# The Heavy-Tailed Normal

## Answers to Questions

Donald J. Wheeler

Last month I showed how the normal distribution is the distribution of maximum uncertainty. In this column I will expand on that theme and answer the questions generated by that column.

Last month I demonstrated that the middle 91 percent of the normal distribution is spread out to the maximum extent possible for any unimodal probability model and that the outer 9 percent of the normal distribution is as far, or further, away from the mean than the outer nine percent of any unimodal probability model. This property of the normal distribution is not a fluke, but is part of a systematic pattern where the outer tails of the normal distribution dominate the outer tails of other unimodal distributions.

In Figure 1 we see the central intervals that bracket 88%, 89%, 90%, 91%, and 92% for a standard normal distribution. These five intervals respectively define outer tail areas of 12%, 11%, 10%, 9%, and 8%. For any central interval between ± 1.55 and ±1.75, the standard normal distribution will have a smaller percentage within that interval, and a larger percentage outside that interval, than virtually any other standardized unimodal probability model.

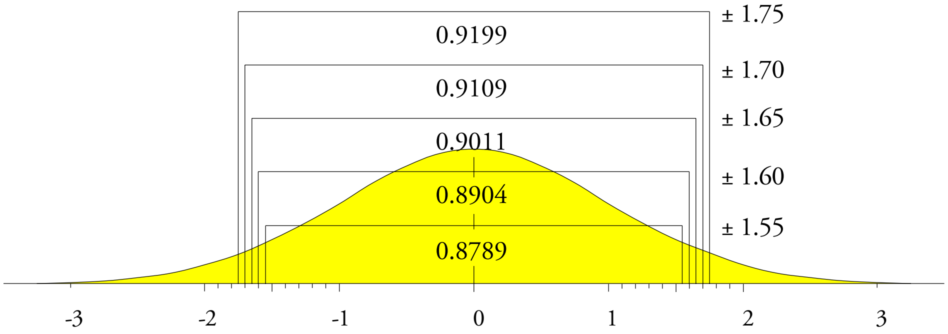

**Figure 1: Five Central Intervals for a Normal Distribution**

In support of the statement above I computed the percentages within and outside each of these five central intervals using 3366 different probability models. (This set of probability models was chosen to give uniform coverage to that region of the shape characterization plane where kurtosis is less than 12. This is substantially the same as the set of distributions that was used in my September article.) Figure 2 summarizes the results of this study. There each box-plot shows the outer tail areas for the 3366 distributions for each of the central intervals. The dots show the outer tail areas for the standard normal.

For a central interval of ± 1.55 the 3366 outer tail areas ranged from 7.0% to 12.7% with a median value of 8.1%. The standard normal distribution has 12.1% outside this interval. In this case seventeen distributions had slightly heavier outer tails than the normal.

For a central interval of ± 1.60 the 3366 outer tail areas ranged from 6.5% to 11.2% with a median value of 7.6%. The standard normal distribution has 11.0% outside this interval. Here eleven distributions had slightly heavier outer tails than the normal.

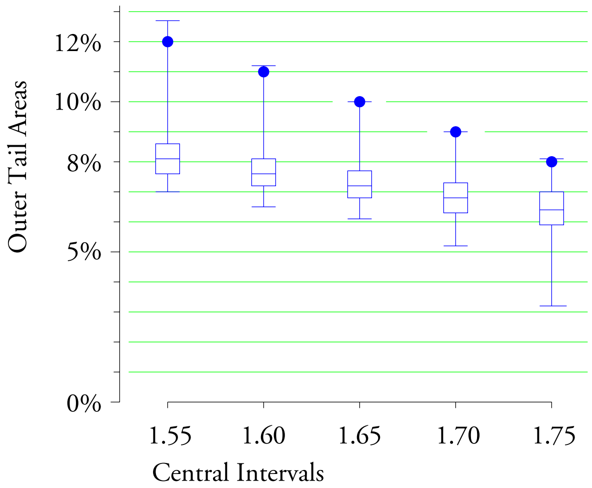

**Figure 2: Outer Tail Areas for 3366 Distributions for the Five Central Intervals**

For a central interval of ± 1.65 the 3366 outer tail areas ranged from 6.1% to 10.0% with a median value of 7.2%. The standard normal distribution has 9.9% outside this interval. Here nine distributions had slightly heavier outer tails than the normal.

For a central interval of ± 1.70 the 3366 outer tail areas ranged from 5.2% to 9.0% with a median value of 6.8%. The standard normal distribution has 8.9% outside this interval. Here six distributions had slightly heavier outer tails than the normal

And for a central interval of ± 1.75 the 3366 outer tail areas ranged from 3.3% to 8.1% with a median value of 6.4%. The standard normal distribution has 8.0% outside this interval. In this case nine distributions had slightly heavier outer tails than the normal.

While in each case a small number of distributions were found to have slightly heavier outer tails than the normal, as we move from case to case these distributions change. None of the most extreme distributions were more extreme than the normal in all five of the cases considered. Thus, while the normal may not be the most extreme in any one case, it is the most extreme over all the cases combined. This is the essence of what is meant by the normal distribution having maximum entropy.

## STUDENT'S T-DISTRIBUTION

"But I thought the t-distribution had heavier tails than the normal. This is the way all the pictures in the books are all drawn."

The traditional picture for a Student's t-distribution does show it with heavier tails than a normal distribution. This happens because the units used in drawing the traditional picture are multiples of the standard deviation of the average statistic, SD(X-bar). If we redraw the traditional picture in standardized form we will have to use the standard deviation of the random variable T. This standard deviation depends upon the degrees of freedom:

$$SD(T) \;=\; \sqrt{\frac{\textit{degrees of freedom}}{\textit{degrees of freedom} - 2}}$$

Figure 3 shows three of these standardized t-distributions along with their proportions within an interval of ± 1.65 SD(T). These proportions drop as the degrees of freedom increase.

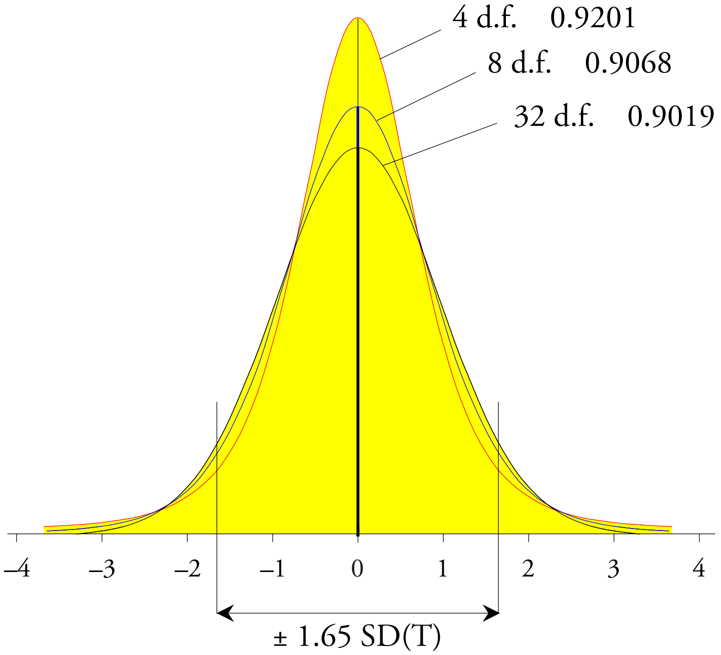

**Figure 3: Four Standardized Student's t-Distributions**

Specifically, for a t-distribution with 4 degrees of freedom, SD(T) = 1.414 and 1.65 SD(T) is 2.333. The area between -2.333 and +2.333 is 0.9201 and the outer tail area is 0.0799.

For a t-distribution with 8 degrees of freedom, SD(T) = 1.155 and 1.65 SD(T) is 1.905. The area between –1.905 and +1.905 is 0.9068 and the outer tail area is 0.0932.

For a t-distribution with 32 degrees of freedom, SD(T) = 1.033 and 1.65 SD(T) is 1.704. The area between –1.704 and +1.704 is 0.9019 and the outer tail area is 0.0981.

In contrast with these values, the area between –1.650 and +1.650 for the standard normal distribution is 0.9011 and the outer tail area is 0.0989. Thus, the standard normal distribution has the smallest area within the interval ±1.65 SD, and consequently the largest area outside this interval. All of the t-distributions have less than 9.89 percent outside the interval ± 1.65 SD(T), hence they all have lighter outer tails than the normal distribution.

## THE FOUNDATIONS OF STATISTICS

"Does this result undermine the foundations of statistics?"

No, not at all. Statistical techniques and tests work because of the way they are structured. However, this structure has not always been clearly taught and as a result many misconceptions are common.

Consider the t-test: The original data X are summarized using the average and the standard deviation statistic. Next we use these two statistics to construct a t-statistic. Then we use the t-statistic and the critical values from the t-distribution in our statistical analysis. Note that it is not the original data X, but rather the t-statistic that is characterized by the t-distribution. *In fact, as has been rigorously proven in the statistical literature, the t-statistic will be reasonably characterized by the t-distribution regardless of what distribution applies to the original data X.* This property is called robustness.

This structure is present in most of what we do in statistics. We collect our original data, and then we compute some complex statistic from these data. When the procedure is robust, the distribution of the complex statistic will be reasonably well known regardless of what probability model might characterize the original data. Thus, our distributional assumptions are generally applied not to the original data, but to some function of those data. By working, as it were, one step removed from the original data we find that we can use any of several forms of the central limit theorem to stabilize our results and thereby to finesse the question of "How are the original data distributed?" This approach generalizes our techniques while it simplifies their application.

"But why were we taught to check for normality before we do almost anything else in statistics?" This recommendation is not found in the standard textbooks, but is rather something that comes from the large pool of unqualified instructors trying to figure out how to use the software that is available. It is a fact that today most industrial statistics classes have been, and are being taught by people who do not have graduate degrees in statistics. In a field that rests upon the complexity of probability theory, this lack of understanding of the foundations will inevitably result in superstitious nonsense being taught as gospel. So if you were taught to check for normality, or to fit a probability model to your data as the first step in your statistical analysis, you now have permission to simplify your life by skipping over these unnecessary and inappropriate steps.

But the journal articles often start off with the assumption that the data are normally distributed. Isn't this important? What I illustrated in my September column is that the assumption of normally distributed data is essentially a worst case assumption. This assumption allows the middle 90 percent of the data to be spread out to the maximum extent possible and for the outer 10 percent to be as far away from the mean as possible. Moreover, making a worst-case assumption in order to compute the distribution of some complex function of the data is not the same as requiring the data to actually behave in the worst case manner. We obviously hope that our test statistic will be robust, and that we can therefore avoid any question about the distribution of the original data. This is why it is the robust techniques such as t-tests and F-tests for means that get used most often.

So, once you make a distinction between fitting a probability model to the original data, and having a test statistic that has a reasonably well known distribution regardless of the distribution of the original data, the fact that we cannot successfully fit non-normal models to most data sets will not upset the use of the common techniques of statistical analysis.

## DOES THIS PRESENT AN ALTERNATIVE APPROACH TO ANALYSIS?

While knowledge that the normal distribution has heavier outer tails than other unimodal distributions may be used to provide an approximate analysis, there is no advantage to using it in place of a standard analysis. For example consider computing an interval estimate for the mean. A 90% interval estimate for the mean based upon a t-distribution with 4 degrees of freedom will use the t-distribution critical values of ± 2.132. From Figure 3 above, with 4 d.f., we found that ±1.65 SD(T) = ± 2.333. Thus, we could use ± 2.333 to obtain a conservative approximation to the 90% interval estimate for the mean, but the actual t-values will provide a narrower interval estimate with the same uncertainty.

On the other hand, when estimating some parameter via simulation studies, where the underlying distribution for the estimator is unknown, multiple estimates may be obtained, their average and standard deviation computed, and an approximate (and conservative) 90% interval estimate for the value of that parameter can be obtained using:

*Average of Estimates* ± 1.65 *Std. Dev. of Estimates*

Thus, while we may use normal based approximations in those cases where we do not know the appropriate distribution for the estimator, this is no excuse for not using the traditional analysis when such an analysis is available.

## SELECTING A PROBABILITY MODEL

Before you even think about fitting a probability model to a data set you should place those data on a process behavior chart (either an XmR chart, or with rational subgrouping, an average and range chart). Only when this chart shows no evidence of a lack of homogeneity will it make sense to fit a probability model to those data. And the probability model that will impose the fewest constraints upon the situation will be a normal distribution with mean and variance estimated from the process behavior chart. This will be the distribution of maximum uncertainty for the underlying process that generated the data.

If you do the above and still detect a lack of fit between your data and the fitted normal distribution, then, according to Shewhart, in all likelihood your process still contains some assignable causes of exceptional variation. Why would this be the case? Simply because any system that is the result of a large number of common causes of routine variation, where no one cause dominates the others, will almost surely have a histogram that approaches the normal distribution. This is the essence of the central limit theorem of Laplace.

Therefore, to fit any other probability model to a set of homogeneous data you will have to have some sort of additional information that will allow you to determine both that the middle 90 percent of the data is more concentrated than the normal distribution and also that the outer tails are lighter than the normal distribution. Without such additional information it will be a mistake to fit a non-normal distribution to the data.

So where can you find this additional information? As I argued last month, unless your data set is homogeneous and also contains thousands upon thousands of data, you will not be able to obtain this additional information from the data. This leaves context. If the context is one where a certain type of probability model makes sense, such as a binomial, a Poisson, an exponential, or a Weibull, then the homogeneous data may be used to fit this type of distribution. The homogeneity of the data will support the estimation required to make the assumed model work.

If your process behavior chart reveals your data to be nonhomogeneous, any attempt to fit a probability model will be premature. If you proceed to fit a probability model to a nonhomogeneous data set you will end up with a model that will not only have no predictive value but will also mislead and misrepresent the underlying process. Do not do this.

## HOW DOES A PROCESS BEHAVIOR CHART WORK WITHOUT A DISTRIBUTION?

In answer to this Shewhart wrote the following on page 35 of his 1939 book "***Statistical Method from the Viewpoint of Quality Control***."

> "We next come to the requirement that the [chart] shall be as simple as possible and adaptable to a continuing and self-correcting operation. Experience shows that the process of detecting and eliminating assignable causes of variability so as to attain a state of statistical control is a long one. From time to time the chart limits must be revised as assignable causes are found and eliminated.
>
> "A simple procedure is used for establishing the limits without the use of

> probability tables because it does not seem that much is to be gained during the process of weeding out assignable causes by trying to set up exact probability limits upon the basis of assumptions that we know from experience do not hold until the state of statistical control has been reached. This is particularly true since such probabilities do not indicate the probability of detecting assignable causes but simply the probability of looking for such causes when they do not exist, which is of secondary importance until a state of statistical control has been reached. Then too, as already indicated, the design of an efficient [chart] for the important job of indicating the presence of assignable causes depends more upon [rational sampling and rational subgrouping] than it does upon the use of any exact mathematical distribution."

Here Shewhart clearly contradicts the claim that we have to fit a probability model to the data prior to computing the limits for a process behavior chart. If you and Shewhart see things differently, who do you think is right? Continuing in this vein, Shewhart concluded the epilogue of his book with the following:

> "Throughout this monograph care has been taken to keep in the foreground the distinction between the distribution theory of formal mathematical statistics and the use of such theory in statistical techniques designed to serve some practical end. Distribution theory rests upon a framework of mathematics, whereas the validity of statistical techniques can only be determined empirically. … The technique involved in the operation of statistical control [i.e. process behavior charts] has been thoroughly tested and not found wanting, whereas the formal mathematical theory of distribution[s] constitutes a generating plant for new techniques to be tried."

Mathematical theory can only approximate what happens in practice. In order for any technique for data analysis to be useful it will, of necessity, have to be robust to the assumptions of mathematical theory, otherwise it would not work in practice. When we turn the assumptions of mathematical theory into requirements to be satisfied before we use some data analysis technique we do nothing but add unnecessary complexity to our analysis.

Another aspect of the absurdity of turning mathematical assumptions into preconditions for practice lies in the fact that the techniques for checking on the preconditions will generally be much less robust than the analysis techniques being qualified. As the late Francis Anscombe, Fellow of the American Statistical Association, said: "Checking your data for normality prior to placing them on a control chart is like setting to sea in a rowboat to see if the Queen Mary can sail."

## BUT WHAT IF MY DATA ARE LOGNORMALLY DISTRIBUTED?

Even though you may think your data might be modeled by a lognormal distribution, they probably are not—as many as 9 out of 10 uses of the lognormal distribution, if not more, are completely inappropriate. Why is this? First, we note that the most common usage of a lognormal distribution is to fit a probability model to a skewed histogram. Next we observe that the most common cause of a skewed histogram is the collection of data while the process is changing. Finally, when we place our data on a process behavior chart we find at least 9 out of 10 processes turn out to be unpredictable, and are therefore changing. Thus, we arrive at the conclusion that most uses of the lognormal distribution are inappropriate.

## SO WHEN WOULD A LOGNORMAL DISTRIBUTION BE APPROPRIATE?

Before any probability model makes sense you must have a predictable process producing a stream of homogeneous data. When this happens we can think of the routine variation in the data stream to be due to the effects of a large group of common causes where no one cause has a dominant effect. For over 200 years we have found it to be satisfactory to assume that the effects of these common causes are, on the whole, additive. This was the assumption Shewhart used 80 years ago and it has stood the test of time. When this is the case, the classic model for the effects of the constant system of common causes is the normal distribution.

However, in those rare situations where we have reason to believe that the effects of ALL of the common causes combine in a multiplicative manner rather than an additive manner, then the lognormal model is appropriate. This assumption that dozens of common causes interact in a multiplicative manner is a very strong assumption that can only be justified by the context for the data being gathered. It cannot be supported by any empirical argument arising from the shape of the histogram. So, unless you have a homogeneous data set that satisfies the assumption of multiplicative effects for your common causes, you should avoid using a lognormal distribution.

## TRANSFORMING THE DATA

"We only look at the fit to see if we need to transform the data."

Any technique for transforming the data is built upon the implicit assumption that the data are homogeneous. Any lack of homogeneity will completely undermine the transformation. Moreover, the idea of transforming the data "to make it look more normal" is built upon the erroneous idea discussed above that the data "have to be normally distributed." This idea is profoundly wrong simply because robust techniques do not require specific distributional assumptions for the original data. Thus, the idea of transforming the data prior to analysis is not only based on an incorrect presupposition but it also assumes a degree of homogeneity for the data that is rarely found in practice. Once you have transformed your data in a nonlinear manner you are unlikely to discover anything interesting within those data regardless of the sophistication of your subsequent analysis.

## SUMMARY

The normal distribution is the heavy-tailed distribution. Its middle 90 percent is more spread out than the middle 90 percent of virtually all other unimodal distributions, and its outer 10 percent is further from the mean than the outer 10 percent of virtually all other unimodal distributions. If all you know about a distribution is the mean and the variance, then the probability model that will impose the fewest constraints upon the situation will be a normal distribution having that mean and variance.

The popular practices of checking your data for lack of fit, fitting nonnormal distributions to your data, and transforming the data to make them "more normal" are nothing more than triumphs of computation over common sense. These inappropriate exercises in complexity are made possible by the software, encouraged by instructors without degrees in statistics, and practiced by those who do not know any better. Now you know better.