Process Behavior Charts for Non-Normal Data, Part One

A guide for charts for location

Donald J. Wheeler

Whenever the original data pile up against a barrier or a boundary value the histogram tends to be skewed and non-normal in shape. In 1967 Irving W. Burr computed the appropriate bias correction factors for non-normal probability models. These bias correction factors allow us to evaluate the effects of non-normality upon the computation of the limits for process behavior charts. To understand exactly what these effects are, read on.

BACKGROUND

Before we can discuss the computation of limits for non-normal probability models we will need to have a way to quantify the uncertainty in the computed limits. The curve that does this is given in Figure 1.

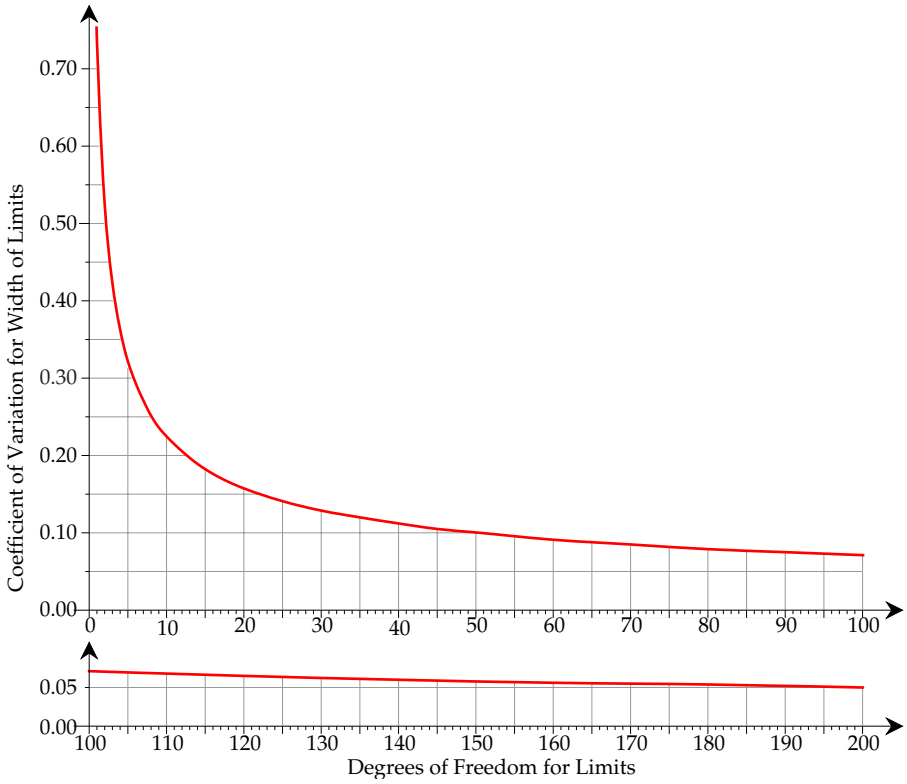

Figure 1: Uncertainty and Degrees of Freedom for Process Behavior Chart Limits

This curve involves two quantities that require some explanation: the first of these is the *degrees of freedom* for the limits and second is the *coefficient of variation* for the limits. Figure 1

shows that as the degrees of freedom increase the coefficient of variation will decrease. However, the relationship is very nonlinear, with the first few degrees of freedom accounting for the greatest reduction in the coefficient of variation. As the names suggest, the degrees of freedom is tied to the amount of data used, while the coefficient of variation is a measure of the uncertainty in the computed limits. The exact natures of these two quantities will be discussed below, but the shape of the curve in Figure 1 tells us that the first 10 degrees of freedom are critical, and that when we have 30 to 40 degrees of freedom the coefficient of variation will have fairly well stabilized.

DEGREES OF FREEDOM

For an Average and Range Chart with limits based on k subgroups of size n the Average Range could be said to be based on a total of nk data. However, because of various properties of subgroup ranges, a slightly different characterization of the amount of data used in the computation is preferred. This characterization is called the *effective degrees of freedom*, and for the average of k subgroup ranges, where each range is based on n data, this quantity is given by:

$$\textit{effective degrees of freedom for } \bar{R} = \frac{k\ d_2^2}{2\ d_3^2}$$

where d_2 and d_3 are the bias correction factors for subgroups of size n. While the expression above is relatively simple, it is common to use some even simpler approximations in practice:

$$\begin{aligned} \textit{effective d.f. for } \bar{R} &\approx 0.9\,k\,(n\text{–}1) && \text{for } n < 7 \\ &\approx 0.85\,k\,(n\text{–}1) && \text{for } n = 7, 8, 9, 10 \end{aligned}$$

The degrees of freedom for the limits will be the same as the degrees of freedom for the Average Range used to compute those limits. So, for example, the limits for an Average and Range Chart based on 25 subgroups of size 2 would be said to have about 22 degrees of freedom.

For an *XmR* chart there is a different formula for degrees of freedom. When using (k–1) two-point moving ranges to compute an Average Moving Range the degrees of freedom will be:

$$\textit{effective d.f. for } \overline{mR} = \frac{(k\text{–}1)^2}{1.6529\,k\ -\ 2.1642}$$

For routine use a simple approximation for this quantity is:

$$\textit{effective d.f. for } \overline{mR} \approx 0.62\,(k\text{–}1)$$

Thus, the limits for an *XmR* chart with a baseline of $k = 50$ data would be said to have about 30 degrees of freedom.

COEFFICIENTS OF VARIATION

When working with variables having a fixed zero point it is common to divide the standard deviation by the mean and to call the result the coefficient of variation. For the Average Range, based on k subgroups of size n:

$$MEAN(\bar{R}) = d_2\,\sigma \qquad\qquad SD(\bar{R}) = \frac{d_3\,\sigma}{\sqrt{k}}$$

$$\text{and} \quad CV(\bar{R}) = \frac{SD(\bar{R})}{MEAN(\bar{R})} = \frac{\frac{d_3\sigma}{\sqrt{k}}}{d_2\sigma} = \frac{d_3}{d_2\sqrt{k}}$$

The coefficient of variation is not affected by linear transformations. This means that the coefficients of variation for the three-sigma limits will be the same as the coefficient of variation for the Average Range used to compute those limits. The coefficient of variation above depends solely upon the bias correction factors d_2 and d_3 and the number of subgroups used, k. Since these bias correction factors depend upon the subgroup size, n, this coefficient of variation can be seen to depend solely upon the amount of data used, n and k. The subgroup size, n, will fix the values of d_2 and d_3, so that, as the number of subgroups, k, increases, these coefficients of variation will decrease in proportion to the square root of k. This result supports the intuitive notion that limits which are based on a greater amount of data should be more trustworthy than limits based on a lesser amount of data.

Comparing the formulas for the coefficient of variation and the effective degrees of freedom it seems that there should be some relationship between these two quantities. There is, and it is the curve plotted in Figure 1:

$$CV(\bar{R}) = \frac{1}{\sqrt{2\ d.f.}}$$

When we compute limits we naturally want to know how much uncertainty exists in those values. While degrees of freedom are the traditional way of characterizing measures of dispersion, it is the coefficient of variation that actually quantifies the uncertainty in the limits. Therefore, the most straightforward way to characterize uncertainty is to convert degrees of freedom into coefficients of variation using the curve in Figure 1 or the equation given above. The relationship shown in Figure 1 holds for every estimator of the standard deviation parameter, σ. Because of this, we can compare the efficiency of different estimators by comparing their degrees of freedom.

HOW MANY DATA ?

Clearly, the first few degrees of freedom will have the greatest impact upon improving the quality of your limits. In every case, to cut the coefficient of variation in half you will have to increase the degrees of freedom four-fold. This means that 32 degrees of freedom will only be twice as good as 8 degrees of freedom. Likewise, 128 degrees of freedom will be twice as good as 32 degrees of freedom. Figure 1 makes two lessons plain: degrees of freedom are a highly nonlinear way of characterizing uncertainty, and diminishing returns set in early.

The curve in Figure 1 has a 45° tangent somewhere in the neighborhood of 10 degrees of freedom. Thus, the elbow point can be taken to be 10 degrees of freedom.

- When you have fewer than 10 degrees of freedom your limits will be very soft, and each additional degree of freedom will give you valuable information.
- Between 10 degrees of freedom and 30 degrees of freedom your limits will coalesce, gel and firm up.
- Beyond 30 degrees of freedom your limits will have, for all practical purposes, solidified.

So, how many degrees of freedom do you need? How many do you have? Shewhart

suggested that, based on his experience, useful limits could be found using as few as six degrees of freedom. Clearly this is minimal. Also, it should be clear the there is little need to continue to update limits once they have been computed using 30 or 40 degrees of freedom. By using the relationship between degrees of freedom and coefficients of variation summarized by the bullet points above you can understand when you have soft limits, when your limits are getting firm, and when your limits are solid. Remember, the objective is not to compute the right number, or even the best estimate of the right value, but to take the right action, and you can often take the right action even when the limits themselves are fairly soft.

SOME HISTORY BEHIND THE COMPUTATION OF LIMITS

During World War II, in the interest of simplicity, the formulas for computing the limits for process behavior charts were written using scaling factors such as E_2, and A_2.

$$\text{Limits for Individual Values} = \text{Grand Average} \pm E_2 \times \text{Average Range}$$

$$\text{Limits for Averages} = \text{Grand Average} \pm A_2 \times \text{Average Range}$$

These scaling factors minimized the number of computations required, which was important when everyone was doing them by hand. These two scaling factors depend upon the subgroup size and the bias correction factor, d_2.

$$E_2 = \frac{3}{d_2} \qquad\qquad A_2 = \frac{3}{d_2\sqrt{n}}$$

The evaluation of these bias correction factors requires the numerical evaluation of a rather messy triple integral. While this is feasible when $n = 2$, it quickly becomes overwhelmingly tedious for larger subgroup sizes. In 1925 H. C. Tippett avoided this problem by using a simulation study to estimate the values for d_2 and d_3. In 1933 A. T. McKay and E. S. Pearson managed to publish the exact bias correction values for the case when $n = 3$. By 1942 E. S. Pearson and H. O. Hartley had published the exact bias correction factors for $n = 2$ to 20. Finally, in 1960 H. L. Harter published the exact bias correction factors out to 10 decimal places for $n = 2$ to 100. Of course all of these computations were carried out under the worst-case scenario: they used a normal distribution for the original data. Thus, it is a fair and correct statement to say that the normal distribution was used to compute the bias correction factors commonly found in textbooks today. The question here is how much of a restriction does this place upon the charts for location?

In 1967 Irving W. Burr decided to compute the exact bias correction values using each of 27 different probability models. These models were all members of the family of Burr distributions, and as such they effectively cover the whole region of mound-shaped probability models. Figure 2 shows six of the models Burr used.

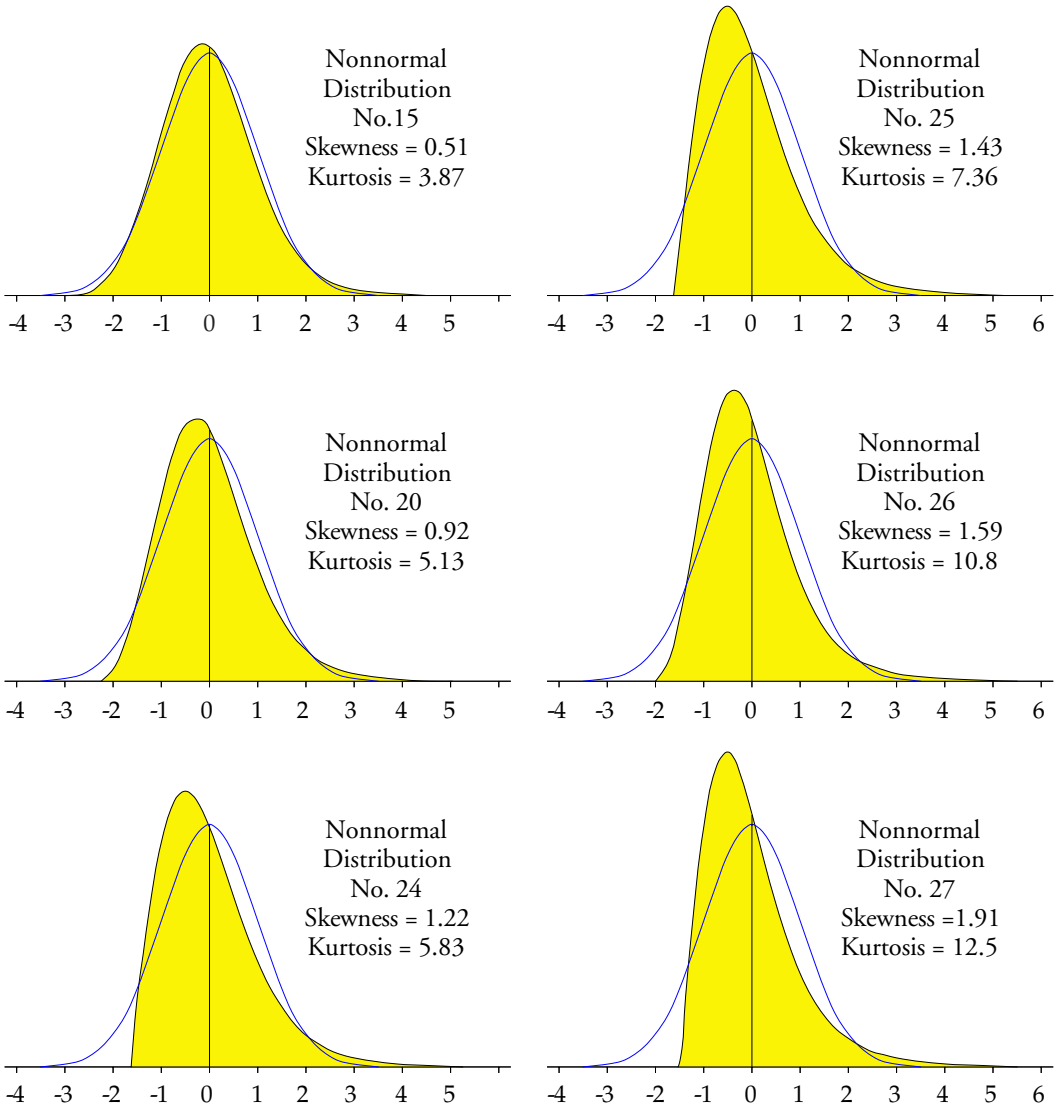

Figure 2: Six of the 27 Non-Normal Distributions Burr Used

Figure 3 gives Burr's exact values of d_2 for each of his 27 non-normal distributions. The first row gives the usual bias correction factors found using the normal distribution. As you go down each column you will see the bias correction factors that are appropriate for each of the 27 non-normal distributions. The non-normal values for d_2 tend to be slightly smaller than the normal theory values. This is as expected simply because the normal distribution has maximum entropy. Other distributions will result in slightly smaller subgroup ranges.

Distributions	skewness α_3	kurtosis α_4	Values of d_2 $n = 2$	3	4	5	8	10
normal	0.00	3.00	*1.13*	*1.69*	*2.06*	*2.33*	*2.85*	*3.08*
-1-	**–.01**	**3.01**	*1.13*	*1.69*	*2.06*	*2.33*	*2.85*	*3.08*
-2-	**0.00**	**3.33**	*1.12*	*1.68*	*2.05*	*2.32*	*2.85*	*3.09*
-3-	**0.04**	**3.65**	*1.11*	*1.67*	*2.04*	*2.31*	*2.86*	*3.10*
-4-	**0.07**	**2.88**	*1.13*	*1.70*	*2.06*	*2.33*	*2.84*	*3.07*
-5-	**0.11**	**3.67**	*1.14*	*1.67*	*2.04*	*2.31*	*2.86*	*3.10*
-6-	**0.12**	**3.19**	*1.12*	*1.69*	*2.05*	*2.32*	*2.85*	*3.08*
-7-	**0.18**	**3.05**	*1.13*	*1.69*	*2.06*	*2.32*	*2.84*	*3.07*
-8-	**0.19**	**3.74**	*1.11*	*1.67*	*2.04*	*2.31*	*2.85*	*3.10*
-9-	**0.28**	**3.48**	*1.12*	*1.68*	*2.05*	*2.31*	*2.85*	*3.09*
-10-	**0.29**	**3.86**	*1.11*	*1.67*	*2.04*	*2.31*	*2.85*	*3.10*
-11-	**0.34**	**3.36**	*1.12*	*1.68*	*2.05*	*2.32*	*2.84*	*3.07*
-12-	**0.35**	**3.04**	*1.13*	*1.69*	*2.06*	*2.32*	*2.83*	*3.05*
-13-	**0.43**	**4.11**	*1.11*	*1.66*	*2.03*	*2.30*	*2.84*	*3.09*
-14-	**0.48**	**3.38**	*1.12*	*1.68*	*2.05*	*2.31*	*2.83*	*3.05*
-15-	**0.51**	**3.87**	*1.11*	*1.67*	*2.04*	*2.30*	*2.83*	*3.07*
-16-	**0.56**	**3.60**	*1.12*	*1.68*	*2.04*	*2.30*	*2.82*	*3.05*
-17-	**0.64**	**4.63**	*1.10*	*1.66*	*2.02*	*2.29*	*2.83*	*3.08*
-18-	**0.68**	**4.04**	*1.11*	*1.67*	*2.03*	*2.29*	*2.82*	*3.05*
-19-	**0.88**	**4.12**	*1.10*	*1.66*	*2.01*	*2.27*	*2.78*	*3.01*
-20-	**0.92**	**5.13**	*1.10*	*1.64*	*2.00*	*2.27*	*2.80*	*3.04*
-21-	**0.96**	**5.94**	*1.09*	*1.64*	*2.00*	*2.26*	*2.80*	*3.05*
-22-	**1.01**	**4.71**	*1.09*	*1.64*	*2.00*	*2.26*	*2.77*	*3.00*
-23-	**1.09**	**5.12**	*1.09*	*1.63*	*1.99*	*2.25*	*2.76*	*3.00*
-24-	**1.22**	**5.83**	*1.08*	*1.62*	*1.97*	*2.23*	*2.75*	*2.99*
-25-	**1.43**	**7.36**	*1.06*	*1.60*	*1.95*	*2.21*	*2.73*	*2.97*
-26-	**1.59**	**10.81**	*1.06*	*1.58*	*1.93*	*2.19*	*2.73*	*2.97*
-27-	**1.91**	**12.46**	*1.03*	*1.55*	*1.89*	*2.15*	*2.67*	*2.91*

Figure 3: Burr's Values of d_2 for 27 Non-normal Distributions

BURR'S APPROACH

Irving Burr's original idea was that we could use the values in Figure 3 to sharpen up the limits by first computing the skewness and kurtosis statistics for our data and then choosing appropriate bias correction factors from his table.

Unfortunately, in practice, the uncertainty in both the skewness and kurtosis statistics is so great that we can never be sure that our estimates for these shape parameters are even remotely correct. As I showed in "Problems with Skewness and Kurtosis, Part Two," ***QDD*** , Aug. 2, 2011, any estimate of skewness will have 2.45 times more uncertainty that your estimates of location and dispersion, and any estimate of kurtosis will have 4.90 times more uncertainty than your estimates of location and dispersion. This unavoidable uncertainty in the statistics for skewness and kurtosis is due to their dependence upon the extreme values in the data set. Regardless of how many data you have, you will always know more about location and dispersion than you will ever know about skewness and kurtosis. Until you have thousands of data collected from a predictable process, any use of the skewness and kurtosis statistics is an exercise in fitting noise. This inherent and unavoidable problem undermines Burr's approach.

LIMITS FOR CHARTS FOR LOCATION

So, even though Burr's idea does not quite work out as intended in practice, we can use the values in Figure 3 to assess the effects of non-normality upon the computation of the limits. Up to this point we have treated the bias correction factors as constants. However, since the problem Burr was considering was how to determine the values of these constants, we shall instead look at the bias correction factors as unknown variables that must be approximated. With this change in perspective the question becomes one of assessing the impact of not knowing the exact value for these fundamental constants.

We begin by observing that the d_2 bias correction factor always occurs in the denominator of the formulas for the scaling factors above. Thus, we begin by inverting the values for d_2 from Figure 3. These new values are shown in Figure 4.

To characterize how the values in Figure 4 vary, we compute the coefficient of variation for each column. These coefficients of variation will summarize how our not knowing the exact values for d_2 will affect the computation of the limits. The first six columns show the impact upon limits for charts for location (that is, limits for *X* charts or for average charts).

Distributions		Values of $1/d_2$					
skewness	kurtosis						
α_3	α_4	$n = 2$	3	4	5	8	10
0.00	3.00	0.885	0.592	0.485	0.429	0.351	0.325
–.01	**3.01**	0.885	0.592	0.485	0.429	0.351	0.325
0.00	**3.33**	0.893	0.595	0.488	0.431	0.351	0.324
0.04	**3.65**	0.901	0.599	0.490	0.433	0.350	0.323
0.07	**2.88**	0.885	0.588	0.485	0.429	0.352	0.326
0.11	**3.67**	0.877	0.599	0.490	0.433	0.350	0.323
0.12	**3.19**	0.893	0.592	0.488	0.431	0.351	0.325
0.18	**3.05**	0.885	0.592	0.485	0.431	0.352	0.326
0.19	**3.74**	0.901	0.599	0.490	0.433	0.351	0.323
0.28	**3.48**	0.893	0.595	0.488	0.433	0.351	0.324
0.29	**3.86**	0.901	0.599	0.490	0.433	0.351	0.323
0.34	**3.36**	0.893	0.595	0.488	0.431	0.352	0.326
0.35	**3.04**	0.885	0.592	0.485	0.431	0.353	0.328
0.43	**4.11**	0.901	0.602	0.493	0.435	0.352	0.324
0.48	**3.38**	0.893	0.595	0.488	0.433	0.353	0.328
0.51	**3.87**	0.901	0.599	0.490	0.435	0.353	0.326
0.56	**3.60**	0.893	0.595	0.490	0.435	0.355	0.328
0.64	**4.63**	0.909	0.602	0.495	0.437	0.353	0.325
0.68	**4.04**	0.901	0.599	0.493	0.437	0.355	0.328
0.88	**4.12**	0.909	0.602	0.498	0.441	0.360	0.332
0.92	**5.13**	0.909	0.610	0.500	0.441	0.357	0.329
0.96	**5.94**	0.917	0.610	0.500	0.442	0.357	0.328
1.01	**4.71**	0.917	0.610	0.500	0.442	0.361	0.333
1.09	**5.12**	0.917	0.613	0.503	0.444	0.362	0.333
1.22	**5.83**	0.926	0.617	0.508	0.448	0.364	0.334
1.43	**7.36**	0.943	0.625	0.513	0.452	0.366	0.337
1.59	**10.81**	0.943	0.633	0.518	0.457	0.366	0.337
1.91	**12.46**	0.971	0.645	0.529	0.465	0.375	0.344
	Average	.905	.603	.495	.438	.356	.328
	Std. Dev.	.021	.013	.011	.009	.006	.005
	Coeff. of Var.	.0233	.0223	.0219	.0204	.0175	.0161

Figure 4: Uncertainties in Scaling Factors

The formulas for E_2 and A_2 show that limits for individual values and limits for subgroup averages will suffer some uncertainty due to our not knowing the exact value for d_2. The first six columns of Figure 4 show this uncertainty will vary from 2.3 percent to 1.6 percent. By combining the uncertainty due to d_2 with the uncertainty inherent in using the Average Range we can see what happens in practice.

Figure 5 shows the incremental uncertainty in limits for charts for location that is introduced by not knowing the exact value for d_2. The bottom curve is the same as that in Figure 1. The upper curve (yes, there are two curves in each of the two panels) shows the impact of not knowing the exact value for d_2. for the worst case scenario of $n = 2$.

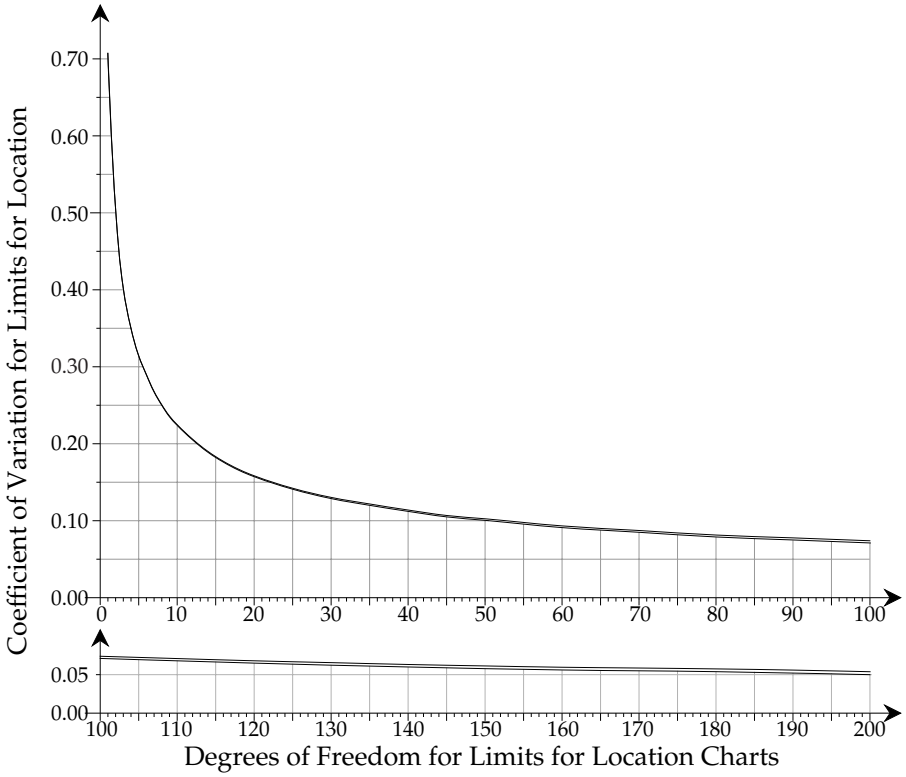

Figure 5: How Variation in d_2 Affects Limits for Location Charts

To illustrate the computations involved in Figure 5 consider what happens if you have an *XmR* chart based on $k = 331$ data. The Average Moving Range would have 200 degrees of freedom, and this would correspond to a coefficient of variation of 5.0 percent. The effective subgroup size for the ranges would be $n = 2$. Thus, if we consider the value of d_2 to be a random variable with a coefficient of variation of 2.33 percent as shown in Figure 4, we can combine the uncertainty due to d_2 with the uncertainty of the Average Moving Range. Ignoring the effects of correlation this would result in an approximate coefficient of variation of at most:

$$\text{CV}\left(\bar{R}/d_2\right) \approx \sqrt{0.0500^2 + 0.0233^2} = 0.0552$$

Thus, any uncertainty we might have about the exact value of d_2 will have a minimal impact upon our overall uncertainty. Here it causes the coefficient of variation for these limits to go

from 5 percent to 5.5 percent, which is shown by the two values in Figure 5 for 200 degrees of freedom.

For the case of an Average and Range Chart with limits based on 25 subgroups of size 2 the limits for the averages will have 22 degrees of freedom. The uncertainty due to the Average Range is therefore 15.1 percent. Include the uncertainty due to not knowing d_2 exactly and we have a coefficient of variation of at most:

$$\text{CV}\left(\bar{R}/d_2\right) \approx \sqrt{0.1508^2 + 0.0233^2} = 0.1526$$

So here the inherent uncertainty due to the Average Range statistic is 15.1 percent, and when we include the uncertainty in d_2 this goes up to 15.3 percent.

Figure 5 shows that in practice our not knowing the exact value for d_2 does not have an appreciable impact upon the limits for charts for location. For all intents and purposes the normal theory values are sufficiently close to the exact values to work with all types of data. For charts for location the fine-tuning the computations as Burr envisioned will only reduce the uncertainty in the limits by a trivial amount.

Another way of looking at the uncertainty introduced by not knowing the exact value for d_2 is to consider how many degrees of freedom would be needed before the uncertainty due to the Average Range is the same size as the uncertainty due to d_2. For $n = 2$ this is 921 degrees of freedom. This corresponds to a baseline of 1485 points on an *XmR* chart, or 1023 subgroups of size 2 for an Average and Range Chart. Until your baseline becomes this large, the uncertainty in the Average Range will dominate the uncertainty due to not knowing the exact value for d_2. Figure 6 contains the degrees of freedom for which the two sources of uncertainty reach parity for subgroup sizes greater than 2.

	Limits for Individual Values and Limits for Averages					
Subgroup Size	2	3	4	5	8	10
Degrees of Freedom	921	1005	1043	1201	1633	1929

Figure 6: Baseline Degrees of Freedom Needed for Parity Between Uncertainties

Since baselines with 1000 to 2000 degrees of freedom are rare (and are rarely rational when they do occur) we do not need to be concerned about fine-tuning the limits for the charts for location due to any potential lack of normality for the original data.

SO WHAT DO WE DO IN PRACTICE?

Remember, the objective is to take the right action. The computations are merely a means to characterize process behavior. The objective is not to compute the right number, or even to find the best estimate of the right value. You only need numbers that are good enough to allow you to separate the potential signals from the probable noise so you can take the right action. The limits on a process behavior chart are a statistical axe—they work by brute force. Just as there is no point in putting too fine of an edge on an axe, we also do not need to compute the limits with high precision. The generic three-sigma limits of a process behavior chart are sufficient to separate *dominant* cause-and-effect relationships from the run-of-the-mill routine variation. This

is why you can take the right action even when the limits are based on a few degrees of freedom.

So while Irving Burr built a more complex mousetrap, the difficulties of using his approach in practice make it less useful than the traditional approach. Instead of fine-tuning the bias correction factors to make small adjustments to the limits on a process behavior chart, it is simpler, easier, and better to use the traditional scaling factors. This will not only save you from becoming lost in the details of the computations, but also allow you to get on with the job of discovering the assignable causes that are preventing your process from operating up to its full potential.

In "Are You Sure We Don't Need Normally Distributed Data?" (***QDD*** Nov. 1, 2010.) I showed that three sigma limits will filter out virtually all of the routine variation regardless of the shape of the histogram for the original data. In "Don't We Need to Remove the Outliers" (***QDD*** Oct. 6, 2014) I illustrated the robustness of the computations. And in "Myths About Process Behavior Charts" (***QDD*** Sept. 1, 2011) I described the origin of the myth regarding normally distributed data and also showed how Shewhart's approach to the analysis of data is profoundly different from the statistical approach. Here I have shown how the traditional bias correction factors do not impose a requirement that the data be normally distributed.

The best analysis is the simplest analysis that allows you to discover what you need to know. And in this regard, the simple process behavior chart with its three-sigma limits computed using the traditional scaling factors is the undisputed champion.

In part two we will look at what happens to limits for range charts.