What is Leptokurtophobia?

And why does it matter?

Donald J. Wheeler

Three years ago this month I published "Do You Have Leptokurtophobia?" Based on the reaction to that column it contained a message that was needed. In this column I would like to explain the symptoms of leptokurtophobia and the cure for this pandemic affliction.

Leptokurtosis is a Greek word that literally means "thin mound." It was used to describe those probability models that have a central mound that is narrower than that of a normal distribution. In reality, due to the mathematics involved, a leptokurtic probability model is one that has heavier tails than the normal distribution. By a wide margin, most leptokurtic distributions are also skewed, and most skewed distributions will be leptokurtic.

The fear of leptokurtosis can be traced back to the surge in training in SPC 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 of the surge, of necessity, many of the SPC instructors of the 1980s were 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). Over the years this simple but incorrect idea has grown and mutated into a prohibition on doing any statistical analysis without first testing the data for normality or defining a reference probability model for the data.

Therefore, you may have leptokurtophobia if you have an irrational fear of using non-normal data in your analysis. Symptoms include asking if your data are normally distributed, transforming your data to make them more "mound-shaped," or fitting a probability model to your data as the first step in your analysis. This phobia was originally held in check by the complexity of the remedies, such as performing a non-linear transformation or computing a lack-of-fit statistic. However, due to the availability of software that will perform these complex operations, today we find leptokurtophobia to be truly pandemic, with outbreaks occurring around the world. People are fitting probability models and transforming data with a few keystrokes, and as a result they are unknowingly suffering undesirable side-effects. Insidiously, while these side-effects have few symptoms, they tend to completely undermine your analysis and your predictions.

Let's begin with the problem of fitting a probability model to your data. Figure 1 shows a histogram of the number of major hurricanes per year in the North Atlantic for 1940 through 2007. These 68 counts have an average of 2.59. Using this value as the mean value for a Poisson distribution a lack-of-fit test will fail to find any detectable lack of fit. Therefore, we might well conclude that a Poisson probability model with a mean of 2.59 is a reasonable model to use. From this we might then characterize the likelihood of various numbers of major hurricanes in a given year. Specifically, the probability of getting seven or more major hurricanes in a single year is found to be 0.017. Thus, in 68 years we should expect to find about one year with seven or more major hurricanes.

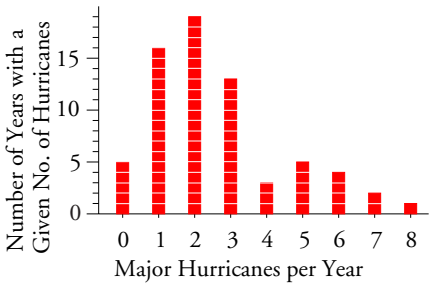

Figure 1: North Atlantic Major Hurricanes

However, NOAA researchers think that these data represent two different weather patterns. They call the change between these patterns the "multi-decadal tropical oscillation." They break this time period of 1940 to 2007 into four segments. In the time period used here the era of lower activity includes 1940 to 1947 and 1970 to 1994. The era of higher activity includes 1948 to 1969 and 1995 to 2007. The histograms for these two eras are seen in Figure 2.

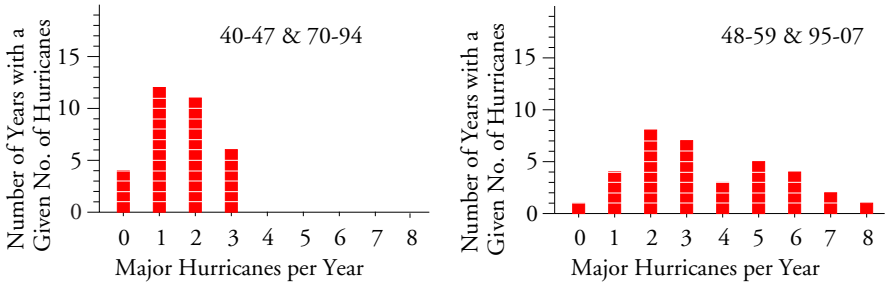

Figure 2: North Atlantic Major Hurricanes

During the era of low activity the average number of major hurricanes per year was 1.58. During the era of high activity this average doubled to 3.54 per year. So, which years would you say are characterized by the average of 2.59 major hurricanes per year? Clearly, this average does not apply to the era of low activity, neither does it characterize the era of high activity. While your model based on Figure 1 predicts one year with seven or more major hurricanes, the data show three years with seven or eight major hurricanes.

Whenever you fit a model to your data you are assuming that those data are homogeneous. If they are not homogeneous, all of your statistics, all of your models, and all of your predictions are going to be wrong.

Well, if fitting a probability model is not the answer, what about transforming the data?

When you transform the data you are reshaping it to fit your preconceived notions. This is always a dangerous thing to do. Figure 3 shows the histogram of 141 hot metal transit times. These values are the times (to the nearest five minutes) between the call alerting the steel furnace that a load of hot metal was on the way and the actual arrival time of that load at the steel furnace ladle house. The average delivery time is 60 minutes. The standard deviation is 30 minutes. The skewness is 1.70, and the kurtosis is 6.0. (Anything above 3.0 is

leptokurtic.) As they stand they form a very skewed and heavy tailed histogram.

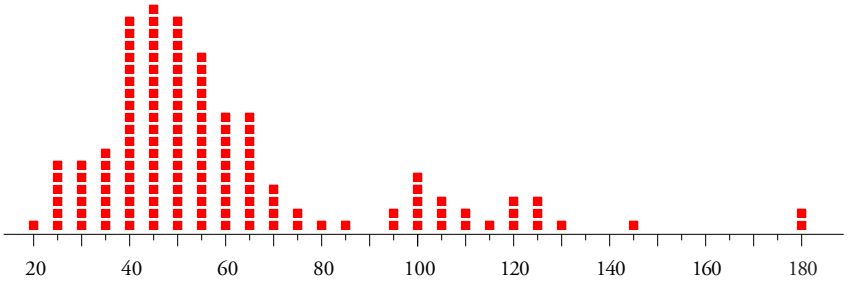

Figure 3: Hot Metal Transit Times

Some software packages would suggest a logarithmic transformation for these data. Taking the natural logarithm of each of these transit times results in the histogram in figure 4. There the horizontal scales show both the original and the transformed values. The logarithmic transformation has spaced out the values on the left and has crowded the values on the right together so that the overall shape of the histogram is much more "mound shaped" than before. But is this an improvement? Now the "distance" from 20 minutes to 25 minutes is about the same size as the "distance" from 140 minutes to 180 minutes. How are you going to explain this to your boss? While the original histogram clearly showed a histogram with two and possibly three humps, the transformed histogram blurs this important feature of the data.

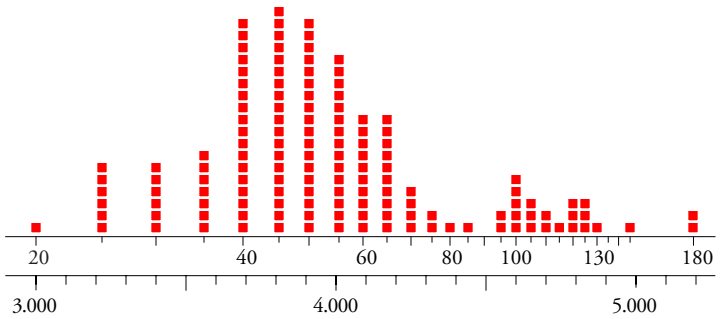

Figure 4: Logarithms of the Hot Metal Transit Times

By itself, this distortion of the data should be sufficient to make you want to avoid the practice of transforming the data to achieve statistical properties. However, the impact of non-linear transformations is not confined to the histograms.

One of the major reasons for analyzing data is to detect signals buried within those data. And when we go looking for signals, the premier technique will be the process behavior chart. Figure 5 shows the *X* Chart for the original hot metal transit times. 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 do not have any idea about when the hot metal will arrive in the ladle house.

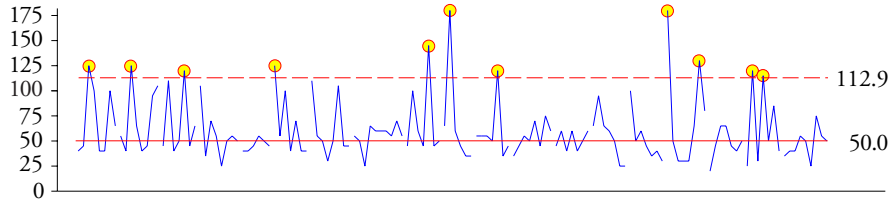

Figure 5: *X* Chart for the Hot Metal Transit Times

However, if we use a non-linear transform on the data prior to placing them on a process behavior chart we end up with the *X* chart shown in Figure 6. There we find no points outside the limits!

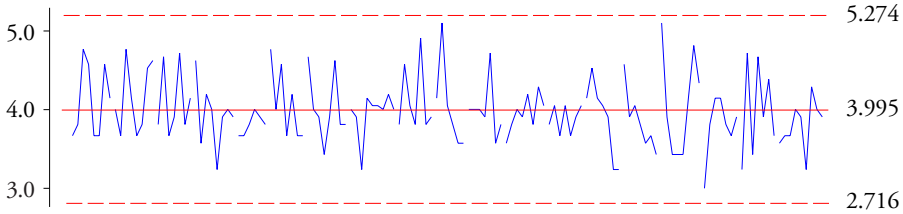

Figure 6: *X* Chart for the Logarithms of the Hot Metal Transit Times

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 on computations that presume there are no signals within in the data.*

(For more on the hurricane data see my ***Quality Digest*** columns for February and March of 2009. For more on the problems of transforming the data see my ***Quality Digest Daily*** column of August 5, 2009. For an explanation of how three-sigma limits work with non-normal data see my ***Quality Digest Daily*** column of November 1, 2010.)

So, what should be the first question of data analysis? Should you try to accommodate to the shape of the histogram by fitting a probability model? Should you seek to reshape the histogram by using some non-linear transformation? Or should you check the data for evidence of a lack of homogeneity? Since a lack of homogeneity will undermine the fitting of a probability model, and since it will invalidate the rationale for the transformation of the data, it is imperative that we begin by checking for possible nonhomogeneity.

So how can we determine when a data set is homogeneous? That is what the process behavior chart was created to do! This is why it is essential to begin any analysis by organizing your data in a logical manner and placing them on a process behavior chart. If you do not have the requisite homogeneity, anything else you might do will be flawed.

When you fit a probability model to your data you are making a strong assumption that the data are homogeneous. If they are not homogeneous, then your model, your analysis,

and your predictions will all be wrong. 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.

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.