**Transforming Variables to Reduce Skewness**

 Many traditional inferential analyses assume that the variables are normally distributed. When planning to employ such techniques, one should always screen the data to detect any problems. Sometimes the problem may be due to the presence of outliers. Investigation of outliers might reveal that they are bad data, and deleting them from the dataset might reduce the skewness to an acceptable level. With most analyses, I am good with the normality assumption if the absolute value of the *g1* skewness statistic does not exceed 1. High values of kurtosis indicate that outliers are probably present and should be investigated.

 After dealing with outliers, if the scores are still unacceptably skewed, consider transforming them to reduce skewness. If the skewness is positive, the following transformations should be tried (it is OK to try several and chose the one that most reduces skewness). One should re-evaluate homogeneity of variance assumptions after applying a transformation, as these transformations can alter differences in within-group variances.

**Positive Skewness**

* Any transformation that preserves the order of the scores (that is, an ordinal (aka monotonic) transformation and reduces the value of large scores more than of small scores might help
* Transform the scores to their square roots. Other roots may be tried to, such as cube roots. The square root of Y is Y.5, that is, Y\*\*.5. The cube root is Y1/3, that is, Y\*\*.33333. One might also consider exponents such as 1/2.5 – that would be Y\*\*.4. If your data contains negative values, you should, before applying the root transformation, add a constant to every score such that the lowest score is no longer negative.
* Log transformations may be useful, and generally reduce skewness more than a square root transformation.
* A negative reciprocal transformation, that is, T = -1/Y, generally reduces positive skewess quite a bit.
* Any of these transformation might eliminate the problem with positive skewness but result in a problem with negative skewness.

**Negative Skewness**

 What if the skewness is negative? A traditional method of dealing with is to first reflect the scores and then transform. Reflecting the scores will change the negative skewness into positive skewness – the absolute value of *g1* will not change, but its sign will change from negative to positive. After reflection, a transformation is applied to reduce the positive skewness. When interpreting the results of the subsequent analysis, one must remember the reflection. For example, suppose you were dealing with scores from a measure of social conservatism and you found that they had a distinctive negative skewness. You reflect them and then apply a log transformation, successfully normalizing them. High scores now indicate high social liberalism, not high social conservatism. To avoid that confusion, you could re-reflect the scores after transformation.

 It is possible to transform to reduce negative skewness without reflection. For example, you might try exponentiating the scores. Exponentiated Y is , where e is the base of the natural logarithm, 2.718. Here you may also need to reduce the values of the scores (before exponentiation) by dividing by a constant, to avoid producing transformed scores so large that they exceed the ability of your computer to record them. If your highest raw score were a 10, your highest transformed score would be a 22,026. A 20 would transform to 485,165,195, and so on. Just as root transformations reduce positive skewness, transforming to Yexp (when exp > 1) will reduce negative skewness. Again, you may need to divide by a constant before transforming. As with log and root transformations, if you have any negative scores, add a constant to all scores such that the lowest score is no longer negative.

**Back-Transformation**

 Suppose you were comparing the relative effectiveness of two weight-loss programs. Body-weight change (pre minus post) was measured in pounds. The data suffered from positive skewness (in both groups), so you used a square root transformation to normalize the data. A t-test on the transformed data showed a significant difference in weight-loss, *MA* = 2 and *MB* = 4. The unit of measure for these means is not pounds but rather the square root of pounds. To get back to the original unit of measure, just square the transformed means, so *MA* = 4 and *MB* = 16.

**Nonparametric and Resampling Statistics**

 These procedures do not assume that the populations are normally distributed. The classic nonparametric procedures involve transforming the scores to ranks. In their usual application nonparametric procedures do involve the assumption that the populations are of the same shapes and dispersions. Resampling statistics come very close to being assumption free – the only assumption is that the scores in the populations are distributed like those in the samples.

**Example of Use of Log Transformation**

 From [this article](https://royalsocietypublishing.org/doi/full/10.1098/rspb.2018.1520) on the effects of hormones on body scents.

“The reported estimates in the multilevel models are unstandardized regression coefficients. Because examination of hormonal data revealed that the distributions were skewed, we log transformed the hormone values to achieve normal distributions. We report analyses performed with log-transformed data, but whether we used raw or normalized data did not change the results.”



Links of Interest on this Topic

* [Handbook of Biological Statistics](http://www.biostathandbook.com/transformation.html)
* [Examples Using R](http://rcompanion.org/handbook/I_12.html)
* [Bates University](http://abacus.bates.edu/~ganderso/biology/bio270/homework_files/Data_Transformation.pdf)
* [J Pharmacol Pharmacother](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3043340/)

[Karl L. Wuensch](http://core.ecu.edu/psyc/wuenschk/klw.htm), January, 2019