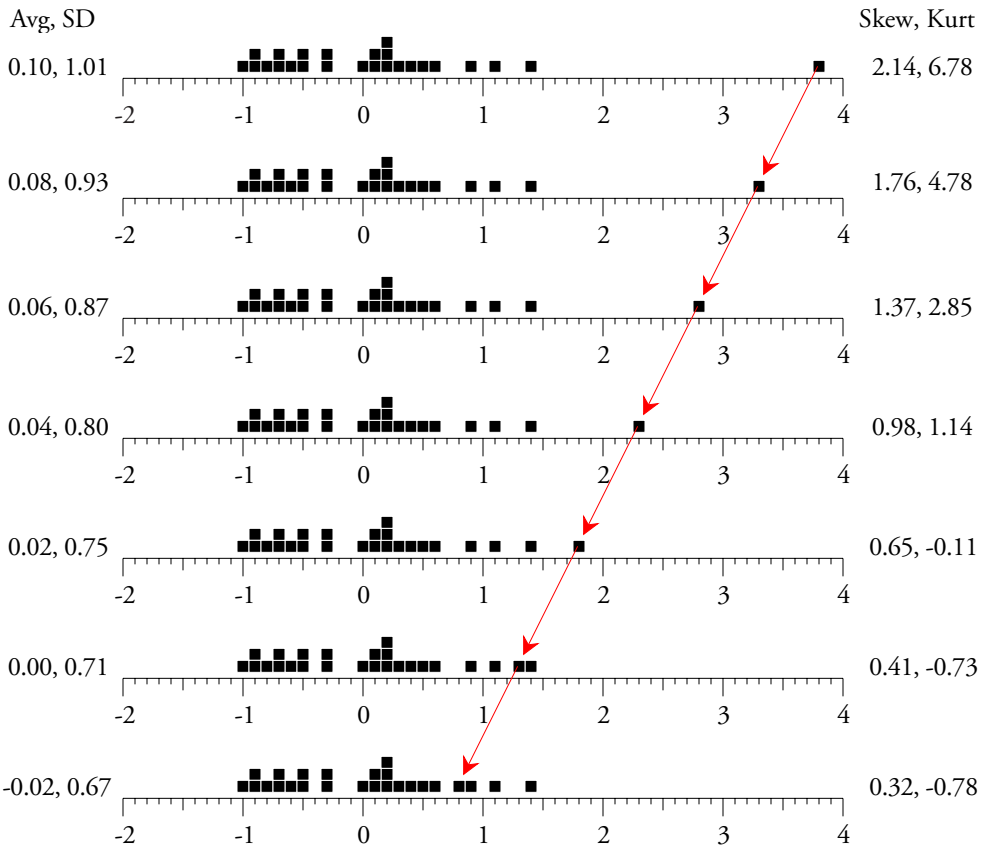

Figure 7: How the Skewness and Kurtosis Statistics Change with the Extreme Value

Notice how the skewness and kurtosis statistics barely change between the bottom two data sets, unlike all preceding changes.

Figure 8 shows fitted probability models that match each of the histograms in Figure 7. Without going into the details for each model fitted, the information in Figure 8 identifies each fitted model and gives the skewness and kurtosis parameters for that model. Each model was then location and scale shifted to match the average and standard deviation for the fitted data.

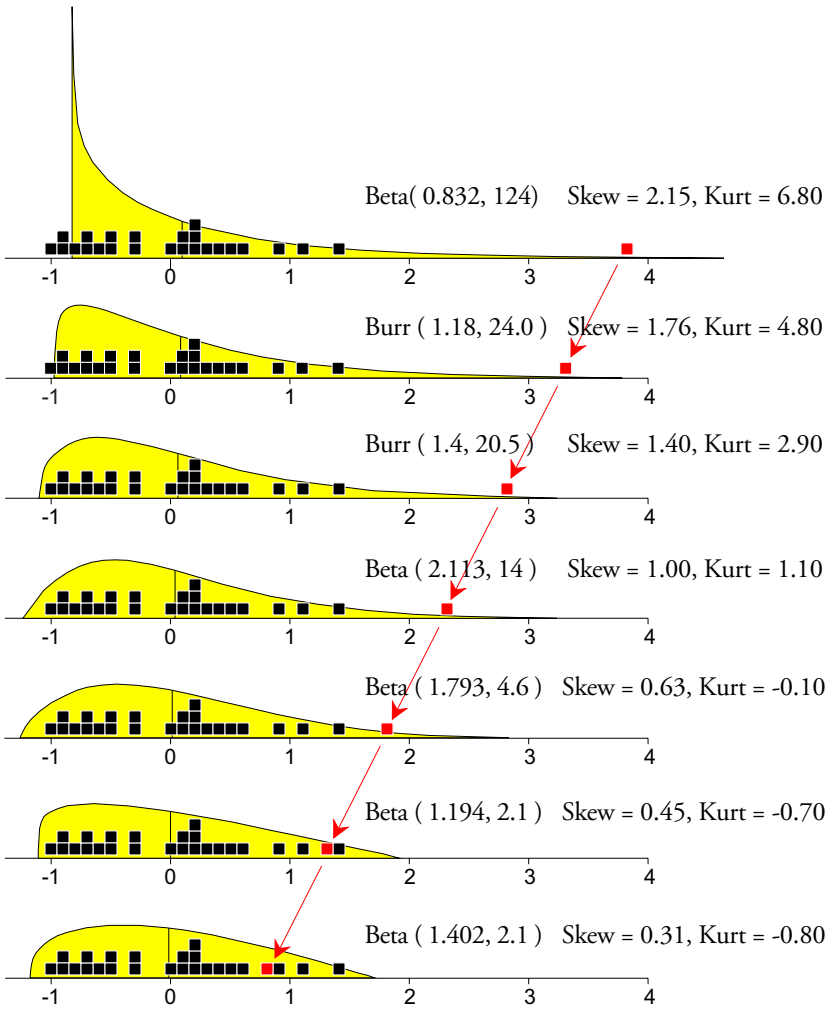

Figure 8: How the Extreme Value Determines the Probability Model

Clearly, both the skewness and kurtosis statistics are heavily dependent upon the extreme value(s) in your data set. As a result, the shape of your fitted model will also be heavily dependent upon the extreme values rather than upon the overall "shape" of the histogram. The seven models shown in Figure 8 do their best to accommodate the largest value, and they do this with very little regard for the other 24 values. (The other 24 values primarily determine the location and dispersion, but they have little effect upon the shape statistics.)

It is this heavy dependency of both skewness and kurtosis statistics upon the extreme values of your data that effectively undermines any attempt to obtain a meaningful fit between a probability model and a data set. Your algorithm may get a model that fits your data very nicely, like we did seven times in Figure 8, but more than anything else, your model will be almost certainly fitting the extreme values in the tails of your data.

Remember, the first histogram is the actual data set in this case. So, is the model for the first histogram correct? The model has the right mean; it has the right standard deviation; it has the right skewness; and it has the right kurtosis; and yet the histogram has three values that would be impossible to observe if the fitted model was correct. Since your data always trumps your model,

we have to conclude that the J-shaped model is incorrect.

"So what can we do?" Whether you try to use the shape statistics directly, or indirectly through some algorithm in your software, you will end up fitting the most extreme values. If you restrict yourself to using only the location and dispersion statistics, then the generic, worst-case model is the normal distribution with that location and dispersion. Compare Figure 9 with the models in Figure 8. The normal distribution does a good job on the 24 observations, and it reveals the largest value to be an outlier.

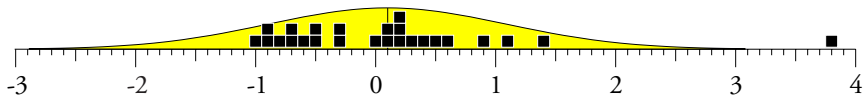

Figure 9: A Normal Distribution fit to the Data of Figure 3

While the 25 values here were all obtained from one and the same probability model (a standard normal distribution), this set of 25 values was the most extreme set out of 10,000 such sets generated by the random number generator. So Figure 9 tells the correct story here. The value of 3.81 is simply one of those very rare values from a standard normal distribution that fall in the region beyond three sigma.

SUMMARY

Thus, the notion of probability limits is based upon the assumption that we can fit a probability model to our data and then find the *exact* critical values **A** and **B** that will yield a predetermined value for **P**. However, as we have seen in Figure 8, any model that we end up fitting to our data will be highly dependent upon the extreme value(s) in our data. This will, in turn, severely affect the critical values, **A** and **B**, which will affect both the results and the interpretation of those results. Ultimately, the problem here is that all of the models in Figure 8 *assume* that the 25 values are homogeneous. The model in Figure 9 shows that the extreme value is likely to be an outlier, and this outlier will always skew any probability model fitted to these data.

"So why don't we simply eliminate the outliers before fitting the model?" Beautifully simple, yet as soon as we adopt this approach the question becomes, "How do you identify an outlier?" The process behavior chart, with its generic, three-sigma limits, is an operational definition of what constitutes an outlier! (It *defines* an outlier, it gives us a *procedure* for detecting outliers, and it allows us to *judge* whether a specific point is, or is not, likely to be an outlier.) All other definitions of outliers end up being more conservative than the process behavior chart simply because they are based on the total variation within the data set. So, if we have to delete the outliers in order to fit a model to our data before we can compute the "correct" probability limits for our process behavior chart, then we are indeed without hope.

Moreover, the outliers that get deleted are exactly those signals of changes in our process that we want to detect. Removing outliers from the analysis changes the focus of the analysis from finding and fixing problems to getting a pretty picture from our data. (For those who wish to fit a model to the data in order to estimate the fraction nonconforming, that question is discussed in my article "Estimating the Fraction Nonconforming," ***QDD***, June 1, 2011.)

Thus, we return to Shewhart's statement that "For the most part, however, we never know [the probability model] in sufficient detail to set up such [probability] limits." While software may blind us to this insufficiency, it does not remove it. *This lack of information is not a problem that can be cured by computations.* Instead of trying to determine probability limits, why not use Shewhart's proven approach? As we saw in Figures 2 and 3, symmetric, three-sigma limits are sufficiently conservative to work with all types of probability models. Moreover, they are robust enough to work with data that are not homogeneous. They are not unduly disturbed by the extreme values in your data.

In the words of my friend Bill Scherkenbach, "The only reason to collect data is to take action." You need to separate the probable noise from the potential signals, and the symmetric three-sigma limits of process behavior charts will do this with sufficient generality and robustness to let you take appropriate action. Computing probability limits is all about getting exactly the right false alarm rate. Using process behavior charts is all about detecting the signals of process changes. Since there are generally many more signals to be found than there are false alarms to be avoided, the use of probability limits is focused on the wrong aspect of the decision problem regarding when to take action.

So, if your data happen to come from a process that is being operated predictably (a rare thing), and if you have hundreds of data without any unusual values, then your probability limits *might* work as well as Shewhart's generic, symmetric three-sigma limits. But if not, then your probability limits can result in you taking the wrong actions by missing signals or reacting to noise. Working harder, to implement a more complex solution, that will only occasionally work as well as a simpler solution, does not make sense.

Caveat computor.

POSTSCRIPT

While there are many processes out there that are very nicely *modeled* using Gamma and Weibull and Lognormal distributions, etc., this is no excuse for using these models in *analysis*. The primary question of analysis is "How is your process behaving?" To answer this question you will need to actually examine your process behavior, rather than acting like the mother of the defendant and claiming that your process "wouldn't dream of misbehaving." When we model a process we may well use an appropriate probability distribution. But when we analyze data we need to listen and let the data speak for themselves. In this world, your data are *never* generated by a probability model; they are always generated by some process. And those processes, like everything in this world, are always subject to change. "Has a change occurred?" is a question that can never be answered by "Assume a model…"