流通科学大学論集―経済・情報・政策編―第 22 巻第 2 号，87-99（2014）

Examination of Transformations to Normality

Takafumi Isogai*

The idea of an R(x) plot in Tarter and Kowalski (1970), which is defined by the ratio of densities with a non-normal distribution and a normal, is developed to examine several problems in normalizing transformation theory. A particular structure of the function R(x) enables us to introduce a new diagnostic method for the existence problem of a transformation independent of the population parameter (see Efron (1982)). Performance and sensitivity of three diagnostic methods, including the Efron's method, are examined by several examples. It is shown that our new method has the most stable performance among three methods.

Keywords: normalizing transformation, R(x) plot, 2nd order differential equation, Runge-Kutta algorithm

Ⅰ. Introduction

To transform non-normal data sets to normality often makes us get more understanding and easier interpretation of the original data sets. Let Z be a random variable distributed as a normal and let X be a random variable distributed as another distribution. Let us consider the relationship Z=T(X), where T denotes the transformation to normality. Basically there are two currents in normalizing transformation theory. One is to construct a transformation family of Z systematically through given T (for example, Box and Cox (1964), Johnson (1949) and Isogai (1999, 2005)) and the other is to find a transformation T for a given distribution of X (for example, Wilson and Hilferty (1931)).

In reference to the second standpoint, Tarter and Kowalski (1970) proposed an R(x) $(=1/\frac{dz}{dx}=1/\frac{dT}{dx})$ plotting method to detect the functional form of a transformation T. Also, Kaskey, Kolman, Steinberg and Krishnaiah (1980) considered the problem of transforming Pearson type distributions to normality by using a 2nd order differential equation, of which solution curve defines a transformation T, and they evaluated the transformation function T numerically by Runge-Kutta algorithm. On the other hand, for one-parameter families Efron (1982) has considered the existence problem of a transformation T independent of the original parameter and also has given a diagnostic tool for the existence of T.

Our objective here is to develop the idea of the R(x) plotting method (see Tarter and Kowalski (1970))

*流通科学大学商学部、〒651-2188　神戸市西区学園西町 3-1

（2013 年 9 月 4 日受理）

©2014 UMDS Research Association

by use of the numerical approach provided by Kaskey et al. (1980), and to discuss several uses of R(x): a detective use and derivation of a diagnostic tool in the meaning of Efron (1982).

Thus the organization of the paper is as follows. In Section 2 we shall briefly review the numerical method provided by Kaskey et al. (1980). In Section 3 the Efron's (1982) diagnostic method is discussed. A particular factorization of the function R(x) derives a new diagnostic tool. It is shown that two diagnostic methods are equivalent, but they have different aspects for diagnostics. A generalization of the diagnostic methods to a multiparameter case is also considered. Finally, in Section 4, performance and sensitivity of our new method are examined and compared to the Efron's method by several examples.

Ⅱ. Transformation and a differential equation

Let $F_\theta(x)$ be a distribution function having the density $f_\theta(x)$ which is positive in some interval [L, U] and zero outside of this interval, and continuously differentiable with respect to x and θ. Suppose that the end points L and U do not depend on θ. Let $\Phi(z)$ be the standard normal distribution function with the density $\phi(z)$. The transformation function z = T(x) is defined by

$$z = T(x) = \Phi^{-1}\left(F_\theta(x)\right) \tag{1}$$

or equivalently defined by

$$\int_{-\infty}^{z} \phi(u)\,du = \int_{L}^{x} f_\theta(t)\,dt\,.$$

Differentiating the equation (1) twice with respect to x, we get the following 2nd order differential equation

$$\ddot{z} = z(\dot{z})^2 + \left\{\frac{\partial}{\partial x}\log f_\theta(x)\right\}\dot{z} \tag{2}$$

where we put

$$\begin{cases} \dfrac{dz}{dx} = \dot{z}, \\ \dfrac{d^2z}{dx^2} = \ddot{z} \end{cases}.$$

The transformation function z = T(x) is given as a solution z = z(x) of the equation (2) with suitable initial conditions. We shall take medians of both distributions $\Phi(z)$ and $F_\theta(x)$ as initial conditions of (2), which are given by

$$\begin{cases} z(x_{0.5}) = 0, \\ \dot{z}(x_{0.5}) = \sqrt{2\pi} f_\theta(x_{0.5}) \end{cases} \tag{3}$$

where 100α $(0<\alpha<1)$ percent points z_α and x_α are defined by

$$\Phi(z_\alpha)=\alpha,$$
$$F_\theta(x_\alpha)=\alpha$$

and the solution z = z(x) of (2) satisfies the relationship:

$$z(x_\alpha)=z_\alpha .$$

Note that the numerical solution z = z(x) is evaluated over the interval $[x_{0.001}, x_{0.999}]$.

Figure 1 shows the numerical solution z = T(x) for F distribution with (3, 6) degrees of freedom. Figure 1 also shows an R(x) plot for the F distribution. Its almost linear configuration $R(x)\approx x$ indicates that a log transformation is appropriate as a normalizing transformation in this case.

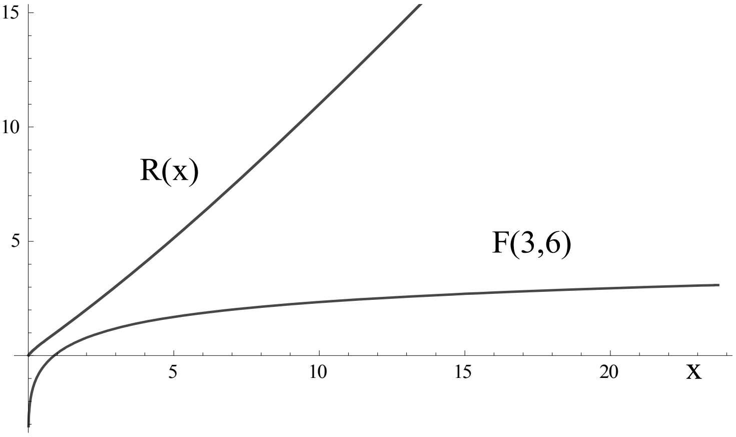

Figure 1. Transformation function z = T(x) and R(x) plot for F(3, 6).

Ⅲ. Diagnostics for the existence of a normalizing transformation

Efron (1982) considered the existence problem of a transformation function z = T(x) which can be expressed as

$$z=z(x,\,\theta)=\frac{g(x)-\nu_\theta}{\sigma_\theta} \tag{4}$$

with some function g(x) independent of the population parameter θ. Here ν_θ is the median of g(x) and σ_θ the standard deviation of g(x).

If such a function g(x) exists, then the relationship (4) can be written as

$$\nu_\theta+\sigma_\theta\, z(x,\theta)=g(x) \tag{5}$$

and after differentiation of both sides with respect to θ, the equation (5) becomes

$$\partial_\theta z = -\frac{\frac{d}{d\theta}\sigma_\theta}{\sigma_\theta} z - \frac{\frac{d}{d\theta}v_\theta}{\sigma_\theta} \tag{6}$$

where we put

$$\frac{\partial z}{\partial \theta} = \partial_\theta z \ .$$

Using the initial conditions (3) to eliminate the constant term in (6), we have

$$\frac{\partial_\theta z}{\partial_\theta z\left(x_{0.5}\right)} = 1 + \frac{\frac{d}{d\theta}\sigma_\theta}{\frac{d}{d\theta}v_\theta} z \ . \tag{7}$$

The Efron's (1982) diagnostic method consists in plotting z versus $\partial_\theta z / \partial_\theta z\left(x_{0.5}\right)$. The linearity of this graph means the existence of a function g(x) independent of θ. Here we should remark that if $\partial_\theta z\left(x_{0.5}\right) = 0$, we cannot use the Efron's method. We have to modify the Efron's method and evaluate

$$\partial_\theta z - \partial_\theta z\left(x_{0.5}\right) = -\frac{\frac{d}{d\theta}\sigma_\theta}{\sigma_\theta} z \ . \tag{8}$$

The linearity of the graph $\left(z,\ \partial_\theta z - \partial_\theta z\left(x_{0.5}\right)\right)$ means the existence of a function g(x) independent of θ.

1. Diagnostics for ż

Tarter and Kowalski (1972) considered the problem of finding a normal transformation function z = T(x) through examination of a functional form of $\dot{z}\ \left(= dz/dx\right)$ (actually they used its reciprocal $R\left(x\right) = 1/\dot{z}$). From their results the functional form of $\dot{z}$ is generally expressed as

$$\dot{z} = h\left(x,\ \theta\right) \tag{9}$$

where a functional form of h is written in terms of polynomials or elementary functions.

Our objective here is to check the structure of dependency of $h\left(x,\ \theta\right)$ on the original parameter θ. Now suppose that $h\left(x,\ \theta\right)$ is divided into two factors α_θ and h(x), and that $\dot{z}$ is expressed as

$$\dot{z} = \alpha_\theta\, h\left(x\right) \quad \alpha_\theta > 0\,,\ \ h\left(x\right) > 0 \tag{10}$$

where α_θ is a function of only θ and h(x) is a function of only x.

If the relationship (10) is true, the transformation function is obtained as

$$z = T\left(x\right) = \alpha_\theta \int_{x_{0.5}}^{x} h\left(u\right) du$$

which is equivalent to the form of (4). Thus we may put $h\left(x\right) = \dot{g}\left(x\right)$ and $\alpha_\theta = 1/\sigma_\theta$ in (10).

To diagnose the existence of such a function h(x), we differentiate both sides of (10) with respect to

θ. Then we have

$$\frac{\partial_\theta \dot{z}}{\dot{z}} = -\frac{\frac{d}{d\theta}\sigma_\theta}{\sigma_\theta}. \tag{11}$$

The constancy of the graph $\left(z, \frac{\partial_\theta \dot{z}}{\dot{z}}\right)$ means the existence of a function h(x).

Furthermore, substitution of (11) into (6) leads to

$$\partial_\theta z - \frac{\partial_\theta \dot{z}}{\dot{z}} z = -\frac{\frac{d}{d\theta}v_\theta}{\sigma_\theta} \tag{12}$$

which gives us a supplementary graph $\left(z, \partial_\theta z - \frac{\partial_\theta \dot{z}}{\dot{z}} z\right)$. This plotting should be constant under our assumption.

After all, our method is related to that of Efron (1982). While the Efron's method examines the linearity between z and $\partial_\theta z$ in the relationship (6), our approach focuses on the examination of the slope coefficient as well as the intercept term in the equation (6).

2. Multiparameter case

Even when the population parameter θ of $F_\theta(x)$ is a p-dimensional vector $\boldsymbol{\theta}' = (\theta_1, \theta_2, \cdots, \theta_p)$, the preceding discussion goes through straightforwardly.

The Efron's method is to plot

$$\left(z, \frac{\partial_i z}{\partial_i z(x_{0.5})}\right),\ i = 1, \cdots, p$$

where we put

$$\frac{\partial z}{\partial \theta_i} = \partial_i z,\ i = 1, \cdots, p.$$

Similarly, the modified Efron's method is to plot

$$(z, \partial_i z - \partial_i z(x_{0.5})),\ i = 1, \cdots, p.$$

The linearity of all configurations in each graphical display means the existence of a function g(x) independent of the multiparameter $\boldsymbol{\theta}' = (\theta_1, \theta_2, \cdots, \theta_p)$.

Our new diagnostic method is to plot

$$\left(z, \frac{\partial_i \dot{z}}{\dot{z}}\right),\ i = 1, \cdots, p$$

and the supplementary diagnostic method is to plot

$$\left(z,\ \partial_i z - \frac{\partial_i \dot{z}}{\dot{z}} z \right),\ \ i = 1, \cdots, p\ .$$

The constancy of all configurations means the existence of a function h(x) independent of the multiparameter $\boldsymbol{\theta}$, which suggests indirectly the existence of a function g(x) independent of $\boldsymbol{\theta}$.

3. Another differential equation

To use diagnostic methods mentioned above, we need the numerical calculation of $\partial z / \partial \theta$ and $\partial \dot{z} / \partial \theta$. These terms satisfy the following differential equation.

The differentiation of both sides of the equation (2) with respect to θ leads to

$$\frac{\partial \ddot{z}}{\partial \theta} = \frac{\partial z}{\partial \theta} \left(\dot{z} \right)^2 + 2\, z\, \dot{z} \frac{\partial \dot{z}}{\partial \theta} + \left(\frac{\partial}{\partial x} \log f_\theta \left(x \right) \right) \frac{\partial \dot{z}}{\partial \theta} + \left(\frac{\partial}{\partial \theta} \frac{\partial}{\partial x} \log f_\theta \left(x \right) \right) \dot{z}\ . \tag{13}$$

Our required solution $\left(z,\ \partial z / \partial \theta \right)$ is obtained as the solution of the simultaneous differential equation composed of (2) and (13) with suitable initial conditions. In addition to (3), the initial conditions we need for the equation (13) are

$$\partial_\theta z\left(x_{0.5} \right) = \sqrt{2\pi} \int_L^{x_{0.5}} \partial_\theta f_\theta \left(x \right) dx \ \text{ and } \ \partial_\theta \dot{z}\left(x_{0.5} \right) = \sqrt{2\pi}\, \partial_\theta f_\theta \left(x_{0.5} \right). \tag{14}$$

For a multiparameter case we have only to replace the differential operator $\partial \theta \ \left(= \partial / \partial \theta \right)$ in (13) and (14) by $\partial_i \ \left(= \partial / \partial \theta_i \right), i = 1, \cdots, p$.

Ⅳ. Examples

In this section we shall examine performance of three diagnostic methods, that is, the Efron's method, the modified Efron's method and our new method through concrete statistical models.

1. Log-normal distribution

First we shall consider a typical case, where there exists a transformation function g(x) independent of population parameters. One of such distributions is the Log-normal distribution $LN\left(\mu, \sigma^2 \right)$. Put $\boldsymbol{\theta}' = \left(\mu,\ \sigma \right)$. The distribution function and its density of $LN\left(\mu, \sigma^2 \right)$ is respectively given by

$$F_{\boldsymbol{\theta}}(x) = \Phi\left(\frac{\log x - \mu}{\sigma} \right) \ \text{ and } f_{\boldsymbol{\theta}}\left(x \right) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{ -\frac{\left(\log x - \mu \right)}{2\sigma^2} \right\}.$$

Figure 2 gives an R(x) plot and a solution curve z = T(x) for $\mu = 0$ and $\sigma = 1$. Clearly, R(x) is nearly equal to x.

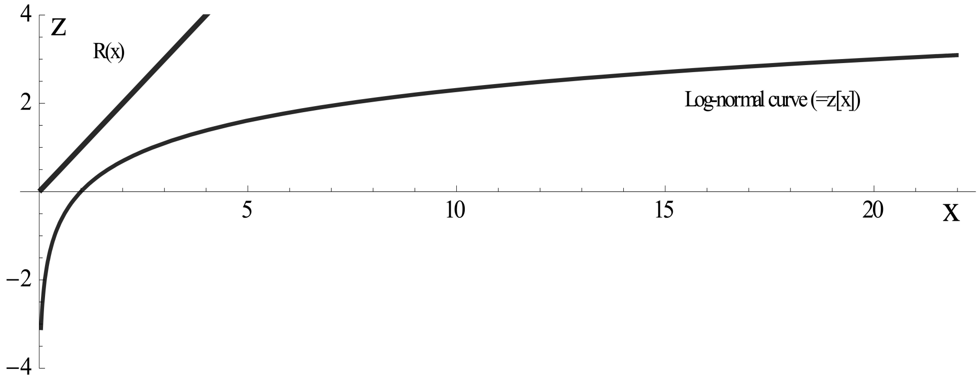

Figure 2. R(x) plot and the solution curve z=T(x) for $LN\left(\mu,\sigma^2\right)$ with $\mu=0$ and $\sigma=1$.

Figure 3 shows diagnostic plots of the Efron's and our new methods with respect to μ. Constancy of both curves means that there exists a transformation function g(x) independent of μ. In Figure 3 the curve of our modified Efron's method coincides with that of our new method.

Furthermore, Figure 4 gives diagnostic plots of the modified Efron's and our new methods with respect to σ. In this case we cannot use the Efron's method, because $\partial_\sigma z\left(x_{0.5}\right)=0$ for any μ and σ of $LN\left(\mu,\sigma^2\right)$. Linearity of the modified Efron's plot and constancy of our new diagnostic plot mean that there exists a transformation function g(x) independent of σ. After all, there exists a transformation function g(x) = log(x) independent of μ and σ.

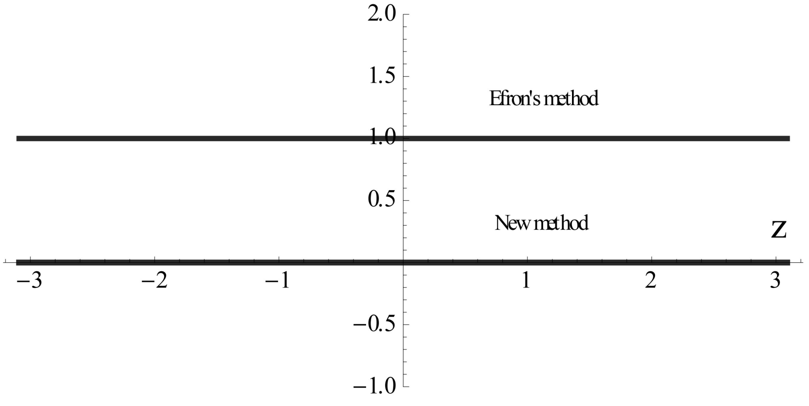

Figure 3. Diagnostic plots of Efron's and our new methods with respect to μ of $LN\left(\mu,\sigma^2\right)$ in case of $\mu=0$ and $\sigma=1$

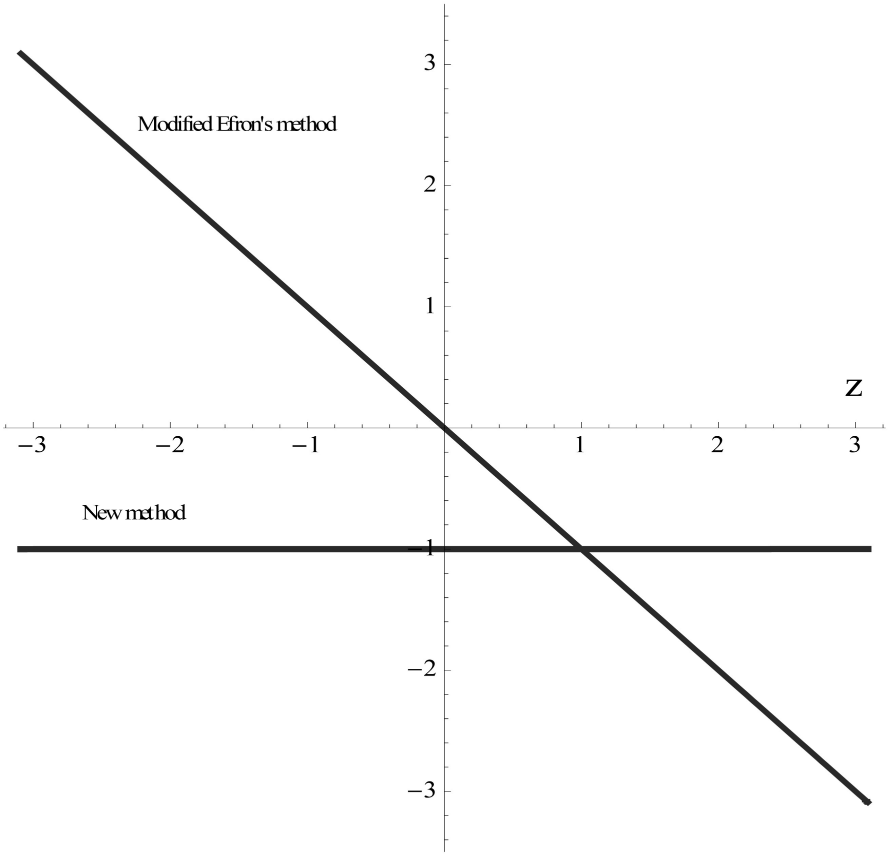

Figure 4. Diagnostic plots of modified Efron’s and our new methods with respect to σ of $LN(\mu,\sigma^2)$ in case of $\mu=0$ and $\sigma=1$.

2. Student’s t distribution

Next we shall consider a typical symmetric non-normal distribution, that is, Student’s t distribution denoted by $t(\theta)$, where $\theta\ (>0)$ means the degrees of freedom. The density of $t(\theta)$ is given by

$$f_\theta(x)=\frac{1}{\sqrt{\theta\pi}}\frac{\Gamma\left(\dfrac{\theta+1}{2}\right)}{\Gamma\left(\dfrac{\theta}{2}\right)}\left(1+\frac{x^2}{\theta}\right)^{-\frac{\theta+1}{2}}.$$

Figure 5 displays an R(x) plot and a solution curve z = T(x) of $t(4)$. R(x) may be approximated by some concave function. Here we do not go further into this subject.

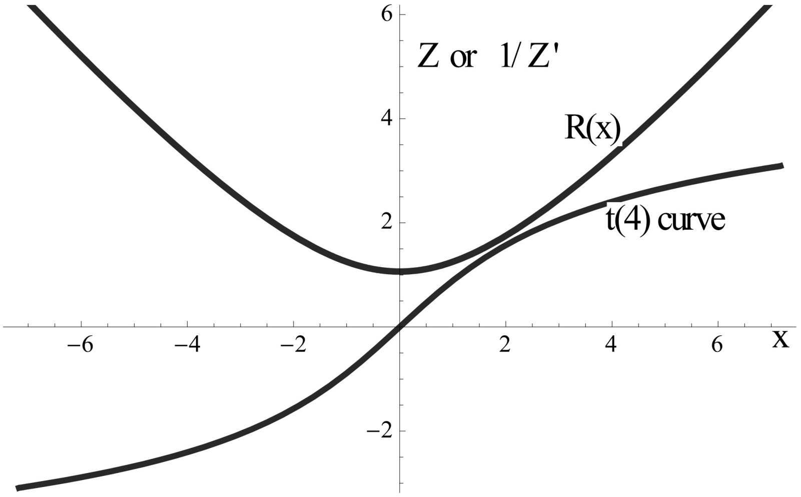

Figure 5. R(x) plot and the solution curve z = T(x) for $t(4)$.

We cannot use the Efron's method, because $\partial_\theta z\left(x_{0.5}\right)=0$ for any θ of $t(\theta)$. Figure 6 shows three plots of the modified Efron's method for $t(1)$, $t(4)$ and $t(10)$. Even when θ becomes large, it is difficult for us to discriminate the degree of linearity among three plots. There is similar tendency in all plots.

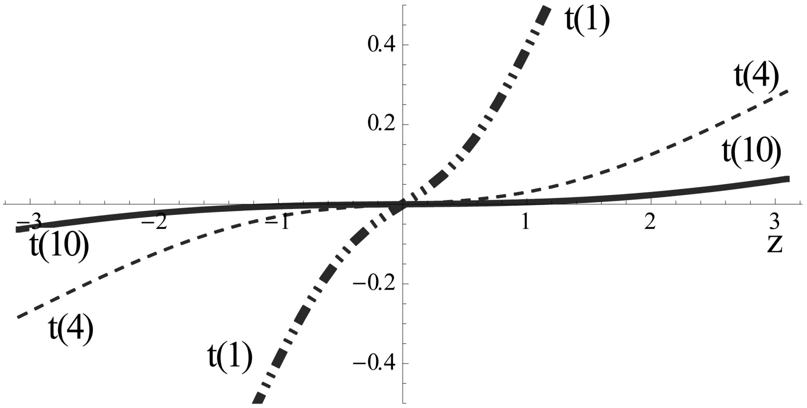

Figure 6. The modified Efron's plots for $t(1)$, $t(4)$ and $t(10)$.

Figure 7 displays three plots of our new method for $t(1)$, $t(4)$ and $t(10)$. Clearly the degree of constancy increases as θ becomes large. It seems that our new method works well.

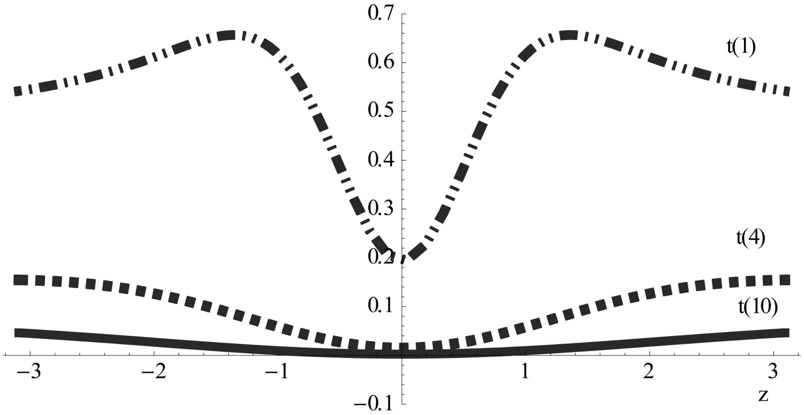

Figure 7. Plots of our new method for $t(1)$, $t(4)$ and $t(10)$.

3. Distribution of sample correlation coefficient

For a random sample of size n from a bivariate normal distribution, the density of the sample correlation coefficient $\hat{\rho}$ (= X, say) is given by

$$f_\theta(x) = \frac{(n-2)\left(1-\theta^2\right)^{(n-1)/2}\left(1-x^2\right)^{(n-4)/2}}{\sqrt{2}\,(n-1)B\left(\frac{1}{2},\,n-\frac{1}{2}\right)(1-\theta x)^{n-3/2}}\;{}_2F_1\left(\frac{1}{2},\frac{1}{2},n-\frac{1}{2};\frac{1+\theta x}{2}\right),$$

where we put $\theta = \rho$ (the population correlation coefficient, $-1 < \rho < 1$) and ${}_2F_1(\cdots)$ is the Gauss hypergeometric function (see Johnson, Kotz and Balakrishnan (1995), p.549).

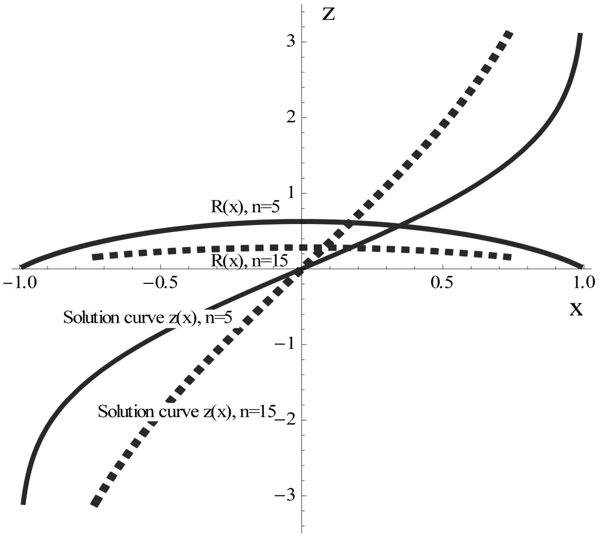

Figure 8. R(x) plots and solution curves z = z(x) in two cases of n=5 (solid line) and n=15 (dashed line) with $\theta = 0\;(\rho = 0)$ for the sample correlation coefficient.

Figure 8 shows R(x) plots and solution curves z = z(x) for two cases of sample sizes n=5 and n=15 with the population parameter $\theta = 0\left(\rho = 0\right)$. Clearly convexity of R(x) plot and an anti-S-shaped configuration of the solution curve z = z(x) for n=5 indicate large deviation from normality, but constancy of R(x) plot and linearity of solution curve z = z(x) for n=15 suggest that the distribution of X (= $\hat{\rho}$) is close to normal.

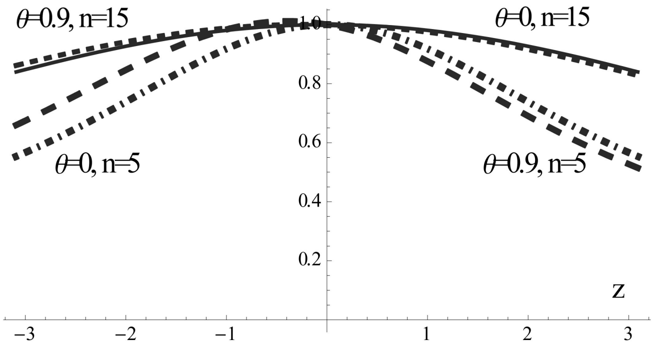

Figure 9. Plots of the Efron's method in four cases: (1) $\theta = 0$, n=15 (solid line), (2) $\theta = 0.9$, n=15 (short dashed line), (3) $\theta = 0$, n=5 (dashed-dotted line), (4) $\theta = 0.9$, n=5 (long dashed line) for the sample correlation coefficient.

Figure 9 displays plots of the Efron's method in cases of $\theta = 0$ and $\theta = 0.9$ for n=5 and n=15. Strange behaviors of plots appear in Figure 9. Configurations of plots in cases of $\theta = 0$ and $\theta = 0.9$ for n=15 are overlapping. The case $\theta = 0$ (n=15) is close to normal, but the other case $\theta = 0.9$ (n=15) is extremely negatively-skewed non-normal one. The Efron's method cannot distinguish these cases. Performance and sensitivity of the Efron's method may decrease as sample sizes become large.

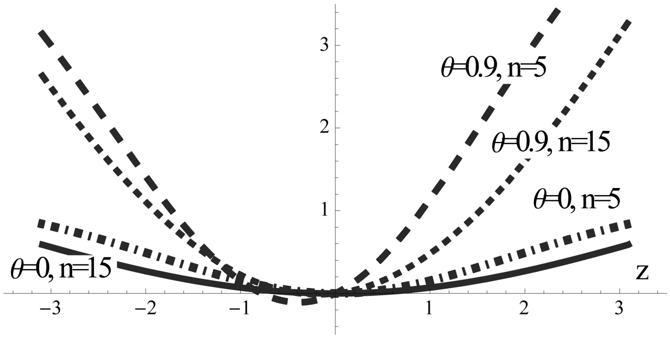

Figure 10. Plots of the modified Efron's method in four cases: (1) $\theta = 0$, n=15 (solid line), (2) $\theta = 0.9$, n=15 (short dashed line), (3) $\theta = 0$, n=5 (dashed-dotted line), (4) $\theta = 0.9$, n=5 (long dashed line) for the sample correlation coefficient.

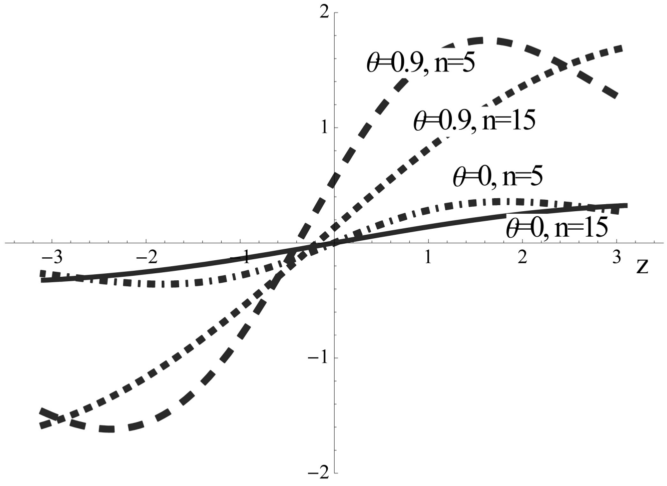

Figure 11. Plots of our new method in four cases: (1) $\theta = 0$, n=15 (solid line), (2) $\theta = 0.9$, n=15 (short dashed line), (3) $\theta = 0$, n=5 (dashed-dotted line), (4) $\theta = 0.9$, n=5 (long dashed line) for the sample correlation coefficient.

Figures 10 and 11 give plots of the modified Efron's and our new methods in cases of $\theta = 0$ and $\theta = 0.9$ for n=5 and n=15. Concerning those methods, sensitivity for non-normality increases as $|\theta|$ becomes large, or as sample sizes decrease. The modified Efron's and our new methods have good performance.

V. Conclusions

In the present paper we have examined and compared performance and sensitivity of the Efron's, the modified Efron's and our new methods by using some of typical non-normal distributions. Our conclusions are as follows:

(1) The Efron's method is sometimes unavailable due to $\partial_\theta\, z\left(x_{0.5}\right) = 0$, which leads us to introduce the modified Efron's method.

(2) Performance and sensitivity of the Efron's method are less effective than the modified Efron's and our new methods.

(3) It is difficult for us to evaluate the degree of linearity in using the Efron's and modified Efron's methods.

(4) Performance and sensitivity of our new method are the most stable among three methods.

Acknowledgement

The author would like to express his great thanks to Grant-in-Aid for Scientific Research (C) No.24500266 by JSPS.

References

[1] Box, G. E. P. and Cox, D. R. (1964), "*An analysis of transformation*", Journal of Royal Statistical Society, Ser. B26, pp. 211-252.

[2] Efron, B. (1982), "*Transformation Theory: How normal is a family of distributions*", Annals of Statistics, Vol.10, pp. 323-339.

[3] Johnson, N. L. (1949), "*Systems of frequency curves generated by methods of translation*", Biometrika, Vol.36, pp. 149-176.

[4] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995), "*Continuous Univariate Distributions*", Vol.2, 2nd Edition, Wiley-Interscience.

[5] Isogai, T. (1999), "*Power transformation of the F distribution and a power normal family*", Journal of Applied Statistics, Vol.26, pp.351-367.

[6] Isogai, T. (2005), "*Applications of a New Power Normal Family*", Journal of Applied Statistics, Vol.32, No. 4, pp.421-436.

[7] Kaskay, G., Kolman, B., Krishnaiah, P. R. and Steinberg, L. (1980), "*Transformations to normality*", Handbook of Statistics, Vol.1, North-Holland, pp.321-341.

[8] Tarter, M. E. and Kowalski, C. J. (1972), "*A new test for and class of transformations to normality*", Technometrics, Vol.14, pp.735-744.

[9] Wilson, E. B. and Hilferty, M. M. (1931), "*The distributions of chi-square*", Proceedings of the National Academy of Sciences, Vol.17, pp.684-688.