Power Transformations for Families of Statistical Distributions to Satisfy Normality

Ozer OZDEMIR

***Abstract*—**Data transformations are an important tool for the proper statistical analysis of data from various disciplines such as biological, ecological, medical studies. The requirement for the data transformations is normality which is the one of main important central assumptions in these statistical analyses. A Monte Carlo simulation study is made for controlling the power transformation methods to achieve normality in this study. Log-normal, Beta, Gamma, Weibull and Rayleigh probability distributions are simulated with different parameters in order to transform them to be normal. The interpretations of the results are made and the convenient transformations for the each specified distribution is determined.

***Keywords*—**Normality, Power transformation, Simulation, Statistical distributions.

I. Introduction

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. For instance, the usual regression model techniques are applied by assuming that $Y = X\beta + \varepsilon$, where $Y = \left(Y_1,\ldots,Y_n\right)^T$ is the response vector to be estimated and $\varepsilon = \left(\varepsilon_1,\ldots,\varepsilon_n\right)^T$ is a normal random vector having mean 0 and covariance matrix $\sigma^2 I$. In essence, the distribution of ε_i in reality usually does not reflect the normality, because they often don't have equal variances. Therefore various transformation methods were defined to handle these problems and to transform the data to be almost normal. 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. Linear, log, square root, inverse, quadratic, cubic, and similar transformations are all special cases of Box–Cox formulations [6]. Box–Cox transformations attempt to transform the variable to a normally distributed one. Asymptotic theory for Box-Cox transformations in linear models are investigated by [7]. The efficiency of *t*-Test and Hotelling's T^2-Test after a Box-Cox transformation are studied by [8]. Moreover, [9] investigated the effects of skewed and leptokurtic multivariate data on the Type I error and power of Hotelling's T^2 were examined by manipulating distribution.

In this study, we aimed to compare special transformations on simulated data from different statistical families, i.e. Lognormal, Beta, Gamma, Weibull, and Rayleigh. The considered transformations are log, inverse, square root, partial entropy, geometric mean*log, Box-Cox modified transformation. We make Monte Carlo simulations and compare the results after transformation via normality test which is known to be powerful for the corresponding distribution.

II. Power Transformations for Normality

In this study we considered Log-normal, Beta, Gamma, Weibull and Rayleigh probability distributions which probability density functions are presented in Table 1(a) and Table 1(b).

Table 1(a): Probability Density Functions of Distributions

Probability Distribution	**Probability Density Functions**
Log-normal distribution	$f\left(x;\mu,\sigma\right) = \dfrac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{\left(\ln(x)-\mu\right)^2}{2\sigma^2}}$
Beta distribution	$f\left(x;\alpha,\beta\right) = \dfrac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} x^{\alpha-1}\left(1-x\right)^{\beta-1}$
Gamma distribution	$f\left(x;k,\theta\right) = x^{k-1}\dfrac{e^{-x/\theta}}{\theta^k\Gamma\left(k\right)}$ or $g\left(x;\alpha,\beta\right) = \dfrac{\beta^\alpha}{\Gamma\left(\alpha\right)} x^{\alpha-1} e^{-\beta x}$
Weibull distribution	$f\left(x;\lambda,k\right) = \begin{cases} \dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1} e^{-\left(x/\lambda\right)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Rayleigh distribution	$f\left(x;\sigma\right) = \dfrac{x}{\sigma^2}\exp\left(\dfrac{-x^2}{2\sigma^2}\right)$

Ozer OZDEMIR is with the Department of Statistics, Anadolu University Eskisehir 26470 TURKEY (corresponding author to provide phone: 902223350580-4668; e-mail: ozerozdemir@anadolu.edu.tr).

Table 1(b): Probability Density Functions of Distributions

Probability Distribution	**Details**
Log-normal distribution	for $x>0$, where μ and σ are the mean and standard deviation of the variable's natural logarithm
Beta distribution	where Γ is the gamma function. The beta function, B, appears as a normalization constant to ensure that the total probability integrates to unity.
Gamma distribution	for $x>0$ and $k,\theta>0$. for $x>0$, a shape parameter $\alpha=k$ and an inverse scale parameter $\beta=\dfrac{1}{\theta}$, called a rate parameter.
Weibull distribution	where $k>0$ is the shape parameter and $\lambda>0$ is the scale parameter of the distribution.
Rayleigh distribution	for $x\in[0,\infty)$.

The log-normal distribution is commonly used to model failure times in reliability applications. The Weibull distribution is related to a number of other probability distributions; in particular, it interpolates between the exponential distribution (*k* =1) and the Rayleigh distribution (*k* =2). The Weibull distribution and the Rayleigh distribution are very important for modelling the distribution of wind speed and significant wave height. Weibull distribution is also the one that is famous in survival analysis.

Unfortunately, in statistical analysis, we usually assume that the data is from a normal population when in fact it is not. So, we often have to transform the variables before carrying out the analysis. In this study we used several types of transformations (such as the reciprocal transformation, 1/Y, the square root transformation, $\sqrt{Y}$, and the logarithmic transformation, $\log Y$ and Box-Cox transformations).

Box-Cox transformations proposed by [4] in which $x_i>0$,

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

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

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

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

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

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

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

Additionally, we considered the geometric mean $\underline{x}$ *log and the partial entropy $x\log_e\left(x\right)$ for data transformation which are new considerations for data transformation.

III. Simulation Study

In this part of the study, we simulated data from different statistical families, i.e. Lognormal, Beta, Gamma, Weibull, and Rayleigh. 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)), Weibull distribution with parameters ((1, 0.5), (1, 1) ,(1, 1.5), (1, 5) ,(5, 10) ,(10, 5), (10, 25) ,(25, 10)) and Rayleigh distribution with parameters (0.5, 1.0, 2.0, 3.0, 4.0, 10, 25) 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, Weibull and Rayleigh distributions 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 2(a) and Table 2(b).

Table 2(a): Simulation results for Beta distribution

BETA (Anderson-Darling test)			
Parameters	**X**	**Log**	**Inverse**
(0.5, 0.5)	0.1433	0.0046	0
(5, 1)	0.2111	0.0645	0.0177
(1, 3)	0.3724	0.3732	0.0003
(2, 2)	0.9255	0.3111	0.0063
(2, 5)	0.7823	0.6005	0.0131
(5, 25)	0.8226	0.8489	0.1824
(25, 5)	0.8222	0.6971	0.5455
(0.5, 25)	0.0018	0.3568	0

Table 2(b): Simulation results for Beta distribution

BETA (Anderson-Darling test)			
Parameters	**Square Root**	**Partial Entropy**	**Geo. Mean*Log**
(0.5, 0.5)	0.1882	0.1667	0.0046
(5, 1)	0.1174	0.5167	0.0645
(1, 3)	0.9382	0.053	0.3732
(2, 2)	0.8036	0.0315	0.3111
(2, 5)	0.956	0.0089	0.6005
(5, 25)	0.9508	0.7431	0.8489
(25, 5)	0.7666	0.9207	0.6971
(0.5, 25)	0.4734	0.066	0.3568

According to Table 2(a) and Table 2(b), it is obvious that if we have random sample from Beta distribution then only the square root transformation is appropriate to achieve normality. However, it works well only for some of the Beta distributions (i.e. with parameters (1, 3) ,(2, 5) and (5, 25)). In general for the simulations from Beta distribution, we can say that while location parameter is increasing and scale parameter is decreasing, the best data transformation is partial entropy. For example, the random samples from Beta(5, 1) are 21.11% normally distributed before transformation and after the partial entropy transformation that proportion increases to 51.67%). In contrast, while location parameter is decreasing and scale parameter is increasing, the best data transformation is square root. If both location and scale parameter are smaller than 1, data transformation may not be significant for beta distribution.

The simulation results for Lognormal distribution are shown in Table 3(a) and Table 3(b).

Table 3(a): Simulation results for Lognormal distribution

LOGNORMAL (Jarque-Bera)			
Parameters	**X**	**Log**	**Inverse**
(0, 10)	0	0.9712	0
(0, 1.5)	0.0135	0.9672	0.0133
(0, 1)	0.09	0.9683	0.0902
(0, 0.5)	0.4689	0.9687	0.4672
(0, 0.25)	0.8084	0.9702	0.8091
(0, 0.125)	0.9315	0.97	0.9247

Table 3(b): Simulation results for Lognormal distribution

LOGNORMAL (Jarque-Bera)			
Parameters	**Square Root**	**Partial Entropy**	**Geo. Mean*Log**
(0, 10)	0	0	0.9712
(0, 1.5)	0.2129	0.0007	0.9672
(0, 1)	0.4751	0.0078	0.9683
(0, 0.5)	0.8113	0.1078	0.9687
(0, 0.25)	0.9248	0.4946	0.9702
(0, 0.125)	0.963	0.8241	0.97

Table 4(a) and Table 4(b) shows the simulation results for Gamma distribution.

Table 4(a): Simulation results for Gamma distribution

GAMMA (Jarque-Bera)			
Parameters	**X**	**Log**	**Inverse**
(1, 2)	0.2803	0.6662	0.0068
(2, 2)	0.5253	0.8049	0.0614
(3, 2)	0.6476	0.8545	0.1418
(5, 1)	0.7719	0.8966	0.2959
(9, 0.5)	0.8522	0.9283	0.507

Table 4(b): Simulation results for Gamma distribution

GAMMA (Jarque-Bera)			
Parameters	**Square Root**	**Partial Entropy**	**Geo. Mean*Log**
(1, 2)	0.8818	0.0355	0.6662
(2, 2)	0.9329	0.0521	0.8049
(3, 2)	0.9488	0.2014	0.8545
(5, 1)	0.9618	0.5417	0.8966
(9, 0.5)	0.9628	0.7596	0.9283

Anderson-Darling test is used for the normality checking after the data simulated from Weibull distribution is transformed. The results are given in Table 5(a) and Table 5(b).

Table 5(a): Simulation results for Weibull distribution

WEIBULL (Anderson-Darling test)			
Parameters	**X**	**Log**	**Inverse**
(1, 0.5)	0.069	0.6099	0.0003
(1, 1)	0.0669	0.6072	0.0002
(1, 1.5)	0.0613	0.6131	0.0003
(1, 5)	0.0633	0.6022	0.0003
(5, 10)	0.9352	0.6023	0.2032
(10, 5)	0.8223	0.6035	0.3769
(10, 25)	0.8171	0.5972	0.3734
(25, 10)	0.6986	0.6069	0.5107

Table 5(b): Simulation results for Weibull distribution

WEIBULL (Anderson-Darling test)			
Parameters	**Square Root**	**Partial Entropy**	**Geo. Mean*Log**
(1, 0.5)	0.8115	0.0032	0.6099
(1, 1)	0.8172	0	0.6072
(1, 1.5)	0.8153	0	0.6131
(1, 5)	0.8121	0.0013	0.6022
(5, 10)	0.8183	0.955	0.6023
(10, 5)	0.728	0.8786	0.6035
(10, 25)	0.7146	0.8559	0.5972
(25, 10)	0.6525	0.7258	0.6069

Simulation results for Rayleigh distribution is obtained by applying Anderson-Darling test and is summarized in Table 6(a) and Table 6(b).

Table 6(a): Simulation results for Rayleigh distribution

RAYLEIGH (Anderson-Darling test)			
Parameters	**X**	**Log**	**Inverse**
0.5	0.812	0.6026	0.0212
1.0	0.8079	0.6037	0.0232
2.0	0.8155	0.6053	0.0217
3.0	0.8181	0.6085	0.0223
4.0	0.8088	0.6115	0.0247
10	0.8132	0.6074	0.0223
25	0.8125	0.6016	0.023

Table 6(b): Simulation results for Rayleigh distribution

RAYLEIGH (Anderson-Darling test)			
Parameters	**Square Root**	**Partial Entropy**	**Geo. Mean*Log**
0.5	0.947	0.0037	0.6026
1.0	0.9531	0.0629	0.6037
2.0	0.9528	0.253	0.6053
3.0	0.9559	0.3523	0.6085
4.0	0.9525	0.4097	0.6115
10	0.9548	0.5251	0.6074
25	0.9504	0.5947	0.6016

IV. Results and Conclusion

In this study, we consider continuous probability functions (i.e. Lognormal, Beta, Gamma, Weibull, and Rayleigh) 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.

Anderson-Darling test is used for testing the normality of the data transformed from Beta distribution. Results denote the proportion of random samples which are found to be normal distributed after the each transformations (i.e. log, inverse, square root, partial entropy, geometric mean*log). First column of the Table 2(a) and Table 2(b) gives the proportion of random samples which are normally distributed without transformations. So, we can compare the proportion of normally distributed samples before transformation and after transformation.

Jarque-Bera test is used to check the normality of the data transformed from Lognormal distribution in Table 3(a) and Table 3(b). The simulation results showed that the Log and Geometrical Mean*Log transformation are perfect for Lognormal distribution. They are enough for the Log-normal distribution with any parameter and other transformations are not necessary. Decreasing of value of scale parameter is meaningful and effective for inverse, square root and partial entropy data transformations.

Table 4(a) and Table 4(b) shows the simulation results for Gamma distribution. In this case, we used Jarque-Bera test to check the normality of the random samples. The results denote that the best results for normality test are obtained by square root transformation for Gamma distribution. We can express that while location parameter is increasing and scale parameter is decreasing, all results are better. In these results, square root transformation is sufficient for every kind of parameters and has the most random samples for normal distribution.

According to Table 5(a) and Table 5(b), the best results are obtained by square root transformation when location parameter is equal to 1, and best transformation is partial entropy when location parameter is greater than 1. The results of Log and Geometrical Mean*Log transformation are not affected by parameter variation.

According to Table 6(a) and Table 6(b), we can state that the results of data transformation except partial entropy are not affected by parameter variation. Increasing of value of parameter is important for using of inverse data transformation. The square root transformation has the successful results for Rayleigh distribution and is the only appropriate one.

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.

[6] Refaat, M., Data Transformations, *Data Preparation for Data Mining Using SAS*, 2007, pp. 115-140.

[7] Cho, K., Yeo, I-K., Johnson, R. A. and Loh, W. Y., Asymptotic theory for Box-Cox transformations in linear models, *Statistics & Probability Letters*, Vol. 51, 2001, pp. 337-343.

[8] Freeman, J. and Modarres, D., Efficiency of *t*-Test and Hotelling's *T*2-Test After Box-Cox Transformation, C*ommunications in Statistics—Theory and Methods*, Vol. 35, 2006, pp. 1109–1122.

[9] Kirisci, L., Al-Subaihi, A. A. and Tarter, R., Effects of the generalized Box–Cox transformation on Type I error rate and power of Hotelling's T^2, *Journal of Statistical Computation and Simulation,* 75(3), 2005, pp. 199–206.

Ozer Ozdemir was born in Turkey in 1982. He received his B.Sc., M.Sc. and Ph.D. degrees in statistics in the Department of Statistics at Anadolu University, Turkey, respectively in 2005, in 2008 and in 2013. He has worked as a Research Assistant from 2006-2008, as a Lecturer from 2008-2014 and as an Assistant Professor from 2014 in the Department of Statistics at Anadolu University, Turkey.

He has published over 50 international conference papers and journals in his research areas. His research interests include Applied Statistics, Simulation, Artificial Neural Networks, Fuzzy Logic, Fuzzy Modeling, Time Series, Computer Programming, Statistical Software and Computer Applications in Statistics.