A Comparison Study of Data Transformation Methods to Achieve Normality

OZER OZDEMIR
Department of Statistics
Anadolu University
Anadolu University, Department of Statistics, 26470, Eskisehir, Turkey
TURKEY
ozerozdemir@anadolu.edu.tr

Abstract: - Normality is the one of main important central assumptions in statistical studies. Since in reality this is not the fact, transformation of random variables are required to achieve specified purposes i.e. stability of variance, the additivity of effects and the symmetry of the density. In this study, we make a comparison study in order to check the power of the transformations method for satisfying the normality. We simulated Log-normal, Beta and Gamma probability distributions with various parameters in order to transform them to be normal. The statistical hypothesis tests that are well known to be powerful are used in order to examine the performance of the transformation methods.

Key-Words: - Data transformations, Monte Carlo simulation, Box-Cox transformation, Normality comparison.

1 Introduction

Data transformations are an important tool for the proper statistical analysis of data from various disciplines such as biological, ecological, medical studies. There are a great variety of possible data transformations, from adding constants to multiplying, squaring or rising 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 [1],[2],[3]. The most frequently used transformation method through others is Box–Cox transformations, also known as power transformations. In this study, we aimed to compare special transformations on simulated data from different statistical families, i.e. Lognormal, Beta, Gamma, Weibull, and Rayleigh.

2 Non-normal Probability Distributions and Data Transformations for Normality

In this study we considered Log-normal, Beta and Gamma probability distributions. Box-Cox transformations proposed by [4] in which $x_i > 0$,

$$y_i(\lambda) = \begin{cases} \left(x_i^{\lambda} - 1\right)/\lambda, & if\ \lambda \neq 0 \\ \log_e(x_i), & if\ \lambda = 0 \end{cases} \quad (1)$$

where coefficient λ can be the maximum likelihood estimation. Another form of power transformation that is frequently used is given by

$$y_i(\lambda) = \begin{cases} x_i^{\lambda} & if\ \lambda \neq 0 \\ \log(x_i) & if\ \lambda = 0. \end{cases} \quad (2)$$

In 1982, Box and Cox [5] gave a modification of formulation by

$$y(\lambda) = \begin{cases} \dfrac{x^{\lambda} - 1}{\underline{x}^{(\lambda - 1)}} & if\ \lambda \neq 0 \\ \underline{x}\log_e(x) & if\ \lambda = 0, \end{cases} \quad (3)$$

where $\underline{x}$ is the geometric mean of all observations.

These transformations have been chosen based on theoretical or empirical evidence to achieve normality.

3 Monte Carlo Simulations

In this part of the study, we simulated data from different statistical families, i.e. Lognormal, Beta and Gamma. Lognormal distributions with location parameter 0 and scale parameters (10, 1.5, 1, 0.5, 0.25, 0.125), Beta distribution with parameters ((0.5, 0.5), (5, 1), (1, 3),(2, 2),(2, 5),(5, 25),(25, 5),(0.5, 25)), Gamma distribution with parameters ((1, 2), (2, 2), (3, 2), (5, 1), (9, 0.5)) are considered for simulation study. 10.000 random samples with 30 units are generated for each specified probability distribution.

Anderson-Darling test is used for the data transformed from Beta distribution and Jarque-Bera test which uses skewness and kurtosis is used for the transformed data from Lognormal and Gamma distributions.

The simulation results for Beta distribution are shown in Table 1.

Table 1. Simulation results for Beta distribution

Parameters	X	Square Root	Geo. Mean*Log
(0.5, 0.5)	0.1433	0.1882	0.0046
(5, 1)	0.2111	0.1174	0.0645
(1, 3)	0.3724	0.9382	0.3732
(2, 2)	0.9255	0.8036	0.3111
(2, 5)	0.7823	0.956	0.6005
(5, 25)	0.8226	0.9508	0.8489
(25, 5)	0.8222	0.7666	0.6971
(0.5, 25)	0.0018	0.4734	0.3568

According to Table 1, it is obvious that if we have random sample from Beta distribution then only the square root transformation is appropriate to achieve normality. The simulation results for Lognormal distribution are shown in Table 2.

Table 2. Simulation for Lognormal distribution

Parameters	X	Square Root	Geo. Mean*Log
(0, 10)	0	0	0.9712
(0, 1.5)	0.0135	0.2129	0.9672
(0, 1)	0.09	0.4751	0.9683
(0, 0.5)	0.4689	0.8113	0.9687
(0, 0.25)	0.8084	0.9248	0.9702
(0, 0.125)	0.9315	0.963	0.97

The simulation results showed that the Log and Geometrical Mean*Log transformation are perfect for Lognormal distribution. Table 3 shows the simulation results for Gamma distribution.

Table 3. Simulation results for Gamma distribution

Parameters	X	Square Root	Geo. Mean*Log
(1, 2)	0.2803	0.8818	0.6662
(2, 2)	0.5253	0.9329	0.8049
(3, 2)	0.6476	0.9488	0.8545
(5, 1)	0.7719	0.9618	0.8966
(9, 0.5)	0.8522	0.9628	0.9283

The results denote that the best results for normality test are obtained by square root transformation for Gamma distribution.

4 Conclusion

In this study, we consider continuous probability functions (i.e. Lognormal, Beta and Gamma) which have importance in applications of several disciplines as engineering, biology, medical sciences etc. These distributions are considered with different parameters in order to make appropriate comparison to detect the best transformation for normality. Monte Carlo simulation results showed that the square root transformation is the only one that success to achieve normality in all different cases.

References:

[1] Osborne, Jason, Notes on the use of data transformations, *Practical Assessment, Research & Evaluation*, Vol. 8, No.6, 2002, Retrieved March 21, 2009.

[2] Hoyle, M.H., Transformations: an introduction and a bibliography, *International Statistical Review*, Vol.41, No.2, 1973, pp. 203–223.

[3] Tan W.D., Gan F.F., Chang T.C., Using normal quantile plot to select an appropriate transformation to achieve normality, *Computational Statistics & Data Analysis*, Vol.45, 2004, pp.609 – 619.

[4] Box, G.E.P., Cox, D.R., An analysis of transformations, *J. Roy. Statist. Soc. Ser. B*, Vol.26, 1964, pp.211-252.

[5] Box, G.E.P. and Cox, D.R., An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, Vol.77, 1982, pp.209–210.