Kasetsart J. (Nat. Sci.) 45 : 366 - 383 (2011)

Comparison of Four Data Transformation Methods for Weibull Distributed Data

Thunyaporn Chortirat*, Boonorm Chomtee and Juthaphorn Sinsomboonthong

ABSTRACT

The objective of this research was to compare four data transformation methods: the error function transformation, the dual power transformation, the exponential transformation of Manly, and the Box-Cox transformation. The criterion used for the study was the ratio of the percentage of acceptances of the null hypothesis H_0 to the data having a normal distribution, after the four data transformation methods were applied to Weibull distributed data. The approaches were evaluated using both real and simulated data. For the simulated data, Weibull distributed datasets were generated for skewness and kurtosis levels using MATLAB version 7.0 with three levels of sample size (n): small (10, 30), medium (50, 70) and large (100, 120). Each situation was repeated 500 times and the significance level was set at 0.05.

The results consisted of two parts: part I presented the simulated data and part II the real data.

With the simulated data with right-skew distribution, and n=10, for skewness (0.3, 0.6], the Box-Cox and exponential transformation methods were the best methods, for skewness (0.6, 1.2], the Box-Cox method was the best and for skewness (1.2, 2.1], the Box-Cox and exponential transformation methods were the best methods. When n=30, 50, 70, 100 and 120, the Box-Cox method was the best. When the data had left-skew distribution, for small and medium sample sizes, the exponential transformation method was the best method for almost all situations. However, for a large sample size, the Box-Cox method was generally the best method.

For the real data, the *P-values* and the histogram of the empirical data were presented. It was also found that the best transformation method was the Box-Cox method.

Keywords: data transformation, error function transformation, dual power transformation, exponential transformation, Box and Cox transformation

INTRODUCTION

Nowadays, statistics are widely used both in research and academia. Statistical methods consist of organizing data, analyzing data, presenting information and drawing conclusions. In addition, researchers must provide insight on their studies, as well as having the knowledge to apply any statistical methods accurately. In general, statistics are divided into two types: parametric and nonparametric statistics. Using either parametric or nonparametric statistic depends on the assumptions. An important assumption of parametric statistics involves the population distribution. Most parametric statistical methods require a normally distributed population.

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand.

* Corresponding author, e-mail: fonnew56@hotmail.com

Received date : 26/08/10 Accepted date : 09/12/10

In practice, if data are not normal distributed, nonparametric statistics may be used in the analysis. Otherwise, data transformation methods are used to transform non normally distributed data into normally distributed data.

This research compared four data transformation methods: the error function transformation, the dual power transformation, the exponential transformation of Manly and the Box-Cox transformation. The criterion used for the study was the ratio of the percentage of acceptances of the null hypothesis H_0 to the data having a normal distribution, where the higher the percentage, the better the data transformation method.

MATERIALS AND METHODS

Materials and equipment

In this research, MATLAB version 7.0 (MathWorks, Natick, Massachusetts, USA) and Microsoft Excel 2003 (Microsoft, Redmond, Washington, USA) software were used.

Scope of research

2.1 The data were positive values.

2.2 The Anderson-Darling statistic (A^2; Anderson and Darling, 1952) was used to test the normality of the data both before and after using transformation methods, using Equation 1:

$$A^{*2}=A^2\left(1+\frac{4}{n}-\frac{25}{n^2}\right) \quad (1)$$

When $$A^2=-\frac{\left\{\sum_{i=1}^{n}(2i-1)\left[\ln z_i+\ln\left(1-z_{n+1-i}\right)\right]\right\}}{n}-n$$

$$z_i=\phi\left(\frac{\left(x_{(i)}-\bar{x}\right)}{s}\right) \quad (2)$$

where: $\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$ and $s=\sqrt{\frac{\sum_{i=1}^{n}\left(x_1-\bar{x}\right)^2}{n-1}}$

ϕ (.) is the probability density function of N(0,1).

2.3 The datasets were Weibull distributions (Weibull, 1951) in which the probability density function is shown in Equation 3:

$$f(x)=\begin{cases} ab^{-a}x^{a-1}e^{-(x/b)^a} & ;x>0 \\ 0 & ;otherwise \end{cases} \quad (3)$$

where: the shape parameter is a ($a>0$) and the scale parameter is b, ($b>0$).

2.4 The four transformation methods used in this study were:

2.4.1 The error function transformation method (EF; van Albada and Robinson, 2006) is shown in Equation 4:

$$y_e(x)=\sqrt{2}erf^{-1}\left[1-2e^{-(x/b)^a}\right] \quad (4)$$

where: x is untransformed data, $y_e(x)$ is the transformed data by the error function and a, b are shape and scale parameters of the Weibull distribution, respectively.

The inverse error function (MathWorld, 1999) is Equation 5:

$$erf^{-1}(x)=\sum_{n=0}^{\infty}\frac{C_n}{2n+1}\left(\frac{\sqrt{\pi}}{2}x\right)^{2n+1} \quad (5)$$

where $C_0=1$ and $C_n=\sum_{k=0}^{n-1}\frac{C_kC_{n-1-k}}{(k+1)(2k+1)}$

For 2.4.2 to 2.4.4, the pf value which is a real number in the range (˘ ∞, ∞) is the power of the transformation methods.

2.4.2 The dual power transformation method (DP; Yang, 2006) is shown in Equation 6:

$$y_D(x)=\begin{cases} \frac{x^p-x^{-p}}{2p} & ;p\neq 0 \\ \log x & ;p=0 \end{cases} \quad (6)$$

where: $y_D(x)$ is the transformed data by the dual power transformation and p is the power of the transformation methods.

2.4.3 The exponential transformation method (EP; Manly, 1976) is shown in Equation 7:

$$y_E(x)=\begin{cases} [\exp(px)-1]/p & ;p\neq 0 \\ x & ;p=0 \end{cases} \quad (7)$$

where: $y_E(x)$ is the transformed data by Exponential transformation and p is the power of transformation methods.

2.4.4 The Box and Cox transformation method (BC; Box and Cox, 1964) is shown in Equation 8:

$$y_B(x)=\begin{cases}\frac{(x^p-1)}{p} & ;p\neq 0, x>0\\ \log_{10}(x) & ;p=0, x>0\end{cases} \quad (8)$$

where: $y_B(x)$ is the transformed data by the Box and Cox transformation and p is the power of the transformation methods.

2.5 There were three levels of sample size (n): small (10, 30), medium (50, 70) and large (100, 120).

2.6 In this research, for right skew distribution, there were six intervals of skewness: (0.3, 0.6], (0.6, 0.9], (0.9, 1.2], (1.2, 1.5], (1.5, 1.8] and (1.8, 2.1]. For left skew distribution, there were six intervals of skewness: (-0.6, -0.3], (-0.9, -0.6], (-1.2, -0.9], (-1.5, -1.2], (-1.8, -1.5] and (-2.1, -1.8].

2.7 In this research, the kurtosis values for right and left skew distribution were (1, 2], (2, 3], (3, 4], (4, 5], (5, 6], (6, 7], (7, 8], (8, 9], (9, 10], and (10, 11].

Therefore, in total, 285 situations were studied in this research.

2.8 The significance level (α) was set at 0.05.

2.9 Each situation was repeated 500 times.

Process

For the simulated datasets:

Step 1. The Weibull distributed data were simulated for each skewness and kurtosis level.

Step 2. The data in step 1 were transformed by the four transformation methods.

Step 3. The transformed data were checked for a normal distribution using A^2 at the 0.05 significance level.

For each situation (285 situations), Steps 1 to 3 were repeated 500 times. Then, the percentages of acceptance of H_0 : the data having a normal distribution were calculated using Equation 9:

$$\frac{100\times(\text{The number of acceptance } H_0)}{500} \quad (9)$$

For the real datasets:

Step 1. The 60 real datasets were checked by MATLAB version 7.0 software to ensure that they had Weibull distributions.

Step 2. The 60 real datasets of Weibull distributions of sample sizes 10, 30, 50, 70, 100 and 120 were then used in the study.

Step 3. The real data were checked by A^2 at the 0.05 significance level to ensure that they did not have normal distributions.

Step 4. For each dataset, the shape and scale parameters of the Weibull distribution were estimated.

Step 5. The real datasets were transformed by the four methods.

Step 6. The transformed real data were rechecked for normality again using A^2 at the 0.05 significance level.

The criterion for the study of the real datasets was the *P-values* of normality testing of A^2 in which a larger *P-value* indicated a better transformation method.

RESULTS AND DISCUSSION

The results were divided into two parts, with Parts I and II showing the results for the simulated and real datasets, respectively.

Simulated data

For right-skew distributed data, some of the results are shown in Tables 1˘7.

In Table 2, for n=10 and skewness (0.3, 0.6], the BC and EP transformation methods were the best methods, for skewness (0.6, 1.2], the BC transformation method was the best, and for skewness (1.2, 2.1], the BC and EP transformation

methods were the best methods. For n=30, 50, 70, 100 and 120 and for all skewed intervals, the BC transformation method was the best.

In overview, BC was the best transformation method in the studied cases.

For left-skew distributed data, some of the results are shown in Tables 8˜14.

In overview, BC was a good method for skewness (-0.6, -0.3] and (-2.1, -1.8] in almost all of the sample sizes. In addition, BC was a good method for skewness (-1.5, -0.6] and n=120. However, EP was a good method for skewness (-1.8, -0.6] in almost all sample sizes except for the large sample sizes.

In summary, BC and EP were recommended for data transformation methods.

In the present research, the results were not consistent with past research, especially with regard to the error function transformation method which was not considered to be good enough as a transformation method.

Table 1 Summary of the best data transformation methods for simulated data with right-skew distribution.

Sample size	Skewness level		
	(0.3, 0.6]	(0.6, 1.2]	(1.2, 2.1]
10	BC, EP	BC	BC, EP
30-120	BC	BC	BC

Table 2 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=10 and right-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(0.3, 0.6]	(1, 2]	45.2	47.8	81.2	**81.8**
	(2, 3]	38.8	46.4	**78.2**	76.6
	(3, 4]	38.6	44.8	70.4	**73.6**
	(4, 5]	32.8	42.6	**50.6**	50.2
(0.6, 0.9]	(1, 2]	46.8	49.8	90.6	**93.2**
	(2, 3]	44.8	55.6	94.2	**94.8**
	(3, 4]	46.4	56.4	93.8	**96.2**
	(4, 5]	46.8	54.8	92.2	**94.2**
(0.9, 1.2]	(2, 3]	45.2	56.9	**97.0**	96.8
	(3, 4]	48.4	56.8	97.2	**97.8**
	(4, 5]	40.6	41.0	74.0	**75.0**
	(5, 6]	46.2	51.2	92.2	**96.0**
(1.2, 1.5]	(2, 3]	46.8	63.2	**94.6**	**94.6**
	(3, 4]	42.6	57.8	**98.6**	98.4
	(4, 5]	48.2	52.8	96.6	**98.6**
	(5, 6]	52.0	52.8	**91.0**	**91.0**
(1.5, 1.8]	(3, 4]	42.4	58.8	**98.8**	98.0
	(4, 5]	44.6	58.6	**98.6**	98.4
	(5, 6]	46.2	54.6	97.2	**98.6**
(1.8, 2.1]	(4, 5]	50.2	58.4	98.6	**99.8**
	(5, 6]	49.6	56.4	98.2	**99.2**
	(6, 7]	60.8	53.8	**97.0**	96.2

Note : The highest percentage for each dataset is in bold.

Table 3 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=30 and right-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(0.3, 0.6]	(1, 2]	52.8	59.4	50.8	**90.0**
	(2, 3]	60.3	73.4	76.6	**93.2**
	(3, 4]	59.4	29.4	**68.2**	62.4
	(1, 2]	32.6	23.0	64.6	**98.6**
(0.6, 0.9]	(2, 3]	40.2	52.8	64.6	**98.8**
	(3, 4]	48.3	60.6	80.4	**95.8**
(0.9, 1.2]	(2, 3]	39.4	48.6	85.6	**99.6**
	(3, 4]	55.2	49.6	84.6	**99.4**
	(4, 5]	56.8	52.8	81.4	**97.6**
	(5, 6]	45.8	52.8	81.2	**98.0**
(1.2, 1.5]	(2, 3]	32.2	87.6	63.2	**99.4**
	(3, 4]	45.8	89.2	74.6	**99.4**
	(4, 5]	53.0	94.4	69.0	**99.4**
(1.5, 1.8]	(3, 4]	45.2	89.2	77.4	**99.2**
	(4, 5]	54.2	89.8	78.6	**99.6**
	(5, 6]	55.4	58.4	76.0	**99.4**
	(6, 7]	60.2	56.0	75.6	**99.6**
(1.8, 2.1]	(4, 5]	48.2	59.0	73.4	**98.6**
	(5, 6]	59.0	54.2	77.0	**98.8**
	(6, 7]	59.2	53.2	75.0	**100.0**
	(7, 8]	62.8	56.8	82.2	**99.2**

Note : The highest percentage for each dataset is in bold.

Table 4 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=50 and right-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(0.3, 0.6]	(1, 2]	35.2	43.4	35.8	**78.2**
	(2, 3]	38.4	43.4	82.0	**95.0**
	(3, 4]	36.2	40.6	77.0	**77.2**
	(4, 5]	38.6	40.4	**69.0**	68.8
(0.6, 0.9]	(1, 2]	45.8	48.8	69.6	**97.4**
	(2, 3]	46.0	46.8	72.0	**97.8**
	(3, 4]	46.2	60.0	92.0	**99.0**
	(4, 5]	49.2	36.0	84.6	83.8
(0.9, 1.2]	(2, 3]	30.1	26.6	86.0	**97.2**
	(3, 4]	38.0	25.8	59.2	**99.6**
	(4, 5]	35.4	46.2	85.2	**98.4**
	(5, 6]	38.4	53.0	93.2	**95.6**
(1.2, 1.5]	(3, 4]	21.8	27.8	80.0	**99.6**
	(4, 5]	31.2	31.0	82.2	**99.6**
	(5, 6]	31.8	32.4	63.8	**99.2**
	(6, 7]	31.8	36.4	86.4	**97.6**
(1.5, 1.8]	(4, 5]	29.5	35.4	38.2	**100.0**
	(5, 6]	32.0	31.2	53.8	**99.6**
	(6, 7]	32.5	34.2	67.4	**98.4**
	(7, 8]	39.4	49.4	85.6	**96.2**
(1.8, 2.1]	(5, 6]	28.2	36.2	42.6	**99.2**
	(6, 7]	31.0	31.2	45.0	**99.8**
	(7, 8]	31.2	29.5	38.6	**99.6**
	(8, 9]	28.5	28.3	44.0	**99.8**
	(9, 10]	34.2	28.8	49.2	**98.4**

Note : The highest percentage for each dataset is in bold.

Table 5 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=70 and right-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(0.3, 0.6]	(1, 2]	30.2	32.2	52.6	**85.8**
	(2, 3]	45.3	44.4	58.6	**96.2**
	(3, 4]	49.0	36.4	67.6	**85.2**
(0.6, 0.9]	(1, 2]	44.5	26.0	62.2	**98.4**
	(2, 3]	43.9	26.2	62.4	**97.2**
	(3, 4]	46.8	32.0	73.8	**98.8**
	(4, 5]	43.2	39.8	69.2	**93.0**
(0.9, 1.2]	(2, 3]	28.4	27.8	28.2	**95.0**
	(3, 4]	32.2	26.6	27.8	**97.8**
	(4, 5]	34.2	39.6	65.8	**98.6**
	(5, 6]	34.8	36.2	74.8	**98.4**
(1.2, 1.5]	(2, 3]	29.2	13.9	29.6	**99.0**
	(3, 4]	28.6	24.6	32.4	**99.6**
	(4, 5]	35.4	22.2	30.4	**99.0**
	(5, 6]	38.2	25.2	38.8	**98.8**
	(6, 7]	41.2	42.2	61.8	**97.2**
(1.5, 1.8]	(3, 4]	34.4	28.8	35.2	**99.4**
	(4, 5]	47.2	21.2	35.8	**99.8**
	(5, 6]	40.6	20.8	38.2	**100.0**
	(6, 7]	40.8	20.2	38.6	**99.8**
	(7, 8]	38.2	24.2	39.0	**99.8**
(1.8, 2.1]	(5, 6]	32.2	43.2	39.6	**98.8**
	(6, 7]	34.6	30.8	40.4	**99.4**
	(7, 8]	33.0	28.4	43.4	**99.6**
	(8, 9]	34.8	28.2	45.2	**98.4**

Note : The highest percentage for each dataset is in bold.

Table 6 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=100 and right-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(0.3, 0.6]	(1, 2]	24.0	19.0	64.2	**86.8**
	(2, 3]	24.4	18.2	69.0	**92.2**
	(3, 4]	28.6	16.0	56.8	**97.2**
	(4, 5]	53.0	10.0	**94.0**	93.4
(0.6, 0.9]	(1, 2]	30.0	21.2	52.6	**98.0**
	(2, 3]	34.2	20.0	51.0	**97.2**
	(3, 4]	24.2	18.0	43.0	**99.6**
	(4, 5]	40.0	12.6	65.0	**97.0**
	(5, 6]	43.0	13.2	62.0	**93.8**
(0.9, 1.2]	(2, 3]	28.8	20.0	45.0	**93.8**
	(3, 4]	20.0	16.0	35.0	**95.4**
	(4, 5]	18.6	14.0	46.2	**94.6**
	(5, 6]	24.0	10.2	55.6	**96.8**
(1.2, 1.5]	(2, 3]	28.0	11.0	38.0	**95.2**
	(3, 4]	27.2	9.8	43.2	**94.2**
	(4, 5]	18.0	9.2	45.0	**97.4**
	(5, 6]	29.4	8.8	58.0	**97.2**
(1.5, 1.8]	(3, 4]	28.6	28.4	42.0	**96.6**
	(4, 5]	34.0	27.8	40.4	**98.2**
	(5, 6]	41.2	26.3	41.0	**98.2**
	(6, 7]	52.8	12.4	51.2	**98.4**
(1.8, 2.1]	(6, 7]	27.8	27.0	47.4	**98.6**
	(7, 8]	30.0	27.0	52.0	**99.8**
	(8, 9]	35.0	19.2	53.6	**98.2**
	(9, 10]	49.8	20.0	53.0	**95.4**

Note : The highest percentage for each dataset is in bold.

Table 7 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=120 and right-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(0.3, 0.6]	(1, 2]	28.8	21.0	46.0	**95.6**
	(2, 3]	26.0	19.0	44.6	**96.0**
	(3, 4]	29.2	23.4	52.0	**94.4**
	(4, 5]	32.2	23.8	52.6	**95.6**
(0.6, 0.9]	(1, 2]	34.0	33.2	46.2	**97.8**
	(2, 3]	30.2	33.0	47.6	**98.2**
	(3, 4]	21.0	18.0	55.0	**99.2**
	(4, 5]	19.0	13.6	56.0	**99.2**
(0.9, 1.2]	(2, 3]	32.8	17.0	53.8	**98.9**
	(3, 4]	30.8	20.0	53.5	**92.0**
	(4, 5]	34.0	20.0	54.6	**91.4**
	(5, 6]	35.0	23.2	56.6	**96.2**
(1.2, 1.5]	(2, 3]	43.2	18.6	55.0	**97.0**
	(3, 4]	39.0	18.0	64.4	**97.8**
	(4, 5]	39.8	17.8	64.0	**93.8**
	(5, 6]	37.0	15.0	65.6	**96.6**
(1.5, 1.8]	(4, 5]	29.8	30.6	59.0	**96.8**
	(5, 6]	30.2	26.4	58.2	**97.8**
	(6, 7]	29.0	25.0	60.4	**97.8**
	(7, 8]	24.0	25.8	67.0	**97.8**
(1.8, 2.1]	(6, 7]	29.2	36.4	57.2	**98.6**
	(7, 8]	27.0	32.0	54.0	**98.4**
	(8, 9]	28.8	30.8	60.0	**99.0**
	(9, 10]	23.0	29.8	65.4	**99.8**

Note : The highest percentage for each dataset is in bold.

Table 8 Summary of the best data transformation methods for simulated data with left-skew distribution.

Sample size	Skewness					
	(-0.6, -0.3]	(-0.9, -0.6]	(-1.2, -0.9]	(-1.5, -1.2]	(-1.8, -1.5]	(-2.1, -1.8]
10	BC	BC, EP	EP	EP	EP	EP
30	BC, EP	EP	EP	EP	EP	BC, EP
50	BC, EP	EP	EP	EP	BC, EP	EP
70	BC	EP	EP	EP	BC, EP	BC, EP
100	EP	EP	EP	BC	BC	BC, EP
120	BC	BC, EP	BC	BC	EP	BC, EP

Table 9 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=10 and left-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(-0.6, -0.3]	(1, 2]	39.2	46.2	**77.8**	48.0
	(2, 3]	40.0	36.0	77.4	**78.0**
	(3, 4]	28.0	32.0	67.8	**87.6**
	(4, 5]	27.8	29.2	66.0	**75.0**
(-0.9, -0.6]	(1, 2]	38.0	55.8	**85.6**	58.2
	(2, 3]	35.6	70.0	**90.8**	73.8
	(3, 4]	36.0	60.2	72.6	**73.0**
	(4, 5]	32.4	37.2	45.4	**53.8**
(-1.2, -0.9]	(2, 3]	40.8	74.4	**96.0**	76.4
	(3, 4]	40.0	84.4	**96.0**	90.4
	(4, 5]	39.2	74.4	**84.2**	83.2
	(5, 6]	37.8	80.0	**84.2**	**84.2**
(-1.5, -1.2]	(2, 3]	26.0	58.4	79.6	**80.2**
	(3, 4]	29.0	65.2	**93.8**	68.2
	(4, 5]	39.8	91.0	**98.0**	95.6
	(5, 6]	38.2	75.6	**84.0**	81.6
(-1.8, -1.5]	(3, 4]	29.0	64.0	**60.8**	55.2
	(4, 5]	35.0	76.4	**95.6**	79.4
	(5, 6]	40.8	89.2	**96.8**	92.2
	(6, 7]	43.0	89.8	**96.8**	96.0
(-2.1, -1.8]	(3, 4]	28.8	65.2	73.4	**75.4**
	(4, 5]	34.6	66.0	75.0	**76.0**
	(5, 6]	39.0	70.8	**84.0**	80.0
	(6, 7]	39.8	72.2	85.4	**86.2**

Note : The highest percentage for each dataset is in bold.

Table 10 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=30 and left-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(-0.6, -0.3]	(1, 2]	32.0	51.4	**74.4**	55.0
	(2, 3]	31.0	82.6	**91.0**	86.6
	(3, 4]	34.4	62.4	73.8	**80.2**
	(4, 5]	28.2	58.6	74.0	**79.6**
(-0.9, -0.6]	(1, 2]	30.2	40.4	**70.0**	64.2
	(2, 3]	41.4	86.0	**98.0**	86.2
	(3, 4]	37.6	92.2	**97.8**	95.6
	(4, 5]	29.0	69.8	78.6	**82.6**
(-1.2, -0.9]	(2, 3]	34.4	44.4	**89.0**	46.6
	(3, 4]	43.8	86.4	**99.2**	86.6
	(4, 5]	42.8	95.2	**98.0**	96.0
	(5, 6]	45.0	80.2	87.4	**88.8**
(-1.5, -1.2]	(2, 3]	35.8	31.8	**81.2**	35.4
	(3, 4]	44.0	76.4	**98.0**	78.2
	(4, 5]	54.4	91.0	**97.6**	95.4
	(5, 6]	54.6	86.2	**94.4**	93.2
(-1.8, -1.5]	(3, 4]	47.0	87.0	**97.4**	91.2
	(4, 5]	52.0	28.6	**85.6**	82.6
	(5, 6]	52.2	58.6	**92.2**	61.2
(-2.1, -1.8]	(5, 6]	28.8	28.6	**85.6**	82.6
	(6, 7]	34.8	65.0	**91.4**	69.0
	(7, 8]	40.6	87.8	**95.8**	94.0
	(8, 9]	40.0	88.4	90.8	**94.4**
	(9, 10]	51.0	43.4	65.8	**78.2**

Note : The highest percentage for each dataset is in bold.

Table 11 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=50 and left-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(-0.6, -0.3]	(1, 2]	57.8	82.6	**91.8**	86.0
	(2, 3]	55.0	84.4	**93.2**	88.0
	(3, 4]	46.8	86.0	89.0	**89.8**
	(4, 5]	42.0	86.4	86.2	**89.8**
(-0.9, -0.6]	(1, 2]	62.8	85.0	**98.0**	85.4
	(2, 3]	59.2	82.0	**98.4**	82.4
	(3, 4]	66.0	95.2	**99.4**	95.8
	(4, 5]	45.8	83.2	91.4	**93.2**
(-1.2, -0.9]	(2, 3]	67.2	81.4	**99.2**	81.6
	(3, 4]	68.8	83.6	**99.4**	83.8
	(4, 5]	69.8	94.2	**98.6**	94.4
	(5, 6]	54.0	91.4	94.8	**96.0**
(-1.5, -1.2]	(2, 3]	53.2	60.4	**77.6**	69.8
	(3, 4]	65.4	60.4	**94.0**	63.2
	(4, 5]	66.2	65.6	**94.6**	68.2
	(5, 6]	67.0	88.6	**98.8**	89.4
(-1.8, -1.5]	(3, 4]	44.0	89.6	68.0	**99.0**
	(4, 5]	38.8	87.8	86.0	**99.4**
	(5, 6]	45.0	89.0	90.8	**92.0**
	(6, 7]	45.6	89.4	**96.6**	96.0
	(7, 8]	46.2	90.0	**95.0**	94.0
(-2.1, -1.8]	(3, 4]	49.2	69.8	**84.4**	78.8
	(4, 5]	49.0	69.0	**84.8**	82.0
	(5, 6]	50.2	69.6	**88.6**	75.2
	(6, 7]	55.0	79.2	**92.2**	83.0

Note : The highest percentage for each dataset is in bold.

Table 12 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=70 and left-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(-0.6, -0.3]	(2, 3]	50.2	86.2	**95.2**	91.2
	(3, 4]	46.6	80.6	93.2	**95.8**
	(4, 5]	41.8	85.4	92.8	**95.2**
	(5, 6]	40.0	85.2	92.2	**96.0**
(-0.9, -0.6]	(2, 3]	52.2	82.0	**97.6**	82.6
	(3, 4]	58.0	95.0	**99.2**	95.2
	(4, 5]	48.2	90.2	95.6	**95.8**
(-1.2, -0.9]	(3, 4]	39.6	73.2	**98.8**	74.0
	(4, 5]	56.8	90.6	**99.0**	90.6
	(5, 6]	42.8	96.2	**99.0**	98.8
	(6, 7]	46.5	96.8	**99.0**	**99.0**
(-1.5, -1.2]	(3, 4]	35.0	86.0	**88.2**	80.8
	(4, 5]	38.8	89.4	**89.8**	**89.8**
	(5, 6]	47.0	78.6	**97.2**	80.6
	(6, 7]	56.4	76.0	**98.0**	80.0
(-1.8, -1.5]	(4, 5]	37.0	79.2	80.4	**84.0**
	(5, 6]	39.2	88.0	86.2	**86.6**
	(6, 7]	42.2	89.4	**90.4**	89.8
(-2.1, -1.8]	(6, 7]	39.0	72.2	82.6	**84.0**
	(7, 8]	40.2	72.0	**86.6**	**86.6**
	(8, 9]	39.0	76.0	**90.6**	82.4
	(9, 10]	44.8	86.0	**94.6**	93.6

Note : The highest percentage for each dataset is in bold.

Table 13 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=100 and left-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(-0.6, -0.3]	(1, 2]	39.6	76.6	**82.6**	74.0
	(2, 3]	47.0	76.0	**94.2**	91.6
	(3, 4]	34.4	82.0	94.0	**96.4**
	(4, 5]	32.2	87.0	67.4	**81.8**
(-0.9, -0.6]	(1, 2]	29.8	82.2	**95.0**	78.0
	(2, 3]	44.8	84.8	**96.2**	79.2
	(3, 4]	62.0	87.0	**99.4**	94.0
	(4, 5]	56.2	87.4	97.8	**98.0**
(-1.2, -0.9]	(2, 3]	33.6	79.8	**85.0**	79.4
	(3, 4]	35.0	85.4	**98.6**	78.0
	(4, 5]	43.4	86.6	**98.0**	94.6
	(5, 6]	50.2	87.2	96.2	**98.0**
(-1.5, -1.2]	(2, 3]	32.0	78.0	**87.8**	80.2
	(3, 4]	36.8	86.7	94.2	**94.8**
	(4, 5]	49.2	86.0	90.0	**95.2**
	(5, 6]	56.2	89.8	89.8	**97.0**
(-1.8, -1.5]	(3, 4]	39.2	72.2	85.0	**90.8**
	(4, 5]	46.8	75.8	78.4	**96.4**
	(5, 6]	47.8	76.6	87.9	**97.2**
	(6, 7]	54.0	78.0	86.0	**99.0**
(-2.1, -1.8]	(3, 4]	35.8	75.4	87.4	**96.8**
	(4, 5]	43.0	78.8	**97.8**	96.0
	(5, 6]	52.2	80.8	**97.2**	**97.2**
	(6, 7]	58.0	87.0	93.0	**99.4**

Note : The highest percentage for each dataset is in bold.

Table 14 Percentage of acceptance of null hypothesis H_0 : the data having a normal distribution, for n=120 and left-skew distributed data.

Skewness	Kurtosis	Percentage of acceptance H_0			
		EF	DP	EP	BC
(-0.6, -0.3]	(1, 2]	38.0	83.0	82.8	**86.2**
	(2, 3]	42.4	80.8	**93.0**	89.0
	(3, 4]	42.0	89.0	92.6	**95.2**
	(4, 5]	45.6	89.4	95.0	**97.8**
(-0.9, -0.6]	(1, 2]	35.0	76.8	**89.8**	89.0
	(2, 3]	38.8	84.0	**93.0**	89.4
	(3, 4]	47.0	84.6	97.2	**97.8**
	(4, 5]	49.8	88.0	97.8	**99.0**
(-1.2, -0.9]	(2, 3]	46.4	83.4	81.6	**85.4**
	(3, 4]	53.0	84.4	89.0	**96.0**
	(4, 5]	53.8	82.0	91.0	**96.8**
	(5, 6]	57.2	89.2	89.4	**98.0**
(-1.5, -1.2]	(2, 3]	42.2	76.6	86.6	**87.2**
	(3, 4]	46.0	79.2	89.0	**95.6**
	(4, 5]	59.0	81.4	92.2	**97.0**
	(5, 6]	49.8	85.2	93.0	**96.4**
(-1.8, -1.5]	(3, 4]	43.2	75.8	**86.6**	84.4
	(4, 5]	47.2	79.8	85.8	**91.0**
	(5, 6]	57.8	85.4	**93.0**	92.4
	(6, 7]	42.0	86.2	**94.8**	94.2
(-2.1, -1.8]	(3, 4]	41.8	71.8	83.0	**85.4**
	(4, 5]	45.0	74.6	**84.2**	84.0
	(5, 6]	59.8	79.0	**87.0**	86.6
	(6, 7]	50.2	80.8	89.0	**91.6**

Note : The highest percentage for each dataset is in bold.

Real datasets

To illustrate the results of the four transformation methods, sixty real datasets which had Weibull distributions for small, medium and large sample sizes were collected. After the data had been transformed, the *P-values* of A^2 for normality testing were considered as criteria for comparisons of the four transformation methods, where a larger *P-value* indicated a better method. The results are shown in Tables 15 and 16.

Table 15 indicates that for almost all the studied cases, BC was the best method with *P-values* ranging from 0.0652 to 0.8694 for small sample sizes (10, 30), from 0.1451 to 0.8219 for medium sample sizes (50, 70) and from 0.2902 to 0.9807 for large sample sizes (100, 120).

Table 16 indicates that for almost all the studied cases, BC was the best method with *P-values* ranging from 0.0585 to 0.8299 for small sample sizes (10, 30), from 0.3443 to 0.9655 for medium sample sizes (50, 70) and from 0.1809 to 0.9641 for large sample sizes (100, 120).

Some of the histograms for the real datasets are shown in Figures 1 and 2 and they provide corresponding results with the *P-values* of A^2.

DISCUSSION

The result summarized above were consistent with the results of Kerdsawang (2005) and Rotwirat (2008)˘namely, that the Box-Cox data transformation method was the best method for right skew distribution. In contrast, for left skew distribution, Kerdsawang (2005) concluded that the exponential transformation method was the best and Rotwirat (2008) concluded that the dual power data transformation method was the best, while in the present study, the exponential and the Box-Cox data transformation methods were suitable and provided similar values for the percentage of acceptance of the null hypothesis and so were considered to be egually good.

CONCLUSION

For the right-skew distribution, using the BC transformation method, the percentages of acceptance of H_0 tended to increase when the sample size increased. However, for methods EF, DP and EP, the percentages tended to decrease when the sample size increased. Some example cases are presented in Figures 3˘6.

For the left-skew distribution, the EP and BC transformation methods were suitable and provided similar values of the percentage of acceptance of the null hypothesis for each situation.

Based on the scope of this research, BC gave the best percentages of acceptance of the null hypothesis for simulated data and gave the best *P-values* for the real datasets. However, in other situations there might be other methods that could be beneficial, such as EP, which performed well for right-skew distributed data when the sample size was small.

Table 15 *P-value* of the Anderson-Darling test for the real data with right-skew distribution after the data had been transformed.

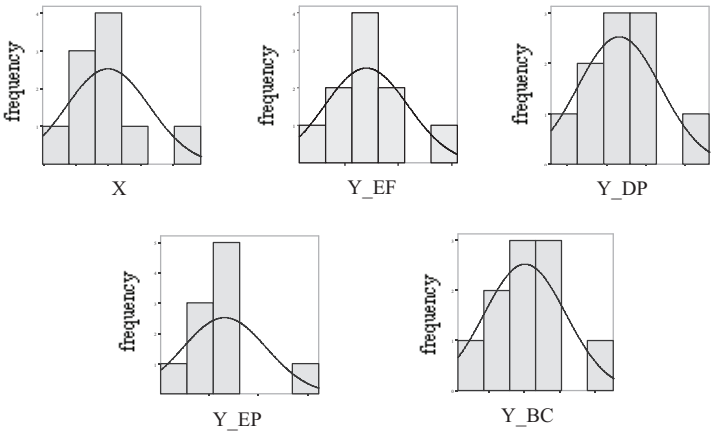
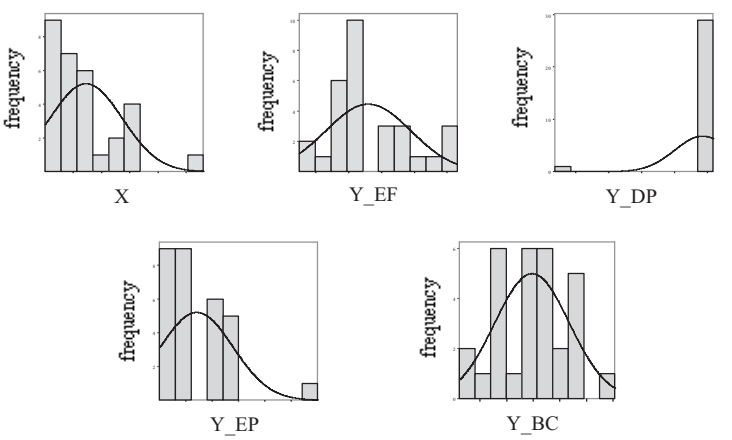
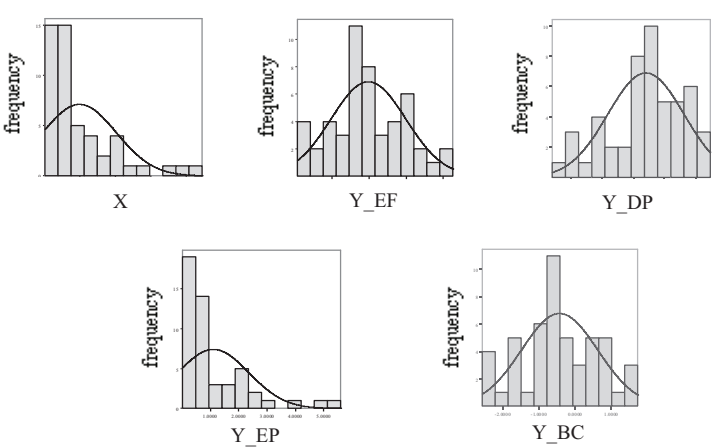
Dataset	Sample size	Transformation method				Dataset	Sample size	Transformation method			
		EF	DP	EP	BC			EF	DP	EP	BC
1	10	0.2065	**0.2255**	0.1546	**0.2255**	16	70	**0.0505**	0.0000	0.0028	0.0232
2	10	**0.1092**	0.0000	0.0016	0.1049	17	70	0.4149	0.0432	0.0001	**0.5176**
3	10	0.2152	0.0607	0.0023	**0.2331**	18	70	0.5488	0.0887	0.0185	**0.6692**
4	10	0.0819	0.5738	0.0022	**0.8694**	19	70	**0.3482**	0.0022	0.0000	0.2620
5	10	0.0552	**0.0652**	0.0391	**0.0652**	20	70	0.0239	0.0000	0.0000	**0.5250**
6	30	0.0005	0.0000	0.0012	**0.7394**	21	100	**0.4092**	0.0027	0.0000	0.2902
7	30	0.5066	0.0000	0.0001	**0.6682**	22	100	0.0004	0.0066	0.0000	**0.6783**
8	30	0.0415	0.2826	0.0000	**0.7732**	23	100	0.1726	0.0088	0.0112	**0.3335**
9	30	0.5349	0.0391	0.0000	**0.7063**	24	100	0.0500	0.0000	0.0016	**0.3003**
10	30	0.4202	0.0182	0.0002	**0.6089**	25	100	0.5043	0.0000	0.0000	**0.9807**
11	50	0.3003	0.0543	0.0069	**0.4308**	26	120	**0.8521**	0.0001	0.0000	0.7615
12	50	0.4624	0.0878	0.0000	**0.5466**	27	120	0.3102	0.0407	0.0000	**0.8603**
13	50	0.4525	0.0002	0.0001	**0.8219**	28	120	**0.6734**	0.0003	0.0000	0.3941
14	50	0.0769	0.0000	0.0000	**0.4195**	29	120	0.7482	0.1300	0.0186	**0.9218**
15	50	0.1226	**0.1451**	0.0728	**0.1451**	30	120	0.0000	0.0000	0.0000	**0.9331**

Note : The highest *P-value* for each dataset is in bold.

Table 16 *P-value* of the Anderson-Darling test for real data with left-skew distribution after the data had been transformed.

Dataset	Sample size	Transformation method				Dataset	Sample size	Transformation method			
		EF	DP	EP	BC			EF	DP	EP	BC
31	10	0.0298	0.1278	0.1374	**0.2359**	46	70	0.0008	0.0003	0.0713	**0.8847**
32	10	0.0028	0.0025	0.0060	**0.4440**	47	70	0.7360	0.4783	0.8006	**0.8268**
33	10	0.013	0.4291	**0.5282**	0.5272	48	70	0.9377	0.9608	0.9556	**0.9655**
34	10	0.0113	0.1428	0.1615	**0.8299**	49	70	0.0049	0.0018	0.1473	**0.4292**
35	10	0.0177	0.0271	0.0413	**0.0585**	50	70	0.3689	0.5774	0.6227	**0.7103**
36	30	0.0097	0.0679	0.1576	**0.2016**	51	100	0.0000	0.3695	0.0136	**0.4110**
37	30	0.0061	0.4837	0.1103	**0.4884**	52	100	0.0320	0.5336	0.7647	**0.9641**
38	30	0.0000	0.0181	0.0001	**0.1284**	53	100	0.0001	0.4249	0.0182	**0.4866**
39	30	0.0000	0.0056	0.0004	**0.0814**	54	100	0.0000	0.2523	0.0523	**0.4881**
40	30	0.0002	0.174	0.0042	**0.4928**	55	100	0.0001	0.0001	0.0022	**0.8052**
41	50	0.0149	0.1626	0.7533	**0.9061**	56	120	0.0010	0.0004	0.0531	**0.5780**
42	50	0.0346	0.1919	0.4358	**0.5045**	57	120	0.0008	0.0003	0.0507	**0.9212**
43	50	0.0091	0.1174	0.5536	**0.8094**	58	120	0.0019	0.0014	0.0110	**0.2515**
44	50	0.0107	0.5224	0.4022	**0.6506**	59	120	0.0011	0.0002	0.1454	**0.1809**
45	50	0.2041	0.3248	0.3653	**0.3443**	60	120	0.0509	0.0297	0.3416	**0.4634**

Note : The highest *P-value* for each dataset is in bold.

Dataset	Sample size	Histograms and curve
1	n=10	
6	n=30	
12	n=50	

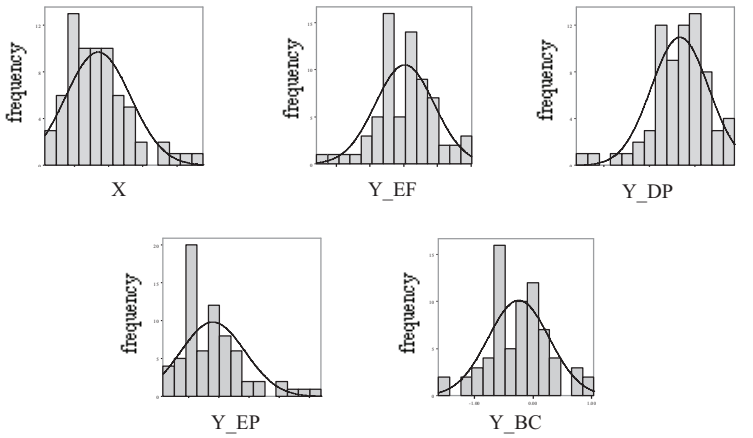
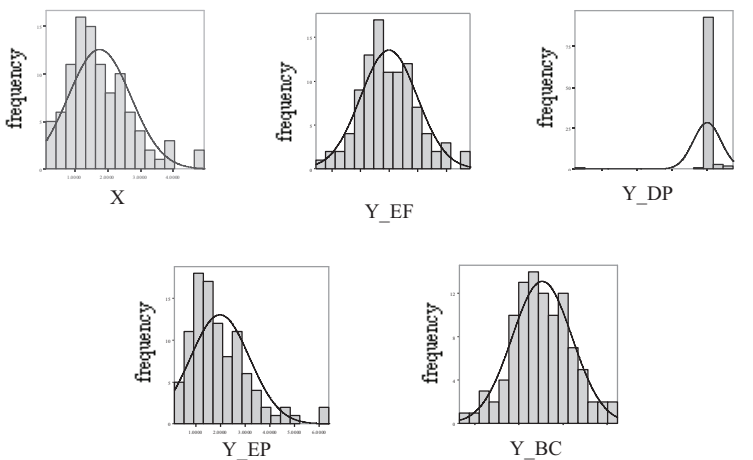
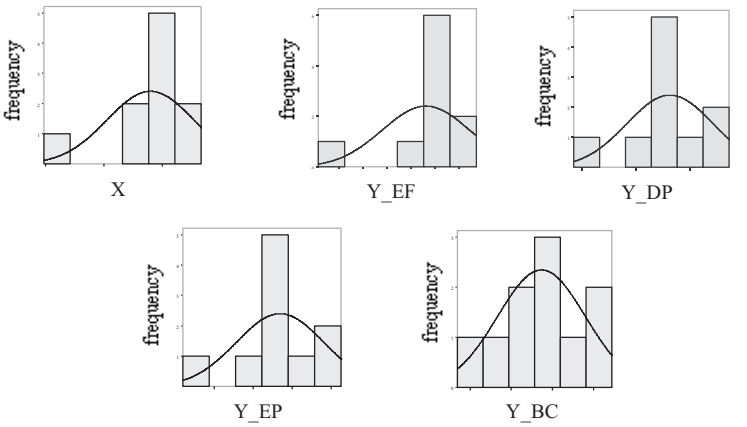
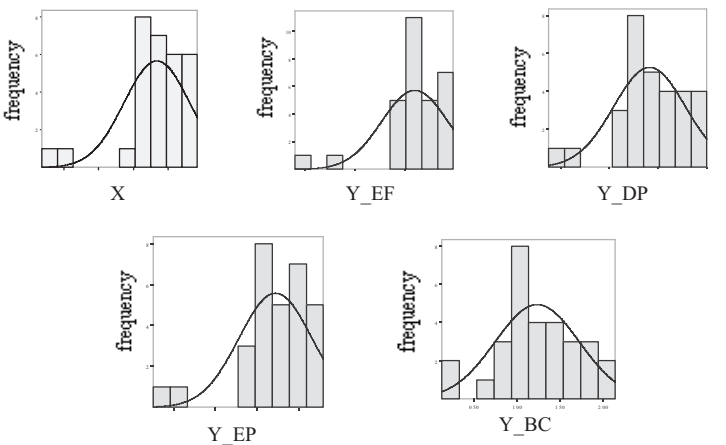
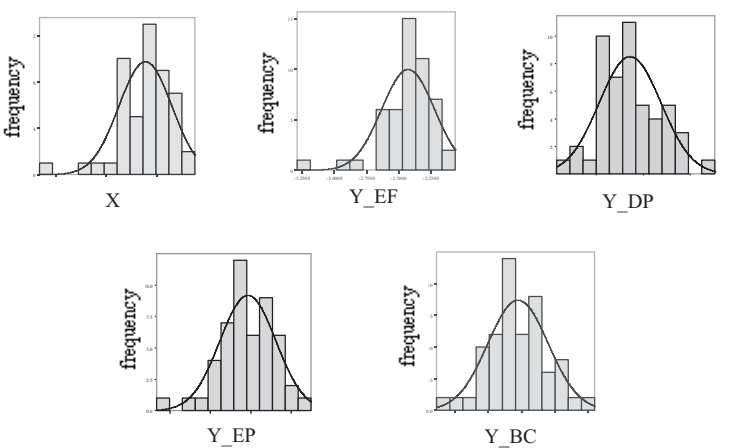
Dataset	Sample size	Histograms and curve
17	n=70	
25	n=100	

Figure 1 Histograms and curve for empirical right-skew data before and after each transformation method. X = untransformed dataset; Y_EF = EF transformed dataset; Y_BC = BC transformed dataset; Y_EP = EP transformed dataset; and Y_DP = DP transformed dataset.

Dataset	Sample size	Histograms and curve
31	n=10	
40	n=30	
44	n=50	

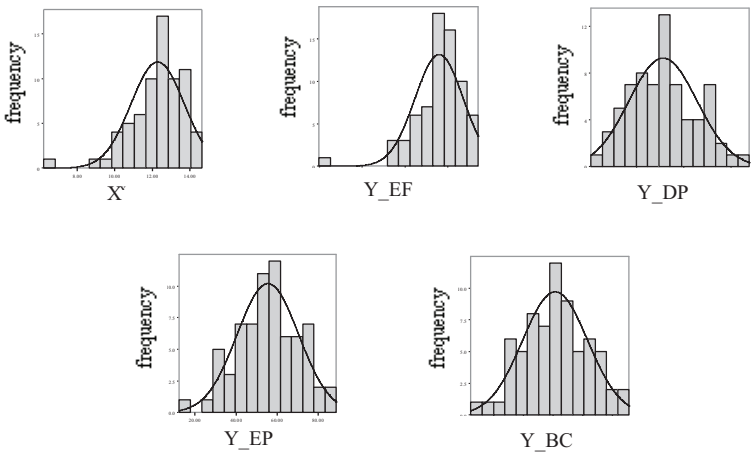
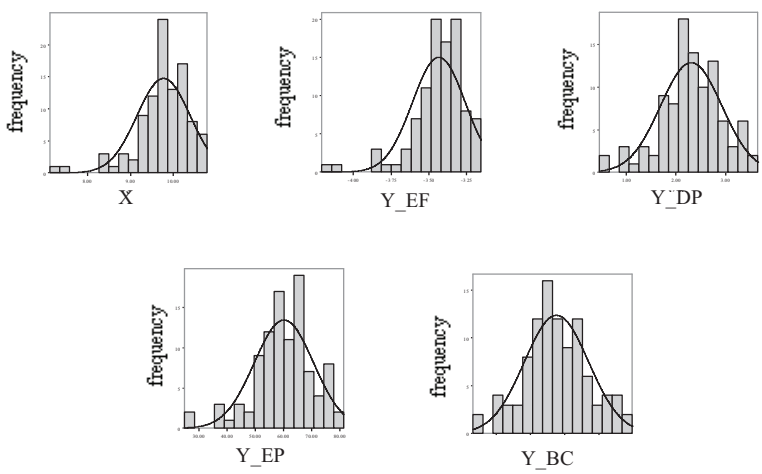
Dataset	Sample size	Histograms and curve
46	n=70	
54	n=100	

Figure 2 Histograms and curve for empirical left-skew data before and after each transformation method. X = untransformed dataset; Y_EF = EF transformed dataset; Y_BC = BC transformed dataset; Y_EP = EP transformed dataset; and Y_DP = DP transformed dataset.

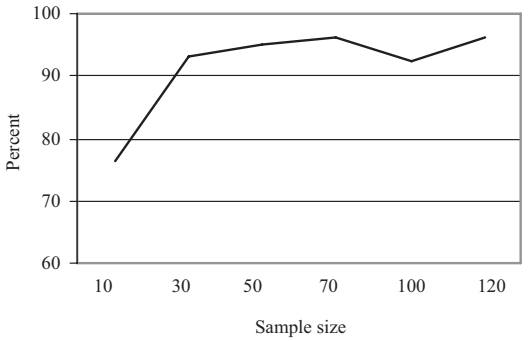


Figure 3 Trend of percentage acceptance of null hypothesis for BC transformation with skewness (0.3, 0.6] and kurtosis (2, 3] data.

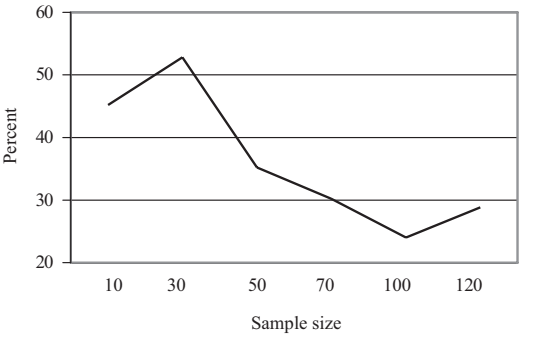


Figure 4 Trend of percentage acceptance of null hypothesis for EF transformation with skewness (0.3, 0.6] and kurtosis (1, 2] data.

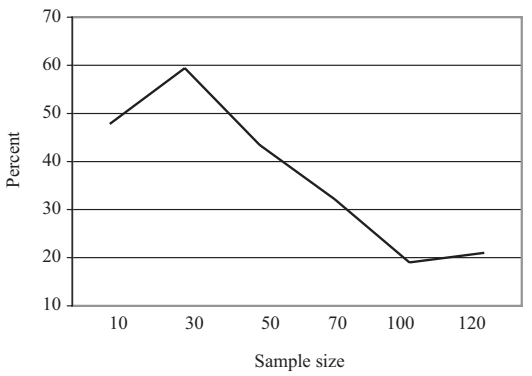


Figure 5 Trend of percentage acceptance of null hypothesis for DP transformation with skewness (0.3, 0.6] and kurtosis (1, 2] data.

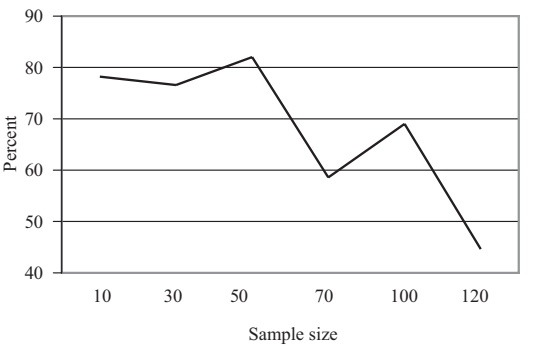


Figure 6 Trend of percentage acceptance of null hypothesis for EP transformation with skewness (0.3, 0.6] and kurtosis (2, 3] data.

ACKNOWLEDGEMENTS

The authors thanks the Siam Business Administration College, Nonthaburi for the real database (score database) used in the study.

LITERATURE CITED

Anderson, T.W. and D.A. Darling. 1952. Asymptotic theory of certain goodness-of-fit*f* criteria based on stochastic processes. **Ann. Math. Stat.** 23: 193˘212.

Box, G.E.P. and D.R. Cox, 1964. An analysis of transformations. **J. Roy. Statist. Soc. Series B.** 26: 211˘252.

Kerdsawang, S.T. 2005. **Comparison of Data Transformation Technigues for Normality.** M. Sc. Thesis Chulalongkorn University. Bangkok. Thailand.

Manly, B.F.J. 1976. Exponential data transformations. **J. Roy. Statist. Soc.** 25: 37˘42.

MathWorld. 1999. About the Inverse Erf, [cited: 18 August 2010]. [Available from: http://mathworld. wolfram. com/ Inverse Erf.html].

Rotwirat, A.W. 2008. **Correction of Non Normality for Response Observation in Randomized Complete Block Design**. Master. Science. Chulalongkorn University. Bangkok. Thailand.

van Albada, S.J. and P.A. Robinson. 2006. Transformation of arbitrary distributions to the normal distribution with application to EEG test-retest reliability. **J. Neurosci Methods.** 161: 205˘21 1.

Weibull, W. 1951. A statistical distribution function of wide applicability. **J. Appl. Mech.** 18: 293˘297.

Yang, Z. 2006. A modified family of power transformations. **Econ. Lett.** 92: 14˘19.

Text and Journal Publication Co., Ltd.
158/3 Soi Viphawadi Rangsit 5, Viphawadi Rangsit Road, Chom Phon, Chatuchak, Bangkok 10900
tel. 0-2617-8611-2 fax. 0-2617-8616