ERIC Identifier: ED470204
Publication Date: 2002-08-00
Author: Osborne, Jason W.
Source: ERIC Clearinghouse on Assessment and Evaluation College Park MD.

Normalizing Data Transformations. ERIC Digest.

Data transformations are the application of a mathematical modification to the values of a variable. There are a great variety of possible data transformations, from adding constants to multiplying, squaring, or raising to a power, converting to logarithmic scales, inverting and reflecting, taking the square root of the values, and even applying trigonometric transformations such as sine wave transformations.

The goal of this Digest is to present some of the issues involved in data transformation, with particular focus on the use of data transformation for normalization of variables. The Digest is intended to serve as an aid to researchers who do not have extensive mathematical backgrounds or who have not had extensive exposure to this issue. It will focus on three of the most common data transformations utilized for improving normality as discussed in texts and the literature: square root, logarithmic, and inverse transformations. Readers looking for more information on data transformations might refer to Hartwig and Dearing (1979) or Micceri (1989).

DATA TRANSFORMATION AND NORMALITY

Many statistical procedures assume that the variables are normally distributed. A significant violation of the assumption of normality can seriously increase the chances of the researcher committing either a Type I (overestimation) or Type II (underestimation) error, depending on the nature of the analysis and the non-normality.. However, Micceri (1989) points out that true normality is exceedingly rare in education and psychology. Thus, one reason (although not the only reason) that researchers utilize data transformations is to improve the normality of variables. Additionally, authors such as Zimmerman (1998) have pointed out that nonparametric tests (where no explicit assumption of normality is made) can suffer as much, or more, than parametric tests when normality assumptions are violated, confirming the importance of normality in all statistical analyses, not just parametric analyses.

There are multiple options for dealing with non-normal data. First, the researcher must make certain that the non-normality is due to a valid reason (real observed data points). Invalid reasons for non-normality include mistakes in data entry and missing data values not declared missing.

Not all non-normality is due to data entry error or nondeclared missing values. Two other reasons for non-normality are the presence of outliers (scores that are extreme relative to the rest of the sample) and the nature of the variable itself. There is great debate in the literature about whether outliers should be removed or not. Judd and McClelland (1989) argue that outlier removal is desirable, honest, and important; Orr, Sackett, and DuBois (1991) are among those holding the opposing view. Should a researcher remove outliers and find substantial non-normality, or choose not to remove outliers, data transformationis a viable option for improving normality of a variable.

IDENTIFYING A NORMALITY ASSUMPTION VIOLATION

There are several ways to tell whether a variable is substantially non-normal. While researchers tend to report favoring "eyeballing the data," or visual inspection (Orr, Sackett, and DuBois, 1991), researchers and reviewers are often more comfortable with a more objective assessment of normality, which can range from simple examination of skew and kurtosis to examination of P-P plots (available through most statistical software packages) and inferential tests of normality, such as the Kolmorogov-Smirnov test and adaptations of this test.

THREE DATA TRANSFORMATION

While many researchers in the social sciences are well trained in statistical methods, not many of us have had significant mathematical training, or if we have, it has often been long forgotten. This section is intended to give a brief refresher on what really happens when one applies a data transformation.

Square root transformation. Most readers will be familiar with this procedure. When one applies a square root transformation, the square root of every value is taken. However, as one cannot take the square root of a negative number, if there are negative values for a variable, a constant must be added to move the minimum value of the distribution above 0, preferably to 1.00 (the rationale for this assertion is explained below). Another important point is that numbers of 1.00 and above behave differently than numbers between 0.00 and 0.99. The square root of numbers above 1.00 always become smaller, 1.00 and 0.00 remain constant, and number between 0.00 and 1.00 become larger (the square root of 4 is 2, but the square root of 0.40 is 0.63). Thus, if you apply a square root to a continuous variable that contains values between 0 and 1 as well as above 1, you are treating some numbers differently than others, which is probably not desirable in most cases.

Log transformation(s). Logarithmic transformations are actually a class of transformations, rather than a single transformation. In brief, a logarithm is the power (exponent) to which a base number must be raised in order to get the original number. Any given number can be expressed as y to the x power in an infinite number of ways. For example, if we were talking about base 10, 1 is 100, 100 is 102, 16 is 101.2, and so on. Thus, log10(100)=2 and log10(16)=1.2. However, base 10 is not the only option for log transformations. Another common option is the Natural Logarithm, where the constant e (2.7182818) is the base. In this case the natural log 100 is 4.605. As the logarithm of any negative number or number less than 1 is undefined, if a variable contains values less than 1.0, a constant must be added to move the minimum value of the distribution, preferably to 1.00.

There are good reasons to consider a range of bases. Cleveland (1984) argues that base 10, 2, and e should always be considered at a minimum. For example, in cases in which there are extremes of range, base 10 is desirable, but when there are ranges that are less extreme, using base 10 will result in a loss of resolution, and using a lower base (e or 2) will serve. (Higher bases tend to pull extreme values in more drastically than lower bases). Readers are encouraged to consult Cleveland (1984) for more details.

Inverse transformation. To take the inverse of a number (x) is to compute 1/x. What this does is essentially make very small numbers very large, and very large numbers very small. This transformation has the effect of reversing the order of your scores. Thus, one must be careful to reflect, or reverse the distribution prior to applying an inverse transformation. To reflect, one multiplies a variable by -1, and then adds a constant to the distribution to bring the minimum value back above 1.0. Then, once the inverse transformation is complete, the ordering of the values will be identical to the original data.

In general, these three transformations have been presented in the relative order of power (from weakest to most powerful). A good guideline is to use the minimum amount of transformation necessary to improve normality.

Positive vs. Negative Skew. There are, of course, two types of skew: positive and negative. All of the above-mentioned transformations work by compressing the right side of the distribution more than the left side. Thus, they are effective on positively skewed distributions. Should a researcher have a negatively skewed distribution, the researcher must reflect the distribution, add a constant to bring it to 1.0, apply the transformation, and then reflect again to restore the original order of the variable.

ISSUES SURROUNDING THE USE OF DATA TRANSFORMATIONS

Data transformations are valuable tools, offering man benefits. However, they should be used appropriately, in an informed manner. Too many statistical texts gloss over this issue, leaving researchers ill-prepared to utilize these tools appropriately. All of the transformations examined here reduce non-normality by reducing the relative spacing of scores on the right side of the distribution more than the scores on the left side.

However, the very act of altering the relative distances between data points, which is how these transformations improve normality, raises issues in the interpretation of the data. If done correctly, all data points remain in the same relative order as prior to transformation. This allows researchers to continue to interpret results in terms of increasing scores. However, this might be undesirable if the original variables were meant to be substantively interpretable (e.g., annual income, years of age, grade, GPA) because the variables become more complex to interpret due to the curvilinear nature of the transformations. Researchers must therefore be careful when interpreting results based on transformed data.

CONCLUSIONS AND OTHER DIRECTIONS

Four recommendations for researchers grow out of this discussion:

1. Always examine and understand data prior to performing analyses. To do less is to increase the chance of drawing incorrect conclusions.

2. Know the requirements of the data analysis technique to be used. As Zimmerman (1998) and others have pointed out, even nonparametric analyses, which are generally thought to be "assumption-free," can benefit from examination of the data.

3. Utilize data transformations with care--and never unless there is a clear reason. Data transformations can alter the fundamental nature of the data, such as changing the measurement scale from interval or ratio to ordinal and creating curvilinear relationships, thereby complicating interpretation. As discussed above, there are many valid reasons for utilizing data transformations, including improvement of normality, variance stabilization, and conversion of scales to interval measurement

4. Ensure that the variable is anchored at a place where the transformation will have the optimal effect. In the case of the three transformations discussed, the anchor point should be 1.0.

REFERENCES

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. The American Statistician, 38 (4): 270-280.

Hartwig, F. and Dearing, B.E. (1979). Exploratory Data Analysis. Newberry Park, CA: Sage Publications, Inc.

Judd, C. M. and McClelland, G.H. (1989). Data Analysis: A Model-Comparison Approach. San Diego, CA: Harcourt Brace Jovanovich.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105: 156-166.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. Personnel Psychology, 44: 473- 486.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. Journal of Experimental Education, 67: 55-68.