Transformation To Normality Based On Empirical Distribution Functions

[1]Mark Borres and [2]Efren Barabat

Abstract

The paper examines an efficient alternative to the Box-Cox and Yeo-Johnson's transformation to normality procedures which works under very general conditions. The method hinges on two fundamental results : the fact that the cumulative distribution function F(x) of a random variable X always has a U(0,1) distribution and the Box-Mueller transformation of uniform random variables to standard normal random variables. Given two observations x and y, we computed Fn(x) and Fn(y) , which for large n, are approximately uniform random variables. These values are then inputted into the Box-Mueller transformations. Bounds for the Kolmogorov-Smirnov statistic between the distribution of the transformed observations and the normal distribution are provided through numerical simulation and by appealing to the Dvoretzky-Kiefer-Wolfowitz inequality.

Keywords: transformation to normality, Box-Cox method, Johnson method, inequalities, Dvoretzky-Kiefer-Wolfowitz

1.0 Introduction

Most parametric statistical tests of hypotheses in statistical inference rely on the assumption that the data on hand are normally distributed. In fact, for some of these statistical tests, departure from this assumption can lead to serious consequences in terms of either the power of the tests or the level of significance of the tests while others can be quite robust to departures from the normality assumption (Huber, 1981). Since such statistical tests are often used , it is a good practice to transform the data to one which obeys the normal distribution prior to their use in data analysis.

The most popular method used to transform data to normality is the Box-Cox transformation technique. Thus, if X is any non-negative random variable whose distribution is not normal, then the Box-Cox technique finds an exponent α such that:

$$(1)\quad Y = (X^{\alpha} - 1)/\alpha\,, if\ \alpha \neq 0\ or\ Y = log(X)\ if\ \alpha = 0$$

is normal. If α = 1, then no transformation is needed; if α = -1, then an inverse transformation is required; if α = ½, a square root transformation may be appropriate. By convention, α = 0 will refer to a logarithmic data transformation. The usual range for the values of α is between -2 to 2 and the process is by trial and error. The trial and error procedure involved in using the family of Box-Cox transformations makes it unpopular in practice. A recent addition to the methodologies for transforming data distribution to normal is the Yeo- Johnson (2000) transformation which generalizes the Box-Cox methodology for negative random variables but which also suffers from the same analytic problems as the Box-Cox method:

$$\psi(x,y) = \begin{cases} \left[(y+1)^{\lambda} - 1\right]/\lambda, & \lambda \neq 0, y \geq 0 \\ \log(y+1) & \lambda = 0, y \geq 0 \\ -\left[(-y+1)^{2-\lambda}\right]/2 - \lambda, & \lambda \neq 2, y < 0 \\ -\log(-y+1) & \lambda = 2, y < 0 \end{cases}$$

[1]College of Arts and Sciences [2]College of Engineering

For some obvious distributions, for instance, when data are obtained from a uniform distribution on 0 to 1, the Johnson's method is unable to find an appropriate transformation to normality for the data even if some well-established procedures for transforming U(0,1) random variables to normally distributed random variables are available.

The search for better and more efficient methods for transforming non-normal data to normal ones continues to date. This paper proposes a more general approach to data transformation which does not require trial and error and which can be easily implemented with today's faster and more efficient computing power. The proposed method is surprisingly simple and is based on the well-known inverse transform theorem in probability and the popular Box-Mueller transformation to normality for uniform , U(0,1), random numbers. While there may be other reasons for transforming data , we restrict our concern to the objective of transforming observations so that they become normally distributed. Section 2 discusses the basic concepts needed to understand the implementation of the proposed procedure.

2.0 Basic Concepts

The uniform distribution on [0,1] whose density is given by:

(2) $g(u) = 1, \ 0 \leq u \leq 1$

is the basis for generating random numbers from other distributions. We now state and prove the inverse-transform theorem.

Theorem 1: Let X be a random variable with density f(x) and cumulative distribution function F(x), then F(x) is uniformly distributed on [0,1]. That is,

(3) $U = F(x)\ has\ a\ U[0,1]\ distribution.$

Proof:

Let x have the cdf F(x). Then,

$$P(U \leq u) = P(F(x) \leq u) = P(x \leq F^{-1}(u)) = F(F^{-1}(u) = u$$

which is the cdf of a uniform random variable. It follows that F(x) is uniformly distributed on [0,1]. ■

It follows that $x = F^{-1}(U)$. If we can generate a uniform random number U, then we can always generate a random number x from a distribution f(x) by simply following this inversion formula.

Theorem 2: (Box-Mueller Theorem) Suppose U_1 and U_2 are independent random variables that are uniformly distributed in the interval (0, 1]. Let

$$Z_1 = R\cos(\Theta) = \sqrt{-2\ln U_1}\cos(2\pi U_2)$$

and

$$Z_2 = R\sin(\Theta) = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

Then Z_1 and Z_2 are independent random variables with a normal distribution of mean zero standard deviation 1.

Proof

Let $u_1, u_2 \underset{\sim}{iid}\ U(0,1)$. Then:

$$g(u_1, u_2) = 1, \qquad 0 \leq u_1 \leq 1 \ , \ 0 \leq u_2 \leq 1.$$

Let:

$$Z_1 = \sqrt{-2\ln u_1} \quad \cos 2\pi\ u_2$$

$$Z_2 = \sqrt{-2\ln u_1} \quad \sin 2\pi\ u_2$$

It follows that;

$$u_1 = e^{-\frac{1}{2}(Z_1^2 + Z_2^2)}$$

$$u_2 = \frac{1}{2\pi}\ tan^{-1}\left(\frac{Z_2}{Z_1}\right).$$

The Jacobian of this transformation is:

$$J = \begin{vmatrix} \frac{\partial^2 u_1}{\partial Z_1^2} & \frac{\partial^2 u_1}{\partial Z_1 \partial Z_2} \\ \frac{\partial^2 u_2}{\partial Z_1 \partial Z_2} & \frac{\partial^2 u_2}{\partial Z_2^2} \end{vmatrix} = \frac{1}{2\pi} e^{-\frac{1}{2}(Z_1^2 + Z_2^2)}$$

Hence:

$$g(Z_1, Z_2) = 1.|J| = \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2+z_2^2)}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_2^2}$$

The last line states that $g(Z_1, Z_2)$ is the product of the densities of two standard normal variates■

The Box-Mueller transformation is the transformation used in most computer-generated normal random variates. It is quite efficient in that it provides one normal variate for every uniform variate given. Other transformation approaches to generate normal random variates employ the Central Limit Theorem.

A combination of these two standard results provides a way of transforming non-normal observations into normally distributed random variables. Roughly, if X is distributed F(x), then we know that F(x) has a U(0,1) distribution. Let Y be independently drawn from F(.) so that F(y) will also have the uniform distribution on (0,1). Define Z = g(F(X),F(Y))) be the Box-Mueller transformation provided above. Our ability to implement this algorithm depends to a large extent on the availability of a closed-form expression for the cumulative distribution function F(x):

(4) $$F(x) = \int_{-\infty}^{x} f(t)dt$$

Even for well-known probability densities f(.), a closed-form expression for (4) may not be easily obtained e.g. normal densities, the family of beta densities and others. In order to circumvent this problem, we assume that we have a sufficiently large number of observations $x_1, x_2, ..., x_n$ iid F(.), where F(.) is unknown. We estimate F(x) by the empirical distribution function $F_n(x)$ given by:

(5) $$F_n(x) = \frac{1}{n} \sum I(x_i \leq x)$$

where I(.) is the indicator function. In effect, the empirical distribution function puts a mass of 1/n to each of the observations less than or equal to x_i. Each I(xi) is a Bernoulli random variable with p = F(x) so that by the Law of Large Numbers, we know that $F_n(x)$ converges to F(x) in probability. A stronger result was established independently by Glivenko and Cantelli showing that the convergence to F(x) in fact is uniform. The Glivenko-Cantelli Theorem states that

(6) $$Sup \left| F_n(x) - F(x) \right| \rightarrow 0 \; as \; n \rightarrow \infty.$$

Bounds for the approximation have been established in the past, the latest being that of Massart (1990). The more popular bound, however, is the Dvoretsky-Kiefer-Wolfowitz bound .

The Dvoretzky–Kiefer–Wolfowitz inequality bounds the probability that the random function F_n differs from F by more than a given constant $\varepsilon > 0$ anywhere on the real line. More precisely, there is the one-sided estimate

$$\Pr\left(\sup_{x\in\mathbb{R}} \left(F_n(x) - F(x)\right) > \varepsilon\right) \leq e^{-2n\varepsilon^2} \quad for \; every \; \varepsilon \geq \sqrt{\frac{1}{2n} ln2},$$

and the two-sided estimate

$$\Pr\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2} \quad for \; every \; \varepsilon \geq 0$$

This strengthens the Glivenko–Cantelli theorem by quantifying the rate of convergence as n tends to infinity. It also estimates the tail probability of the Kolmogorov–Smirnov statistic.

The D-K-W inequality provides a convenient way for determining the number of observations n needed to estimate F(x) to any desired degree of accuracy with probability 1- α:

(7) $$n \geq -\frac{1}{2\epsilon^2} log \frac{\alpha}{2}$$

We illustrate the sample size requirements for a 95% confidence with various accuracy levels in Table 1:

Table 1: Sample Size Requirement

Epsilon	alpha1	N
0.01	0.05	18444.4
0.02	0.05	4611.1
0.03	0.05	2049.4
0.04	0.05	1152.8
0.05	0.05	737.8
0.06	0.05	512.3
0.07	0.05	376.4
0.08	0.05	288.2
0.09	0.05	227.7
0.1	0.05	184.4

The sample size needed at a fixed significance level increases as the margin of error decreases. In fact, when it is desired to estimate the parent F(x) by $F_n(x)$ with error .01, the sample size needed is more than 18,000. When the available data are small, say, n < 30, a bootstrap resampling procedure can be undertaken. Through bootstrapping, the number of samples can be increased to any desired number.

3.0 The Proposed Procedure

We formalize the proposed procedure in this section.

Main Theorem: Let $x_1, x_2, ..., x_n$ be iid F(x). We assume that F(x) is absolutely continuous with respect to a Lebesgue measure Let $F_n(x)$ be the empirical distribution function of the random sample. Assume n is large enough so that the maximum difference between the parent distribution and the empirical distribution function is small, say, ε. Let:

(8) $$g(u_1,u_2) = \sqrt{-2lnU_1}cos(2\pi U_2)\,;\; h(u_1,u_2) = \sqrt{-2lnU_1}\,sin(2\pi U_2)$$

where U1 = Fn(x1) and U2 = Fn(x2). Then g(.) and h(.) are approximately independent standard normal random variables.

Proof:

It suffices to prove that g(.) is a standard normal random variable. If $U_1 = F(x_1)$ and $U_2 = F(x_2)$, then by the previous result of Box-Mueller, the result follows. We replace F by its empirical estimate F_n:

$\widehat{U_1} = F_n(x_1)$ and $\widehat{U_2} = F_n(x_2)$.

We measure the difference between $g(\widehat{U_1}, \widehat{U_2})$ and $g(U_1,U_2)$.

Now,

$$|g(U_1,U_2) - g(\widehat{U_1},\widehat{U_2})| = |g(F(X_1),F(X_2)) - g(F_n(X_1),F_n(X_2))|$$

$$= \left|\sqrt{-2lnF(X_1)}cos2\pi F(X_2) - \sqrt{-2lnF_n(X_1)}cos2\pi F_n(X_2)\right|$$

$$= \left|\sqrt{-2lnF(X_1)}cos2\pi F(X_2) - \sqrt{-2lnF(X_1)}cos2\pi F_n(X_2) + \sqrt{-2lnF(X_1)}cos2\pi F_n(X_2) - \sqrt{-2lnF_n(X_1)}cos2\pi F_n(X_2)\right|$$

$$\leq \left|\sqrt{-2lnF(X_1)}(cos2\pi F(X_2) - cos2\pi F_n(X_2))\right| + \left|cos2\pi F_n(X_1)(\sqrt{-2lnF(X_1)} - \sqrt{-2lnF_n(X_1)}\right| \to 0 \quad as\ n \to \infty$$

The last convergence statement follows from the following result in analysis:

Result: If $h_n(x) \to h(x)$ uniformly for all x, then if t(.) is continuous,

$$(h_n(x)) \to t(h(x)) \text{ as } n \to \infty.$$

Proof: Since if $h_n(x) \to h(x)$ uniformly, then $\forall \delta > 0 \quad \exists N > 0$ such that:

$$|h_n(x) \to h(x)| < \delta \qquad \forall\, n \geq N.$$

Let t(.) be a continuous function, then $\forall \delta > 0 \quad \exists N > 0$ such that:

$$|t(a) - t(b)| < \varepsilon \text{ whenever } |a - b| < \delta.$$

Set $a = h_n(x)$ and $b = h(x)$, and the result follows[7].

The result is used in the last statement of the **proof** of the theorem by considering $t_1(\theta) = cos2\pi\theta$ and $t_2(\theta) = \sqrt{-2ln\theta}$ which are continuous on [0, 1]. It follows that g(U_1, U_2) is stochastically close to g($\widehat{U}_1$, $\widehat{U}_2$). Since g(U_1, U_2) is a standard normal variate by the Box-Mueller, it follows that g($\widehat{U}_1$, $\widehat{U}_2$) is approximately normal.■

We now establish the fact that the maximum deviation of the distribution of g($\widehat{U}_1$, $\widehat{U}_2$) from the standard normal distribution φ(x) is bounded by the DKW upper bound. We accomplish this by noting that g(U_1,U_2) is identical to the standard normal distribution while g($\widehat{U}_1$, $\widehat{U}_2$) is quite close to g(U_1,U_2) when the sample size is sufficiently large for the Glivenko-Cantelli theorem to hold i.e. for the uniform convergence of the empirical distribution function to the true distribution function. The norm used in the proof is the infinity norm.

Theorem 3. The maximum deviation of the distribution of g($\widehat{U}_1$, $\widehat{U}_2$) from the standard normal distribution φ(x) is bounded by the DKW upper bound, that is,

$$P\left(sup\left|F_n\left(g(\widehat{U}_1,\widehat{U}_2)\right) - \varphi\left(g(U_1,U_2)\right|\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

Proof:

Let F_n(g($\widehat{U}_1$, $\widehat{U}_2$)) be the empirical distribution function of g($\widehat{U}_1$, $\widehat{U}_2$). Then:

$$\begin{aligned} D &= sup\left|F_n\left(g(\widehat{U}_1,\widehat{U}_2)\right) - \varphi\, g(U_1,U_2)\right| \\ &= sup\left\{\left|F_n\left(g(\widehat{U}_1,\widehat{U}_2)\right) - F(g(U_1,U_2)\right| + \left|\mathrm{F}(\mathrm{g}(U_1,U_2)) - \varphi(\mathrm{g}(U_1,U_2)\right|\right\} \\ &\leq sup\left\{\left|F_n\left(g(\widehat{U}_1,\widehat{U}_2)\right) - F(g(U_1,U_2)\right| + sup\left|\mathrm{F}(\mathrm{g}(U_1,U_2)) - \varphi(\mathrm{g}(U_1,U_2)\right|\right\} \\ &\to 0 \quad \text{as } n \to \infty. \end{aligned}$$

It follows that:

$$P\left(sup\left|F_n\left(g(\widehat{U}_1,\widehat{U}_2)\right) - \varphi\left(g(U_1,U_2)\right|\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

The larger the sample size n is , the better is the approximation of the sampling distribution of the statistic g(.) by a standard normal distribution.■

4.0 Simulation Results

We wish to compare the proposed procedure with the Yeo-Johnson transformation technique using numerical simulations. We simulated 500 observations from the family of beta densities, Gamma densities and Laplace distribution. For each set of observations, we performed both the proposed procedure and the Yeo-Johnson transformation to transform them into normally distributed random numbers. The results of the transformations were compared using the Kolmogorov-Smirnov deviance statistics.

The following distributions were used as base distributions for generating the random observations:

Beta: B(1,2), B(1,3), B(2,1), B(2,2), B(2,3), B(3,1), B(3,2), B(3,3)

Gamma: G(1,2), G(1,3), G(2,1), G(2,2), G(2,3), G(3,1), G(3,2), G(3,3)

Laplace: L(1,2), L(1,3), L(2,1), L(2,2), L(2,3), L(3,1), L(3,2), L(3,3)

In order to implement the proposed procedure, we followed the algorithm below:

Algorithm:

1. Input random data
2. Arrange random data from smallest to highest
3. Assign a weight of 1/n, 2/n, 3/n..., n-1/n, 1 to the smallest , second lowest, third lowest up to the highest data respectively.
4. Put the appropriate weights to the original set of unsorted data

5. Apply Box-Mueller transformation to the weights in step 4.
6. Test the transformed data for normality by the Kolmogorov-Smirnov statistic

Whenever feasible, we apply the Yeo- Johnson transformation in step 4 to the original data set and compute the Kolmogorov-Smirnov statistic for the transformed data by this method. The Yeo-Johnson algorithm may or may not produce the desired transformation, a problem which it shares with the Box-Cox method.

5.0 Results and Discussions

Table 2 shows the summary of the simulation results.

The null hypothesis that the distribution of the transformed data is normal is accepted in all cases for the proposed method. The same observation holds true for the Yeo-Johnson algorithm whenever a transformation is available. We emphasize that the availability of a Yeo-Johnson transformation is dependent on the statistical software used.

Whenever a Yeo-Johnson transformation is available, the computed Kolmogorov statistic or maximum deviation statistic tended to be lower for it than the proposed method . However, the differences observed for the Kolmogorov statistical distances between the proposed method and Yeo-Johnson method are very small indeed showing that the two methods provide equally reliable results. In this sense, the proposed method provides a sensible alternative to the existing data transformation algorithms.

The main advantage of the proposed method over the Yeo-Johnson algorithm (and the Box-Cox

Table 2: Comparison of the Yeo-Johhnson Algorithm and the Proposed Algorithm

Distribution	Proposed Algorithm		YEO-JOHNSON Algorithm	
	P-value	Kolmogorov-Smirnov	P-value	Kolmogorov-Smirnov
Beta (1,3)	> 0.15	0.021	> 0.15	No transformation
Beta (1,2)	>0.15	0.027	>0.15	No transformation
Beta (2,3)	> 0.15	0.016	> 0.15	No transformation
Beta (2,2)	>0.15	0.013	>0.15	No transformation
Beta (2,1)	> 0.15	0.030	> 0.15	No transformation
Beta (3,1)	>0.15	0.020	>0.15	No transformation
Beta (3,2)	> 0.15	0.017	> 0.15	0.019
Beta (3,3)	>0.15	0.026	>0.15	0.024
Gamma (1,3)	>0.15	0.021	> 0.15	0.017
Gamma (1,2)	>0.15	0.028	>0.15	0.016
Gamma (2,3)	> 0.15	0.031	> 0.15	No transformation
Gamma (2,2)	>0.15	0.021	>0.15	No transformation
Gamma (2,1)	> 0.15	0.022	> 0.15	No transformation
Gamma (3,1)	>0.15	0.020	>0.15	No transformation
Gamma (3,2)	> 0.15	0.024	> 0.15	No transformation
Gamma (3,3)	>0.15	0.028	>0.15	No transformation
Laplace(1,3)	> 0.15	0.029	> 0.15	0.020
Laplace (1,2)	>0.15	0.032	>0.15	0.029
Laplace (2,3)	> 0.15	0.026	> 0.15	0.025
Laplace (2,2)	>0.15	0.032	>0.15	0.031
Laplace (2,1)	> 0.15	0.026	> 0.15	0.025
Laplace (3,1)	>0.15	0.019	>0.15	0.015
Laplace (3,2)	> 0.15	0.023	> 0.15	0.031
Laplace (3,3)	>0.15	0.029	>0.15	0.033

method) is that transformations to normality are always possible for the proposed method while the same may not be available for the Yeo-Johnson algorithm. The disadvantage, however, is the fact that the proposed method requires a large number of observations (n > 100) for it to work efficiently. To remedy this limitation, we suggest the use of a bootstrap re-sampling procedure to increase the sample size. The proposed method can easily be coded and incorporated in statistical software packages.

6.0 Conclusions and Recommendations

We have introduced a new method for transforming any set of random observations to normality via the empirical distribution function $F_n(x)$ and the Box-Mueller transformation. The theoretical and statistical properties of the proposed method were discussed. In particular, we showed that the a bound for the probability of a Kolmogorov-Smirnov type statistic is identical to the Dvoretzky-Kiefer-Wolfowitz two-sided bound. The proposed method compares very well with the Yeo-Johnson technique (a generalization of the popular Box-Cox transformation technique) and has the added advantage of being able to transform any set of data to normality which is not always the case for the Yeo-Johnson algorithm. Moreover, the proposed method can be easily incorporated in available statistical softwares.

ACKNOWLEDGMENT

The authors wish to thank the anonymous referees for providing helpful suggestions and comments to improve the paper. In particular, we acknowledge the suggestion of one of the referees to incorporate the newest bound discovered by Paul Massart (1990) in place of the Dvoretzky-Kiefer-Wolfowitz inequality. This work was completed as a class project in Mathematical Statistics.

References

Cook, R. D. & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*, New York: Wiley.

Craig, W. & Hogg , An Introduction to Mathematical Statistics, (Wiley and Sons, New York, 2000)

Dudley, R. M. (1999). "Uniform Central Limit Theorems", Cambridge University Press. ISBN 0 521 46102.

Durrett, R. (1991). *Probability: Theory and Examples*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Esseen, C. (1956). "A moment inequality with an application to the central limit theorem". *Skand. Aktuarietidskr.* **39**: 160–170.

Feller, W. (1972). *An Introduction to Probability Theory and Its Applications, Volume II* (2nd ed.). New York: John Wiley & Sons.

Graybill, J. An Introductory Course in Mathematical Statistics (Wiley Series, New York, 1987)

Huber, P. (1985). Projection pursuit. *The annals of Statistics,* 13(2):435 – 475.

Johnson, R & Wichern, Applied Multivariate Statistical Analysis (Wiley and Sons, New York, 2000)

Manoukian, E. B. (1986). *Modern Concepts and Theorems of Mathematical Statistics*. New York: Springer-Verlag.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Shevtsova, I. G. (2007). "Sharpening of the upper bound of the absolute constant in the Berry–Esseen inequality". *Theory of Probability and its Applications* **51** (3): 549–553.

Shevtsova, I. G. (2008). "On the absolute constant in the Berry-Esseen inequality". *The Collection of Papers of Young Scientists of the Faculty of Computational Mathematics and CyberneticsTheory of Probability and its Applications* (5): 101-110.

Shiganov, I.S. (1986). "Refinement of the upper bound of a constant in the remainder term of the central limit theorem". *Journal of Soviet mathematics* **35**: 109–115.

Shorack, G.R., Wellner J.A. (1986) *Empirical Processes with Applications to Statistics*, Wiley.

Tyurin, I.S. (2009). "On the accuracy of the Gaussian approximation". *Doklady Mathematics* **80** (3): 840-843.

A. W. van der Vaart (1998), Asymptotic Statistics. Cambridge Series in Probabilistic Mathematics.

Vapnik, V.N. and Chervonenkis, A. Ya (1971). On uniform convergence of the frequencies of events to their probabilities. Theor. Prob. Appl. 16, 264-280

Yeo, I. & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954-959.