Problems with Skewness and Kurtosis, Part One

What do the shape parameters do?

Donald J. Wheeler

With the use of statistical software many individuals are being exposed to more than just measures of location and dispersion. In addition to the average and standard deviation, they often find some funny numbers labeled as skewness and kurtosis. Since these numbers appear automatically it is natural to wonder how they might be used in practice. The purpose of part one of this paper is to illustrate what the skewness and kurtosis *parameters* do. In part two I will look at the use of skewness and kurtosis *statistics* provided by software packages.

Since the previous sentence makes a distinction between a statistic and a parameter we should begin there. Statistics are merely functions of the data. We find the value for a statistic by performing a set of arithmetic operations using a set of data. For example, we compute the average for a set of numbers by adding up all the numbers and dividing by the number of values in the sum. Thus, any time we have a collection of numbers we can compute any one of a number of statistics. Data plus arithmetic equals a statistic.

On the other hand, a parameter is a descriptive constant for a probability model. Parameters are used to characterize specific properties of a probability model. This means that, rather than using data, parameters are obtained by performing certain mathematical operations using the probability model. Since probability models must meet certain requirements, parameters are not well defined until the probability model is well defined.

The first four parameters for a probability model are the mean, the variance, the skewness, and the kurtosis. Given a continuoous probability model characterized by the probability density function $f(x)$, the mean of the probability model will characterize the location and is defined as:

$$MEAN(X) = \mu = \int_{all\ x} x\ f(x)\ dx$$

The variance of the probability model will characterize the dispersion and is defined as:

$$VARIANCE(X) = \sigma^2 = \int_{all\ x} (x-\mu)^2\ f(x)\ dx$$

The square root of the variance is commonly known as the standard deviation. It provides an alternative way to characterize the dispersion of the probability model:

$$SD(X) = \sigma = \sqrt{\sigma^2}$$

The skewness and kurtosis are collectively known as the shape parameters for the probability model. The skewness parameter for the probability model is defined to be the third standardized central moment. This means that we begin with the standardized form for the random variable: $[(x-\mu)/\sigma]$, raise it to the third power, multiply by the probability model, and integrate over all x.

$$SKEWNESS(X) = \alpha_3 = \int_{all\ x} \frac{(x-\mu)^3}{\sigma^3}\ f(x)\ dx$$

In a similar manner, the kurtosis parameter for the probability model is defined as the fourth standardized central moment:

$$KURTOSIS(X) = \alpha_4 = \int_{all\ x} \frac{(x-\mu)^4}{\sigma^4}\ f(x)\ dx$$

At this point it should be abundantly clear why you never computed the skewness and kurtosis parameters in your stat class. Moreover, since you do not routinely evaluate integrals it is fairly safe to say that you have probably not computed any parameters since you finished (or dropped out of) your stat class. However, because these parameters characterize various aspects of a probability model they are useful in organizing the zoo of probability models.

To illustrate how the skewness and kurtosis parameters characterize the shape of a probability model we shall use a simple probability model for which the integrals above will be easy to illustrate and evaluate. This probability model is the standardized right triangular distribution. It has a probability density function *f(x)* of:

$$f(x) = \frac{\sqrt{8}-x}{9} \qquad \text{whenever } -\sqrt{2} \le x \le 2\sqrt{2}$$

$$\text{and} \qquad f(x) = 0 \qquad \text{otherwise.}$$

This probability model has a mean of zero, a standard deviation of 1.000, and is shown in Figure 1.

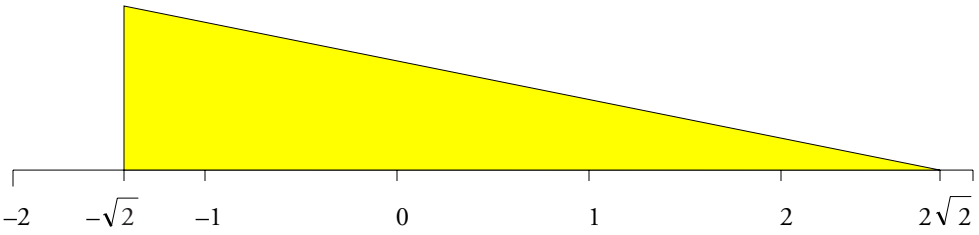

Figure 1: The Standardized Right Triangular Distribution

Since this is a standardized distribution, the standardized form for the random variable reduces down to simply [*x*]. Thus, the formulas for the skewness and kurtosis parameters reduce to the following:

$$SKEWNESS(X) = \alpha_3 = \int_{all\ x} x^3\ f(x)\ dx = \int_{-\sqrt{2}}^{2\sqrt{2}} x^3 \left(\frac{\sqrt{8}-x}{9}\right) dx$$

$$KURTOSIS(X) = \alpha_4 = \int_{all\ x} x^4\ f(x)\ dx = \int_{-\sqrt{2}}^{2\sqrt{2}} x^4 \left(\frac{\sqrt{8}-x}{9}\right) dx$$

Thus, we see that in this case the skewness is the integral of the product of the cubic curve and the density function, while the kurtosis is the integral of the product between the quartic

curve and the density function. Figure 2 shows the density function along with the cubic and quartic curves. Figures 3 and 4 show the resulting product curves.

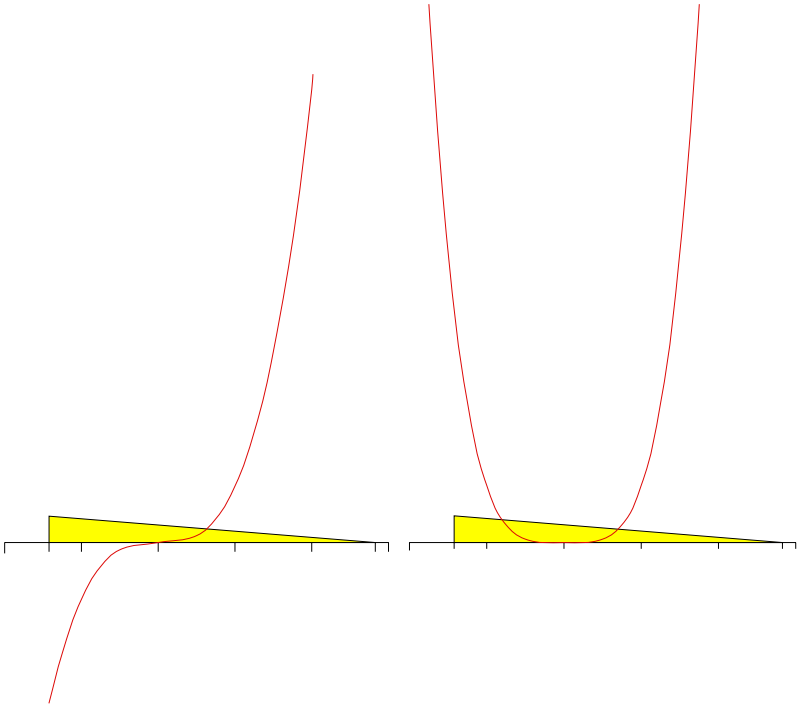

Figure 2: Cubic and Quartic Curves with *f(x)*

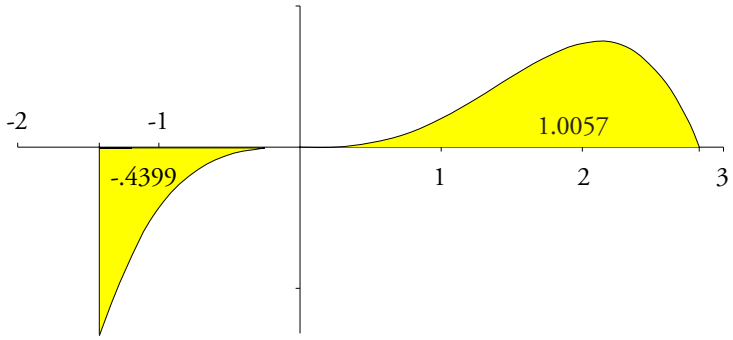

Figure 3: The Areas Which Define the Skewness Parameter

Interpreting the integral as the area between the product curve and the X axis we find that the skewness parameter for this probability model may be interpreted as:

$$SKEWNESS(X) = 1.005663 - 0.439978 = 0.565685.$$

Figure 4 shows the curve that results when we multiply the probability model by the quartic curve. The kurtosis parameter for this probability model may be interpreted as the area under the curve in Figure 4. In this case:

$$KURTOSIS(X) = 1.896296 + 0.503704 = 2.400000.$$

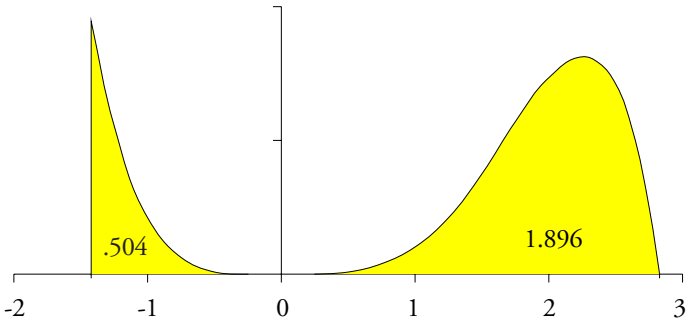

Figure 4: The Areas Which Define the Kurtosis Parameter

The fact that all four regions in Figures 3 and 4 pinch down near zero suggests that the central region of the probability model contributes very little to either of these two parameters. Since the distribution in this example is already in its standardized form, the units on the horizontal axis in Figures 3 and 4 represent the standardized distance from the mean. Thus, the contribution of the central portion of the probability model can be seen by considering how much of the total area under the curves corresponds to X values which fall between –1.0 and +1.0.

While the central portion of this probability model contributes 63 percent of the total area, only 11 percent of the combined areas in Figure 3, and only 5 percent of the area in Figure 4, correspond to the central portion of the probability model. Therefore, we must conclude that both skewness and kurtosis are primarily concerned with characteristics of the tails of the probability model.

The skewness parameter measures the relative sizes of the two tails. Distributions that have tails of equal weight will have a skewness parameter of zero. If the right-hand tail is more massive, then the skewness parameter will be positive. If the left-hand tail is more massive, the skewness parameter will be negative. Moreover, the greater the difference between the two tails, the greater the magnitude of the skewness parameter.

The kurtosis parameter is a measure of the combined weight of the tails relative to the rest of the distribution. As the tails of a distribution become heavier the kurtosis will increase. As the tails become lighter the kurtosis will decrease. As defined here kurtosis cannot be less than 1.00. Probability models with kurtosis values between 1.00 and 3.00 are considered to be light-tailed distributions (platykurtic). Probability models with kurtosis values in excess of 3.00 are considered to be heavy-tailed distributions (leptokurtic).

Kurtosis was originally thought to measure the "peakedness" of a distribution. However, since the central portion of the distribution is virtually ignored by this parameter, kurtosis cannot be said to measure peakedness directly. While there is a correlation between peakedness and kurtosis, the relationship is an indirect and imperfect one at best.

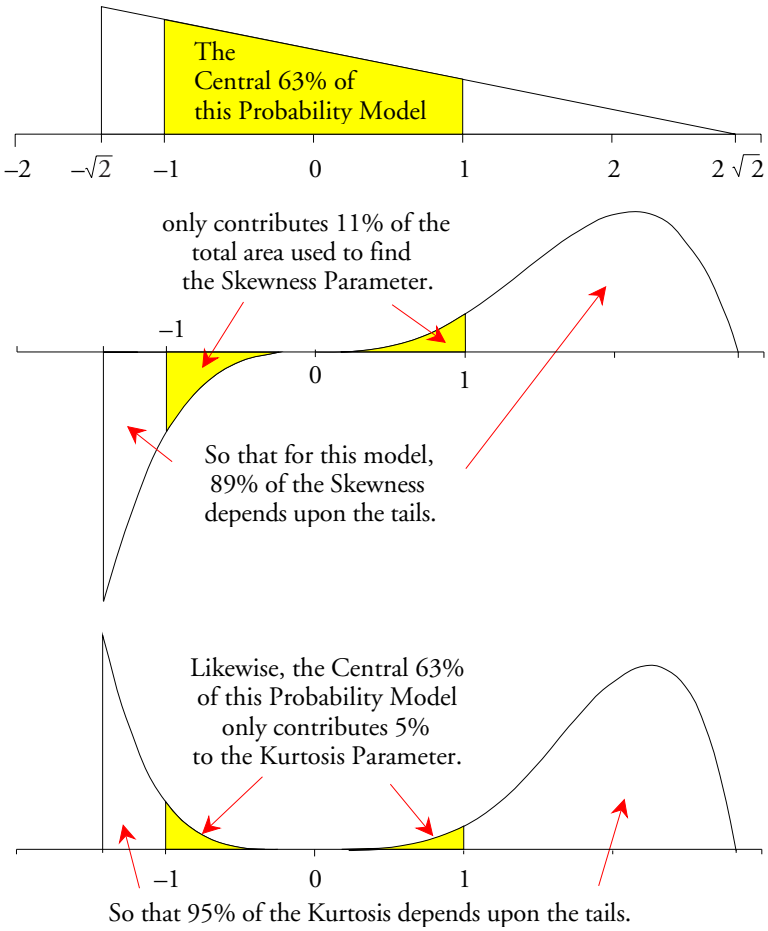

Figure 5: Skewness and Kurtosis Characterize the Tails of a Probability Model

Thus, the shape parameters of skewness and kurtosis actually tell us more about the tails of a probability model than they do about the central portion of that model. At the beginning of the Twentieth Century the shape parameters were used simply because Karl Pearson had developed seven families of probability models that were fully characterized by the first four moments. Of these families the two most important are the Beta Distributions (Pearson Type One) and the Gamma Distributions (Pearson Type Three).

By plotting the values of the shape parameters on Cartesian coordinates Pearson was able to show how these families of probability models were related to each other. This plot is known as the shape characterization plane. In this plane a probability model is represented by a single point, while families of probability models will sometimes fall on a line or fall within in a region of the plane. For example, all normal distributions will have a skewness of zero and a kurtosis of 3.00. In the shape characterization plane the skewness squared defines the X-coordinate, while the kurtosis defines the Y-coordinate. Thus, the family of all normal distributions will be shown on the shape characterization plane by a single point at $(0,3)$. The family of all exponential distributions (skewness = 2, kurtosis = 9) will be shown by a single point at $(4,9)$.

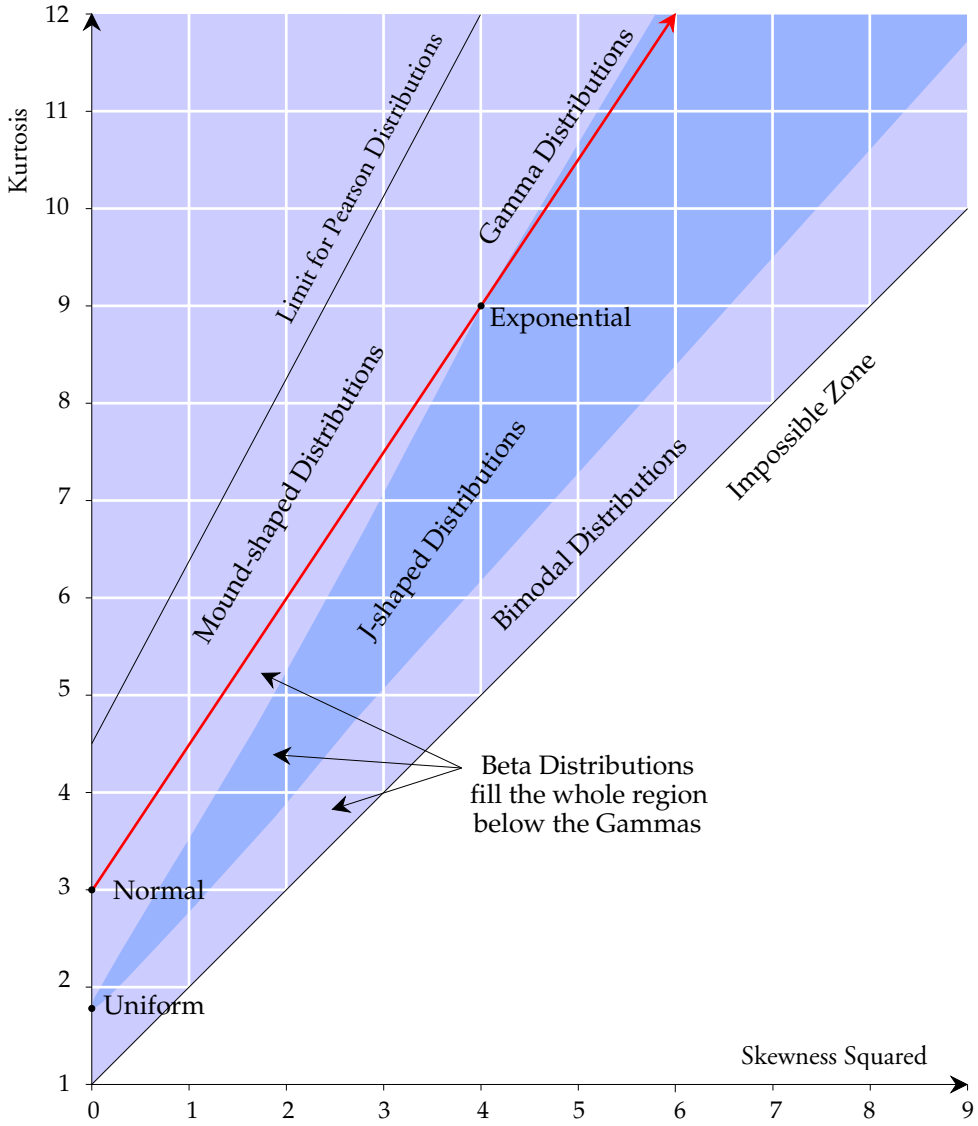

Figure 6: The Beta and Gamma Families of Distributions on the Shape Characterization Plane

Figure 6 shows the heart of the shape characterization plane. The gamma distributions are represented by the line defined by the normal and exponential distributions. All of the chi-square distributions fall on this line. The beta distributions occupy the whole region of the plane below the gamma distribution line. The shape characterization plane can be divided as shown into regions according to whether the probability models are mound-shaped, J-shaped, or bimodal. At the apex of the dividing lines between these three divisions we find the family of uniform distributions, which are neither mound-shaped, J-shaped, nor bimodal.

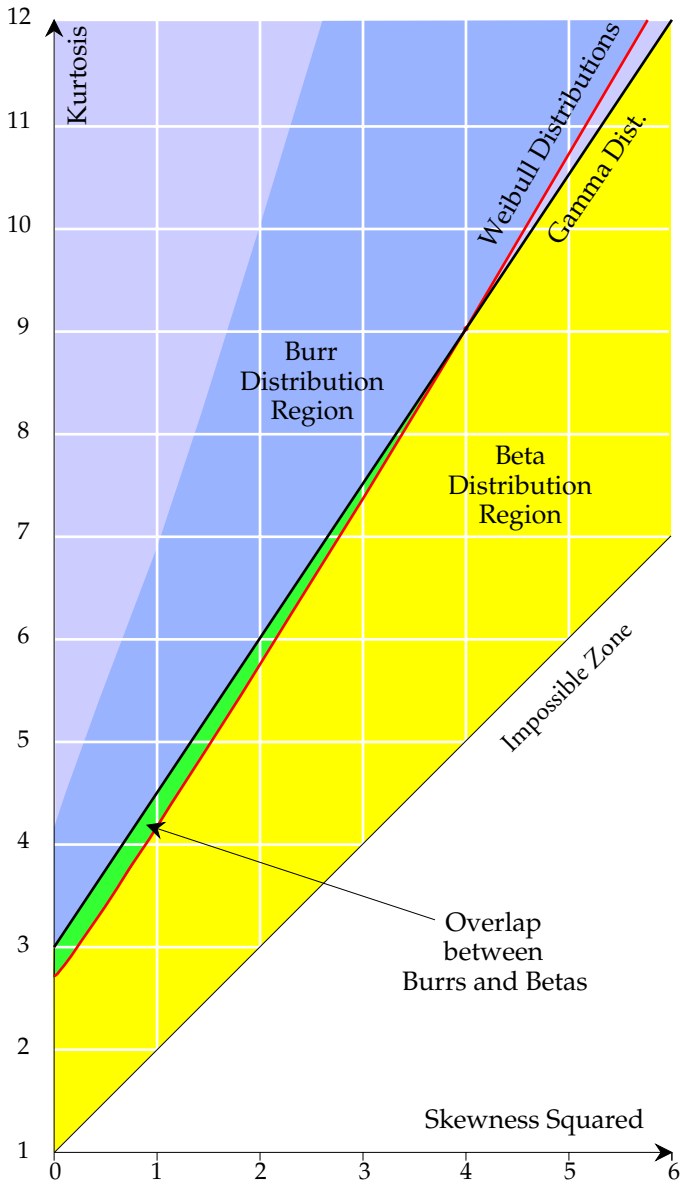

Figure 7: The Burr, Weibull and Beta Families of Distributions on the Shape Characterization Plane

Figure 7 shows the family of positively skewed Weibull distributions as a red line. Above this line we find the family of Burr distributions effectively covering the rest of the region of mound shaped probability models. Thus, skewness and kurtosis parameters are useful because of their ability to characterize and organize the zoo of probability models. Moreover, as seen in Figures 6 and 7, the families of the betas and Burrs, plus their limiting families of the gammas and the Weibulls will effectively cover the whole shape characterization plane. Does this mean that these are the only probability models? By no means. But it does mean that a first order approximation to virtually any probability model can be found among these four families of distributions.

The reason that these distributions will only provide a first order approximation is due to the

fact that the skewness and kurtosis only characterize the tails of the distribution. This is why it is fallacious to think that two distributions having the same mean, standard deviation, skewness and kurtosis will have exactly the same shape. A second, related fallacy is that a distribution with a skewness parameter of zero will be symmetric. That these are indeed fallacies will be illustrated by the following examples.

Figure 8 shows a simple symmetric probability model characterized by the density function, $f(x)$ where:

$$\begin{aligned} f(x) &= 0.6391 + 1.0337\,x && \text{when } -0.0091 < x < 0.5387 \\ &= 1.7527 - 1.0337\,x && \text{when } 0.5387 < x < 1.0864 \\ &= 0 && \text{otherwise} \end{aligned}$$

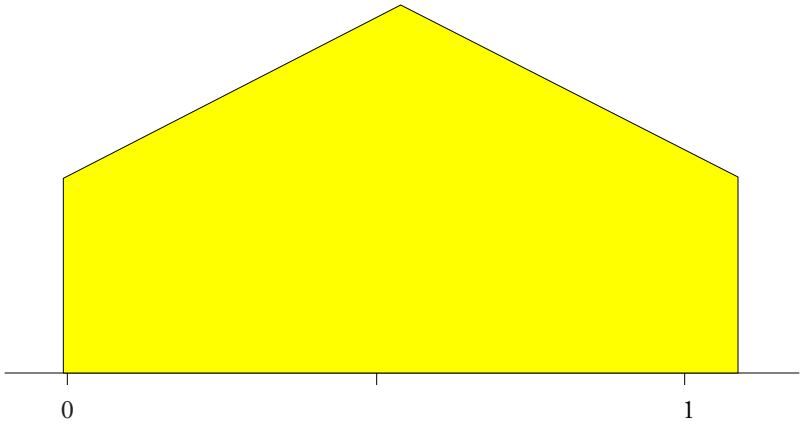

Figure 8: A Simple Symmetric Probability Density Function

The probability model in Figure 8 has a mean of 0.5387, a standard deviation of 0.2907, a skewness of 0.000, and a kurtosis of 2.000. The symmetry of this distribution requires a skewness of zero, and the short tails result in a small value for kurtosis.

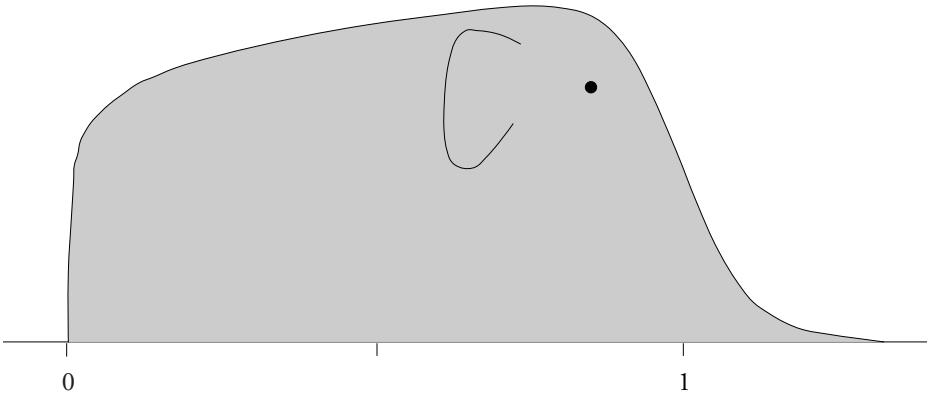

Figure 9: The Elephant Distribution

Figure 9 shows a nonsymmetric probability model from the family of Inverse Burr distributions which are characterized by the density function $g(x)$:

$$g(x) = ck\ x^{-(c+1)}\left[\,1 + x^{-c}\,\right]^{-(k+1)} \qquad \text{when } x > 0.0$$
$$= 0 \qquad \text{otherwise}$$

When we let the value of c be 18.1484, and let the value of k be 0.0629 we get the probability model shown in Figure 9. This probability model has a mean of 0.5387, a standard deviation of 0.2907, a skewness of 0.0000, and a kurtosis of 2.0000. With a couple of extra lines this distribution can be made into a reasonable cartoon of an elephant. While this probability model is definitely not symmetric, it does have a skewness of zero. Moreover, like the probability model in Figure 8, it also has a kurtosis of 2.00.

Figure 10 compares the distributions of Figures 8 and 9. There we find two distributions that have the same mean, the same standard deviation, the same skewness and the same kurtosis, yet they do not look alike. Thus, we may properly conclude that the "shape parameters" of skewness and kurtosis cannot even discriminate between an elephant and the gable end of a house!

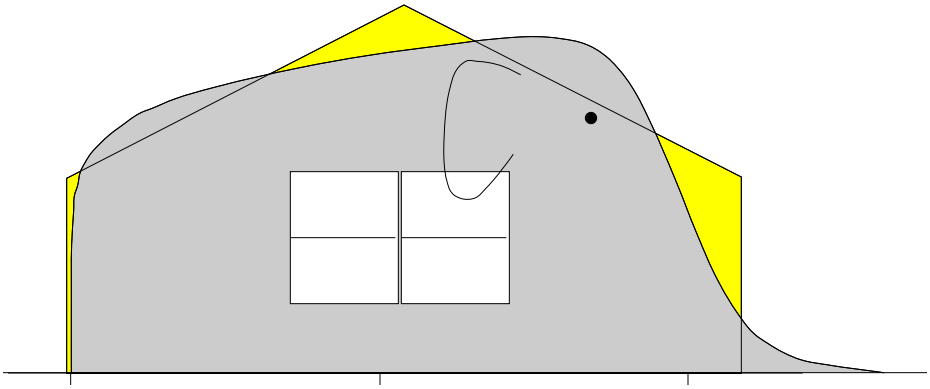

Figure 10: Two Probability Models with the Same Mean, Variance, and Shape Parameters

While probability models having the same shape parameters will display a gross similarity, they do not have to be exactly alike. While most of these differences might be expected to occur in the central portion of the distributions, as we can see in Figure 10, some of these differences can also occur in the tails.

This column has focused on the shape parameters for probability models. In part two we will consider the utility of shape statistics. However, before getting lost in this world of probability models, it is important to note that we cannot begin to use a probability model to approximate reality until we have a predictable process. Any attempt to choose, fit, or otherwise use a probability model to characterize an unpredictable process is a mistake.

Problems with Skewness and Kurtosis, Part Two

What do the shape statistics do?

Donald J. Wheeler

In part one we found that the skewness and kurtosis parameters characterize the tails of a probability model rather than the central portion, and that because of this, probability models with the same shape parameters will only be similar in overall shape, not identical. However, since software packages can only provide shape *statistics* rather than shape *parameters*, we need to look at the usefulness of the shape statistics.

In part one, we saw that the skewness parameter is the third standardized central moment for the probability model. For this reason, a commonly used statistic for skewness is the third standardized central moment of the data:

$$a_3 = \frac{\sum (X_i - \bar{X})^3}{n \; s_n^3}$$

In a similar manner, we shall use the fourth standardized central moment of the data as our statistic for kurtosis:

$$a_4 = \frac{\sum (X_i - \bar{X})^4}{n \; s_n^4}$$

Since both of the formulas above use the root mean squared deviation, s_n, rather than the more common standard deviation statistic, s, there may be slight differences between the statistics listed above and the statistics given by your software. For example, Microsoft Excel uses the formulas:

$$\text{Skewness(Excel)} = \frac{\sqrt{n(n-1)}}{n-2} \; a_3$$

$$\text{Kurtosis(Excel)} = \frac{(n-1)(n+1)}{(n-2)(n-3)} \left[a_4 - \frac{3(n-1)}{(n+1)} \right]$$

Regardless of the different formulas, a_3 contains the essence of all statistics for skewness while a_4 contains the essence of all statistics for kurtosis. For this reason we shall use the simple a_3 and a_4 statistics here.

To illustrate these shape statistics I shall use the 200 observations shown in Table 1. These represent the values logged during a 20 day period of production. The descriptive statistics for these 200 values are: The average is 256.46; the standard deviation statistic is 4.58; the skewness (a_3) is 1.60; and the kurtosis (a_4) is 6.31. (The Excel formulas result in values of 1.61 and 3.42 respectively.) The histogram for these values is shown in Figure 11. This histogram looks like many histograms that come from process related data. The bulk of the values fall in a central mound while some of the values trail off to one side in an elongated tail.

Table 1: Two Hundred Values from Process Log

Day	Values									
1	255	252	250	251	252	253	251	253	251	252
2	257	254	253	257	255	254	254	253	256	252
3	256	251	252	252	254	253	255	253	254	255
4	252	254	257	252	251	253	255	256	254	253
5	251	256	255	253	252	252	255	255	256	257
6	254	254	252	250	253	252	248	249	253	251
7	256	255	255	255	256	258	258	257	255	257
8	255	252	256	257	254	252	252	252	258	254
9	255	258	257	256	254	255	256	252	257	256
10	256	254	256	257	252	257	254	259	252	257
11	261	256	256	258	254	254	257	255	258	258
12	256	256	255	252	256	257	253	254	255	254
13	258	260	257	256	262	260	259	257	258	258
14	254	255	253	256	252	255	253	253	256	256
15	257	259	258	257	259	257	263	259	261	258
16	254	257	257	253	253	252	254	256	255	258
17	258	263	264	264	263	263	262	264	261	265
18	255	260	255	258	256	257	253	255	258	259
19	263	261	258	258	257	263	262	263	258	261
20	270	273	273	268	267	276	271	268	268	271

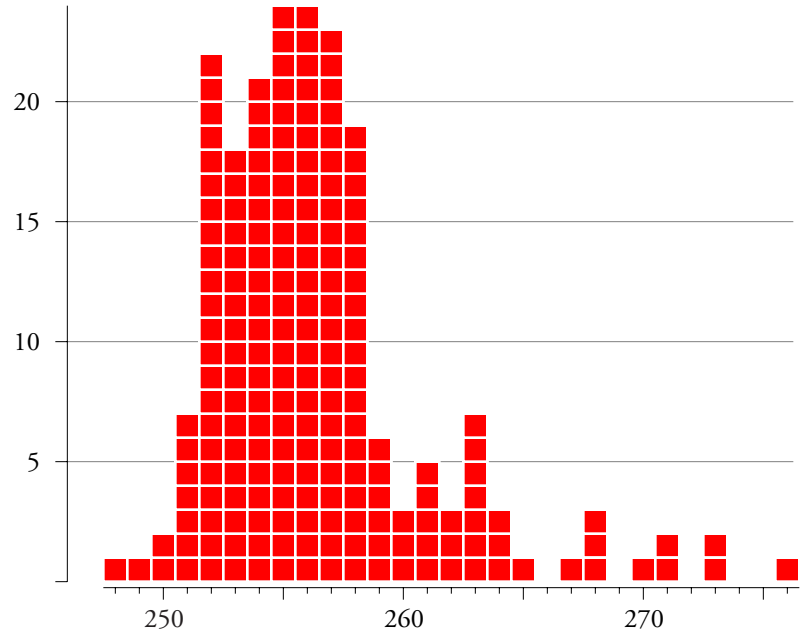

Figure 11: Histogram for the 200 Data of Table 1

While the descriptive statistics do correctly characterize the data shown in the histogram in Figure 11, this is not the same as using these descriptive statistics to estimate the parameters of a probability model. Whenever we use statistics to estimate parameters we will need to take into account the inherent uncertainty attached to our estimates.

When we use the average to estimate the mean of a probability model the uncertainty is a

function of the inverse of the square root of n. That is, the standard deviation of the average statistic is given by:

$$\text{STD. DEV. of Average Statistic} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation parameter of the probability model and n is the number of values used to compute the average.

Next consider using the standard deviation statistic to estimate the standard deviation parameter of a probability model. Here the uncertainty will be a function of the inverse of the square root of twice the number of data. Specifically, the standard deviation of the standard deviation statistic will be:

$$\text{STD. DEV. of a Standard Deviation Statistic} = \frac{\sigma}{\sqrt{2n}}$$

This value will be about 71% as large as the uncertainty in the estimate of the mean parameter. Thus, with any given data set, we will always estimate the location and dispersion parameters with about the same amount of uncertainty.

When we use a skewness statistic to estimate the skewness of a probability model the uncertainty will be 2.45 times the uncertainty in the estimate of the location parameter since:

$$\text{STD. DEV. of a Skewness Statistic} = \sqrt{\frac{6\,\sigma^2}{n}}$$

When we use a kurtosis statistic to estimate the kurtosis of a probability model the uncertainty will be about 4.9 times the uncertainty in the estimate of the location parameter since:

$$\text{STD. DEV. of a Kurtosis Statistic} = \sqrt{\frac{24\,\sigma^2}{n}}$$

These four results mean that we will always estimate the location and dispersion with greater precision than we will ever estimate the shape parameters. For example, if we use 20 data to estimate the mean, and if we then wanted to also estimate the skewness with a similar precision, then we would need to collect and use 120 data to estimate the skewness. Likewise, it would take 480 data to estimate the kurtosis with the same precision that we can achieve when using 20 data to estimate the mean. ***This means that regardless of how many data we have, we will always have much more uncertainty in the shape statistics than we will have in the location and dispersion statistics.*** This limitation on what we can obtain from a collection of data is inherent in the statistics themselves, and must be respected in our analysis of the data.

Using the formulas above we will compute approximate 95% interval estimates for the mean, standard deviation, skewness, and kurtosis of the process that produced the data in Figure 11. These approximate 95% interval estimates will have the form:

$$\text{Point Estimate} \pm 2\,[\,\text{Std. Dev. of Point Estimate}\,]$$

So, with our 200 data, we would estimate the process mean to be about 255.8 to 257.1. We would estimate the process standard deviation to be about 4.6 plus or minus 0.5. Our skewness could be anywhere from 0 to 3.2, and our kurtosis could be anywhere between 3.1 and 9.5!

Table 2: Approximate 95% Interval Estimates

Parameter	Point Estimate	Std. Dev.	Interval Estimate
Mean	256.46	0.32	255.8 to 257.1
Std. Dev.	4.58	0.23	4.12 to 5.04
Skewness	1.60	0.79	0.02 to 3.18
Kurtosis	6.31	1.59	3.14 to 9.48

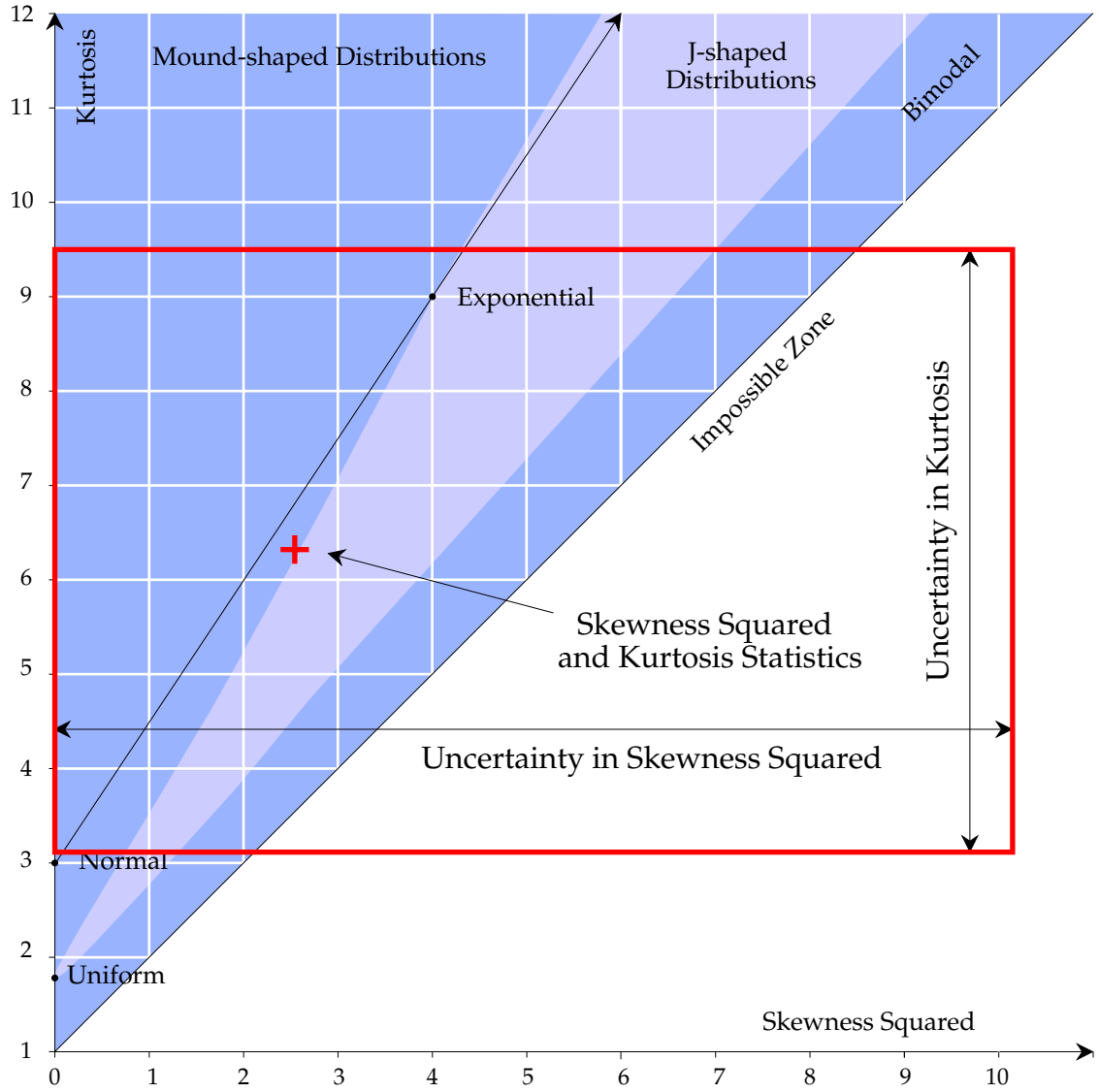

Figure 12: The Region Likely to Contain a Model for the Process of Figure 11

Since the shape characterization plane uses the square of the skewness on the horizontal axis, the 95% interval estimates above refer to points within the red rectangle shown in Figure 12. This rectangle essentially covers the heart of the shape characterization plane.

Based on the uncertainty in our statistics for skewness and kurtosis we can rule out the platykurtic probability models and possibly the normal distribution as well, but very little else.

Thus, with 200 data, our estimates of skewness and kurtosis are simply not sufficient to identify a particular probability model, or even a reasonably small group of probability models, to use in characterizing this production process.

Thus, the first problem with the shape statistics of skewness and kurtosis is simply this: Until thousands of data are involved in the computation, the shape statistics will have so much uncertainty that they will not provide any useful information about which probability models might be reasonable candidates for a process. Any attempt to use the shape parameters to identify which probability model to use will always require more data than you can afford to collect. But even if you could afford enough data, there are two more problems that create a barrier to using the shape statistics.

The second problem is that the shape statistics depend upon the extreme values of the histogram. As we saw in Part One, the shape parameters characterize the tails of a probability model. In a similar manner, the shape statistics will be more dependent upon the extreme values in the data than they will on the data set as a whole.

To illustrate this, Table 3 shows the cumulative values for the shape statistics for Table 1. Here we see how the values for the skewness and kurtosis statistics change as additional data are used in the computation. The first row shows the shape statistics based on Days 1 through 5 (50 data). The second row shows these values based on Days 1 through 10 (100 data). Row three uses Days 1 through 15. Row four uses Days 1 to 17. Row five uses Days 1 to 19. Row six uses all 20 days. There we see that these statistics do not converge to fixed values as the amount of data increases.

Table 3: Cumulative Shape Statistics for Data of Table 1

Days	n	Skewness	Skewness Squared	Kurtosis
1-5	50	0.20	0.04	2.05
1-10	100	–0.12	0.01	2.41
1-15	150	0.22	0.05	3.04
1-17	170	0.71	0.50	3.65
1-19	190	0.60	0.36	3.21
1-20	200	1.60	2.54	6.31

Figure 13 shows the points of Table 3 plotted in the shape characterization plane. We would expect this graph to show a series of points that get closer together, with the distance between successive points getting smaller as the amount of data increases. This is what will happen when a statistic converges to the value of some process parameter. However, in Figure 13 we see the sensitivity to extreme values that is inherent in all shape statistics. Here the jump from the point for $n = 190$ to the point for $n = 200$ is larger than all of the preceding line segments put together. Ten data points out of 200 amount to only 5% of the data, yet they move the point in Figure 13 from (0.36 , 3.21) to (2.54 , 6.31). This is a strong indication that the process represented by the data of Table 1 is changing. And when the process is changing, the notion of process parameters is not well-defined. This leads to the third problem with the use of the shape statistics.

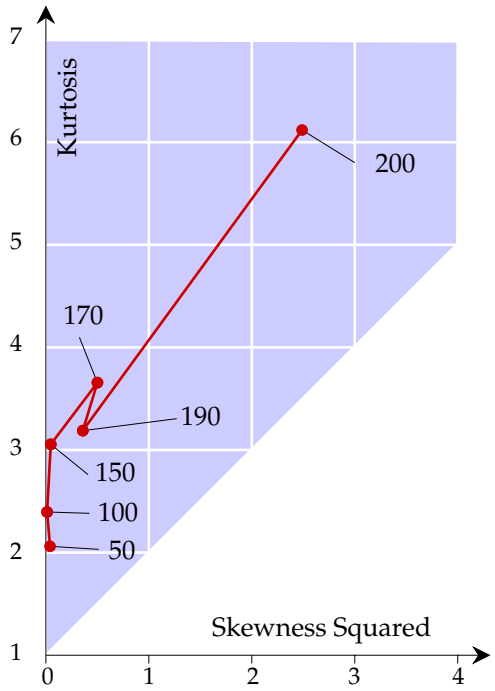

Figure 13: The Random Walk of the Shape Statistics For Table 1

This third problem has to do with the implicit assumptions behind the descriptive statistics. As the name implies, these statistics describe various aspects of the data set for which they were computed. That is, they will characterize the data. Whether these statistics can be used to estimate a process parameter is a much more complex question. Before we can use a statistic to estimate a parameter we will have to have a set of data for which it makes sense to talk about a probability model. And the primary requirement for a probability model to make sense is that the data are homogeneous. *This means that whenever we use a descriptive statistic to estimate a process characteristic we are making a very strong assumption that the data are homogeneous*.

Thus, while we may always compute our descriptive statistics, we cannot begin to use those statistics to estimate process parameters until we know that the data set is homogeneous. And the only completely general technique that can examine suspect data for evidence of a lack of homogeneity is the process behavior chart.

Figure 14 shows the Average and Range Chart for the data of Table 1. There I used the data from each day as a subgroup, resulting in 20 subgroups of size 10. The limits were computed using all of the data. The grand average is 256.46 and the average range is 5.90. Thirteen of the 20 subgroup averages fall outside their limits, which means that this process was operated differently at different times. The first six days seem to show a process that was operated at one level. The next six days show a process that was operated at a slightly higher level. The last eight days show a process that has gone on walkabout. Thus, the data of Table 1 are definitely *not* homogeneous. The histogram in Figure 11 does *not* represent any one process, but instead it represents an unknown mixture of several different processes.

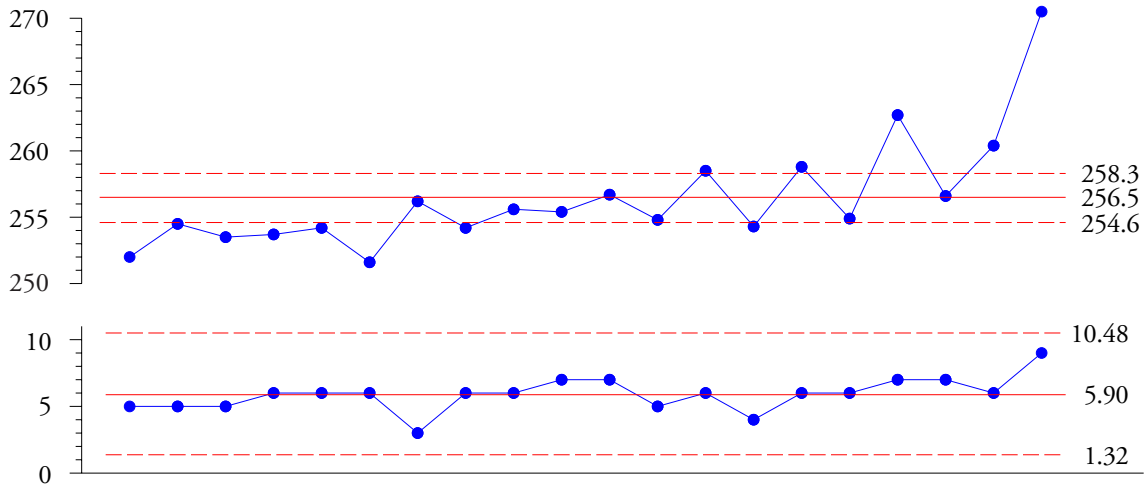

Figure 14: Average and Range Chart for the Data of Table 1

When the data are not homogeneous our descriptive statistics become misleading. While the statistics describe the data, the data are a meaningless blend of values obtained under different conditions. In Figure 14 we used the grand average of 256.5 as the central line for the average chart, yet this chart shows that the process average was detectably different from 256.5 on 13 of the 20 days.

The descriptive standard deviation statistic of 4.58 units is over twice the size of the more robust estimate of 1.92 units obtained from the average range.

The skewness statistic of 1.60 is simultaneously inflated by the random walk of the process and simultaneously deflated by the excessively large value for the standard deviation statistic. In Figure 12 we saw that there is so much uncertainty in this statistic that it does not begin to narrow the possibilities. This statistic simply contains no useful information.

The kurtosis statistic of 6.31 was heavily inflated by the last subgroup. It was also simultaneously deflated by the excessively large value for the standard deviation statistic. In Figure 12 we saw that the uncertainty in this statistic was so great that we could only slightly narrow down the possibilities. The unpredictability of the process makes this statistic unusable.

The company operating the process from which the data of Table 1 was obtained began to use process behavior charts. As they discovered the assignable causes of the unpredictable operation seen in Figure 14 and controlled these assignable causes they were able to operate this process up to its full potential. The histogram for values collected during this period of predictable operation is shown in Figure 15, where it is superimposed on the histogram from Figure 11. While these two histograms represent different amounts of data, they have been adjusted to have equal areas to facilitate the visual comparison.

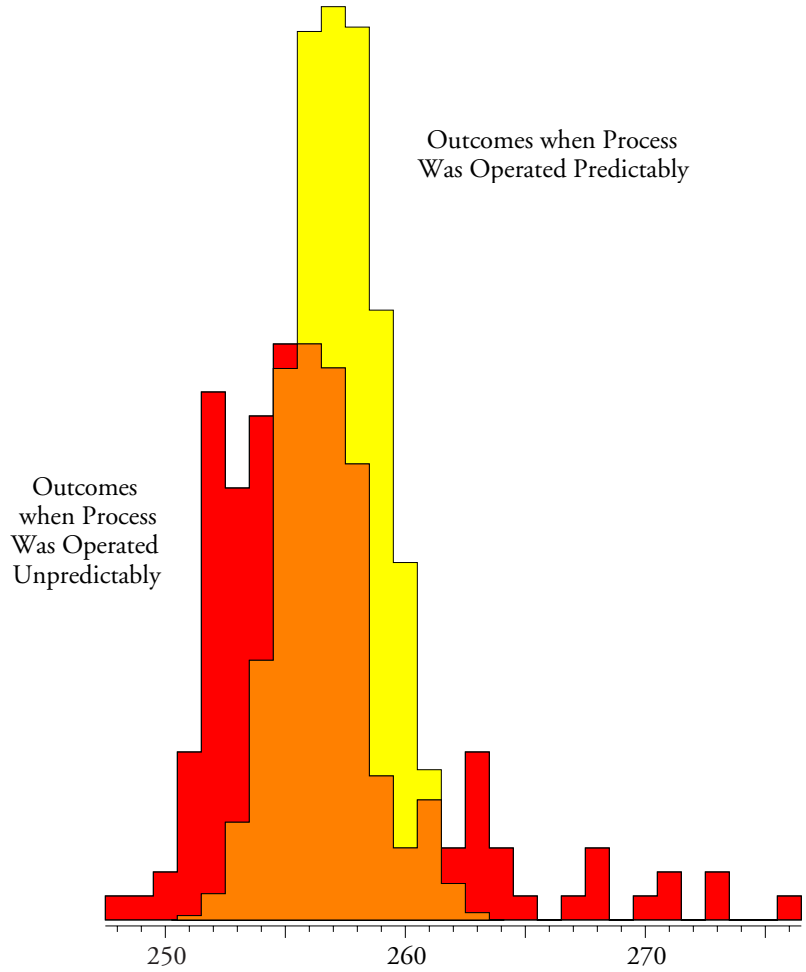

Figure 15: Histograms of a Process when Operated Predictably and Unpredictably

Table 4 compares the values of the descriptive statistics for the two histograms in Figure 15. While the two histograms have about the same average, the rest of the descriptive statistics are wildly different.

Table 4: How a Lack of Homogeneity Undermines Descriptive Statistics

Statistic	Original Values from Unpredictable Process	Values from Predictable Process	95% Interval
n	200	3287	
Average	256.46	257.14	±0.07
Standard Deviation	4.58	1.93	±0.05
a_3 skewness	1.60	0.02	±0.16
a_4 kurtosis	6.31	2.80	±0.33

The original descriptive statistics on the left are nothing but an exercise in computation. Since they came from a non-homogeneous collection of data they provide no useful information about the underlying unpredictable process.

Since the descriptive statistics on the right came from a homogeneous set of data they can be

used to characterize the underlying predictable process. For comparison purposes Figure 16 shows the 95% interval estimate boxes for the two sets of shape parameters in Table 4. The values on the left in Table 4 have the uncertainty shown by the red box, while those on the right have the uncertainty shown by the yellow rectangle. This yellow rectangle is small due to the large number of values used (3287). Since the yellow rectangle represents what this predictable process actually does, it has to be considered to be the correct answer. Note that these two interval estimate boxes do not even overlap.

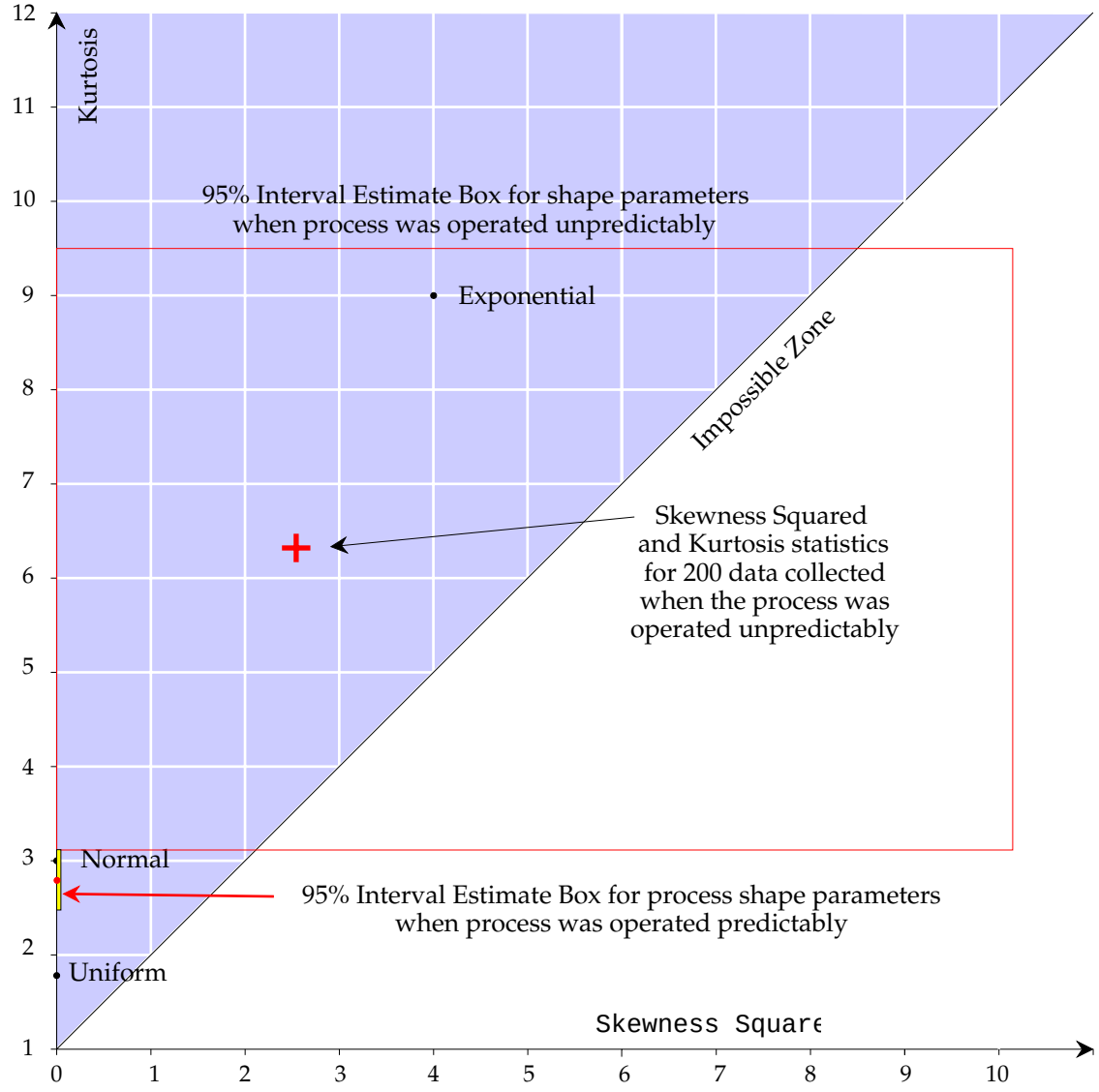

Figure 16: Unpredictable Operation Completely Undermines Descriptive Statistics

So, does this mean that we can use the shape statistics when we have a predictable process? While the shape statistics will converge to the values of the shape parameters for a predictable process, they will still do so very slowly. Over an extended time, as thousands of data are obtained from a predictable process, the skewness and kurtosis statistics will begin to reveal

something about the histogram of the process outcomes, but this information will have no practical utility. In Figure 15, the fact that we have a histogram that has slightly less kurtosis than a normal distribution might interest your local statistician, but it will be of no practical interest otherwise. By the time you could begin to use the shape statistics for a predictable process you will have a histogram consisting of thousands of data, and such a histogram can be used to answer virtually all of the questions of interest regarding the process, making knowledge of the shape parameters moot.

In the first half of the Twentieth Century considerable effort was poured into the problem of how to use the shape statistics to select a probability model to use. In spite of increasing complexity and sophistication, these efforts continually floundered on the huge uncertainties of the shape statistics. Moreover, these efforts were limited to situations where the data were known to be reasonably homogeneous. When these approaches are tried with data sets of suspect homogeneity they simply crash and burn.

SUMMARY

We have seen that the shape parameters characterize the tails of a probability model rather than the central portion. As a consequence of this, two probability models with the same mean, standard deviation, skewness, and kurtosis will have similar shapes, but they do not have to be identical in shape. Moreover, the differences between these two models can occur in both the central portion and in the extreme tails of the two probability models. Nevertheless, the shape parameters allow us to organize the various families of probability models using the shape characterization plane.

On the other hand, the shape statistics have problems. We have seen that when the shape statistics are used to estimate shape parameters they will have so much uncertainty that literally thousands of data are needed to obtain estimates that have any practical utility.

The second problem is the way the shape statistics depend upon the extreme values. With a predictable process the extreme values will stabilize, but with an unpredictable process the extreme values will continually evolve, resulting n drastic and continuing changes in the shape statistics. And this is closely related to the third problem of whether the notion of a single probability model makes sense. While descriptive shape statistics may be computed for any collection of four or more values, these shape statistics can be meaningfully used as estimates of process parameters only when the data are homogeneous. Since large amounts of data will still be required for good estimates, any serious attempt to estimate shape parameters will require a large amount of homogeneous data. Using small amounts of data will not provide estimates with enough precision to be of any practical use (see Figure 12). With large amounts of data the presumption of homogeneity will rarely be correct.

When the data come from a process that is changing it is not the computation of the statistics that is the problem, but rather the assumption that there is a single probability model to be characterized by those statistics. As was illustrated, when the process is changing, the higher order descriptive statistics for dispersion and shape will become inflated. Rather than converging to some specific value as the amount of data increases, these higher order statistics will move around in response to the changes in the underlying process, and the result will be more of a random walk than a convergence (see Figure 13).

So what do these properties of the shape statistics mean in practice? They effectively undermine many of the unnecessary complications that have been taught as elements of data analysis.

Do we need to pick a probability model for our process and then use that model to compute probability limits for our process behavior chart? No, as Shewhart observed, it is not a matter of having an exact probability for a point to fall outside the limits, but rather about using a general set of limits that will give a reasonably small, but unspecified, risk of a false alarm with any and every probability model. The problems with shape statistics completely undermine the idea that we can specify a particular probability model for our data. The shape statistics are simply not specific enough to allow for the selection of a probability model with less than thousands of data (see Figure 16). And even if we could select a probability model, the fact that probability models with the same shape parameters can differ in the extreme tails makes the use of a probability model to compute infinitesimal tail areas into a highly suspect operation that is unlikely to have any contact with reality.

Do we need to test the data to see if they "might be normally distributed?" Once again the answer is no. When we use a lack-of-fit test we are automatically assuming that the data set is homogeneous and that the underlying process is unchanging. When the process is changing you will typically end up with a histogram like Figure 11. The elongated tail will be picked up by the lack-of-fit test, you will decide that your data are not normally distributed, and then you are left trying to figure out what to do next. Some will suggest transforming the data in a non linear manner. However, when the data are not homogeneous it is not the shape of the histogram that is wrong, but the computation and use of descriptive statistics and lack-of-fit tests that is erroneous. When the data are not homogeneous we do not need to transform the data to change the shape of the histogram, but we rather need to stop and question what the lack of homogeneity means in the context of the original observations.

It is important to note that while the 200 data of Figure 11 will probabiliy fail a test for normality, the 3287 data of Figure 15 include the normal distribution within the 95% interval estimate box of Figure 16. If you test the first 200 data, and then transform them prior to further analysis, you are unlikely to ever progress to the point displayed in Figure 15.

Both of the approaches listed above are built on a naive assumption that the data are homogeneous. When the data are not homogeneous all of the computations, all of the lack-of-fit tests, and all of the justifications for transforming the data will break down. Therefore, the first step in any real-world analysis must always be an examination of the data for evidence of a lack of homogeneity. So we return to the one completely general technique we have that can examine suspect data for evidence of a lack of homogeneity—the process behavior chart. Process behavior charts are the premier, general purpose, thoroughly proven and verified technique for examining a collection of data for homogeneity. Simply organize your data in a rational manner, place them on a process behavior chart, and see if you find evidence of a lack of homogeneity. If you do, find out what is causing the process to change. If not, then draw the histogram and proceed to interpret your data in their context.

Since 1935, every attempt to embellish the process behavior chart technique has been built on either flawed assumptions or complete misunderstandings of the theory and purpose of process behavior charts. Process behavior charts are the first step in the analysis of industrial, managerial, and observational data. They do not need to be tweaked or updated. And they do not make any assumptions about the shape of the histogram.

So, in consideration of the many problems with the shape statistics, I have to agree with Shewhart when he concluded that the location and dispersion statistics provide virtually all the useful information which can be obtained from *numerical summaries* of the data. The use of additional statistics such as skewness and kurtosis is superfluous.