Process Behavior Charts for Non-Normal Data, Part Two

What happens to the range chart?

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. Last month we found that this does not appreciably affect the performance of process behavior charts for location. In this article we look at how skewed data affect the charts for dispersion.

In practice, we get appreciable skewness only when the distance between the average and the boundary condition is less than two standard deviations. A careful inspection of the six distributions shown in figure 1 will show that this corresponds to those situations where the skewness is in the neighborhood of 0.90 or larger. When the skewness is smaller than this the departure for normality is minimal, as my be seen with distribution number 15 in Figure 1.

The usual formulas for finding limits for a range chart involve the scaling factors known as D_3 and D_4.

$$\text{Lower Range Limit} = D_3 \text{ x Average Range}$$
$$\text{Upper Range Limit} = D_4 \text{ x Average Range}$$

These scaling factors depend upon the bias correction factors, d_2 and d_3.

$$D_3 = 1 - \frac{3\,d_3}{d_2} \qquad\qquad D_4 = 1 + \frac{3\,d_3}{d_2}$$

As outlined in part one, the traditional values for these bias correction factors were computed using a normal distribution for the original data. About 48 years ago Irving Burr finally found the values for the bias correction factors using 27 different non-normal distributions. Six of these distributions are shown in Figure 1.

As the distributions become more skewed the central portions of each non-normal distribution will become more concentrated. This concentration will result in a slight reduction in the average value for the distribution of the subgroup ranges. Since d_2 characterizes the mean of the distribution of subgroup ranges, we should expect the non-normal d_2 values to be slightly smaller than the normal theory values, and this is exactly what we see in Figure 2.

On the other hand, the elongated tails of the non-normal distributions should create a few more extreme values for the subgroup ranges. These extreme ranges should slightly increase the variation in the distributions of the subgroup ranges. Since d_3 characterizes the dispersion of the distribution of subgroup ranges. we should expect the non-normal d_3 values to be slightly larger than the normal theory values, and this is exactly what we see in Figure 2. Thus, the departures seen in Figure 2 are exactly what we should have expected. The question is how much do these departures affect the computation of the limits for the range chart.

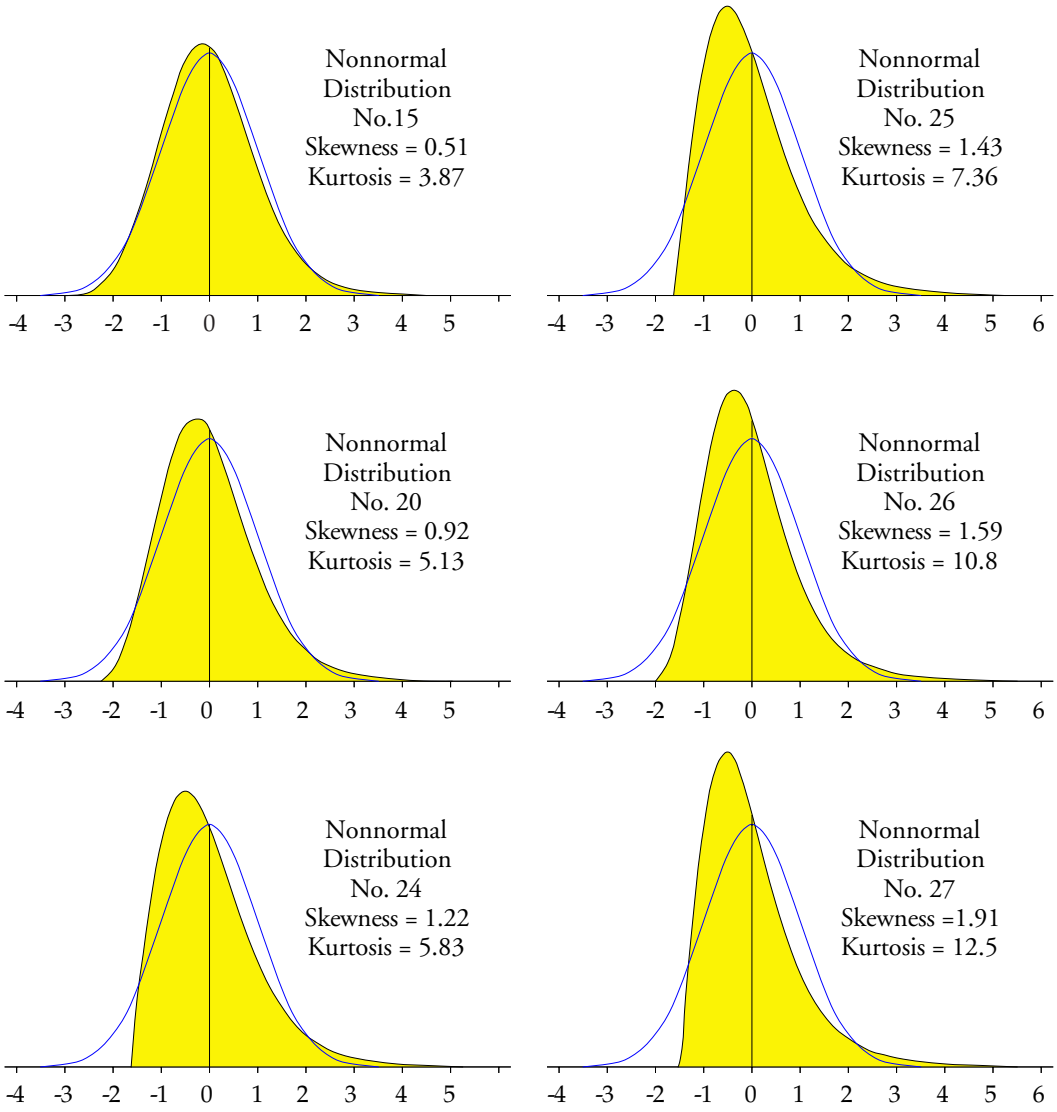

Figure 1: Six of the 27 Non-Normal Distributions Burr Used

BURR'S APPROACH

Irving Burr's original idea was that we could use the values in Figure 2 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. (This is equivalent to today's practice of letting your software fit a probability model to your data.)

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. 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.

Distributions		Values of d_2						Values of d_3					
skewness	kurtosis												
α_3	α_4	$n = 2$	3	4	5	8	10	$n = 2$	3	4	5	8	10
0.00	3.00	*1.13*	*1.69*	*2.06*	*2.33*	*2.85*	*3.08*	*0.85*	*0.89*	*0.88*	*0.86*	*0.82*	*0.80*
–.01	**3.01**	*1.13*	*1.69*	*2.06*	*2.33*	*2.85*	*3.08*	*0.85*	*0.89*	*0.88*	*0.87*	*0.82*	*0.80*
0.00	**3.33**	*1.12*	*1.68*	*2.05*	*2.32*	*2.85*	*3.09*	*0.86*	*0.91*	*0.91*	*0.90*	*0.88*	*0.86*
0.04	**3.65**	*1.11*	*1.67*	*2.04*	*2.31*	*2.86*	*3.10*	*0.87*	*0.93*	*0.94*	*0.93*	*0.92*	*0.91*
0.07	**2.88**	*1.13*	*1.70*	*2.06*	*2.33*	*2.84*	*3.07*	*0.85*	*0.88*	*0.87*	*0.85*	*0.79*	*0.77*
0.11	**3.67**	*1.14*	*1.67*	*2.04*	*2.31*	*2.86*	*3.10*	*0.87*	*0.93*	*0.94*	*0.93*	*0.92*	*0.91*
0.12	**3.19**	*1.12*	*1.69*	*2.05*	*2.32*	*2.85*	*3.08*	*0.86*	*0.90*	*0.90*	*0.88*	*0.85*	*0.83*
0.18	**3.05**	*1.13*	*1.69*	*2.06*	*2.32*	*2.84*	*3.07*	*0.85*	*0.89*	*0.88*	*0.86*	*0.82*	*0.80*
0.19	**3.74**	*1.11*	*1.67*	*2.04*	*2.31*	*2.85*	*3.10*	*0.87*	*0.93*	*0.94*	*0.94*	*0.92*	*0.92*
0.28	**3.48**	*1.12*	*1.68*	*2.05*	*2.31*	*2.85*	*3.09*	*0.86*	*0.91*	*0.91*	*0.91*	*0.88*	*0.87*
0.29	**3.86**	*1.11*	*1.67*	*2.04*	*2.31*	*2.85*	*3.10*	*0.87*	*0.93*	*0.94*	*0.94*	*0.93*	*0.92*
0.34	**3.36**	*1.12*	*1.68*	*2.05*	*2.32*	*2.84*	*3.07*	*0.86*	*0.90*	*0.90*	*0.89*	*0.86*	*0.84*
0.35	**3.04**	*1.13*	*1.69*	*2.06*	*2.32*	*2.83*	*3.05*	*0.85*	*0.89*	*0.87*	*0.86*	*0.81*	*0.78*
0.43	**4.11**	*1.11*	*1.66*	*2.03*	*2.30*	*2.84*	*3.09*	*0.88*	*0.94*	*0.95*	*0.95*	*0.94*	*0.94*
0.48	**3.38**	*1.12*	*1.68*	*2.05*	*2.31*	*2.83*	*3.05*	*0.86*	*0.90*	*0.90*	*0.88*	*0.85*	*0.83*
0.51	**3.87**	*1.11*	*1.67*	*2.04*	*2.30*	*2.83*	*3.07*	*0.87*	*0.92*	*0.93*	*0.92*	*0.91*	*0.90*
0.56	**3.60**	*1.12*	*1.68*	*2.04*	*2.30*	*2.82*	*3.05*	*0.87*	*0.91*	*0.91*	*0.90*	*0.87*	*0.86*
0.64	**4.63**	*1.10*	*1.66*	*2.02*	*2.29*	*2.83*	*3.08*	*0.88*	*0.95*	*0.96*	*0.97*	*0.97*	*0.97*
0.68	**4.04**	*1.11*	*1.67*	*2.03*	*2.29*	*2.82*	*3.05*	*0.88*	*0.93*	*0.93*	*0.93*	*0.91*	*0.90*
0.88	**4.12**	*1.10*	*1.66*	*2.01*	*2.27*	*2.78*	*3.01*	*0.88*	*0.94*	*0.94*	*0.93*	*0.90*	*0.89*
0.92	**5.13**	*1.10*	*1.64*	*2.00*	*2.27*	*2.80*	*3.04*	*0.89*	*0.96*	*0.97*	*0.98*	*0.98*	*0.98*
0.96	**5.94**	*1.09*	*1.64*	*2.00*	*2.26*	*2.80*	*3.05*	*0.90*	*0.97*	*1.00*	*1.01*	*1.03*	*1.03*
1.01	**4.71**	*1.09*	*1.64*	*2.00*	*2.26*	*2.77*	*3.00*	*0.90*	*0.97*	*0.97*	*0.97*	*0.95*	*0.95*
1.09	**5.12**	*1.09*	*1.63*	*1.99*	*2.25*	*2.76*	*3.00*	*0.90*	*0.97*	*0.98*	*0.99*	*0.98*	*0.98*
1.22	**5.83**	*1.08*	*1.62*	*1.97*	*2.23*	*2.75*	*2.99*	*0.91*	*0.99*	*1.01*	*1.02*	*1.02*	*1.03*
1.43	**7.36**	*1.06*	*1.60*	*1.95*	*2.21*	*2.73*	*2.97*	*0.93*	*1.01*	*1.04*	*1.06*	*1.09*	*1.10*
1.59	**10.81**	*1.06*	*1.58*	*1.93*	*2.19*	*2.73*	*2.97*	*0.93*	*1.02*	*1.05*	*1.08*	*1.12*	*1.14*
1.91	**12.46**	*1.03*	*1.55*	*1.89*	*2.15*	*2.67*	*2.91*	*0.95*	*1.05*	*1.09*	*1.12*	*1.16*	*1.19*

Figure 2: Burr's Values of d_2 and d_3 for 27 Non-normal Distributions

So, even though Burr's idea does not quite work out as intended in practice, we can use the values in Figure 2 to assess the effects of non-normality upon the computation of the range chart limits. As we did last month, we 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 scaling factors D_3 and D_4 involve d_3 divided by d_2. Since this ratio accentuates the slight changes in location and dispersion seen in Figure 2, we compute these ratios in Figure 3. Then, for each column in Figure 3 we compute the coefficient of variation. These coefficients of variation will summarize how our not knowing the exact values for d_2 and d_3 will affect the computation of the range chart limits.

Last month we looked at how uncertainty in the bias correction factors affected the limits for charts for location. There we found coefficients of variation in the neighborhood of 2 percent. In Figure 3 we find coefficients of variation ranging from 5 to 13 percent. So clearly the limits for the range chart are not as robust as the limits for charts for location.

As we saw last month, the uncertainty shown in Figure 3 is not the only source of uncertainty in the limits. We need to also consider the uncertainty due to the use of the average range in the computation. Recalling that the coefficient of variation for the average range is the inverse of the square root of twice the number of degrees of freedom, we can do some computations.

Distributions skewness α_3	kurtosis α_4	Values of d_3/d_2 $n=2$	3	4	5	8	10
0.00	3.00	0.752	0.527	0.427	0.369	0.288	0.260
–.01	**3.01**	0.752	0.527	0.427	0.373	0.288	0.260
0.00	**3.33**	0.768	0.542	0.444	0.388	0.309	0.278
0.04	**3.65**	0.784	0.557	0.461	0.403	0.322	0.294
0.07	**2.88**	0.752	0.518	0.422	0.365	0.278	0.251
0.11	**3.67**	0.763	0.557	0.461	0.403	0.322	0.294
0.12	**3.19**	0.768	0.533	0.439	0.379	0.298	0.269
0.18	**3.05**	0.752	0.527	0.427	0.371	0.289	0.261
0.19	**3.74**	0.784	0.557	0.461	0.407	0.323	0.297
0.28	**3.48**	0.768	0.542	0.444	0.394	0.309	0.282
0.29	**3.86**	0.784	0.557	0.461	0.407	0.326	0.297
0.34	**3.36**	0.768	0.536	0.439	0.384	0.303	0.274
0.35	**3.04**	0.752	0.527	0.422	0.371	0.286	0.256
0.43	**4.11**	0.793	0.566	0.468	0.413	0.331	0.304
0.48	**3.38**	0.768	0.536	0.439	0.381	0.300	0.272
0.51	**3.87**	0.784	0.551	0.456	0.400	0.322	0.293
0.56	**3.60**	0.777	0.542	0.446	0.391	0.309	0.282
0.64	**4.63**	0.800	0.572	0.475	0.424	0.343	0.315
0.68	**4.04**	0.793	0.557	0.458	0.406	0.323	0.295
0.88	**4.12**	0.800	0.566	0.468	0.410	0.324	0.296
0.92	**5.13**	0.809	0.585	0.485	0.432	0.350	0.322
0.96	**5.94**	0.826	0.591	0.500	0.447	0.368	0.338
1.01	**4.71**	0.826	0.591	0.485	0.429	0.343	0.317
1.09	**5.12**	0.826	0.595	0.492	0.440	0.355	0.327
1.22	**5.83**	0.843	0.611	0.513	0.457	0.371	0.344
1.43	**7.36**	0.877	0.631	0.533	0.480	0.399	0.370
1.59	**10.81**	0.877	0.646	0.544	0.493	0.410	0.384
1.91	**12.46**	0.922	0.677	0.577	0.521	0.434	0.409
	Average	.795	.565	.467	.412	.329	.301
	Std. Dev.	..043	.039	.038	.039	.039	.039
	Coeff. of Var.	.0540	.0694	.0823	.0951	.1181	.1306

Figure 3: Uncertainties in Scaling Factors for Range Chart Limits

Consider an *XmR* chart based on $k = 50$ data. The Average Moving Range will have approximately $0.62(k–1) = 30$ degrees of freedom which results in a coefficient of variation of 12.9 percent. Thus, when we combine the CV values for our two sources of uncertainty we find that the uncertainty in the limits for the Range Chart will be, at most:

$$\text{CV}\left(d_3\,\bar{R}/d_2\right) \approx \sqrt{0.1291^2 + 0.0540^2} = 0.1399$$

While the impact of the uncertainty in the bias correction factors is larger here than it is for the *X* Chart, the dominant source of uncertainty is still the uncertainty in the average range statistic, rather than the uncertainty due to not having exact values for the computation. Whether the uncertainty in the upper range limit is 13 percent or 14 percent will not greatly affect the interpretation of your *XmR* chart.

For an Average and Range Chart based upon $k = 25$ subgroups of size $n = 2$, the limits will have about $0.9\,k\,(n–1) = 22$ degrees of freedom, so the CV for the average range will be about 15.1 percent and the uncertainty in the upper limit for the Range chart will be, at most:

$$\text{CV}\left(d_3\,\bar{R}/d_2\right) \approx \sqrt{0.1508^2 + 0.0540^2} = 0.1602$$

Once again, while the impact of the uncertainty in the bias correction factors is larger here than it was with the Average Chart, it is still not appreciable. Whether the uncertainty in the upper range limit is 15 percent or 16 percent will not greatly affect the interpretation of your Average and Range Chart.

When we combine the uncertainty in the average range with the uncertainty introduced by not knowing the exact values for the bias correction factors we get the curves shown in Figure 4.

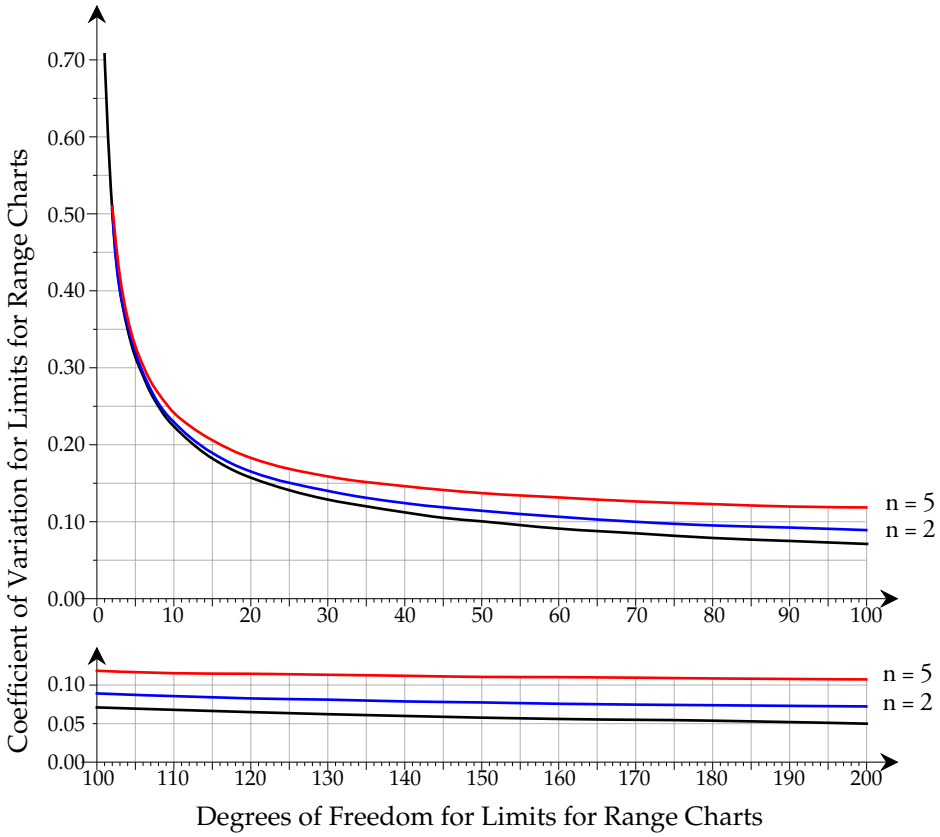

Figure 4: How Variation in Both Bias Correction Factors Affects Limits for Ranges

So, how many degrees of freedom will we need in our baseline before the two sources of uncertainty in the Range Chart limits reach parity? We can use the last row of Figure 3 and the formula relating degrees of freedom to the coefficient of variation to obtain the values in Figure 5.

	Limits for Ranges					
Subgroup Size	2	3	4	5	8	10
Degrees of Freedom	171	104	74	55	36	29

Figure 5: Baseline Degrees of Freedom Needed for Parity Between Uncertainties in Range Limits

So what can we say based on Figures 4 and 5? Initially, when the degrees of freedom are small, there is little need to be concerned about your Range Chart limits. (Here I would define small as less than half the values shown in Figure 5.) The dominant source of variation in the limits will be the uncertainty in the average range statistic and fine tuning the computation of the

upper range limit will add virtually nothing to the interpretation and use of the charts.

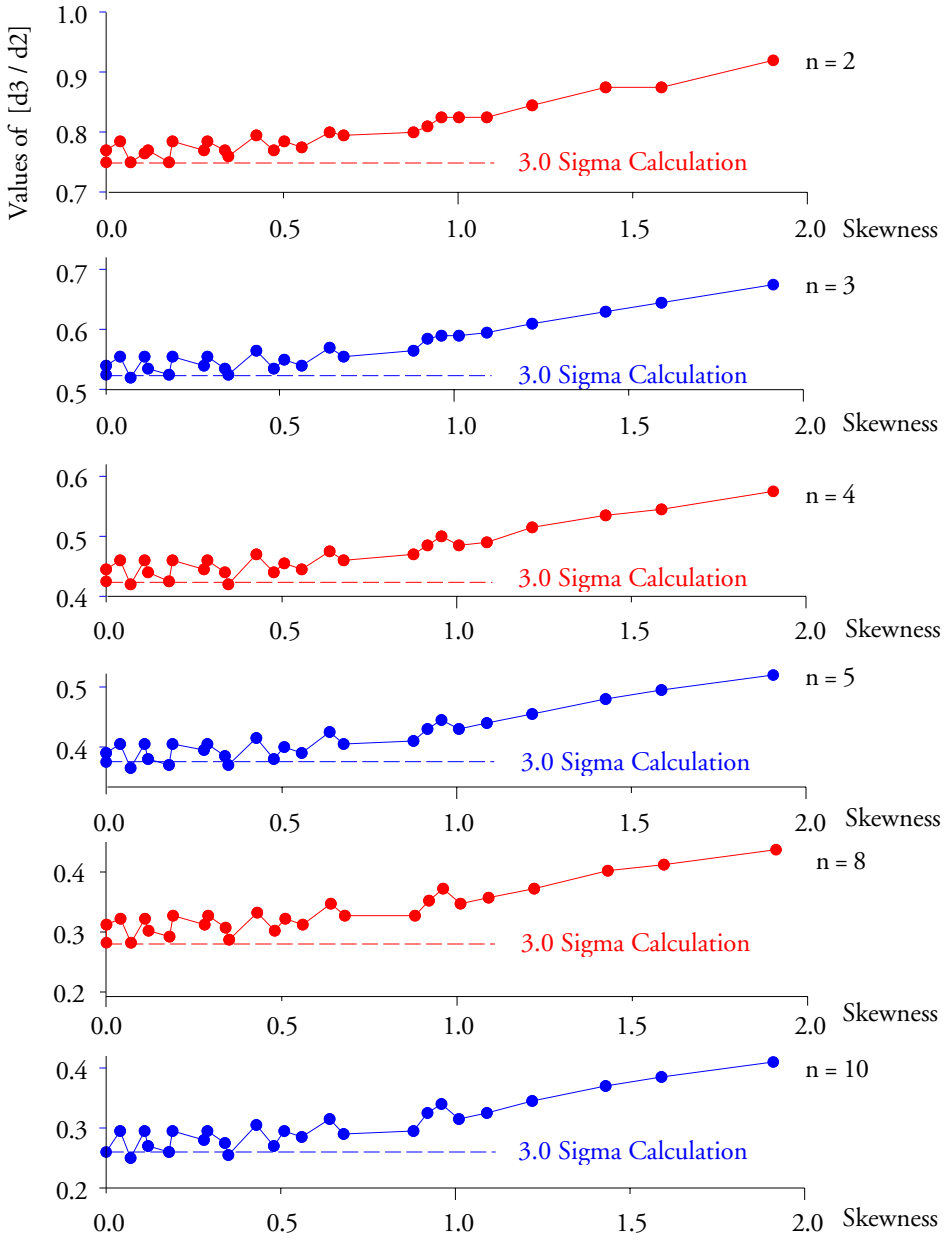

Figure 6: The Ratios of Figure 3

Also, as long as there are signals that your process is being operated unpredictably, any questions about the limits will be moot. With an unpredictable process the emphasis needs to be upon finding and removing the effects of the assignable causes of unpredictable operation. Since nine processes out of ten are operated unpredictably, this will remove the sting of having some lack of robustness for the upper range chart limit. (When you have an abundance of real signals, a few false alarms on the range chart will not matter.) Moreover, since 90 percent of your signals of unpredictable behavior will occur on the chart for location, and since signals on the range chart

are commonly accompanied by signals on the chart for location, we are not likely to be misled by some lack of robustness for the upper range chart limit. Most of the potential signals you will find with your process behavior chart will be real.

So when do you need to be concerned about the upper range limit? If you have a process where the original data pile up against a boundary or barrier condition, and if that process appears to be operating in a reasonably predictable manner, then you might want to fine-tune the upper range chart limit. But how might we go about doing this when we cannot, in practice, reliably identify a particular probability model to use? A clue on how to proceed is found in Figure 6 where the values in Figure 3 are plotted versus the skewness parameters for the different distributions.

In Figure 6 we see how the skewness of the distribution affects the computation of the upper range chart limit. The initial point for each curve shows the normal theory value for the ratio of d_3 to d_2. The horizontal lines attached to these initial values show the value of the traditional calculation for the three-sigma upper range limit. The vertical distances between the plotted points and the horizontal lines will reveal the extent to which the traditional three-sigma upper range limit will be too small for each of the 27 non-normal distributions.

Until the skewness exceeds 0.90 the points of Figure 4 tend to cluster in a horizontal band only slightly above the traditional normal theory value. But when the skewness exceeds 0.90 the curves tend to slope upward. This suggests that it is only when we have pronounced skewness that any adjustment is actually needed in the computations of the upper range chart limit. From Figure 2 we see that we will have pronounced skewness only when the average falls within two standard deviations of the barrier or boundary condition.

So, if you have a reasonably predictable process where the distance from a barrier or boundary condition to the process average is less than twice the within-subgroup estimate of the standard deviation parameter, then you may wish to inflate the upper range limit to avoid a slightly increased false alarm rate.

But how much do we inflate the upper range limit? To identify an exact value for the ratio of d_3 to d_2 we would need to pick out a specific distribution. Since we will never have sufficient data to do this in practice, and since an approximate solution is all that we need, we choose to inflate the upper limit by a fixed amount based upon the subgroup size. As we can see in Figure 7, for $n = 2$, a computed upper 3.7 sigma limit on the range chart will be sufficiently conservative to handle even the most skewed of Burr's distributions. For $n = 3$, a computed 3.8 sigma upper range limit will be sufficiently conservative to work for most of Burr's distributions. For $n = 4$ compute an upper 3.9 sigma limit on the range chart. For $n = 5$ compute an upper 4.0 sigma limit on the range chart. For $n = 10$ compute an upper 4.5 sigma limit on the range chart. By increasing the computed upper range limit by 0.1 sigma for each unit increase in the subgroup size, we obtain reasonably conservative approximations for the actual three-sigma upper range limits that will work with heavily skewed data.

We should note that the adjustments given here are merely adjustments to the computations to allow for the fact that when the original data are excessively skewed the distributions for the subgroup ranges will also become more skewed. The adjusted upper range limits are still approximate three-sigma limits even though they are computed like they are 3.7 to 4.5 sigma limits.

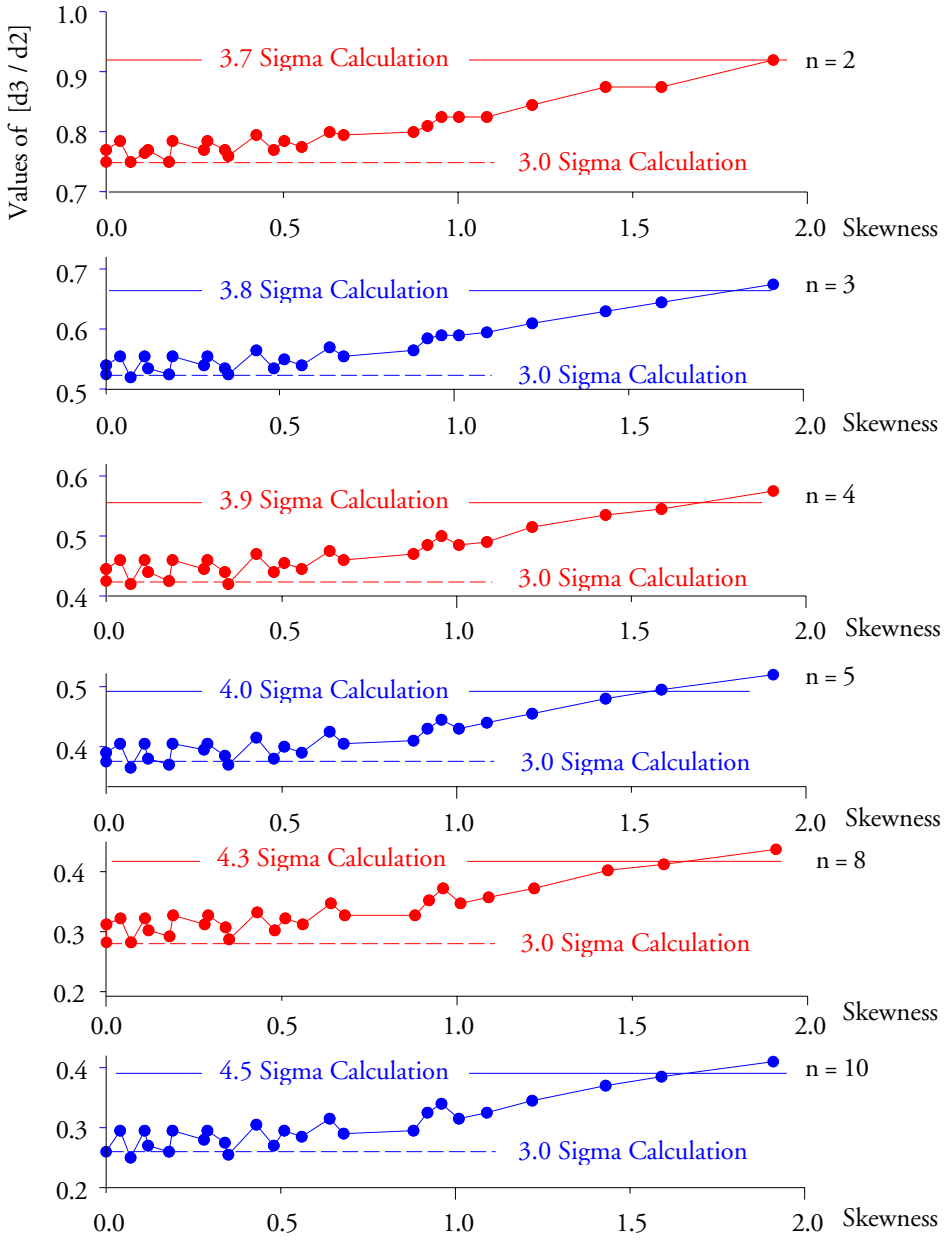

Figure 7: Adjusting the Upper Range Limit for a Predictable but Skewed Process

By computing upper range limits in keeping with the guideline shown in Figure 7 you can minimize the occurrence of false alarms on the range chart even when the original data are severely skewed. Since the bulk of true signals will occur on the charts for location, either with or without accompanying signals on the range chart, this adjustment to the computations is not needed until after the process has been operated in a reasonably predictable manner.

SO WHAT DO WE DO IN PRACTICE?

Remember, the objective is to take the right action. The computations are merely a means to help characterize the process behavior. The objective is not to compute the right number; to find

the best estimate of the right value; or to find limits that correspond to a specific alpha-level. You only need limits 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 high precision when we calculate our limits. 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 without having to specify a reference distribution, or waiting until you have some magic number of degrees of freedom.

In practice, nine times out of ten, your signals will be found on the chart for location. It is rare indeed to find signals on a range chart without accompanying signals on the chart for location. Thus, in practice we generally give more emphasis to the charts for location. This is appropriate. And as we found in Part One, we do not need to know the exact value for d_2 in order for our charts for location to work.

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 to compute the limits. 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.

If your process shows signals of exceptional variation on the chart for location, then do not attempt to assess the skewness of the histogram. When your process is going on walkabout the histogram does not represent a single process, but many different process piled up together. In this case any skewness of the histogram does not represent any inherent property of the process, but rather it characterizes the mixed up nature of the process outcomes. By far, the most common cause of a skewed histogram is a process going on walkabout.

If your process appears to be operating predictably based on the chart for location, and if the original data pile up near a boundary condition or barrier in such a way that the average is within two standard deviations of the boundary value (based on a within subgroup measure of dispersion), then you might want to adjust the upper range chart limit upward according to the guideline shown in Figure 7 in order to avoid false alarms on the range chart due to the effects of skewness.

If the original data do not have the required amount of skewness, no adjustment is needed. (This corresponds to a histogram from a predictable process having an average that is more than two sigma units away from a barrier or boundary condition.)

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. For those situations where the process appears to be operated predictably and yet the data are seriously skewed, a simple adjustment in how we compute the upper range limit can minimize false alarms without unnecessary complexity.