# Transformations with Right Skew Data

Lakhana Watthanacheewakul

***Abstract*—The basic assumption of statistical analysis is that the data have normally distributed. When analyzing data do not match the assumptions of the conventional method of analysis, there are two choices; transform the data to fit the assumptions or develop some new robust methods of analysis. If a satisfied transformation can find, it will almost always be easier to use it rather than to develop a new method of analysis. The well-known Box-Cox transformations often used in previous studies. However, they are not always applicable, they should be used with caution in some cases such as failure time and survival data. Because the some observations in the sets of right skew data may be zero, the Box-Cox transformations are not appropriate. In this paper, the several transformations are investigated for some sets of right skew data. They performs few better than the Box-Cox transformations in sense of normality and homogeneity of variances for several groups of data in some situations.**

***Index Terms*—transformations, Box-Cox transformations, normality, right skew data**

## I. Introduction

IN some parametric test, the basic assumption is that the data are normally distributed or the sample size is large. If the data do not correspond with it, then the nonparametric test are chosen to analyse data. However, the power of nonparametric test is usually less than parametric test. Tukey [1] suggested that when analyzing data that do not match the assumptions of a conventional method of analysis, there are two choices; transform the data to fit the assumptions or develop some new robust methods of analysis. Montgomery [2] suggested that transformations are used for three purposes; stabilizing response variance, making the distribution of the response variable closer to a normal distribution and improving the fit of the model to the data. Choosing an appropriate transformation depends on the probability distribution of the sample data. Moreover, the relationship between the standard deviation and the mean can use for stabilizing variance. Furthermore, it is possible to transform the data using a family of transformations already extensively studied over a long period of time, e.g. Box and Cox [3], Manly [4], and John and Draper [5] . A well-known family of transformations often used in previous studies was proposed by Box and Cox. Doksum and Wong [6] indicated that the Box-Cox transformation should be used with caution in some cases such as failure time and survival data.

Manuscript received February22, 2012; revised March31, 2012.

L. Watthanacheewakul is with the Faculty of Science, Maejo University, Chiang Mai, Thailand (phone: 66-53-873-551; fax: 66-53-878-225; e-mail: lakhana@mju.ac.th).

## II. A Family of Transformations

Let $X$ be a random variable distributed as non-normal, $Y$ the transformed variable of $X$, $x$ the value of $X$, $c$ the range of data and $\lambda$ a transformation parameter.

Box and Cox [3] gave a simple modified form of the power transformation to avoid discontinuity at $\lambda = 0$. They considered

$$Y = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln X, & \lambda = 0 \end{cases} \quad \text{for} \quad x > 0. \tag{1}$$

This has become well known as the Box-Cox transformation.

Manly [4] suggested a one parameter family of exponential transformations

$$Y = \begin{cases} \dfrac{\exp(\lambda X) - 1}{\lambda}, & \lambda \neq 0 \\ X, & \lambda = 0. \end{cases} \tag{2}$$

This is a useful alternative to Box-Cox transformations because negative $x$ values are also allowed. It has been found in particular that this transformation is quite effective at turning skew unimodal distributions into nearly symmetric normal distributions.

The Modified Box and Cox transformation for any sets of right skew data to normality with constant variance proposed here is in this form

$$Y = \begin{cases} \dfrac{[X + c]^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln[X + c], & \lambda = 0. \end{cases} \tag{3}$$

In this paper, the three transformations were investigated in sense of normality and homogeneity of variances.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

## III. Right Skew Data

Exponential and Weibull data were investigated. They have right skew distributed. The Weibull distribution is a continuous probability distribution. It is named after Waloddi Weibull who described it in detail in 1951. The probability density function of a two parameter Weibull random variable *X* is

$$f(x)=\begin{cases}\dfrac{\alpha}{\beta}\left(\dfrac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & x\geq 0;\alpha,\beta>0,\\ 0, & x<0\end{cases} \tag{4}$$

where α is the shape parameter and β is the scale parameter. It's useful in many fields such as survival analysis, extreme value theory, weather forecasting, reliability engineering and failure analysis. Moreover, it is used to describe wind speed distribution, the particle size distribution, and so on. It is related to the other probability distribution such as the Exponential distribution when α=1. The probability density function of one parameter Exponential random variable *X* is

$$f(x)=\begin{cases}\dfrac{1}{\beta}e^{-\left(\frac{x}{\beta}\right)}, & x\geq 0;\beta>0,\\ 0, & x<0\end{cases} \tag{5}$$

where β is the scale parameter. Both distributions are very right long tailed [7].

## IV. Estimation of the Transformation Parameter

For several groups of data, the value of $\lambda$ in (1) , (2) and (3) need to be found so that the transformed variables will be independently normal distribution with homogeneity of variances. The probability density function of each $Y_{ij}$ is in the form

$$f(y_{ij}|\mu_i,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\mu_i)^2\right\}, \tag{6}$$

where $\mu_i$ is the mean of the i*th* transformed population data, $\sigma^2$ the pooled variance of all transformed population data and $y_{ij}$ the observed value of $Y_{ij}$. For (1), the likelihood function in relation to the observations $x_{ij}$ is given by

$$L(\mu_i,\sigma^2,\lambda|x_{ij})=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\frac{x_{ij}^{\lambda}-1}{\lambda}-\mu_i\right]^2\right\}.J(y;x) \tag{7}$$

where $J(y;x)=\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left|\frac{\partial y_{ij}}{\partial x_{ij}}\right|$. For a fixed $\lambda$, the MLE's for $\mu_i$ and $\sigma^2$ are

$$\hat{\mu}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\left[\frac{x_{ij}^{\lambda}-1}{\lambda}\right] \text{ and}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\frac{x_{ij}^{\lambda}-1}{\lambda}-\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{x_{ij}^{\lambda}-1}{\lambda}\right)\right\}^2$$

Substitute $\hat{\mu}_i$ and $\hat{\sigma}^2$ into the likelihood equation (7). Thus for fixed $\lambda$, the maximized log likelihood is

$$\begin{aligned}&\ln L(\lambda|x_{ij})=\\ &-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\frac{x_{ij}^{\lambda}-1}{\lambda}-\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{x_{ij}^{\lambda}-1}{\lambda}\right)\right\}^2-\frac{n}{2}\\ &+(\lambda-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln x_{ij},\end{aligned} \tag{8}$$

except for a constant, the maximum likelihood estimate of $\lambda$ is obtained by solving the likelihood equation

$$\begin{aligned}&\frac{d}{d\lambda}\ln L(\lambda)=\\ &\frac{-n\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^{2\lambda}\ln x_{ij}-\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}x_{ij}^{\lambda}\right)\left(\sum_{j=1}^{n_i}x_{ij}^{\lambda}\ln x_{ij}\right)\right]}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^{2\lambda}-\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}x_{ij}^{\lambda}\right)^2}\\ &+\frac{n}{\lambda}+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln x_{ij}=0.\end{aligned} \tag{9}$$

Similar procedures yield the same results for (2), the maximum likelihood estimate of $\lambda$ is obtained by solving the likelihood equation

$$\begin{aligned}&\frac{d}{d\lambda}\ln L(\lambda)=\\ &\frac{-n\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}e^{2\lambda x_{ij}}x_{ij}-\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}e^{\lambda x_{ij}}\right)\left(\sum_{j=1}^{n_i}e^{\lambda x_{ij}}x_{ij}\right)\right]}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}e^{2\lambda x_{ij}}-\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}e^{\lambda x_{ij}}\right)^2}\\ &+\frac{n}{\lambda}+\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}=0.\end{aligned} \tag{10}$$

Similar procedures yield the same results for (3), the maximum likelihood estimate of $\lambda$ is obtained by solving the likelihood equation

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\frac{d}{d\lambda}\ln L(\lambda) = -n\left[\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}+c_i\right)^{2\lambda}\ln\left(x_{ij}+c_i\right) - \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}\left(x_{ij}+c_i\right)^{\lambda}\right)\left(\sum_{j=1}^{n_i}x_{ij}^{\lambda}\ln\left(x_{ij}+c_i\right)\right)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}+c_i\right)^{2\lambda} - \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}\left(x_{ij}+c_i\right)^{\lambda}\right)^2}\right] + \frac{n}{\lambda} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln\left(x_{ij}+c_i\right) = 0. \quad (11)$$

Since $\lambda$ appears on the exponent of the observations, it is considered to be too complicated for solving it. The maximized log likelihood function is a unimodal function so the value of the transformation parameter is obtained when the slope of the curvature of the maximized log likelihood function is nearly zero [3]. Hence we can also use the numerical method such as bisection for finding the suitable value of $\lambda$ .

## V. Simulation Study

In order to attain the most effective use of the three transformations, we set the values of parameters and the significant value as follows: $k$ = number of the populations = 3, $n_i$ = sample size from the $i$th population =10, 20, 30, $\beta_i$ = scale parameter of the $i$th Weibull and Exponential population = 5, 10, $\alpha_i$ = shape parameters of the $i$ th Weibull population = 1.2 ,1.5, significant level = 0.05. As a numerical study, Weibull populations of size $N_i = 4{,}000$ $(i = 1,2,3)$ are generated for different values of parameters $\beta_i$ , $\alpha_i$ . Then 1,000 random samples, each of size $n_i$ , are drawn. Then we transform each set of the sample data to normality by the Box-Cox transformation, Manly transformation and the Modified Box and Cox transformation. The results of the goodness- of-fit tests and the tests of homogeneity of variances with 1,000 replicated samples of various sizes are shown in Table I and Table II. Similarly, for the Exponential population, the results are shown in Table III and Table IV.

From Table I and Table II, we see that the results from all of three transformations in each situation are small different.

Similarly**,** from Table III and Table IV, we see that the results from all of three transformations in each situation are small different.

TABLE I
Averages of the p-Values for K-S Test of Normality, and of the p-Values for the Levene Test Using Data Transformed by the Three Transformations with Weibull Data when $\alpha_i = 1.2$ , $\beta_i = 5$

| Transformations | $n_i$ | Averages of the p-Values for K-S Test of Transformed Data | | | Averages of the p-Values for the Levene Test |
|---|---|---|---|---|---|
| Box-Cox | 10 | 0.819 | 0.829 | 0.751 | 0.362 |
| Manly | 10 | 0.828 | 0.830 | 0.763 | 0.351 |
| Modified | 10 | 0.823 | 0.829 | 0.758 | 0.362 |
| Box-Cox | 30 | 0.621 | 0.662 | 0.634 | 0.182 |
| Manly | 30 | 0.722 | 0.729 | 0.683 | 0.192 |
| Modified | 30 | 0.665 | 0.706 | 0.671 | 0.187 |
| Box-Cox | 10,20,30 | 0.788 | 0.745 | 0.787 | 0.229 |
| Manly | 10,20,30 | 0.782 | 0.740 | 0.802 | 0.228 |
| Modified | 10,20,30 | 0.787 | 0.743 | 0.800 | 0.229 |

TABLE II
Averages of the p-Values for K-S Test of Normality, and of the p-Values for the Levene Test Using Data Transformed by the Three Transformations with Weibull Data when $\alpha_i = 1.5$ , $\beta_i = 10$

| Transformations | $n_i$ | Averages of the p-Values for K-S Test of Transformed Data | | | Averages of the p-Values for the Levene Test |
|---|---|---|---|---|---|
| Box-Cox | 10 | 0.746 | 0.812 | 0.710 | 0.210 |
| Manly | 10 | 0.757 | 0.810 | 0.680 | 0.195 |
| Modified | 10 | 0.749 | 0.814 | 0.710 | 0.209 |
| Box-Cox | 30 | 0.699 | 0.720 | 0.743 | 0.149 |
| Manly | 30 | 0.771 | 0.697 | 0.726 | 0.153 |
| Modified | 30 | 0.780 | 0.726 | 0.787 | 0.152 |
| Box-Cox | 10,20,30 | 0.760 | 0.697 | 0.677 | 0.256 |
| Manly | 10,20,30 | 0.759 | 0.692 | 0.715 | 0.268 |
| Modified | 10,20,30 | 0.762 | 0.702 | 0.712 | 0.258 |

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE III
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY, AND OF THE P-VALUES FOR THE LEVENE TEST USING DATA TRANSFORMED BY THE THREE TRANSFORMATIONS WITH EXPONENTIAL DATA WHEN $\beta_i = 5$

| Transformations | $n_i$ | Averages of the p-Values for K-S Test of Transformed Data | | | Averages of the p-Values for the Levene Test |
|---|---|---|---|---|---|
| Box-Cox | 10 | 0.753 | 0.753 | 0.759 | 0.228 |
| Manly | 10 | 0.704 | 0.691 | 0.703 | 0.238 |
| Modified | 10 | 0.755 | 0.745 | 0.758 | 0.236 |
| Box-Cox | 30 | 0.661 | 0.678 | 0.673 | 0.190 |
| Manly | 30 | 0.587 | 0.584 | 0.576 | 0.219 |
| Modified | 30 | 0.677 | 0.735 | 0.747 | 0.193 |
| Box-Cox | 10,20,30 | 0.736 | 0.717 | 0.684 | 0.205 |
| Manly | 10,20,30 | 0.674 | 0.565 | 0.564 | 0.227 |
| Modified | 10,20,30 | 0.737 | 0.684 | 0.745 | 0.211 |

TABLE IV
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY, AND OF THE P-VALUES FOR THE LEVENE TEST USING DATA TRANSFORMED BY THE THREE TRANSFORMATIONS WITH EXPONENTIAL DATA WHEN $\beta_i = 10$

| Transformations | $n_i$ | Averages of the p-Values for K-S Test of Transformed Data | | | Averages of the p-Values for the Levene Test |
|---|---|---|---|---|---|
| Box-Cox | 10 | 0.758 | 0.751 | 0.757 | 0.230 |
| Manly | 10 | 0.681 | 0.685 | 0.704 | 0.241 |
| Modified | 10 | 0.756 | 0.754 | 0.763 | 0.236 |
| Box-Cox | 30 | 0.685 | 0.672 | 0.685 | 0.206 |
| Manly | 30 | 0.584 | 0.619 | 0.563 | 0.219 |
| Modified | 30 | 0.736 | 0.718 | 0.736 | 0.210 |
| Box-Cox | 10,20,30 | 0.750 | 0.719 | 0.679 | 0.195 |
| Manly | 10,20,30 | 0.699 | 0.660 | 0.574 | 0.211 |
| Modified | 10,20,30 | 0.764 | 0.732 | 0.737 | 0.196 |

## VI. CONCLUSION

In usual situation, all of three transformations can transform the right skew data to correspond with the basic assumptions. However, in the sets of right skew data, the some observations may be zero, then Box-Cox transformations are not appropriate with them.

## REFERENCES

[1] W. Tukey, "On the comparative anatomy of transformations," *Annals of Mathematical Statistics*, vol. 28, no. 3, pp. 525-540, Sep. 1957.
[2] D. C. Montgomery, *Design and Analysis of Experiments*, 5th ed. New York: Wiley, 2001, pp. 590.
[3] G. E. P. Box and D. R. Cox, "An analysis of transformations (with discussion)," *Journal of the Royal Statistical Society, Ser.B.* vol. 26, no. 2, pp.211-252, Apr. 1964.
[4] B. F. J. Manly, "Exponential Data Transformations," *Statistician*. vol. 25, no. 1, pp.37-42, Mar. 1976.
[5] J. A. John and N. R. Draper, "An alternative family of transformations," *Applied Statistics*, vol. 29, no. 2, pp.190-197, 1980.
[6] K. A. Doksum, and C. Wong, "Statistical tests based on transformed data," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 411-417, Jun. 1983.
[7] N. L. Johnson, S. Kotz, and N.Balakrishnan. *Continuous Univariate Distributions*, 2nd ed. vol. 1. New York: Wiley, 1994, ch.10.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)