Do You Have Leptokurtophobia?

Donald J. Wheeler

The symptoms of leptokurtophobia are (1) routinely asking if your data are normally distributed and (2) transforming your data to make them appear to be less leptokurtic and more "mound shaped." If you have exhibited either of these symptoms then you need to read this article.

The origins of leptokurtophobia go back to the surge in SPC training in the 1980s. Before this surge only two universities in the U.S. were teaching SPC, and only a handful of instructors had any experience with SPC. As a result many of the SPC instructors of the 1980s were, of necessity, neophytes, and many things that were taught at that time can only be classified as superstitious nonsense. One of these erroneous ideas was that you have to have normally distributed data before you can put your data on a process behavior chart (also known as a control chart).

When he created the process behavior chart Shewhart was looking for a way to separate the routine variation from the exceptional variation. Since the exceptional variation, by definition, dominates the routine variation, Shewhart figured that the easiest way to tell the difference would be to filter out the bulk of the routine variation. After looking at several different ways of doing this he found that three-sigma limits will cover all, or almost all, of the routine variation for virtually all types of data.

To show how three-sigma limits do this, Figure 1 contains six different probability models for routine variation. These models range from the uniform distribution to the exponential distribution. (The last three models are leptokurtic.) Each of these models is standardized so that they all have a mean of zero and a standard deviation parameter of 1.00. Figure 1 shows the three-sigma limits and that proportion of the area under each curve that falls within those three-sigma limits.

There are four lessons that can be learned from Figure 1.

The first lesson of Figure 1 is that three-sigma limits will filter out virtually all of the routine variation regardless of the shape of the histogram. These six models are radically different, yet in spite of these differences, three-sigma limits cover 98 percent to 100 percent of the area under each curve.

The second lesson of Figure 1 is that any data point that falls outside the three-sigma limits is a potential signal of a process change. Since it will be a rare event for routine variation to take you outside the three-sigma limits, it is more likely that any point that falls outside these limits is a signal of a process change.

The third lesson of Figure 1 is that symmetric, three-sigma limits work with skewed data. Four of the six models shown are skewed. As we scan down the figure we see that no matter how skewed the model, no matter how heavy the tail becomes, the three-sigma limits are stretched at essentially the same rate as the tail. This means that the length of the elongated tail will effectively determine the three-sigma distance in each case, and that three-sigma limits will cover the bulk of the elongated tail no matter how skewed the data become.

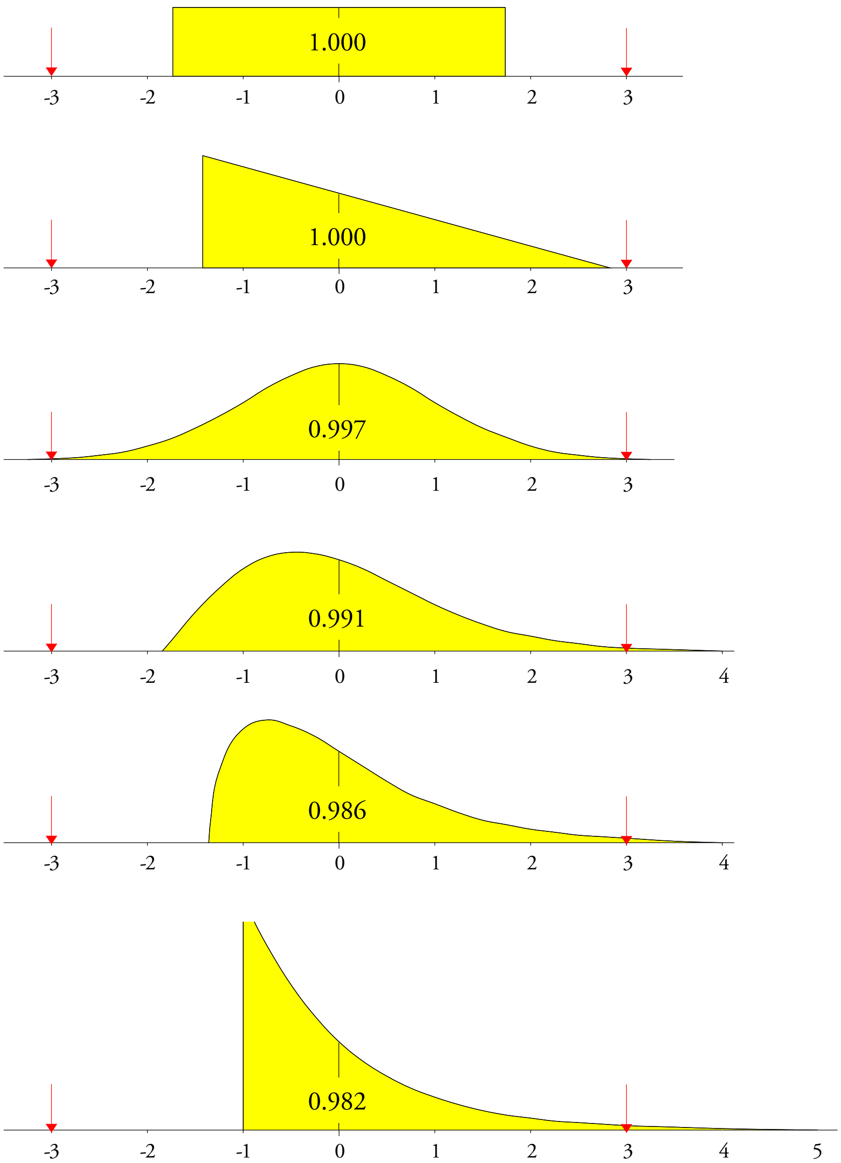


Figure 1: How Three-Sigma Limits Filter Out Virtually All of the Routine Variation Regardless of the Probability Model Used

"But that certainly makes the other limit look silly." Yes, it does. Here we need to pause and think about those situations where we have skewed data. In most cases we skewed data occur when the data pile up against a barrier or boundary condition. Whenever a boundary value falls within the computed limits, the boundary takes precedence over the computed limit, and we end up with a one-sided chart. When this happens the remaining limit covers the long tail and allows us to separate the routine variation from potential signals of deviation away from the boundary. Which is how symmetric, three-sigma limits can work with skewed data.

The fourth lesson of Figure 1 is that any uncertainty in where we draw the three-sigma lines will not greatly affect the coverage of the limits. All of the curves are so flat by the time they reach the neighborhood of the three-sigma limits that any errors we may make when we estimate the limits will have, at most, a minimal impact upon how the chart works.

The six probability models in Figure 1 effectively summarize what was found when this author looked at over 1100 different probability models from seven commonly used families of models. These 1143 models effectively covered all of the shape characterization plane, with 916 mound-shaped models, 182 J-shaped models, and 45 U-shaped models. Eleven hundred and twelve of these models (or 97.3%) had better than 97.5 percent of their area covered by symmetric three-sigma limits.

Thus, three-sigma limits work by brute force. They are sufficiently general to work with all types and shapes of histograms. They work with skewed data, and they work even when the limits are based on few data.

To illustrate this point I used the Exponential probability model from Figure 1 to generate the 100 values shown in rows in Table 1. The histogram for these values is shown in Figure 2. Since such values should, by definition, display only routine variation, we would hope to find almost all of the observations within the limits in Figure 3. We do. Hence, the process behavior chart will work as advertised even with skewed data.

Table 1: 100 Observations from the Standardized Exponential Distribution

-0.53	0.34	-0.49	0.47	0.14	-0.44	-0.90	1.38	0.91	-0.68
-0.75	0.58	-0.48	0.00	-0.85	1.78	-0.86	-0.89	0.31	-0.63
0.95	-0.35	-0.65	-0.84	0.30	-0.54	-0.61	0.08	-0.98	1.00
1.35	-0.03	-0.86	0.55	-0.78	-0.67	0.84	1.28	0.34	-0.07
-0.48	0.07	2.11	-0.65	-0.16	0.90	0.42	-0.25	-0.80	0.33
-0.29	2.36	-0.52	2.21	-0.72	1.27	-0.98	2.53	-0.41	-0.94
-0.44	-0.88	-0.63	-0.90	-0.90	-0.85	0.20	-0.95	-0.66	-0.50
0.58	-0.99	-0.55	-0.61	0.02	-0.79	0.24	-0.20	-0.25	-0.15
-1.00	-0.65	-0.28	-0.31	0.74	-0.70	1.87	-0.81	-0.67	0.43
-0.68	-0.08	0.08	-0.59	-0.87	1.31	0.52	-0.34	0.53	-0.44

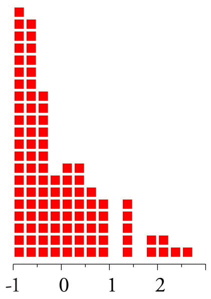


Figure 2: Histogram of 100 Exponential Observations

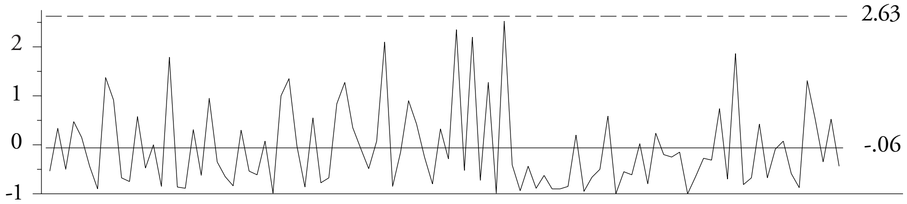


Figure 3: X Chart for 100 Exponential Observations

Therefore, we do not have to pre-qualify our data before we place them on a process behavior chart. We do not need to check the data for normality, nor do we need to define a reference distribution prior to computing limits. Anyone who tells you anything to the contrary is simply trying to complicate your life unnecessarily.

TRANSFORMATIONS OF THE DATA

"But the software suggests transforming the data!" Such advice is simply another piece of confusion. The fallacy of transforming the data is as follows.

The first principle for understanding data is that no data have meaning apart from their context. Analysis begins with context, is driven by context, and ends with the results being interpreted in the context of the original data. This principle requires that there must always be a link between what you do with the data and the original context for the data. Any transformation of the data risks breaking this linkage.

If a transformation makes sense both in terms of the original data and the objectives of the analysis, then it will be okay to use that transformation. Transformations of this type might be things like the use of daily or weekly averages in place of hourly values, or the use of proportions or rates in place of counts to take into account the differing areas of opportunity in different time periods.

Only you as the user can determine when a transformation will make sense in the context of the data. (The software cannot do this because it will never know the context.) Moreover, since these sensible transformations will tend to be fairly simple in nature, they do not tend to distort the data.

A second class of transformations would be those that rescale the data in order to achieve certain statistical properties. (These are the only type of transformations that any software can suggest.) Here the objective is usually to make the data appear to be more "normally distributed" in order to have an "estimate of dispersion that is independent of the estimate of location." Unfortunately, these transformations will tend to be very complex and nonlinear in nature, involving exponential, inverse exponential, or logarithmic functions. (And just what does the logarithm of the percentage of on-time shipments represent?) These nonlinear transformations will distort the data in two ways: at one end of the histogram, values that were originally far apart will now be close together; at the other end of the histogram, values that were originally close together will now be far apart.

To illustrate the effect of transformations to achieve statistical properties we will use the Hot Metal Transit Times shown in rows in Table 2. These values are the times (to the nearest five minutes) between the phone call alerting the steel furnace that a load of hot metal was on the way and the actual arrival of that load at the steel furnace ladle house.

Table 2: The Hot Metal Transit Times in Minutes

40	45	125	100	40	40	100	65	55	40	125	65	40	45	95
105	45	110	40	50	120	45	65	105	35	70	55	25	50	55
50	40	40	45	55	50	45	125	55	100	40	70	40	40	110
55	50	30	50	105	45	45	55	50	25	65	60	60	55	70
55	45	100	60	45	145	45	50	65	180	60	45	35	35	55
55	55	50	120	35	45	35	45	55	50	70	45	75	60	45
60	40	60	40	50	60	65	95	65	60	50	25	25	100	50
60	45	35	40	30	180	50	30	30	30	65	130	80	20	45
65	65	45	40	50	25	120	30	115	50	85	40	35	40	40
55	50	25	75	55	50									

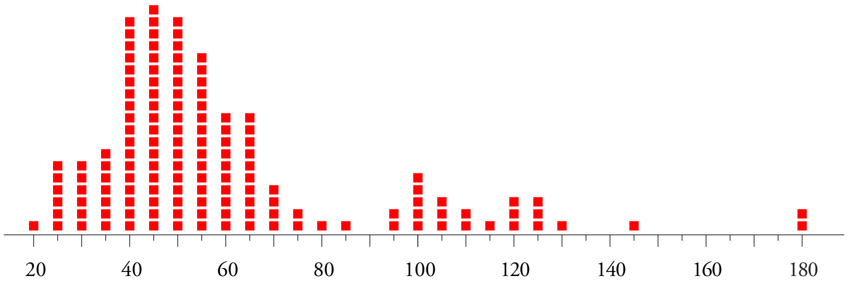


Figure 4: Histogram of the Hot Metal Transit Times

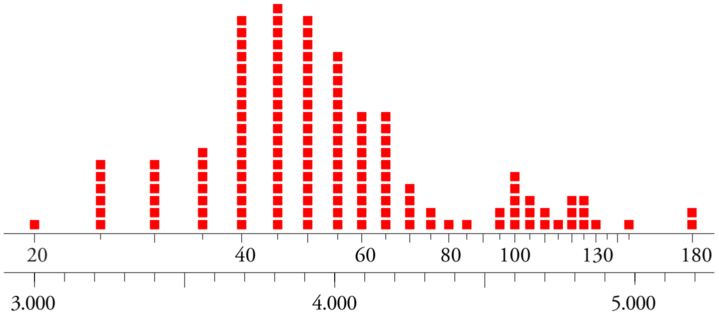


Figure 5: Histogram of the Logarithms of Hot Metal Transit Times

Given the skewed nature of the data in Figure 4 some programs would suggest using a logarithmic transformation. Taking the natural logarithm of each of these transit times results in the histogram in Figure 5. (The horizontal scales show both the original and transformed values.) Notice how the values on the left of Figure 5 are spaced out while those on the right are crowded together. After the transformation the distance from 20 to 25 minutes is about the same size as the distance from 140 to 180 minutes. How could you begin to explain this to your boss?

By itself, this distortion of the data is sufficient to call into question the practice of transforming the data to achieve statistical properties. However, the impact of these non-linear transformations is not confined to the histograms.

Figure 6 shows the *X* Chart for the original, untransformed data of Table 2. Eleven of the 141 transit times are above the upper limit, confirming the impression given by the histogram that these data come from a mixture of at least two different processes. Even after the steel furnace gets the phone call, they still have no idea when the hot metal will arrive at the ladle house.

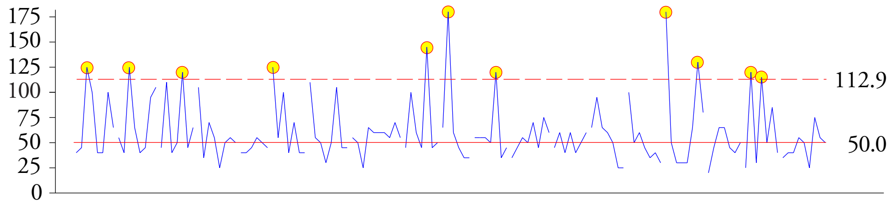


Figure 6: *X* Chart for the Hot Metal Transit Times

However, if we transform the data before we put them on a process behavior chart we end up with Figure 7. There we find no points outside the limits!

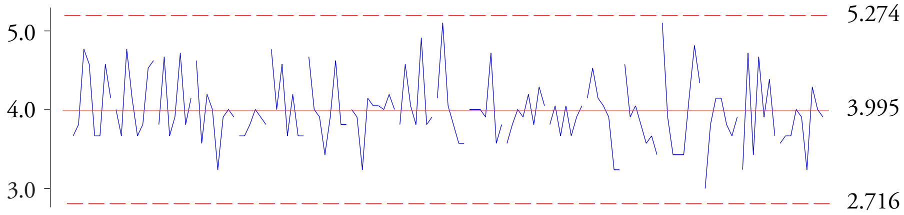


Figure 7: *X* Chart for the Logarithms of the Hot Metal Transit Times

Clearly the logarithmic transformation has obliterated the signals. What good is a transformation that changes the message contained within the data? The transformation of the data to achieve statistical properties is simply a complex way of distorting both the data and the truth.

The results shown here are typical of what happens with nonlinear transformations of the original data. These transformations hide the signals contained within the data simply because they are based upon computations that *presume there are no signals within the data.*

To see how the computations do this we need to pause to consider the nature of the formulas for common descriptive statistics. For a descriptive measure of location we usually use the average, which is simply based upon the sum of the data. However, once we leave the average behind the formulas become much more complex. For a descriptive measure of dispersion we commonly use the global standard deviation statistic, which is a function of the *squared deviations from the average*. For descriptive measures of shape we commonly use the skewness and kurtosis statistics which, respectively, depend upon the *third* and *fourth* powers of the deviations of the data from the average. When we aggregate the data together in this manner and use the second, third, and fourth powers of the distance between each observation and the average value, we are implicitly assuming that these computations make sense. Whether they be measures of dispersion, or measures of skewness, or even measures of kurtosis, *any high-order descriptive statistic that is computed globally is implicitly based upon a very strong assumption that the data are homogeneous*.

When the data are not homogeneous it is not the *shape* of the histogram that is wrong, but the computation and use of the descriptive statistics that is erroneous. We do not need to distort the histogram to make the transformed values more homogeneous, but we need to stop and question what the lack of homogeneity means in the context of the original observations.

So how can we determine when a data set is homogeneous? *That is the purpose of the process behavior chart!* Transforming the data to achieve statistical properties prior to placing them on a process behavior chart is an example of getting everything backwards. It assumes that we need to make the data more homogeneous prior to checking them for homogeneity. Any recommendation regarding the transformation of the data prior to placing them on a process behavior chart reveals a fundamental lack of understanding about the purpose of process behavior charts.

Shewhart's approach, with its generic three-sigma limits computed empirically from the data, does not even require the specification of a probability model. In fact, on page 54 of *Statistical Method from the Viewpoint of Quality Control*, Shewhart wrote *"…we are not concerned with the functional form of the universe* [i.e., the probability model], *but merely with the assumption that a universe exists."* [Italics in the original.]

When you transform the data to achieve statistical properties you deceive both yourself and everyone else who is not sophisticated enough to catch you in your deception. When you check your data for normality prior to placing them on a process behavior chart you are practicing statistical voodoo. Transforming the data prior to using them on a process behavior chart is not only bad advice, it is also an outright mistake.

Whenever the teachers lack understanding, superstitious nonsense is inevitable. Until you learn to separate myth from fact you will be fair game for those who were taught the nonsense. And you may end up with leptokurtophobia without even knowing it.

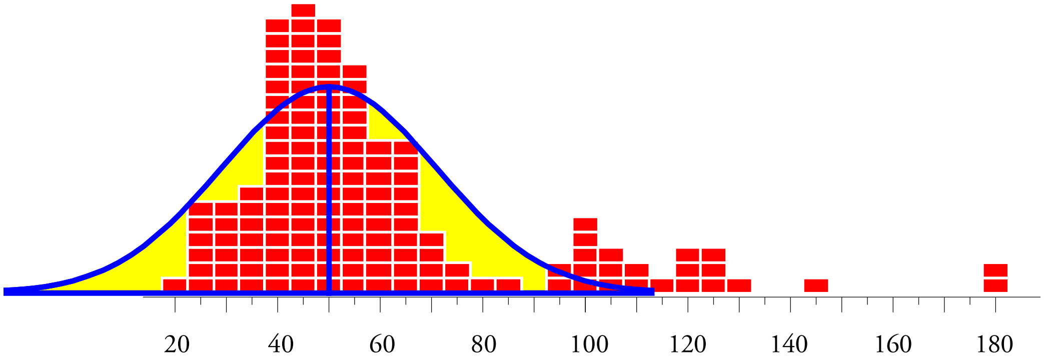
SIDEBAR: DO YOU HAVE LEPTOKURTOPHOBIA?

Leptokurtophobes are those who have an irrational fear of using non-normal data. Symptoms include continually asking if your data are normally distributed and transforming your data to make them appear to be more like a normal distribution prior to using them in a statistical analysis. This phobia was originally held in check by the difficulty of performing the nonlinear transformations usually required. It has recently become epidemic due to the availability of software that will perform the complex transformations for the leptokurtophobe.

Leptokurtosis literally means "thin mound" and refers to probability models that have a central mound that is narrower than that of a normal distribution. In practice, due to the mathematics, leptokurtosis actually refers to those probability models having heavier tails than the normal distribution. By a wide margin, most leptokurtic distributions are also skewed, and most skewed distributions will be leptokurtic.

The origins of leptokurtophobia go back to the surge in SPC training in the 1980s. Before this surge only two universities in the U. S. were teaching SPC, and only a handful of instructors had any experience with SPC. As a result many of the SPC instructors in the 1980s were, of necessity, neophytes, and many things that were taught at that time can only be classified as superstitious nonsense. One of these erroneous ideas was that you have to have normally distributed data before you can put your data on a process behavior chart (also known as a control chart).

For more information on this topic see Chapters Seven and Eight of Dr. Wheeler's newest book, *Twenty Things You Need to Know*.


20
40
60
80
100
120
140
160
180