1 Variable Transformations

Abstract

Variable transformation methods are presented to recover the normality and the equal-standard deviation conditions required by many statistical analysis methods.

Key Words

normal distribution, bell-shaped distribution, histogram, normality, normal probability plot, normal plot, asymmetry, transformation, homoscedastic, heteroscedastic, square root transform, arcsine transform, Freeman-Tukey transform, Box-Cox transform, Johnson transform

Related Methods

- Graphical methods for assessing normality: histograms, histograms with superimposed normal curves, normal probability plots
- Quantitative tests for normality: Kolmogorov-Smirnov test, Lillifors test, Anderson-Darling test
- Quantitative tests for equal standard deviations (i.e. homoscedasticity): F test, Levene's test, Bartlett's test
- Transformations: square root transform, Freeman-Tukey transform, power transform, Box-Cox transform, Johnson transform

Introduction

Many statistical analysis methods assume that the distribution of the population being studied follows the bell-shaped or normal distribution. In addition, if two or more populations are being compared, such as to test for a difference between their means, many analysis methods require that the populations being tested also have equal standard deviations - a condition called *homoscedasticity*. When one or both of these requirements are violated, the distribution normality and/or the homoscedasticity condition can often be recovered by applying a mathematical transformation to the original data. After an appropriate transformation is applied, then the usual analysis methods which require normality and homoscedasticity may be used.

The purpose of this document is to describe the use of variable transformations to enable the use of statistical analysis methods which require normality and homoscedasticity. Methods for transforming non-linear scatter plots to linear scatter plots are outside the scope of this document.

Where the Technique is Used

The transformation method is applied to attribute (i.e. count) data and variable (i.e. measurement) data which are non-normal and/or are from two or more populations with different standard deviations for the purpose of recovering the normality and homoscedasticity conditions so that classical normal-theory statistical methods may be used.

Procedure

The following procedure may be used to apply the variable transformation method:

1. Collect a representative random sample (or samples) from the population(s) of interest.
2. If the issue is distribution shape:

(a) Construct a histogram and/or normal probability plot of the sample data.

(b) Interpret the plots to determine if the data are normal.

(c) A quantitative test for normality (such as the Kolmogorov-Smirnov test, Lillifors' test, or the Anderson-Darling test) may also be helpful.

(d) If the data are not normal either:

i. Identify and apply the correct, theoretical distribution

ii. Apply an appropriate variable transformation. Well known variable transformations are available for certain types of data, but it might be necessary to consult with a subject matter expert on a more challenging problem. Evaluate the transformed data for normality.

3. If the issue is non-normality and/or heteroscedasticity in two or more treatment groups:

(a) Construct histograms and/or normal probability plots of the sample data superimposed on the same graphs.

(b) Interpret the plots to determine if the data are normal and homoscedastic.

(c) A quantitative test for normality (such as the Kolmogorov-Smirnov test, Lillifors' test, or the Anderson-Darling test) and/or heteroscedasticity (such as the two-sample F test or the many-sample Levene's or Bartlett's tests) may be helpful.

(d) If the distributions are not normal and/or homoscedastic, apply an appropriate variable transformation.

(e) Evaluate the transformed distributions for normality and homoscedasticity.

Assessing Normality

A variety of methods for assessing the normality of a sample are available, but they do not all have equal sensitivity for detecting non-normal distributions. This section makes a quick comparison of three methods for assessing normality: histograms, normal probability plots, and quantitative tests for normality. In practice all three methods should be used.

The most common method used to assess normality is the histogram, often with a superimposed normal curve, however, it's difficult to accurately judge the agreement between the compound curvature of the normal curve and the discrete bars of the histogram. This issue is severe enough that other methods for assessing normality are preferred.

A much better graphical method than a histogram for evaluating normality is the normal probability plot or normal plot. A normal plot is a form of two-dimensional scatter plot which plots the observed data values versus their predicted values, where the predicted values are calculated under the assumption that the observed values come from a normal population. Predicted values may be expressed in different ways, but they are usually expressed in terms of normal probability percentage points. Interpretation of a normal plot is simple: if the plotted points fall along a substantially straight line then the normality assumption is probably true but if the plotted points deviate substantially from a straight line, often in a hockey-stick or S-shaped curve, then the normality assumption is probably false.

The interpretation of histograms and normal plots is subjective; it takes much training and practice to learn to use them accurately. Quantitative tests for normality are often used to supplement both graphical methods. There are many quantitative methods for testing normality: the Anderson-Darling test, the Wilk-Shapiro test, the Kolmogorov test, Lillifors' test, the chi-square goodness of fit test, and others. (The Anderson-Darling test is the most popular one in use today.) The test statistics for these tests are all calculated in different ways, but they all involve some measure of deviation from a

normal distribution. To standardize the interpretation of these various test statistics a single common summary statistic called the p value is used. p values are probabilities, so they range from zero to one in value. *The p value for a quantitative normality test indicates the probability of obtaining the observed experimental data set or one even more unusual if the population being sampled is actually normal.* Under this definition, our interpretation of a p value that falls in the interval $0.05 < p < 1$ is that the sample probably comes from a normal population. Our interpretation of a p value that falls in the interval $0 < p \leq 0.05$ is that the sample probably doesn't come from a normal population.

Index of Transformations

Many experimental responses, by their very nature, are non-normal and/or heteroscedastic. The following index of transformations describes some common problematic responses and methods for dealing with them. Transformations appropriate for other situations may be recommended by a subject matter expert. In some cases a response cannot be transformed to normality and other distributions specific to the situation must be employed. When such methods are available, they are generally preferable over the relatively crude method of transformation to approximate normality.

Defect or Poisson-Distributed Count Data

Defect count responses often follow a Poisson distribution. This type of data occurs when the number of events is counted per unit of opportunity. Examples of Poisson-distributed responses are: number of telephone calls per hour, number of transactions processed per day, number of defects per transaction, number of stones thrown from the back of a speeding dump truck per minute, number of computer crashes per month, and number of paint defects per car door.

The Poisson distribution shape and standard deviation change with the distribution mean μ. If fact, the Poisson distribution standard deviation (σ) is equal to the square root of the mean: $\sigma = \sqrt{\mu}$. When the mean count is less than about $\mu = 20$, the Poisson distribution is markedly right-skewed (i.e. with a short left tail and a long right tail). When the mean count is greater than about $\mu = 20$ the Poisson distribution is sufficiently normal and no transformation is usually necessary. The greater the Poisson mean, the closer the normal distribution approximates the Poisson distribution.

The appropriate transformation for Poisson-distributed data is the square root transform which improves the distribution normality and also stabilizes the standard deviation. That is, square root-transformed Poisson data have a standard deviation of about $\sigma' = 1/2$ regardless of the distribution mean (where the prime on σ indicates the standard deviation of the transformed values).

To demonstrate how the square root transformation recovers the approximate normality of Poisson-distributed data, a random sample of 2000 observations from a Poisson distribution with $\mu = 4$ were drawn. Figure 1 shows histograms and normal plots of the original and transformed data. The histogram of the original data (upper left) is markedly skewed right and the corresponding normal plot (upper right) has pronounced curvature. The histogram and normal plot of the square root-transformed data (lower left and right, respectively) show that the transformed data are much more normal than the original data. (The coarseness of the histogram of the transformed counts is due to the discrete nature of the count response and the attempt by the software that created the histogram to employ constant width bins.)

To demonstrate how the square root transform recovers both the normality and the homoscedasticity of Poisson-distributed data, random samples of size $n = 40$ were drawn from Poisson populations with means $\mu = 4, 8, 16, 32$, and 64. The normal plots of the original data in the left panel of Figure 2 show that the normal plots are approximately normal, especially for large μ, however, the non-parallel fitted lines through the plotted points indicate that the Poisson distribution standard deviation increases with μ. The normal plots of the square-root transformed data in the right panel of Figure 2 show that the transformed data are both more normal than the original data and that they are more homoscedastic.

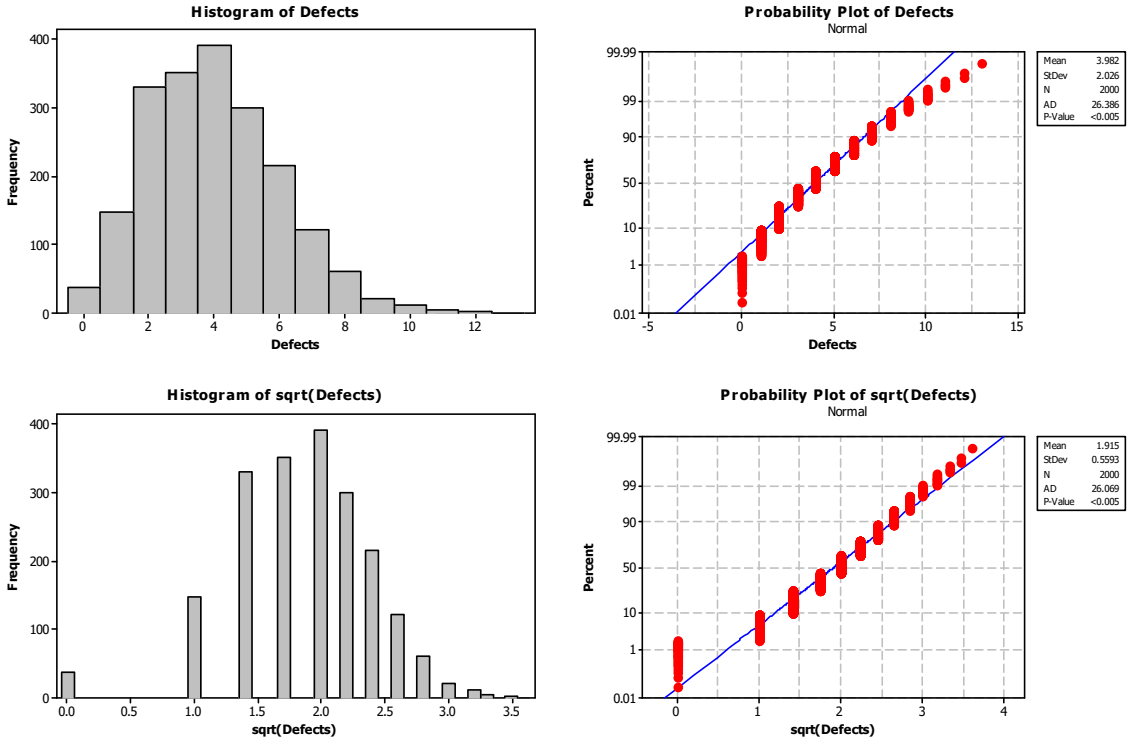

Figure 1: Square root-transformed Poisson data are approximately normal.

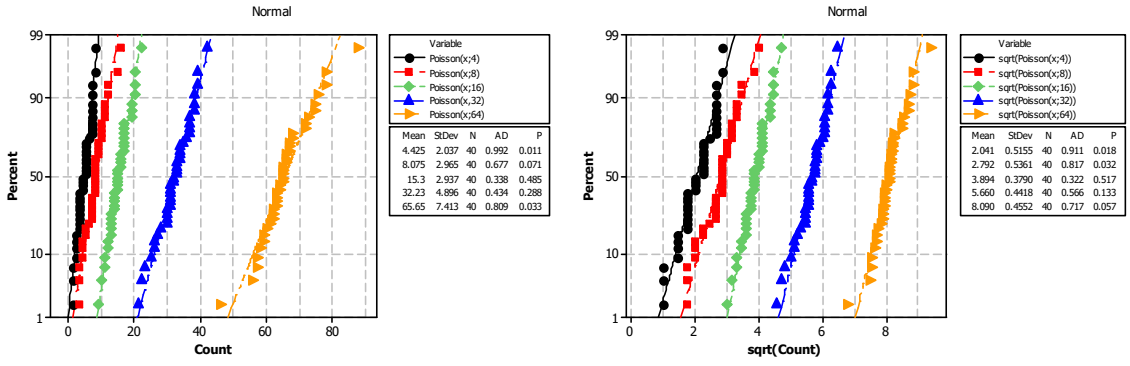

Figure 2: Square root-transformed count data are approximately normal and homoscedastic.

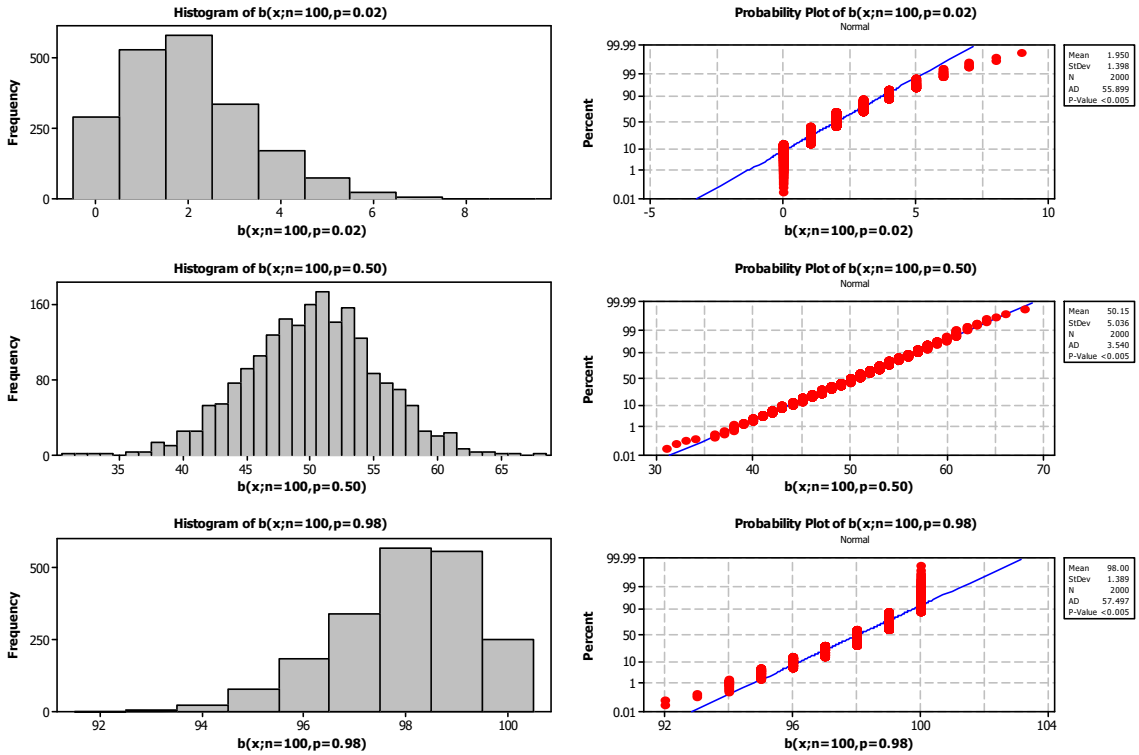

Figure 3: Binomial histograms may be skewed right, approximately normal, or skewed left.

Defective or Binomially-Distributed Count Data

When inspected units are classified into one of two complementary or binary categories, such as defective or not defective, the distribution of the counts often follows a binomial distribution. Examples of binomially distributed data are: the number of defective units observed in a sample, the number of people who fail a drug test, the number of customer transactions that are left incomplete and require follow-up action, the number of traffic lights that you have to stop for on your way home from work, etc. For each example, the complementary counts are also binomially distributed.

It's important but often difficult to distinguish binomially-distributed counts from Poisson-distributed counts. One trick that might be helpful is to recognize that for binomially-distributed counts, the count cannot exceed the sample size, whereas for Poisson-distributed counts, the upper limit on the count is potentially infinite.

The binomial distribution changes its mean, standard deviation, and shape with sample size (n) and population proportion (p). The binomial distribution mean and standard deviation are determined from n and p by $\mu = np$ and $\sigma = \sqrt{np\,(1-p)}$. When n and p simultaneously satisfy the two conditions $np > 5$ and $n\,(1-p) > 5$, the binomial histogram is approximately normal, however, for $np < 5$ the binomial histogram will be skewed right (i.e. have a long right tail) and for $n\,(1-p) < 5$ it will be skewed left (i.e. have a long left tail).

Figure 3 shows binomial histograms and normal plots for three binomial distributions. Each case is constructed from 2000 observations of sample size $n = 100$, however, the binomial proportions are $p = 0.02$, 0.50, and 0.98, respectively. (Note that the distribution means are given by $\mu = np = 2$, 50, and 98, respectively.) The first ($p = 0.02$) and third ($p = 0.98$) cases are markedly skewed, but the second case ($p = 0.50$) is quite symmetric and apparently normal.

The appropriate transformation for binomially-distributed count data which recovers the normality and homoscedasticity requirements is:

$$p' = \arcsin\left(\sqrt{p}\right)$$

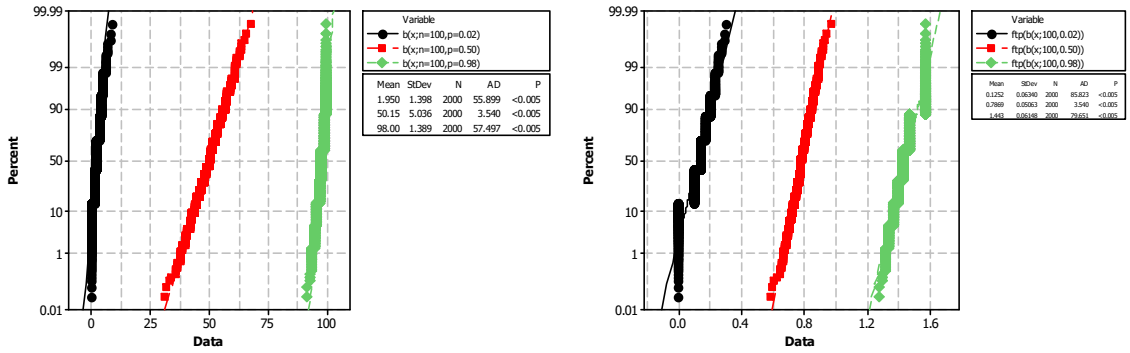

Figure 4: Transformed binomial distributions are approximately normal and homoscedastic.

where p is the ratio of defective counts to sample size: $p = D/n$. The standard deviation of p' is independent of p but dependent on n, so the homoscedasticity condition is only obtained when samples from two or more populations use the same sample size.

Log-Normal Data

The distributions of some common measurement responses are non-normal but the distributions of their logarithm-transformed values are normal. In these cases, any logarithmic transformation such as to base 10 (log), base e (ln), or to any other base, will work. These distributions are called log-normal distributions. Many waiting time and mechanical strength responses are log-normal. While logarithmic transformations recover the normality of log-normal distributions, the transformed distributions from two or more populations may still have different standard deviations.

Figure 5 shows the histogram (upper left corner) and corresponding normal plot (upper right corner) for the amount of time (t) that call center customers were on hold waiting to talk to a service representative. The histogram and normal plot both show that the waiting time distribution is skewed right, i.e. has a long right tail. The histogram (lower left corner) and normal plot (lower right corner) of the log-transformed waiting times show that the transformed time values are normal.

Weibull Distribution

Although waiting time and mechanical strength responses are often log-normally distributed, another common distribution for these types of responses is the Weibull distribution. The Weibull distribution is very flexible because it includes a shape parameter β (*beta*) that allows it to take on a wide variety of histogram shapes. When $\beta = 1$ the Weibull distribution becomes the exponential distribution. When $2.5 < \beta < 4.5$ the Weibull distribution is approximately normal, but outside of this range Weibull-distributed data is markedly non-normal. The characteristic Weibull location parameter, called the scale parameter, is indicated with the symbol η (*eta*).

When a waiting time or mechanical strength response is not fitted well with a normal or log-normal distribution, the Weibull distribution is a likely alternative. The best way to determine whether a sample might come from a Weibull population is to use a Weibull probability plot. Weibull plots are constructed the same way as normal plots except that the Weibull probability distribution is used instead of the normal distribution to determine the predicted values to associate with observed values. When experimental data come from a Weibull distribution, those data will plot on a substantially straight line on a Weibull plot.

Figure 6 shows histograms for three sample Weibull distributions. All three distributions have the same scale parameter, $\eta = 100$, which gives them similar locations, but they have three different shape

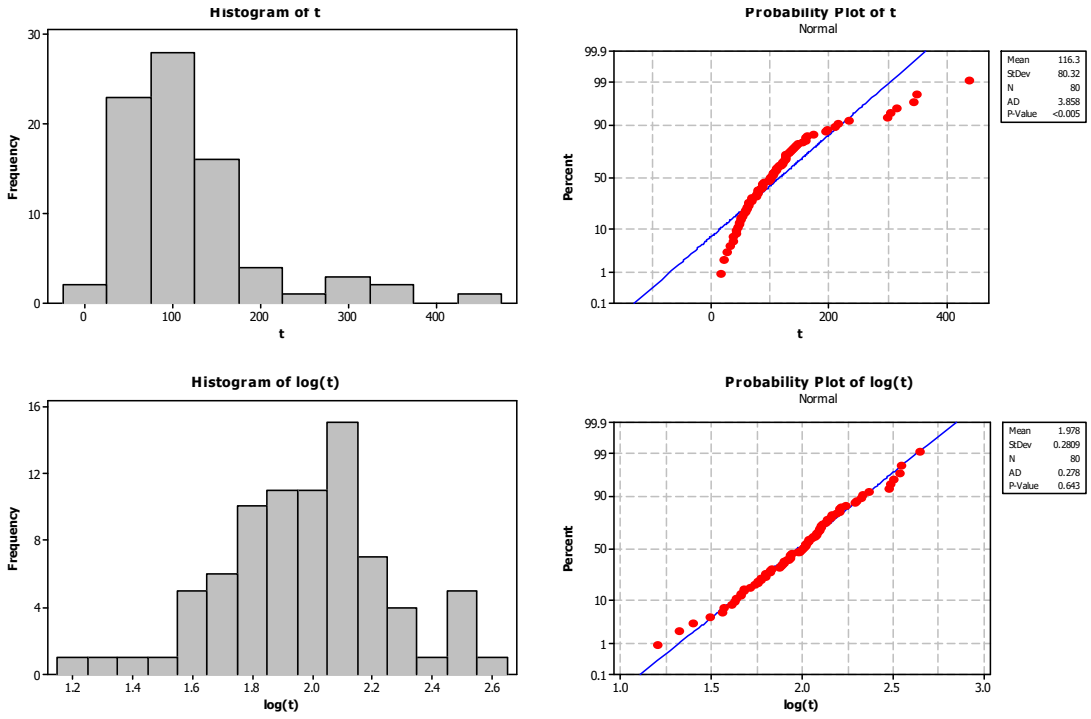

Figure 5: Logarithmic tranformations transform log-normal distributions into normal distributions.

parameters, $\beta = 0.6$, 1.5, and 3.0. The figure clearly shows the wide range of histogram shapes that a Weibull distribution can take on as a function of the shape parameter. The first and second cases in the Figure are clearly right-skewed but the third case appears to be approximately normal.

Figure 7 shows a Weibull plot for the three data sets from Figure 6. All three data sets plot on substantially straight lines and they cross near their common scale parameter value $\eta = 100$. The slopes of the lines in the plot are determined by the shape parameter β.

While normal probabilities must be calculated using appropriate software or a published look-up table, Weibull probabilities can be calculated from the simple formula:

$$R(t;\eta,\beta) = e^{-\left(\frac{t}{\eta}\right)^{\beta}}$$

where t is the waiting time or strength response and $R(t)$ is the probability that the response-limiting event occurs after t, i.e. that the customer's call has not been answered by time t or that the mechanical strength of the sample being tested exceeds strength t. Note that the the *Percent* scale in Figure 7 is the complement of $R(t)$, that is, $Percent = 1 - R(t;\eta,\beta)$.

As an example of how to calculate and interpret the Weibull probability, suppose that the distribution of waiting times for customers requesting service from a call center is Weibull with $\eta = 100$ seconds and $\beta = 1.5$ (The second case in Figure 6 shows a sample histogram taken from a population with these Weibull parameter values.) and that we wish to estimate the fraction of customers who can expect to wait more than $t = 180$ seconds to have their calls answered. From the Weibull reliability equation:

$$\begin{aligned} R(t = 180; \eta = 100, \beta = 1.5) &= e^{-\left(\frac{180}{100}\right)^{1.5}} \\ &= 0.089 \\ &= 8.9\% \end{aligned}$$

or 8.9% of customers will have to wait more than 180 seconds to have their calls answered.

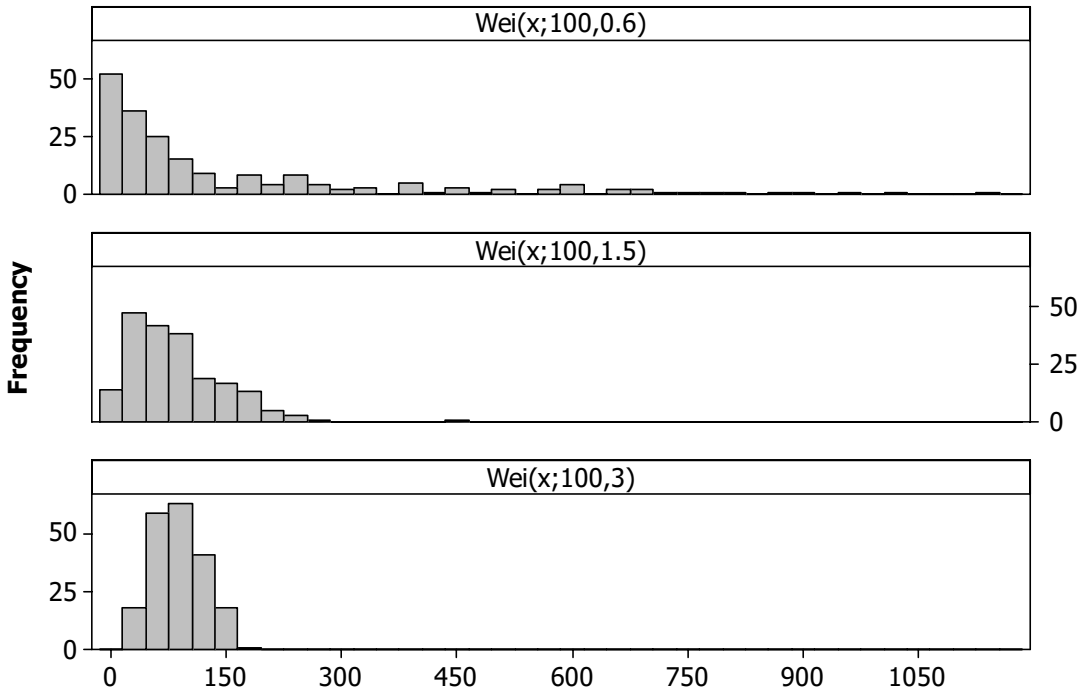

Figure 6: Weibull distributions can take on a variety of shapes.

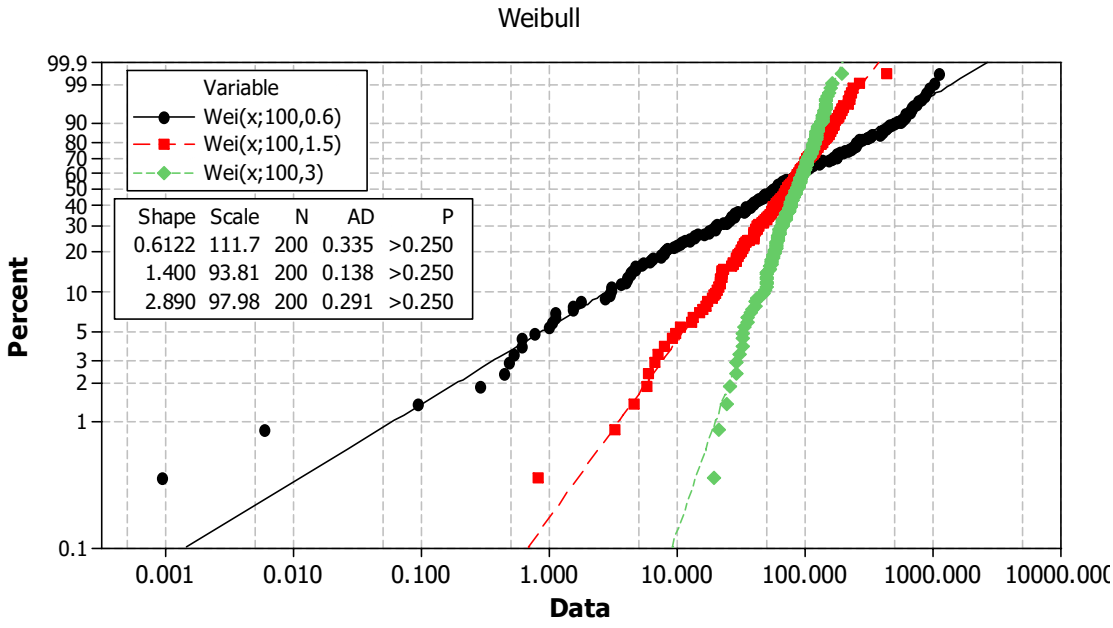

Figure 7: Weibull-distributed response plot as straight lines on a Weibull plot.

Box-Cox and Johnson Transforms

When a variable transformation for measurement data cannot be found by inspection and an appropriate alternative distribution cannot be identified, it may be necessary to attempt to find an empirical transformation that converts the original data to near-normality and/or homoscedasticity. There are many methods available for finding such transformations but most of them are integrated into two ubiquitous methods: Box-Cox transforms and Johnson transforms. Both methods are computationally complex so they are only performed with software. Both methods suffer from difficulties interpreting the transformed data and the empirical nature of the transforms makes them methods of last resort. The usual sequence of steps in applying the methods is to exhaust other methods first, then try the Box-Cox transform, and finally try the Johnson transform.

The Box-Cox method uses a power function transform of the form:

$$y_i' = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} \text{ when } \lambda \neq 0 \\ \log(y_i) \text{ when } \lambda = 0 \end{cases}$$

The optimal value of λ that best transforms the y_i to normality is determined by maximizing the p value of the Anderson-Darling test for normality with respect to λ. The method is restricted to data sets that have all positive values.

The Johnson transformation is a collection of three families of transformations for bounded (S_B), log-normal (S_L), and unbounded (S_U) data. The transformation has flexible coefficients that permit a wide variety of non-normal distributions to be transformed to near-normality. As with the Box-Cox method, the optimal values of those flexible coefficients are determined by maximizing the Anderson-Darling test's p value.

References

Zar (1996) *Biostatistical Analysis*, 3rd Ed., Prentice-Hall.

Hoaglin, Mosteller, and Tukey (1991) *Fundamentals of Exploratory Analysis of Variance*, Wiley..

Snedecor and Cochran (1980) *Statistical Methods, 7th Ed., Iowa State University Press.*

Y. Chou, A.M. Polansky, and R.L. Mason (1998). "Transforming Nonnormal Data to Normality in Statistical Process Control," *Journal of Quality Technology*, 30, April, pp 133-141.