A New Family of Transformations for Lifetime Data

Lakhana Watthanacheewakul

Abstract— **A family of transformations is the one of several methods to analyze the data that do not correspond with the assumption. A well-known family of transformations often used in many studies was proposed by Box and Cox. However, Box-Cox transformation is not always applicable. It should be used with caution in some cases such as failure time and survival data. The simple case, some observations in the set of failure time data may be zero but the value of observation in the condition of Box-Cox transformation is greater than zero. In this case, Manly transformation may be appropriated than Box-Cox transformation because it was proposed as a family of exponential transformations that negative *x* values are also allowed. In this paper, a new family of transformation is proposed to manage with the problem as mentioned and Manly transformation were compared in the lifetime data those have exponential gamma and weibull distribution. They were investigated for some sets of the lifetime data. It is found that the proposed transformation and Manly transformation have not different efficiency in sense of normality. The proposed transformation performs better than Manly transformation in sense of homogeneity of variances for some data set of weibull distributions and exponential distributions when the sample sizes are large.**

***Index Terms*— Manly transformation, proposed transformation, homogeneity of variances, lifetime data, normality**

I. Introduction

IN statistical data analysis, many statistical procedures require data to be approximately normal. If the data are not normally distributed, a transformation that transforms the data set to achieve normality is used. 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. There are several alternatives for transforming such as transformations based on the relationship between the standard deviation and the mean. 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. John and Draper [5] showed that the Box-Cox transformation was not satisfactory even when the best value of transformation parameter had been chosen.

Manuscript received February 24, 2014; revised March 17, 2014.
This work was supported in part by the Faculty of Science, Maejo University, Chiang Mai, Thailand.
L. Watthanacheewakul is with the Faculty of Science, Maejo University, Chiang Mai, Thailand (phone: 66-53-873-551; fax: 66-53-875-205; e-mail: lakhanaw@yahoo.com; lakhana@mju.ac.th).

II. A Family of Transformations

A family of transformations applied over a long period can be used for data from any population so that the transformed data are normally distributed.

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 set and λ a transformation parameter.

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

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

This has become well known as Box-Cox transformation.

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

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

This is a useful alternative to Box-Cox transformations because negative x values are also allowed. It has been

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

found in particular that this transformation is quite effective at turning skew unimodal distributions into nearly symmetric normal distributions.

Yeo and Johnson [8] proposed a family of modified Box and Cox transformation

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

In this paper, the alternative family of transformations for lifetime data is proposed in this form

$$Y=\begin{cases}\dfrac{\left[\sqrt{X+1}\right]^{\lambda}-1}{\lambda} & ,x\geq 0,\ \lambda\neq 0\\ \ln\left[\sqrt{X+1}\right] & ,x\geq 0,\lambda=0.\end{cases} \tag{4}$$

III. Lifetime Data

Lifetime data are important in reliability analysis and survival analysis. It is often of interest to estimate the reliability of the system/component from the observed lifetime data.

Weibull Exponential and Gamma distributions are involved lifetime data. The Weibull distribution is a natural starting point in the modeling of failure times in reliability, material strength data and many other applications. The probability density function of a two parameter Weibull random variable *X* is

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

where α is the shape parameter and β is the scale parameter. 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

$$f(x)=\begin{cases}\dfrac{1}{\beta}e^{-\left(\frac{x}{\beta}\right)} & ,x\geq 0;\beta>0,\\ 0 & ,x<0\end{cases} \tag{6}$$

where β is the scale parameter.

Gamma distribution is the common choices of frailty distribution in lifetime data models.

$$f(x)=\begin{cases}\dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)} & ,x\geq 0;\beta>0\\ 0 & ,x<0\end{cases} \tag{7}$$

where α is the shape parameter and β is the scale parameter.

IV. Estimation of the Transformation Parameter

For several groups of data, the value of λ in (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

$$f(y_{ij}|\mu_i,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\mu_i)^2\right\}, \tag{8}$$

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

$$L(\mu_i,\sigma^2,\lambda|x_{ij})=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\frac{\exp(\lambda x_{ij})-1}{\lambda}-\mu_i\right]^2\right\}.J(y;x) \tag{9}$$

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 λ, the MLE's for μ_i and σ^2 are

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

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

Substitute $\hat{\mu}_i$ and $\hat{\sigma}^2$ into the likelihood equation (9). Thus for fixed λ, the maximized log likelihood is

$$\ln L(\lambda|x_{ij})=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\frac{\exp(\lambda x_{ij})-1}{\lambda}-\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{\exp(\lambda x_{ij})-1}{\lambda}\right)\right\}^2 -\frac{n}{2}+\lambda\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}, \tag{10}$$

except for a constant, the maximum likelihood estimate of λ is obtained by solving the likelihood equation

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

Similar procedures yield the same results for (4), the maximum likelihood estimate of λ is obtained by solving the likelihood equation

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\frac{d}{d\lambda}\ln L(\lambda) = -n\left[\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\sqrt{x_{ij}+1}\right)^{2\lambda}\ln\left(\sqrt{x_{ij}+1}\right) - \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}\left(\sqrt{x_{ij}+1}\right)^{\lambda}\right)\left(\sum_{j=1}^{n_i}\left(\sqrt{x_{ij}+1}\right)^{\lambda}\ln\left(\sqrt{x_{ij}+1}\right)\right)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\sqrt{x_{ij}+1}\right)^{2\lambda} - \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i}\left(\sqrt{x_{ij}+1}\right)^{\lambda}\right)^2}\right] + \frac{n}{\lambda} + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln\left(\sqrt{x_{ij}+1}\right) = 0. \quad (12)$$

Since λ 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 λ.

V. Simulation Study

In order to attain the most effective use of the two 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 ith population is between 10 and 80, β_i = scale parameter of the ith Weibull Exponential and Gamma populations is between 1 and 3, α_i = shape parameter of the ith Weibull and Gamma population is between 2 and 4, the significant level = 0.05. The graph of Weibull Exponential and Gamma distributions are shown in Figure 1 – 7.

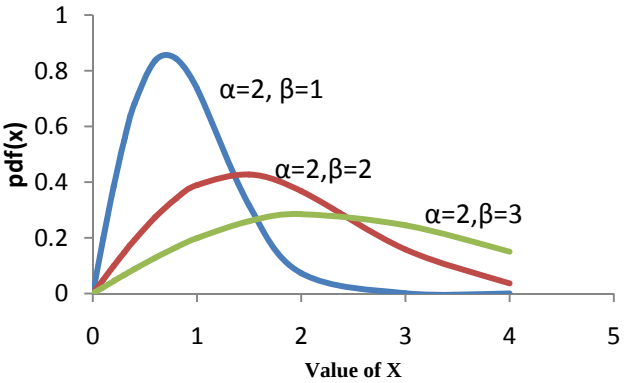

Fig. 1. Graph of Weibull distributions when shape parameters are the same and scale parameters are different.

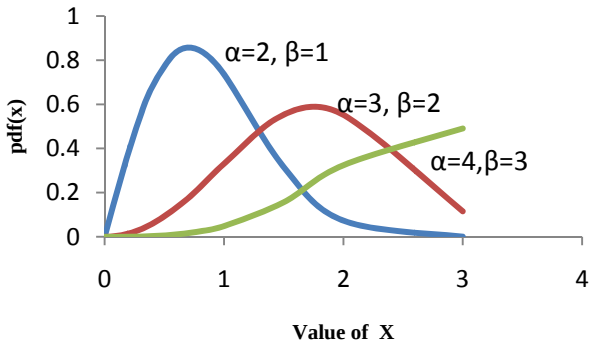

Fig. 2. Graph of Weibull distributions when shape parameters and scale parameters are different.

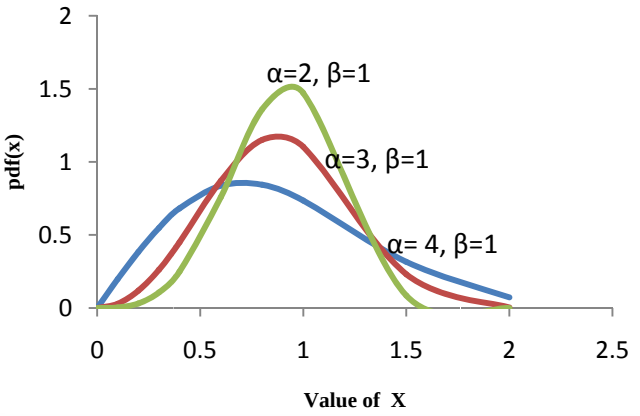

Fig. 3. Graph of Weibull distributions when shape parameters are different and scale parameters are the same.

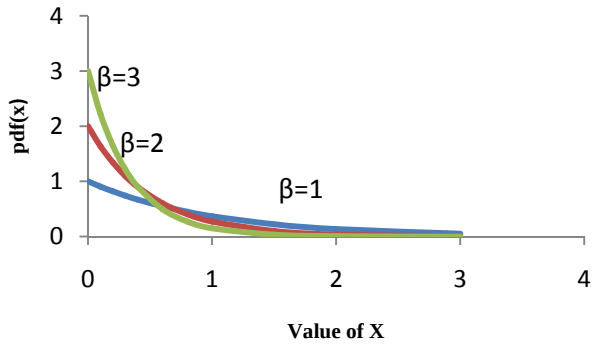

Fig. 4. Graph of Exponential distributions when scale parameters are different.

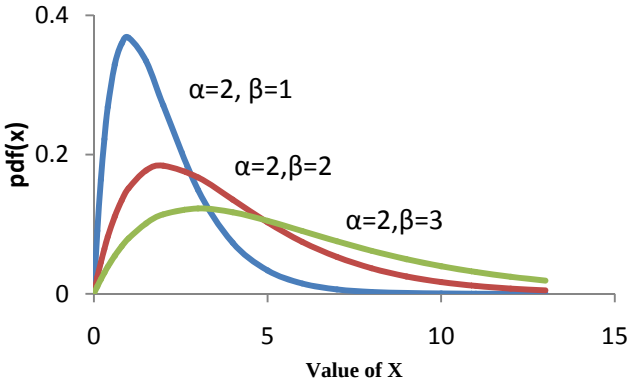

Fig. 5. Graph of Gamma distributions when shape parameters are the same and scale parameters are different.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

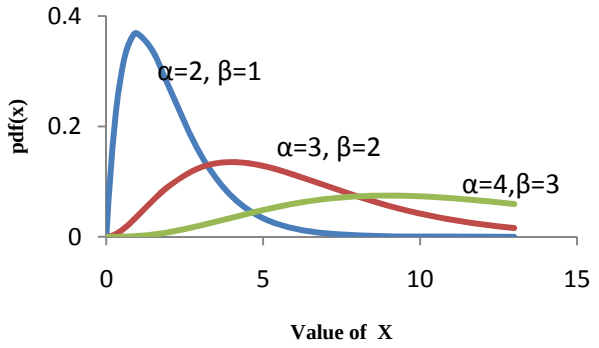

Fig. 6. Graph of Gamma distributions when shape parameters and scale parameters are different.

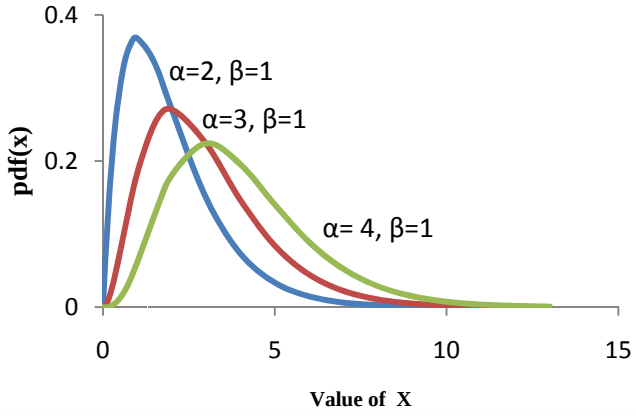

Fig. 7. Graph of Gamma distributions when shape parameters are different and scale parameters are the same.

As a numerical study, Weibull, Exponential and Gamma populations of size $N_i = 5{,}000$ $(i = 1,2,3)$ are generated for different values of parameters β_i, α_i. Then 5,000 random samples, each of size n_i, are drawn. Each set of the sample data was transformed to normality by the proposed transformation and Manly transformation. The results of the goodness-of-fit tests in sense of normality with 5,000 replicated samples of various sizes are shown in Table I – III for Weibull data. Similarly, the results are shown in Table IV for Exponential data and the results are shown in Table V-VII for Gamma data.

TABLE I
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS WITH WEIBULL DATA WHEN $\alpha_i = 2,\ \beta_1 = 1, \beta_2 = 2, \beta_3 = 3$

Transformations	n_i	Averages of the p-Values for K-S Test of Transformed Data		
Manly	10	0.8086	0.8284	0.8038
Proposed	10	0.8142	0.8135	0.8107
Manly	30	0.7487	0.6593	0.5869
Proposed	30	0.7005	0.5390	0.5826
Manly	80	0.6195	0.3482	0.1904
Proposed	80	0.3941	0.1454	0.1620
Manly	10,20,30	0.8180	0.7699	0.6871
Proposed	10,20,30	0.8172	0.6992	0.6440

TABLE II
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS WITH WEIBULL DATA WHEN $\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4, \beta_1 = 1, \beta_2 = 2, \beta_3 = 3$

Transformations	n_i	Averages of the p-Values for K-S Test of Transformed Data		
Manly	10	0.8054	0.8245	0.8079
Proposed	10	0.8129	0.8196	0.8097
Manly	30	0.7712	0.7222	0.5031
Proposed	30	0.7759	0.6936	0.5183
Manly	80	0.7035	0.4382	0.0772
Proposed	80	0.7060	0.3588	0.0866
Manly	10,20,30	0.8126	0.7937	0.5167
Proposed	10,20,30	0.8247	0.7897	0.4950

TABLE III
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS WITH WEIBULL DATA WHEN $\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4, \beta_i = 1$

Transformations	n_i	Averages of the p-Values for K-S Test of Transformed Data		
Manly	10	0.8101	0.8224	0.7832
Proposed	10	0.8331	0.8144	0.7735
Manly	30	0.7029	0.7030	0.5337
Proposed	30	0.7704	0.6181	0.4662
Manly	80	0.4754	0.3530	0.1459
Proposed	80	0.6222	0.1834	0.0898
Manly	10,20,30	0.7857	0.7792	0.6451
Proposed	10,20,30	0.8047	0.7457	0.5954

TABLE IV
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS WITH EXPONENTIAL DATA WHEN $\beta_1 = 1,\ \beta_2 = 2,\ \beta_3 = 3$

Transformations	n_i	Averages of the p-Values for K-S Test of Transformed Data		
Manly	10	0.7302	0.7543	0.8132
Proposed	10	0.8094	0.7990	0.8285
Manly	30	0.5127	0.5478	0.6808
Proposed	30	0.7439	0.6644	0.7134
Manly	80	0.2090	0.2388	0.4282
Proposed	80	0.6049	0.4282	0.4729
Manly	10,20,30	0.7193	0.6882	0.6825
Proposed	10,20,30	0.8079	0.7914	0.7590

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE V
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS WITH GAMMA DATA WHEN $\alpha_i = 2,\ \beta_1 = 1, \beta_2 = 2, \beta_3 = 3$

Transformations	n_i	Averages of the p-Values for K-S Test of Transformed Data		
Manly	10	0.7655	0.7777	0.7919
Proposed	10	0.7688	0.7800	0.7940
Manly	30	0.5820	0.6333	0.6592
Proposed	30	0.5954	0.6408	0.6670
Manly	80	0.2911	0.3585	0.4298
Proposed	80	0.3151	0.3704	0.4416
Manly	10,20,30	0.7814	0.7038	0.6602
Proposed	10,20,30	0.7842	0.7099	0.6683

From Table I to VII, we see that the results from both of two transformations the averages of the p-value of K-S test are small different in each situation. Moreover, the averages of the p-value of K-S test decrease as the sample sizes increase.

TABLE VI
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS WITH GAMMA DATA WHEN $\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4, \beta_1 = 1, \beta_2 = 2, \beta_3 = 3$

Transformations	n_i	Averages of the p-Values for K-S Test of Transformed Data		
Manly	10	0.7507	0.7773	0.7767
Proposed	10	0.7574	0.7760	0.7802
Manly	30	0.5588	0.6026	0.6501
Proposed	30	0.5809	0.5933	0.6583
Manly	80	0.2468	0.3298	0.4063
Proposed	80	0.2812	0.3203	0.4240
Manly	10,20,30	0.7754	0.6837	0.5850
Proposed	10,20,30	0.7810	0.6833	0.5818

TABLE VII
AVERAGES OF THE P-VALUES FOR K-S TEST OF NORMALITY USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS WITH GAMMA DATA WHEN $\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4, \beta_i = 1$

Transformations	n_i	Averages of the p-Values for K-S Test of Transformed Data		
Manly	10	0.7733	0.7749	0.7844
Proposed	10	0.7776	0.7758	0.7842
Manly	30	0.5976	0.6150	0.6398
Proposed	30	0.6048	0.6179	0.6417
Manly	80	0.3073	0.3633	0.3691
Proposed	80	0.3188	0.3682	0.3746
Manly	10,20,30	0.7903	0.7503	0.6347
Proposed	10,20,30	0.7907	0.7521	0.6333

For the check of validity in sense of homogeneity of variance, the results of the Levene test with 5,000 replicated samples of various sizes and data are shown in Table VIII.

TABLE VIII
AVERAGES OF THE P-VALUES FOR LEVENE TEST USING DATA TRANSFORMED BY THE TWO TRANSFORMATIONS

Data	n_i	Manly	Proposed
Weibull (Case I)	10	0.4017	0.4873
$\alpha_i = 2$	30	0.1344	0.2453
$\beta_1 = 1, \beta_2 = 2, \beta_3 = 3$	80	0.0108	0.0571
	10,20,30	0.1944	0.2999
Weibull (Case II)	10	0.5742	0.6013
$\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4$	30	0.4343	0.5032
$\beta_1 = 1, \beta_2 = 2, \beta_3 = 3$	80	0.2384	0.3637
	10,20,30	0.5199	0.5873
Weibull (Case III)	10	0.1901	0.1751
$\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4$	30	0.0093	0.0068
$\beta_i = 1$	80	0.0000	0.0000
	10,20,30	0.0615	0.0547
Exponential (Case IV)	10	0.3304	0.4519
$\beta_1 = 1, \beta_2 = 2, \beta_3 = 3$	30	0.0839	0.2498
	80	0.0025	0.0554
	10,20,30	0.2354	0.3870
Gamma (Case V)	10	0.6602	0.6971
$\alpha_i = 2$	30	0.5596	0.6604
$\beta_1 = 1, \beta_2 = 2, \beta_3 = 3$	80	0.3575	0.5976
	10,20,30	0.5934	0.6639
Gamma (Case VI)	10	0.6357	0.6576
$\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4$	30	0.4823	0.5539
$\beta_1 = 1, \beta_2 = 2, \beta_3 = 3$	80	0.2303	0.3594
	10,20,30	0.5849	0.6088
Gamma (Case VII)	10	0.6033	0.7020
$\alpha_1 = 2,\ \alpha_2 = 3,\ \alpha_3 = 4$	30	0.6696	0.6781
$\beta_i = 1$	80	0.6174	0.6372
	10,20,30	0.6611	0.6627

From Table VIII, for Case I to VII, we see that averages of the p-value of Levene test of proposed transformation are higher than them of Manly transformation in each of sample sizes. In case I and IV when the sample sizes are large, proposed transformation performs better than Manly transformation at significant level 0.05. For Case III, we see that both proposed transformation and Manly transformation work well with only the small sample size. Moreover, the averages of the p-value of Levene test decrease as the sample sizes increase.

VI. CONCLUSION

The efficiency of the proposed transformation is compared with Manly transformation in sense of normality and homogeneity of variance. Both of them can transform the lifetime data to correspond with the basic assumptions in some situation. In sense of normality, it is found that the proposed transformation and Manly transformation have not different efficiency. The proposed transformation performs better than Manly transformation in sense of homogeneity of variances for some data set of weibull distributions and exponential distributions when the sample sizes are large.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

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.
[8] I. Yeo and N. R. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 2, pp.954-959, 2000.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)