Transformation of Weibull Distributed Data
A surprising result

Written By: Rick Haynes, MBB

In most Lean Six Sigma (LSS) work, a limited set of population distributions is used to model historical data as part of the effort to predict the future capability. In most LSS courses, we are taught to use the Normal, Lognormal, and Weibull distributions. These distribution choices are very comfortable since they appear often in LSS projects. Process measurement data often occur with a normal distribution when there are no natural limits to the data. When we record the duration of time required to perform a task we generally find the data to have a lognormal distribution. A Weibull distribution is often found when measuring the time between events that might be found in a reliability analysis.

Of these three distributions, the process stability is easily evaluated for the normal and lognormal distributions using an individuals chart. The lognormal distributed data involves a simple transformation, the natural log, before charting. This transformation makes sense, since the natural log of a lognormal distribution always results in a normal distribution. If you did not realize it, the lognormal distribution parameters (location and scale) are actually the mean and the standard deviation for the natural log of the original data.

The Weibull distribution presents some real problems with respect to a transformation that prepares the data for an individuals control chart. No transformation is available that will always work for the Weibull distribution. In my past training, we were teaching the use of a Box-Cox transformation with a lambda = 0.5 (square root transformation) for the Weibull distribution, which does work sometimes, but so does the natural log transformation. In an effort to identify Weibull distribution transformation guidelines using the @Risk simulation software combined with Minitab software, I found that there is not a simple set of rules to identify when a practitioner should use the square root or the lognormal transformations. What I did find was a a predictive relationship between the Weibull distribution shape parameter and the optimal lambda to use in a Box-Cox transformation of the data.

$$\boxed{\lambda = 10^{-0.6230+1.000*\log10(shape)}}$$

Weibull Distribution background information

The Weibull distribution history is interesting. It was documented by Waloddi Wiebull, from Sweden, in a 1951 paper titled “A Statistical Distribution Function of Wide Applicability” presented to the American Society of Mechanical Engineers (ASME). This paper is easier to read and understand than most statistical publications, I guess because it was written for an engineering journal.

W. Weibull wrote that this distribution is intended to be “applied to a large group of problems, where … an occurrence of an event in any part of an object may be said to have occurred in the object as a whole.” This general quote from the paper means that it applies to a probability of a failure or other event in a system where a component or segment failure in a system will create a failure of the entire system. The example used in the paper uses a chain with multiple links. The probability of any single link failure can be estimated, but the probability of failure for the entire chain is best estimated using his proposed distribution. Other examples of this condition are the yield of a material during tensile testing, the time for a light bulb to fail, or even the death of a human. In each case there are many locations or causes for the event to occur. Each location or component has a probability of failure, so that the probability of system failure is 1 – P(no failures). The Weibull distribution function is derived using only that pre-condition. Waloddi Weibull wrote that his proposed distribution function is quite flexible in its possible shapes and can be used in a lot of conditions. It is this flexibility that makes the Weibull distribution so difficult to transform.

Since the Weibull distribution was proposed, many things have been discovered about the distribution.

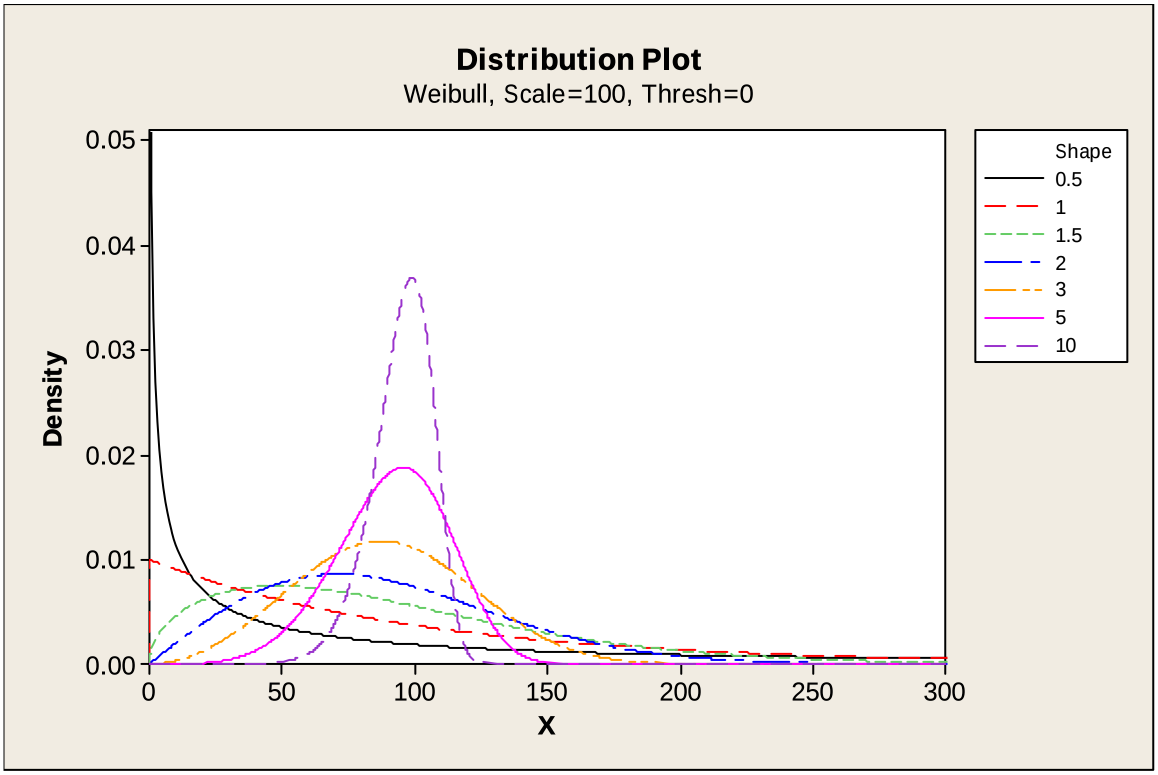
1. The natural log of Weibull distributed data will have a Gumbel distribution. Minitab calls it a smallest extreme value distribution.
2. The shape parameter can be used to describe the failure mode on the bathtub curve.
    a. If the shape parameter is < 1.0, then you are experiencing beginning of life or infancy failures, where the failure rate will drop over time.

   b. If the shape parameter is ~ 1.0, then you are experiencing the constant failure or the flat part of the bathtub curve, where the failure rate will stay about the same over time.
   c. If the shape parameter is > 1.0, then you are in the end-of-life or wear-out failure period, where the rate of failures will increase over time.
3. The exponential distribution is a special case of the Weibull distribution when the shape parameter is equal to 1.0.
4. The Rayleigh distribution is a special case of the Weibull distribution when the shape parameter = 2.0.
5. The scale parameter of a Weibull distribution is the value where the probability of values occurring less than the scale parameter is 63.2%.

After an observation of many years of LSS projects, Weibull distributions with shape parameters from 1 to 3 are most common. This may be due to the types of problems that are addressed with LSS, but there is no concrete evidence to support this belief. In reliability testing of equipment, Weibull distributions with shape parameters from near zero up to 2 or 3 are common, depending on the types of failures that are being experienced.

The Weibull distribution takes many forms as the shape parameter changes.

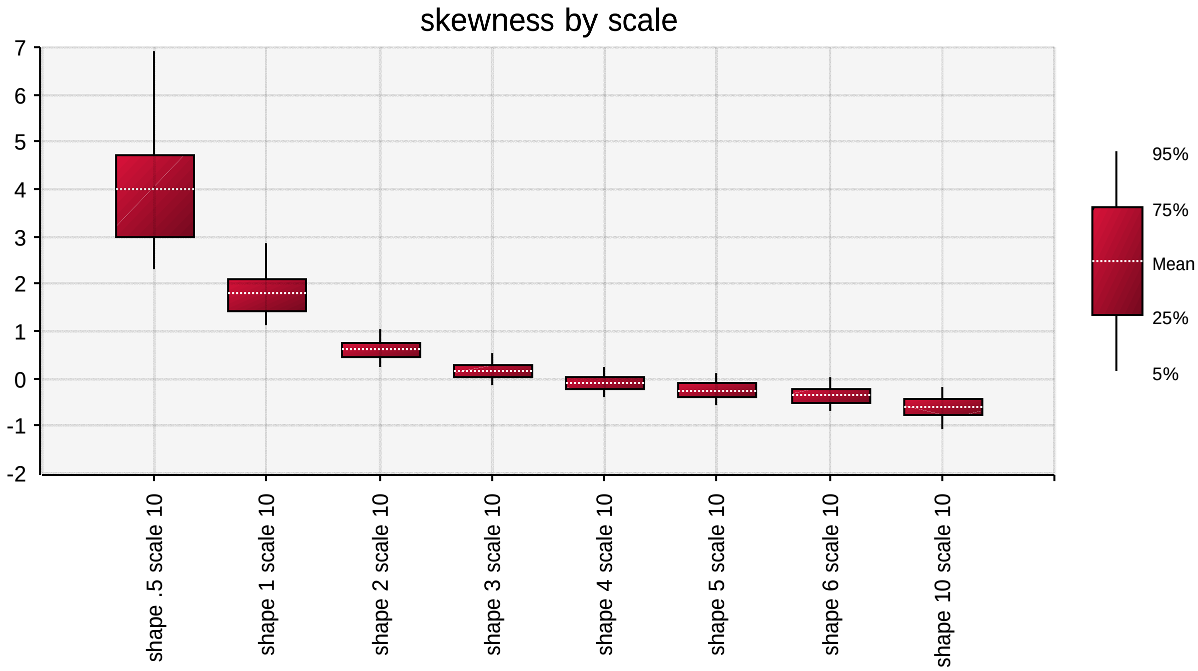
- For shape values less than or equal to 1.0, there is a failure probability that starts high and decreases over time.
- For shape parameters greater than 1.0, the initial failure probability increases to a maximum value and then decreases. These shapes are uni-modal in shape.
    - For shape values from 1.0 to around 3, the distribution is right skewed.
    - For shape values around 3 to 4, the shape of the distribution becomes symmetric and may be modeled as a normal distribution if the sample size is not too large.
    - For shape values greater than the 3 to 4 range, the distribution is left skewed.



The fact that the Weibull distribution can be skewed both to the right and the left makes the identification of a standard transformation of the data difficult.

Simulation Analysis of Weibull distributed data

Using the @Risk simulation add-in to Excel, a series of random Weibull distributed data is generated and evaluated with standard transformations to find the limits of their applicability. A sample size of 100 data points was used with the software to model Weibull distributions having scale parameters ranging from 0.2 to 4. For each set of data, two standard transformations were applied: natural logarithm and a square root. For each transformation and the untransformed data, the skewness was calculated, for each iteration of the simulation, and accumulated using the @Risk software. In this testing it was found that the symmetry (skewness = 0) of the Weibull does occur around a shape parameter of 3 to 4.

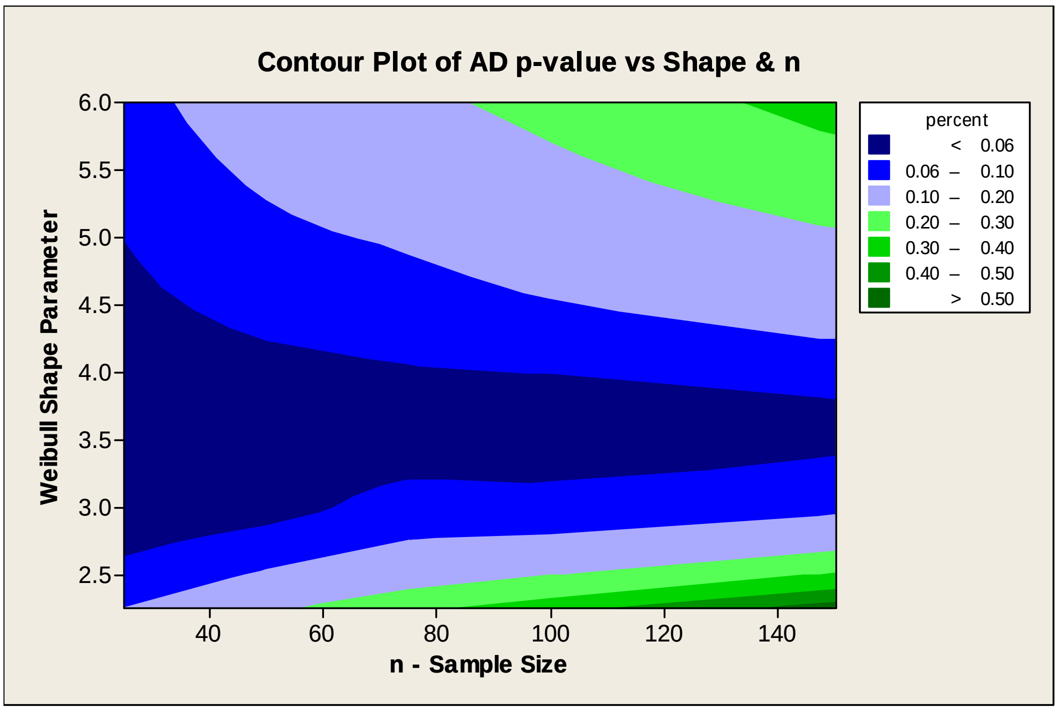


The next simulation involved changing the scale parameter to see if the skewness result is sensitive to changes in the scale. Performing the same simulation with scale values from 1 to 100 showed that the symmetry results appear to be independent of the Weibull distribution scale parameter.

Using the same simulation, the symmetry of the transformed Weibull data was examined. The results for the natural log transformed data fit what the literature stated; every single transformed Weibull distribution has an identical skewness value, -1.05, that corresponds to a Gumbel distribution. For the square root transformation, the symmetry was achieved with shape parameters around 2.0. This result indicates that a Weibull distribution with a shape around 2.0 should allow a successful transformation using a lambda = 0.5 in the Box-Cox transformation. Since shape values around 2.0 are commonly found when analyzing equipment reliability, the square root may be a good choice for transformations within a reliability project.

Since data analysis is easier when there is no need for a transformation of historical data, what could be learned about the Weibull characteristics that may allow the use of the data in an individuals chart without a transformation? To assess this concept, a second simulation was created that collected a random sample of Weibull distributed data and then the Anderson-Darling (AD) goodness of fit test was applied for a normal distribution. The accumulation of the p-values from the AD test easily identifies data conditions that would not test as normally distributed. A true normal distribution will provide p-value data that fits a uniform distribution, where exactly 5% of the results occur with a p-value less than 5%. Using this knowledge, the simulation collects the p-values and then reported the percent of the simulated data p-values that were less than a 5% value. As the percent of the p-values exceeds 5% of the total data, the data can be considered as exceedingly non-normal.

New information came from this simulation: the symmetry and subsequent AD test for normality result was a function of the sample size of the simulated data. This result fits the experiences of most LSS belts, where smaller sample sizes are more apt to fit many different distributions. The results of the simulation are shown in the chart below.



The contour plot shows that the optimal shape parameter may be around 3.6 and that Weibull data may be treated as normally distributed (no transform needed for a control chart) across both dark blue (%<.05) and medium blue (0.5<%<.10) areas. As n increases, the range of shape parameters that may be used un-transformed is smaller and smaller.

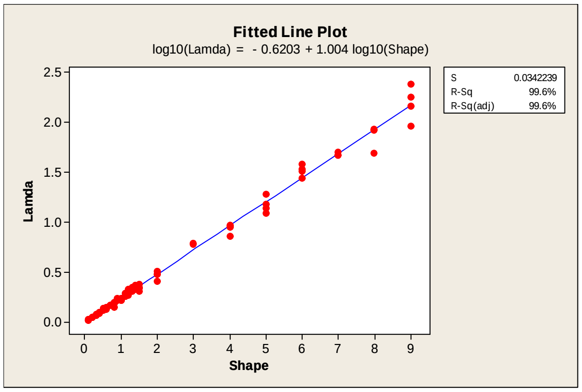
The original question, what is the best lambda to transform a Weibull distributed data set for control charting, has still not been fully answered. It has been demonstrated that there are some shape parameters that are transformable with a square root function (around shape equals 2) and some shape parameters that do not need a transform (shape parameters around

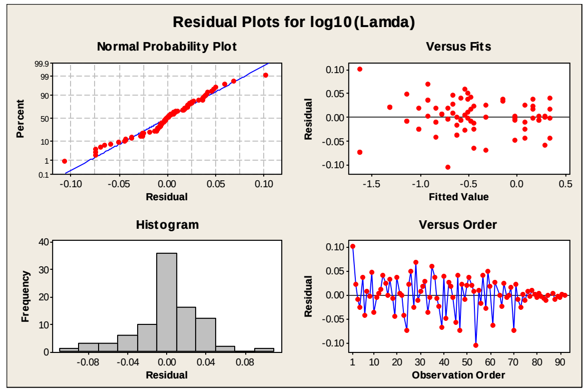
3.6). It appears that no standard transform will be possible. If it is not possible, could there be a relationship between the shape parameter and the lambda? Well, there is!

Using all of the simulated Weibull distributed data with different shape parameters and different sample sizes, the Box-Cox transformation was applied to each set of data and the optimal lambda was recorded. This analysis showed that the optimal lambdas were changing as the shape parameter changed, but it also showed that the optimal lambda was independent of the scale and the data sample size. The results are shown below.

Shape	Mean Lambda	SE Mean Lambda
0.1	0.0225	0.00250
0.2	0.0500	0.00000
0.3	0.0725	0.00250
0.4	0.0925	0.00250
0.5	0.1300	0.00408
0.6	0.1400	0.00408
0.7	0.1700	0.00000
0.8	0.1825	0.01109
0.9	0.2275	0.00479
1.0	0.2275	0.00479
1.1	0.2675	0.00750
1.2	0.2950	0.01258
1.3	0.3300	0.00913
1.4	0.3450	0.00957
1.5	0.3450	0.01443
2.0	0.4700	0.02121
3.0	0.7850	0.00500
4.0	0.9350	0.02533
5.0	1.1725	0.04029
6.0	1.5150	0.02901
7.0	1.6775	0.00750
8.0	1.8725	0.06088
9.0	2.1975	0.08826

Performing a regression on the lambda values as a function of the shape parameter demonstrated a clear relationship, but the residual analysis indicated that the variance of the lambda estimate was increasing as the shape parameter increased, which voids the regression results. The use of a variance stabilization transformation, log10, for both the scale and the lambda value, provided a near perfect relationship and residual result.





The resulting predictive model: $\boxed{\log 10(\lambda) = -0.6203 + 1.004 * \log 10(shape)}$

will allow for the optimal transformation lambda value to be estimated from only the shape parameter. Algebra provides a form of this equation that can be used.

$$\boxed{\lambda = 10^{-0.6203+1.004*\log 10(shape)}}$$

As a verification of this relationship, all of the generated data sets were plotted with an individuals control chart using the generated lambda and all of the charts behaved as expected with no skewness or natural limit issues that are seen with most un-transformed Weibull data charts. The lambda values provided by Minitab all have only two significant figures after the decimal place, so that there will be a known error in prediction due to the rounding of the true lambda value to what is reported.

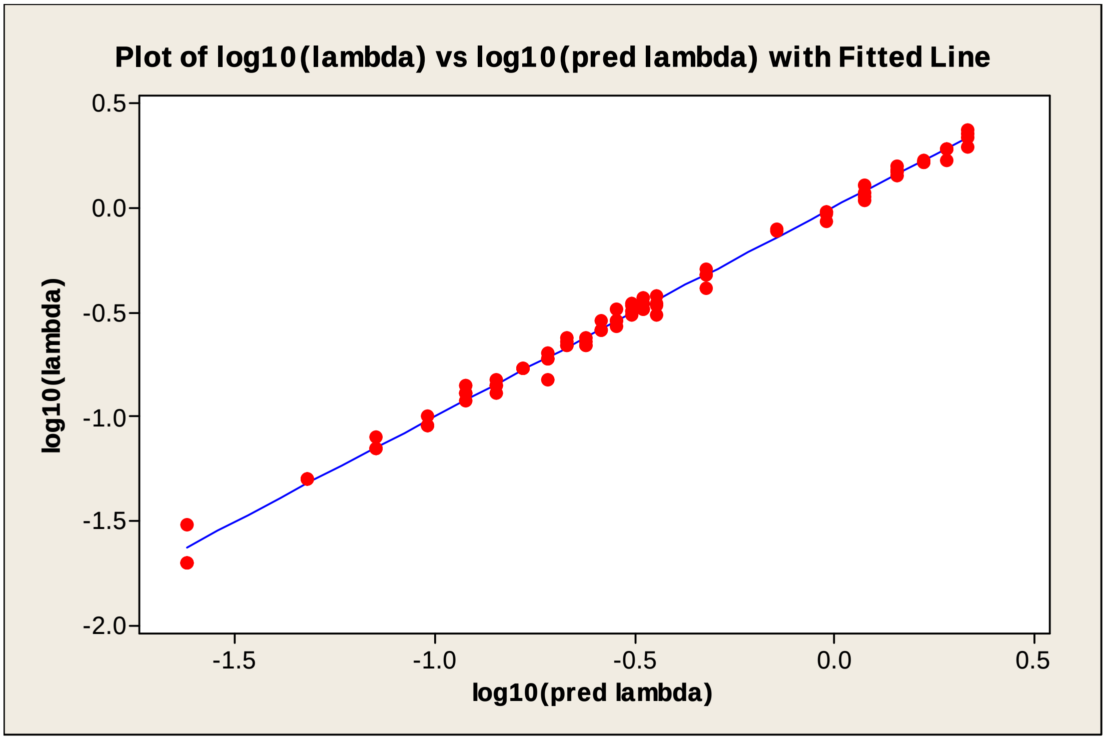
Simulations with sample sizes up to 10000 data points were also performed to see if the regression parameters converge to common values, which was not observed, but it came close. With one coefficient so close to 1.0000 and the other coefficient close to the 0.623 value (which is the percent of Weibull data less than the scale parameter), it was hypothesized that the true parameters of this equation should be -0.623 and 1.000. To test this hypothesis, a set of predicted lambdas was generated using the hypothesized equation.

$$\boxed{\lambda = 10^{-0.6230+1.000*\log 10(shape)}}$$

A comparison of these results to the actual optimal lambdas provided by Minitab was performed using the Orthogonal Regression function, showing that the hypothesized equation is acceptable. This regression method is used to compare to values that should have the same value, such as two measurements of the same item. The hypothesis is validated if the slope of the line equals 1.0 and the intercept of the line equals zero. When both conditions are found, it is reasonable to consider both measurement systems to be equivalent. Due to the prior need for a Log10 variance transformation, the log10 variance transformation was applied before running the regression. This analysis showed that the null hypothesis was validated; the assumed coefficients provided an equal predictivity of the generated lambda values.

Orthogonal Regression Analysis: log10(lambda) versus log10(pred lambda)

```
Error Variance Ratio (log10(lambda)/log10(pred lambda)): 1

Regression Equation
log10(lambda) = 0.007 + 1.007 log10(pred lambda)

Coefficients

Predictor               Coef    SE Coef         Z      P     Approx 95% CI
Constant             0.00656  0.0050987    1.2865  0.198  (-0.003434, 0.01655)
log10(pred lambda)   1.00651  0.0069973  143.8431  0.000  ( 0.992798, 1.02023)

Error Variances

Variable              Variance
log10(lambda)        0.0005759
log10(pred lambda)   0.0005759
```



Given this result, the common value coefficient based equation is recommended.

The discovery of an equation that relates the Weibull distribution parameters to an optimal lambda value is not, in itself, surprising. Finding that the equation contains coefficients that can also be derived from the Weibull distribution may be coincidental, but if true, it would be a extraordinary finding. Since these results have been derived from using simulations rather than the use of a closed form solution, it does not provide proof of the relationship. The derivation of a mathematical proof is the domain of PhD statisticians, and possibly something that this Master degreed statistician is capable of completing. For now, this is empirical evidence of an equation to predict an appropriate Box-Cox lambda value to transform Weibull data, so go ahead and use it.