The Efficiency of the Three Data Transformation Methods for Converting Beta-Distributed Data to Normal-Distributed Data

Mallika Chanaphai*, Boonorm Chomtee and Ampai Thongteeraparp

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

* Corresponding author. E-mail address: mallika.c@ku.th

Received: 30 October 2025; Revised: 12 January 2026; Accepted: 15 January 2026; Available online: 23 January 2026

Abstract

The beta distribution is widely used to model rates and proportions in fields such as epidemiology, ecology, and quality control. However, relatively few studies have systematically compared data transformation methods specifically for beta-distributed data under different shapes and sample sizes. The research aimed to study and compare the efficiency of three data transformation methods: the Johnson transformation method, the Box-Cox transformation method, and the Yeo-Johnson transformation method for transforming beta-distributed data to normally distributed data. The study was conducted using the Monte Carlo simulation technique with 1,000 replications for each situation. In each scenario, the two shape parameters α and β of the beta distribution were defined as 5, 10, and 20 in combinations that generated three types of distributions: right-skewed $(\alpha < \beta)$, left-skewed $(\alpha > \beta)$, and symmetric $(\alpha = \beta)$ respectively; the 3 levels of sample size (n) were: small (n = 10), medium (n = 30, 50), and large (n = 70) and the power transformation (λ) were 0.2, 0.8, 1, and 2. The Anderson-Darling statistic was employed to examine the data distribution. The evaluation criterion was based on the percentage of acceptance (POA) of the null hypothesis for the transformed data to follow a normal distribution. A higher POA indicated a more effective transformation method. The results showed that, for right-skewed beta distributions, the Johnson transformation yielded the POA of normality, especially when the sample size was small. In contrast, the Box-Cox transformation became competitive for medium-sized samples. For left-skewed beta distributions, the Johnson transformation also performed best overall, with Box-Cox yielding comparable results as the sample size increased. For symmetric beta distributions, the Yeo-Johnson and Box-Cox transformations usually performed better than the Johnson transformation when the sample size was large. These findings suggest that selecting the appropriate transformation method based on the shape of the distribution and sample size is critical for improving the accuracy and validity of statistical analyses, especially when normality is a key assumption in methods such as ANOVA. Failure to apply the correct transformation may lead to mistakes or misunderstandings, which can ultimately result in inaccurate conclusions and compromise the validity of statistical tests.

Keywords: Beta Distribution, Normalizing Transformation, Johnson Transformation, Box-Cox Transformation, Yeo-Johnson Transformation

Introduction

Parametric statistical analyses, such as t-tests, analysis of variance (ANOVA), and regression analysis, are widely used as tools in statistical techniques. These methods are favored because they generally provide accurate and reliable results when compared to nonparametric statistics. However, parametric statistics rely on several key assumptions that must be strictly met, including normality of the data distribution, independence of observations, and homogeneity of population variances. Violation of these assumptions may lead to misleading results and reduce the credibility of the analysis.

In practice, datasets in the real world often do not meet these assumptions. For example, the data may be skewed or have a distribution that deviates from normality, making parametric statistics inappropriate. To address this issue, many researchers have proposed various data transformation techniques to modify the data distribution

so that it better aligns with the assumptions required by parametric methods. One commonly used technique is the Box-Cox transformation, which helps make data more normally distributed.

Previous studies have found that certain transformation methods can effectively adjust data to be more normally distributed. For instance, Chortirat et al. (2011) compared four transformation methods on data following a Weibull distribution. Their study evaluated various techniques for normalizing non-normal data. The four considered methods were the error function method, the dual power transformation, the exponential transformation by Manly, and the Box-Cox transformation. The results showed that, for large sample sizes, the Box-Cox transformation method yielded the highest percentage of normality acceptance. Similarly, Watthanacheewakul (2021) studied the eight transformation methods: reflect then logarithm base 10, reflect then square root, Box-Cox, reflect then Box-Cox, Manly, reflect then Manly, Yeo-Johnson, and reflect then Yeo-Johnson of left-skewed data for Weibull and beta distributions. The study found that the reflect then Manly transformation and the reflect then Yeo-Johnson transformation provided the highest acceptance rates in several scenarios. (Phureejarurot et al., 2020) compared the efficiency of data transformation methods for Gamma distributed data, specifically the Box-Cox transformation, power transformation, fourth root transformation, and exponential transformation of the Manly method. The results of the study revealed that the Box-Cox and power transformation methods provided the highest efficiency in most situations. The Manly transformation was most effective for small sample sizes; however, its efficiency declined as the sample size increased. Sumranhun et al. (2023) proposed "The Reduction of the Error in Forecasting Demand for Industrial Products with High Variation by Data Transformation Techniques." Their study focused on using data transformation techniques to reduce forecasting errors for industrial products with high variability. The study found that logarithmic transformation methods significantly reduced forecasting errors, especially when the data was highly skewed. This aligns with the current research, which also explores data transformations to improve the accuracy and reliability of statistical analyses. The findings of Sumranhun et al. (2023) further emphasize the importance of applying appropriate data transformation techniques, particularly when dealing with skewed data, to enhance the robustness of forecasting models and the validity of statistical results.

Building on these studies, this research focuses on the Johnson transformation due to its availability in widely used statistical software, such as Minitab, which is popular among researchers in both scientific and social sciences fields. The Johnson transformation offers a flexible approach for transforming a variety of distributions into normality. While previous studies have predominantly concentrated on left-skewed beta distributions, there has been limited exploration of symmetric and right-skewed beta distributions. This study aims to fill this gap by examining the effectiveness of the Johnson transformation, alongside the Box-Cox transformation, for a range of beta distribution shapes. Both methods are available in standard statistical software, such as SAS and Minitab, ensuring accessibility for a wide range of researchers. Additionally, the Yeo-Johnson transformation, a generalization of Box-Cox, is included in this study to explore its potential. While previous research has explored various transformation techniques, this study is the first to encompass the full spectrum of Beta distributions—including right-skewed, symmetric, and left-skewed—while providing a direct comparison between the Johnson transformation and other established methods. This unique approach highlights the contribution of this study to the field.

However, there are relatively few studies on beta distribution transformations. Beta distribution is commonly used in research involving proportions, probabilities, or percentages. In practice, beta distributions are frequently

used to model real-world data such as rates of disease occurrence, the proportion of land use, voting preferences, success probabilities, or percentages in quality control. Moreover, the beta distribution is highly flexible because its two shape parameters allow it to represent a wide variety of distributional forms, including right-skewed, left-skewed, and symmetric patterns. While this flexibility makes the beta distribution one of the most important tools for modeling bounded data within the interval (0, 1), it also creates analytical challenges when the data deviates strongly from normality.

Hence, this research aims to investigate the effectiveness of the three data transformation methods: Johnson transformation method, Box-Cox transformation method, and Yeo-Johnson transformation method in transforming a beta distribution to better approximate a normal distribution based on percentage of normality acceptance for right-skewed, left-skewed, and symmetric shapes under different parameter values and various sample sizes.

Materials and Methods

The objective of this research is to compare the effectiveness of three data transformation methods in addressing the issue of beta-distributed data by transforming it to approximate a normal distribution. The comparison is conducted through data simulations under various scenarios using R Studio version 4.3.3, as presented below:

1. The dataset x $(x>0)$ represents simulated population data from a beta distribution, with size N where the shape parameters α and β are defined as 5, 10, 20, and N is 10,000. Then, the skewness value was calculated by $\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$. The distributional characteristics of a beta distribution depend on the relationship between the two parameters (α, β), as follows:

right-skewed beta distribution when $\alpha<\beta$,

left-skewed beta distribution when $\alpha>\beta$,

symmetric beta distribution when $\alpha=\beta$.

The skewness values of a beta-distributed population data are presented in Table 1.

Table 1 The α, β and skewness values of the beta distributed population data of the data simulation

Shape	α	β	Skewness value
Right-skewed	5	10	0.333
	5	20	0.567
	10	20	0.747
Left-skewed	10	5	- 0.333
	20	5	- 0.567
	20	10	- 0.747
Symmetric	5	5	0
	10	10	0
	20	20	0

2. For the beta distributed population data, it was tested using the Anderson-Darling statistic (A) with the following steps:

2.1 The dataset x is sorted in ascending order.

2.2 Hypothesis Testing:

H_0 : The dataset x follows a beta distribution.

H_1 : The dataset x does not follow a beta distribution.

2.3 Calculate the Anderson-Darling statistic :

$$A = -\frac{\sum_{i=1}^{N}(2i-1)((\ln(F(x_i)) + \ln(1-F(x_{N+1-i})))}{N} - N ,$$

when $F(x_i)$ denotes the cumulative distribution function of the beta distribution,

N denotes the population size.

The PDF and CDF of the Beta distribution are used in the calculation:

PDF of Beta distribution:

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & ;0 < x < 1; \quad \alpha,\beta > 0 \\ 0 & ;otherwise \end{cases}$$

CDF of Beta distribution :

$$F(x) = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)}$$

2.4 At a significance level of 0.05, the null hypothesis is rejected if the test statistic A exceeds the critical value of 2.492 (D'Agostino & Stephens, 1986).

3. Randomly sample data of size n from the beta distributed population data for 3 levels: small (n = 10), medium (n = 30, 50), and large (n = 70).

4. Define the power transformation parameters (λ) as 0.2, 0.8, 1, and 2 for the Box-Cox and Yeo-Johnson transformation methods.

5. The data were transformed using the three methods, where the details are as follows:

5.1 Johnson transformation method

The Johnson transformation method was originally developed by Norman L. Johnson in 1949 and later discussed in detail by Slifker and Shapiro (1980). It was designed for distribution normalization using a family of transformation functions. The data were classified into different types, and an appropriate transformation was applied based on the distributional characteristics. Since the data in this study follow a beta distribution with defined bounds, the Johnson bounded system (SB) was applied. In cases where $\frac{hl}{c^2} < 1$, based on the calculated values of h, l, and c, the bounded system is appropriate. The formula is given as follows:

$$h = k_{3z} - k_z ,$$
$$l = k_{-z} - k_{-3z} ,$$
$$c = k_z - k_{-z} .$$

The values $k_{-3z}, k_{-z}, k_z, k_{3z}$ represent the positions (percentile points) in the ordered dataset that correspond to cumulative probabilities associated with $-3Z, -Z, Z, 3Z$, respectively. In this study, the significance level was set at 0.05. Therefore, the distribution covers a probability area of 0.95. For a confidence level of 95%, the corresponding Z -scores under the standard normal distribution are approximately -1.96 and

+1.96 which together cover 95% of the area under the curve. As a result, the transformed Z values used in the Johnson transformation are approximately Z = 0.98 and $-Z$ = − 0.98 (Slifker & Shapiro, 1980). The transformation steps are as follows:

$$y_{John} = \gamma + \eta \ln\left(\frac{x-\varepsilon}{\omega+\varepsilon-x}\right),$$

when y_{John} denotes the post-transformation data using Johnson transformation method,

x denotes the pre-transformation data.

$$\gamma = \eta \sinh^{-1}\left[\frac{\left(\frac{c}{l}-\frac{c}{h}\right)\left\{\left(1+\frac{c}{h}\right)\left(1+\frac{c}{l}\right)-4\right\}^{\frac{1}{2}}}{2\left(\frac{c}{h}\frac{c}{l}-1\right)}\right],$$

$$\eta = \frac{z}{\cosh^{-1}\left(\frac{1}{2}\left[\left(1+\frac{c}{h}\right)\left(1+\frac{c}{l}\right)\right]^{\frac{1}{2}}\right)},$$

$$\varepsilon = \frac{k_z + k_{-z}}{2} - \frac{\omega}{2} + \frac{c\left(\frac{c}{l}-\frac{c}{h}\right)}{2\left(\frac{c}{h}\frac{c}{l}-1\right)},$$

and $$\omega = \frac{c\left[\left\{\left(1+\frac{c}{h}\right)\left(1+\frac{c}{l}\right)-2\right\}^{2}-4\right]^{\frac{1}{2}}}{\frac{c^2}{hl}-1}.$$

5.2 Box-Cox transformation method

Box and Cox (1964) addressed violations of the normality assumption in parametric statistical analysis by introducing the Box-Cox transformation method. Its flexibility and key advantage lie in the fact that it is not a fixed transformation but is controlled by a power transformation parameter λ. This allows it to adapt precisely to the specific characteristics of each dataset. Furthermore, this method has integrated a commonly used historical transformation, the Log transformation, as an integral part of its system for the case where $\lambda = 0$. The Box-Cox transformation, when the data values are greater than zero, is computed as follows:

$$y_{BC} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & ;\lambda \neq 0, x > 0 \\ \log_{10}(x) & ;\lambda = 0, x > 0. \end{cases} \quad (1)$$

when y_{BC} denotes the post-transformation data using Box-Cox transformation method,

x denotes the pre-transformation data,

λ denotes the power transformation parameter.

The λ value is estimated from the data, typically using the maximum likelihood estimation method (MLE). In this approach, λ values are chosen to minimize the residual sum of squares (RSS) under the assumption of a linear model. The MLE estimator of λ is obtained by:

$$\hat{\sigma}(\lambda) = \frac{S(\lambda)}{n},$$

$$L_{\max}(\lambda) = -\frac{n}{2}\hat{\sigma}(\lambda) + \log\prod_{i=1}^{n}\left|\frac{dy_{BC_i}}{dx_i}\right|,$$

when $S(\lambda)$ denotes the residual sum of squares of the transformed variable y_{BC},

n denotes the sample size.

5.3 Yeo-Johnson transformation method

Yeo and Johnson (2000) proposed a power transformation method as an extension of the Box-Cox transformation. This method can be applied to data containing both positive and negative values. This transformation is designed to reduce skewness and improve the approximation of data to a normal distribution or symmetry. The key principle is the use of different functional forms depending on the sign of the data. For non-negative values $(x \geq 0)$, the method applies a modified form of the Box-Cox transformation. For negative values $(x < 0)$, it uses a different power parameter as shown in equation (2). This approach allows the method to handle both positive and negative data smoothly, while retaining properties similar to the Box-Cox transformation for positive observations. The transformation is computed as follows:

$$y_{YJ} = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda} & ;\lambda \neq 0, x \geq 0 \\ \ln(x+1) & ;\lambda = 0, x \geq 0 \\ \dfrac{-(-x+1)^{2-\lambda}-1}{2-\lambda} & ;\lambda \neq 2, x < 0 \\ -\ln(-x+1) & ;\lambda = 2, x < 0, \end{cases} \quad (2)$$

when y_{YJ} denotes the post-transformation data using Yeo-Johnson transformation method,

x denotes the pre-transformation data,

λ denotes the power transformation parameter.

6. After the data were processed with the three transformation methods, the adjusted Anderson-Darling statistic (A^*) was used to examine the normality with the following steps:

6.1 The post-transformation dataset y are sorted in ascending order.

6.2 Hypothesis testing:

H_0: The post-transformation dataset y follows a normal distribution.

H_1: The post-transformation dataset y does not follow a normal distribution.

6.3 Estimate the mean and variance of the normal distribution.

6.4 Calculate the test statistic A^*:

$$A = -\frac{\sum_{i=1}^{n}(2i-1)[(\ln(F(y_i)) + \ln(1-F(y_{n+1-i}))]}{n} - n,$$

when $F(y_i)$ denotes the cumulative distribution function of the normal distribution,

n denotes the sample size.

In this study, since the data were transformed and the population parameters μ_y and σ_y^2 are unknown, the Anderson–Darling test needs to be adjusted as follows:

$$A^* = A(1+\frac{0.75}{n}+\frac{2.25}{n^2}).$$

6.5 At a significance level of 0.05, for a normal distribution with unknown μ_y and σ_y^2 the null hypothesis is rejected if the adjusted Anderson–Darling test A^* exceeds the critical value of 0.752 (D'Agostino & Stephens, 1986).

7. Repeat the simulation 1,000 times for each scenarios.

8. Calculate the percentage of null hypothesis acceptance (POA):

$$POA = \frac{100 \times Number\ of\ times\ for\ the\ null\ hypothesis\ was\ accepted}{1,000}.$$

9. Compare the POA among the three data transformation methods.

10. The results were summarized and discussed.

Results

This study aimed to investigate and compare the efficiency of the three data transformation methods: Johnson, Box–Cox, and Yeo–Johnson transformations applied to the beta-distributed dataset to achieve normality. The performance of these methods was evaluated based on the percentage of acceptance (POA) of the null hypothesis of normality, using the adjusted Anderson–Darling test. The data were categorized into three types according to their distribution shape: right-skewed, left-skewed, and symmetric beta distributions. The results of POA values for the right-skewed, left-skewed, and symmetric beta distributions are shown in Tables 2 to 4, respectively. Since the Box–Cox and Yeo–Johnson transformation methods require specific conditions of the power parameter λ, as shown in equations (1) and (2), the POA values corresponding to these two methods are reported in parentheses. The POA values in bold indicate the highest POA among the methods compared.

The results are presented in two main sections:

1. Simulation

1.1. Explanation Based on Sample Size: This section discusses the results of the tests across different sample sizes, focusing on how the performance (POA) of each transformation method varies as the sample size changes.

1.2. Explanation Based on Distribution Characteristics This section explains the results according to the distributional characteristics of the data, which are influenced by the values of the α and β parameters in the beta distribution. It compares how each transformation method performs under various combinations of these parameters.

2. Application on Real Dataset

1. Simulation

For Table 2, the right-skewed beta distribution, when the sample size is 10, the Johnson transformation method provides the highest POA values across all parameter values, followed by the Box-Cox and Yeo-Johnson transformation methods, which exhibit very similar efficiencies. At moderate sample sizes such as 30 and 50, the Box-Cox transformation method generally outperforms the Johnson and Yeo-Johnson transformation methods, showing slightly higher POA values across most parameter values. At the largest sample size of 70, the Johnson transformation method maintains relatively stable performance for certain parameter combinations, but the Box-Cox transformation method demonstrates consistently higher efficiency compared to the Yeo-Johnson transformation method, which exhibits a noticeable decline, particularly in more skewed distributions.

Table 2 The POA values for normality under the right skewed beta distribution for sample size n = 10, 30, 50, 70.

(α, β)	n	Transformation Methods		
		Johnson	**Box-Cox**	**Yeo-Johnson**
(5, 10)	10	**98.7**	96 (0.8)	96 (0.2)
	30	93.6	**94.9** (0.8)	94.7 (0.2)
	50	90.9	**94.4** (0.8)	93.9 (0.2)
	70	89.5	**92.9** (0.8)	92.1 (0.2)
(5, 20)	10	**98.6**	94.2 (0.8)	94 (0.2)
	30	92.3	**93.2** (0.2)	90.3 (0.2)
	50	90.5	**90.6** (0.2)	83.3 (0.2)
	70	**90**	86.3 (0.2)	77.6 (0.2)
(10, 20)	10	**98.1**	94.9 (1/0.8)	95 (0.8)
	30	92.9	**95.9** (0.8)	95.8 (0.2)
	50	91	**95** (0.8)	94.9 (0.2)
	70	89.7	94.4 (0.8)	**94.5** (0.2)

From the perspective of the beta distribution parameters, when the two parameters are (5, 10), the Johnson transformation method is the most efficient at small sample sizes, while the Box-Cox transformation method becomes more effective as the sample size increases, with Yeo-Johnson following closely. For more skewed cases such as (5, 20) and (10, 20), the Johnson transformation method clearly outperforms the others at sample size is 10 and 70, but both the Box-Cox transformation method and the Yeo-Johnson transformation method experience reduced efficiency as the sample size grows, with the Yeo-Johnson transformation method showing the sharpest decline.

For Table 3, the left-skewed beta distribution, when the sample size is 10, the Johnson transformation method consistently provides the highest POA values across all parameters, with the Box-Cox transformation method and the Yeo-Johnson transformation method showing slightly lower but comparable efficiencies. At sample sizes of 30 and 50, the Box-Cox transformation method and the Johnson transformation method generally outperform the Yeo-Johnson transformation method, with higher POA values in most cases. At the largest sample size of 70, the Yeo-Johnson transformation method achieves the highest POA values, while the Johnson transformation method and the Box-Cox transformation method remain comparable but exhibit slightly lower efficiency in some parameter settings.

Table 3 The POA values for normality under the left skewed beta distribution for sample size n = 10, 30, 50, 70.

(α,β)	n	Transformation Methods		
		Johnson	**Box-Cox**	**Yeo-Johnson**
(10, 5)	10	**98.7**	96.6 (2)	96 (2)
	30	93.6	**95.5** (2)	94.6 (2)
	50	90.9	**94.1** (2)	93.7 (2)
	70	89.5	91.9 (2)	**92** (2)
(20, 5)	10	**98.6**	94.9 (2)	93.7 (2)
	30	92.3	**93** (2)	89.3 (2)
	50	**90.5**	88.9 (2)	81.8 (2)
	70	**90**	87.3 (2)	76.9 (2)
(20, 10)	10	**98.1**	95.9 (2)	94.9 (0.8/1/2)
	30	92.9	**97** (2)	95.8 (2)
	50	91	**95.6** (2)	95 (2)
	70	89.7	94.1 (2)	**94.4** (2)

From the perspective of the beta distribution parameters, when the parameters are unbalanced, such as (10, 5) and (20, 10), the Box-Cox transformation method demonstrates the highest POA values, followed by the Johnson and Yeo-Johnson transformations method. However, when the parameters are more highly unbalanced, as in the case of (20, 5), the Johnson transformation method provides superior performance, while the Box-Cox transformation method ranks second. Overall, the Johnson transformation exhibits greater stability across different parameter and sample sizes.

For Table 4, the symmetric beta distribution, n = 10, the Johnson transformation method provides the highest POA values across all parameter combinations, followed by the Box-Cox transformation method and the Yeo-Johnson transformation method, which perform at very similar levels. At sample sizes of 30 and 50, the Box-Cox transformation method and the Yeo-Johnson transformation method generally outperform the Johnson transformation method, with nearly identical POA values across most parameter settings. At the largest sample size of 70, the Johnson transformation method achieves the highest POA values in some parameter combinations, while the Box-Cox transformation method and the Yeo-Johnson transformation method remain competitive but show slightly lower efficiency in those cases.

Table 4 The POA values for normality under the symmetric beta distribution for sample size n = 10, 30, 50, 70

(α,β)	n	Transformation Methods		
		Johnson	**Box-Cox**	**Yeo-Johnson**
(5, 5)	10	**99.1**	95.3 (0.8)	95.4 (0.8)
	30	91.9	**94.9 (0.8)**	94.5 (1)
	50	92.9	93.7 (0.8/1)	**94.3 (0.8)**
	70	92.7	93.7 (1)	**93.8 (0.8)**
(10, 10)	10	**98.1**	95.6 (1)	95.6 (1)
	30	92.3	**95.4 (1)**	**95.4 (1)**
	50	91.3	**94.8 (1)**	**94.8** (0.8/1)
	70	90.8	**95.6 (1)**	**95.6 (1)**

Table 4 (Cont.)

(α, β)	n	Transformation Methods		
		Johnson	**Box-Cox**	**Yeo-Johnson**
(20, 20)	10	**98.5**	94.2 (0.8)	94.2 (0.2/0.8)
	30	92.2	**96.2 (1)**	**96.2 (2/1)**
	50	92.1	**97 (1)**	**97 (1)**
	70	90.9	**96.9 (1)**	**96.9 (1)**

From the perspective of the beta distribution parameters, when the two parameters are small, such as (5, 5), the Johnson transformation method dominates at the smallest sample size, while the Box-Cox transformation method and the Yeo-Johnson transformation method become slightly more efficient at larger sample sizes. At moderate to larger parameters such as (10, 10) and (20, 20), the Box-Cox transformation method and the Yeo-Johnson transformation method exhibit higher overall efficiency, while the Johnson transformation method occasionally achieves the highest POA values depending on the sample size and parameter combination.

The results discussed above are clearly illustrated in Figs. 1 to 3. The graphs display the POA values for each transformation method, making it easy to see trends across different sample sizes and parameter values. The graphs are presented as follows.

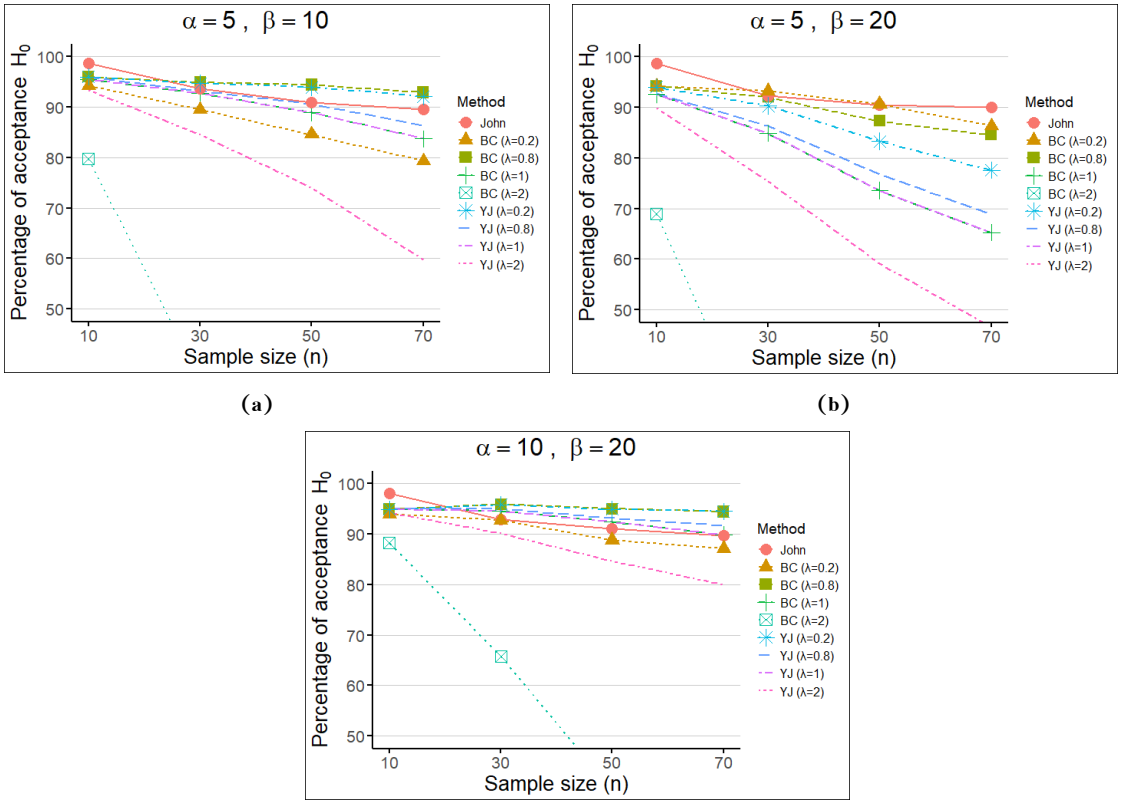


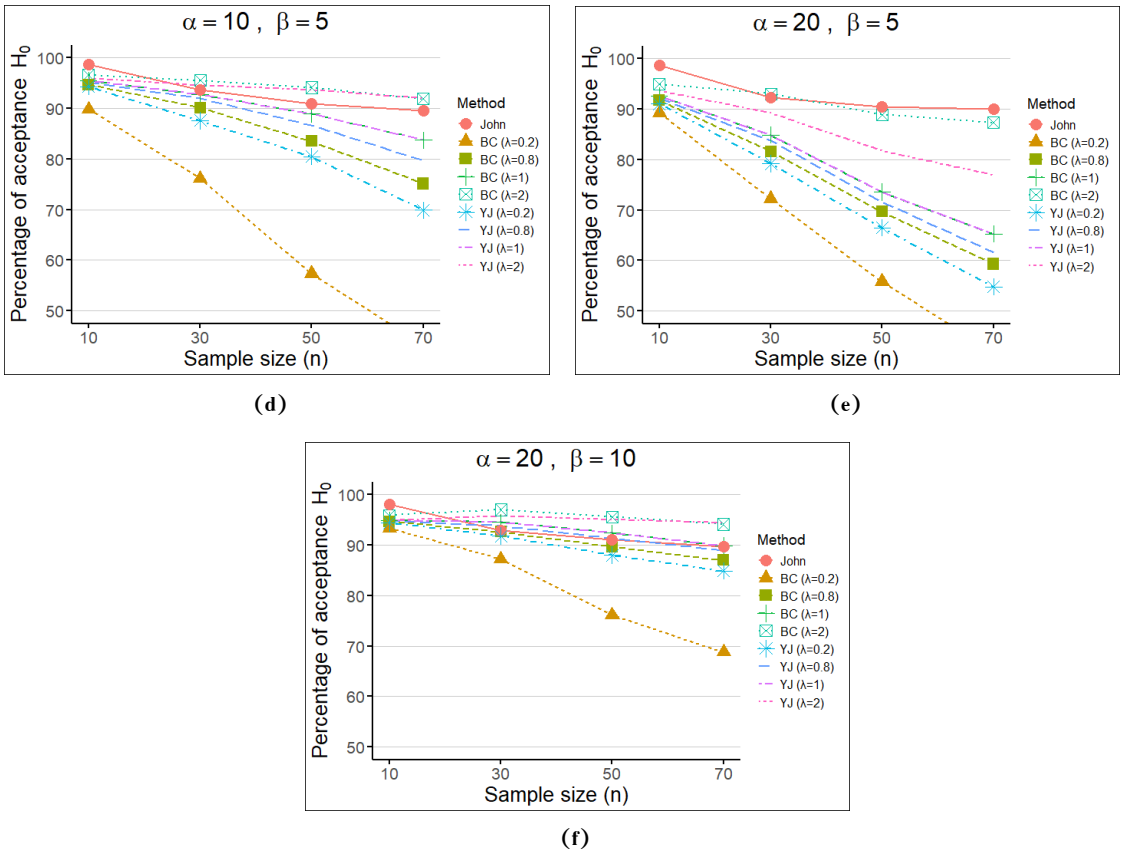
Figure 1 POA values for the right-skewed beta distributions (a-c)

Note: Only POA values greater than 50 are shown

The Fig. 1 shows the POA for beta distributions under right-skewed conditions, evaluated across different sample sizes (n) and transformation methods. The Johnson transformation method consistently achieved the highest and most stable POA values across all parameters, with only minimal decline as the sample size increased.

The Box–Cox transformation method demonstrated competitive performance, often surpassing the Johnson transformation method at moderate to larger sample sizes, particularly when suitable λ values were applied. In contrast, the Yeo–Johnson transformation method showed greater sensitivity to λ, performing comparably to the Box–Cox transformation method in some cases but declining noticeably at larger sample sizes, especially under stronger skewness. Overall, Johnson transformation method offered the most stability, Box–Cox provided efficiency gains with increasing sample size, and the Yeo–Johnson transformation method was the least robust to skewness and parameter selection.

The Fig.2 shows the POA values for beta distributions under left-skewed conditions. The Johnson transformation method consistently produced the highest POA values and demonstrated strong stability across all parameter values, with only slight reductions as the sample size increased. The Box–Cox transformation method exhibited more variability, while it performed comparably to the Johnson transformation method under mild skewness. Its performance declined substantially when the degree of skewness increased, particularly at larger sample sizes. The Yeo–Johnson transformation method also showed sensitivity to parameter choices and skewness, generally performing worse than both Johnson and Box–Cox methods, with sharper decreases in POA as the sample size grew. Overall, the Johnson transformation method maintained the most robust performance, Box–Cox was effective under certain conditions but less stable under stronger skewness, and the Yeo–Johnson transformation method was the least consistent across parameter values.



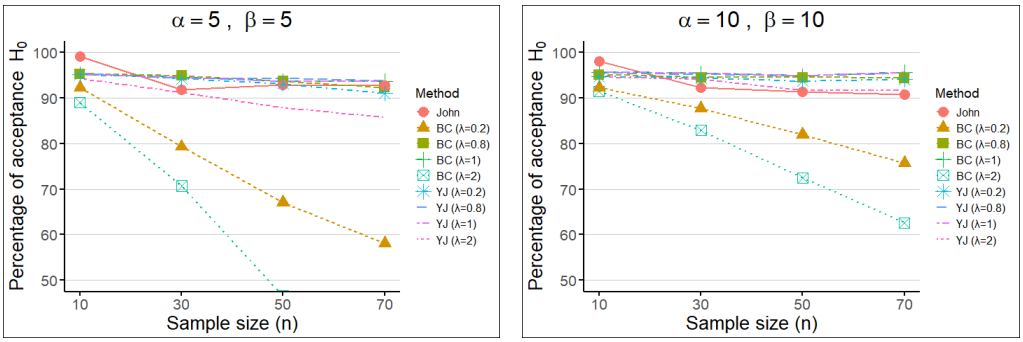
(d) (e)

(f)

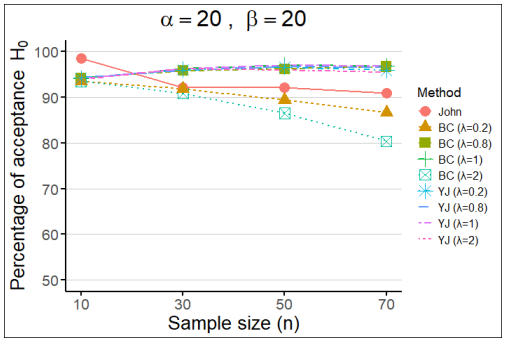
Figure 2 POA values for the left-skewed beta distributions (d–f)

Note: Only POA values greater than 50 are shown

The Fig.3 shows the POA values for beta distributions under symmetric. The Johnson transformation method maintained stable and consistently high POA values across all sample sizes. The Box-Cox and Yeo-Johnson transformation methods also performed reliably when λ values = 0.8 and 1 were applied, though λ value = 2 resulted in noticeable declines in POA as the sample size increased. Overall, under symmetric beta distributions, differences between the transformation methods were minimal, with Johnson showing slightly greater stability.



(g)



(i)

Figure 3 POA values for the symmetric beta distributions (g-i)

Note: Only POA values greater than 50 are shown

For Figure 4 in this study, simulated data under a total of 144 scenarios, varying the parameters α = 5, 10, 20, β = 5, 10, 20, λ = 0.2, 0.8, 1, 2, and n = 10, 30, 50, 70. It was found that the Johnson transformation method yielded the highest number of scenarios with the greatest acceptance rate of the null hypothesis, totaling 83 cases. The Yeo-Johnson transformation method ranked second with 31 scenarios showing the highest acceptance rates, mostly observed when the data were symmetric. The Box-Cox transformation method had a comparable number of scenarios with the highest acceptance rate, totaling 30 cases.

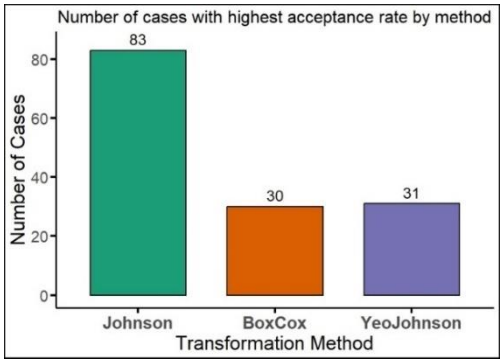


Figure 4 Number of cases with the highest acceptance rate

2. Application on Real Dataset

In this study, a real dataset on the morbidity rates of hypertension per population in Region Health 3 (Health Data Center, 2025) of Thailand during the years 2015-2024 (n = 10) were analyzed. The data were categorized into five age groups: under 15 years, 15-39 years, 40-49 years, 50-59 years, and over 60 years. Preliminary analysis showed that the data in each age group could be described by the beta distribution at the 0.05 significance level. Most age groups exhibited right-skewed distributions, particularly the group under 15 years, which showed strong right skewness (skewness = 0.755). The skewness values were calculated from $\frac{2(\hat{\beta}-\hat{\alpha})\sqrt{\hat{\alpha}+\hat{\beta}+1}}{(\hat{\alpha}+\hat{\beta}+2)\sqrt{\hat{\alpha}\hat{\beta}}}$. In contrast, the group over 60 years exhibited a nearly symmetric distribution, with a skewness value of −0.005, indicating slight left skewness. The parameters of the beta distribution were estimated using the method of moments estimation (MME), as shown in Table 5.

Table 5 The beta parameters estimate, skewness, mean, variance, and Anderson-Darling test statistics for pre-transformation of the morbidity rates of hypertension data

Age	$\hat{\alpha}$	$\hat{\beta}$	skewness	mean	variance	A	p-value (A)
<15	6.99	5711.96	0.755	0.001	2.14×10^{-7}	0.6303	0.6155
15-39	319.62	17291.15	0.389	0.018	1.01×10^{-6}	0.1682	0.9972
40-49	281.92	1908.16	0.095	0.129	5.12×10^{-5}	0.2279	0.9812
50-59	296.05	762.05	0.060	0.280	1.90×10^{-4}	0.3035	0.9343
>60	100.84	97.50	- 0.005	0.508	2.14×10^{-3}	0.444	0.8001

Since the results from the simulated data indicated that the Johnson transformation method is the most effective based on POA, the Johnson transformation method was therefore applied to the real dataset as an example. The Anderson-Darling test (A) results confirmed that at a 0.05 significant level, each age group was a beta distribution (Table6). To evaluate group differences in mean morbidity rates, ANOVA was performed on both pre- and post-transformation data. However, ANOVA requires the assumption of normality. The ANOVA results from the pre-transformation data indicated a statistically significant difference in mean morbidity rates among the age groups ($F = 1490$, p-value < 0.001). On the other hand, after applying the Johnson transformation method, which aims to normalize the data, the ANOVA test revealed no significant difference ($F = 0.066$, p-value = 0.992). Moreover, the pre-transformation data exhibited substantial heterogeneity of variances across age groups, violating the homogeneity of variance assumption. In contrast, the Johnson transformation effectively stabilized the variances, with post-transformation values confined to a narrow range, thereby improving both normality and variance homogeneity, and ensuring the validity of ANOVA inference. In addition, the Box-Cox and Yeo-Johnson transformation methods were also evaluated. Both approaches successfully transformed the data toward approximate normality; however, they exhibited limited effectiveness in stabilizing the variance across groups when compared with the Johnson transformation. This outcome provides further statistical evidence supporting the superior performance of the Johnson transformation in simultaneously achieving distributional normality and homogeneity of variances.

Table 6 The mean, variance and the adjusted Anderson-Darling test statistics for post-transformation of the morbidity rates of hypertension data transformed using the Johnson transformation method

Age	mean	variance	A^*	P-value (A^*)
<15	- 0.272	2.00	0.5032	0.2050
15-39	0.015	2.52	0.2701	0.6774
40-49	- 0.038	2.60	0.2596	0.7129
50-59	0.001	2.85	0.3772	0.4096
>60	0.052	2.67	0.2946	0.5984

The comparison of ANOVA results before and after the Johnson transformation method is summarized in Table 7. These findings highlight the importance of checking the distributional properties of real data and applying suitable transformation methods before conducting statistical analyses that assume normality, such as ANOVA. Appropriate data transformation helps ensure valid and reliable inference.

Table 7 Comparing the ANOVA results between pre and post transformation of the real data

ANOVA	F	p-value
Pre-transformation data	1490	<0.001
Post-transformation data	0.066	0.992

Discussion

This study demonstrates the efficiency of the three data transformation methods, namely the Johnson transformation method, the Box-Cox transformation method, and the Yeo-Johnson transformation method in transforming beta-distributed data to approximate a normal distribution. The comparison was based on the percentage of null hypothesis acceptance under various beta parameters and sample sizes, specifically: α = 5, 10, 20 , β = 5, 10, 20 , λ = 0.2, 0.8, 1, 2, and n = 10, 30, 50, 70. This analysis provides important insights into the appropriateness of each method under different conditions.

In cases of symmetric distribution, the Johnson transformation method typically yielded the highest POA value when the sample size was small. This highlights the method's strong ability to handle limited data effectively, as the Johnson transformation method does not require a specifically defined power parameter, unlike the Box-Cox transformation method and the Yeo-Johnson transformation method, both of which can produce unstable results with small sample sizes. This finding aligns with the study of Watthanacheewakul (2021). As the sample size increased, the performance of both the Box-Cox and Yeo-Johnson transformation methods improved. However, for sample sizes = 30, 50, 70, the POA value for the Johnson transformation method tended to be lower than for the other methods. This may be because the high flexibility of the Johnson method, which is advantageous in small sample situations, becomes less beneficial when more data are available.

The Box-Cox transformation method was found to be highly effective when the sample size was medium (n = 30, 50), which is consistent with the findings of Chortirat et al. (2011). Its performance was especially strong when the data were only moderately skewed, and when the power transformation parameter could be properly adjusted. In particular, for skewness λ was approximately 0.2 or 0.8 for right-skewed distributions. These results are consistent with the theoretical principles underlying the Box-Cox transformation.

In contrast, the Yeo-Johnson transformation method offers a notable advantage: it can be applied to data with zero or negative skewness. In this study, the Yeo-Johnson transformation method produced results comparable to, or in some cases equal to, those of the Box-Cox transformation method, especially when the power transformation parameter (λ) was suitably chosen, such as λ = 0.2 or 0.8. However, for highly skewed data, the Yeo-Johnson transformation exhibited reduced performance. Conversely, as the sample size increased, its effectiveness tended to improve, although POA remained lower than that of the other methods. This pattern may reflect the method's limitations in handling strongly skewed data, even though it performs well for symmetric distributions. Additionally, the optimal power transformation parameter was found to be [0.8, 1] for symmetric distributions, 0.2 for right-skewed distributions, and exactly 2 for left-skewed distributions.

In the data transformation process, the Johnson transformation method incorporates the boundaries of the data, which enhances its ability to handle data with symmetric distributions that are closer to a normal distribution, especially when the sample size is small or the data resembles a normal distribution. In contrast, the Box-Cox and Yeo-Johnson transformations rely solely on power transformation, a process that does not handle symmetric data or data close to a normal distribution as effectively as the Johnson method. As a result, the outcomes from both of these methods are unable to compete with the Johnson transformation in such scenarios.

Conclusion and Suggestions

Previous studies have explored data transformation methods for various distributions to compare the efficiency of each method, such as transformations for normal distributions or slightly skewed distributions. However, research comparing data transformation methods for right-skewed, left-skewed, and symmetric distributions remains limited. Therefore, this study aims to fill this gap by comparing the three transformation methods in handling data with different types of skewness, including right-skewed, left-skewed, and symmetric data. This expands the understanding and applicability of appropriate data transformation methods for various scenarios.

This study aimed to compare the efficiency of the three data transformation methods: the Johnson transformation method, the Box-Cox transformation method, and the Yeo-Johnson transformation method, in transforming beta-distributed data into a normal distribution. The findings indicate that when the sample size is small, the Johnson transformation method yields the highest performance. For medium sample sizes, the Box-Cox transformation method tends to result in the highest percentage of null hypothesis acceptance in many scenarios, particularly when the beta parameters indicate high skewness. For large sample sizes, the Yeo-Johnson transformation method performs best overall. Moreover, when the data are symmetric or have zero skewness, the Yeo-Johnson transformation method performs well in normalizing the data.

The Johnson transformation method determines the boundaries from the data pre-transformation, providing high flexibility. The results can be both positive and negative, allowing it to handle data with diverse characteristics more effectively. In contrast, the Box-Cox and Yeo-Johnson transformations method depend on the power transformation and the restriction that x must lie between 0 and 1, which forces the data to be transformed in a single direction. As a result, the Johnson transformation method tends to be more stable and suitable in some cases compared to the Box-Cox and Yeo-Johnson transformation methods.

Based on these findings, the following suggestions are proposed:

1. The choice of transformation method should be based upon the sample size and the initial distribution characteristics of the data, especially the beta distribution parameters. Small sample size (n = 10): Johnson transformation is recommended. Medium sample size (n = 30, 50): Box-Cox transformation is recommended. For data with a symmetric distribution and a large sample size (n = 70), the Yeo-Johnson transformation is the most appropriate method.

2. The power transformation parameter (λ) should be appropriately specified for the dataset to achieve optimal performance. An appropriate method should be used to estimate the optimal λ value.

3. Future research should explore the performance of these transformation methods on other distributions, such as geometric (commonly used in reliability studies and social sciences) and exponential (widely applied in engineering and survival analysis), to evaluate the effectiveness of transformation methods across various practical contexts.

4. It is recommended to incorporate multiple normality tests, such as the Shapiro-Wilk test or the Kolmogorov-Smirnov test, to enhance the accuracy and robustness of the evaluation.

Acknowledgments

Sincere appreciation is first extended to the Development and Promotion of Science and Technology Talents Project Scholarship (DPST) for its financial support. The researcher is also deeply grateful to the Department of Statistics, Faculty of Science, Kasetsart University, Bangkhen, for academic guidance and support, and to HDC Service (Health Data Center), Ministry of Public Health, Thailand, for providing access to a real-world dataset that was crucial to the analysis and overall success of this research.

Author Contributions

Author 1: Collected data; performed data analysis and interpretation; contributed to manuscript writing, review, and editing.

Author 2: Contributed to conceptualization and methodology design; interpreted data; involved in manuscript writing, review, and editing.

Author 3: Participated in methodology design; reviewed and edited the manuscript.

Conflict of Interests

All authors declare that they have no conflicts of interest.

Funding

This study was supported by the Development and Promotion of Science and Technology Talents Project (DPST) and by the Department of Statistics, Faculty of Science, Kasetsart University, Bangkhen, Thailand.

References

Chortirat, T., Chomtee, B., & Sinsomboonthong, J. (2011). Comparison of four data transformation methods for weibull distributed data. *Agriculture and Natural Resources*, *45*(2), 366-383.

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. Marcel Dekker.

Health Data Center, M. o. P. H. (2025). *Morbidity Rates of Hypertension per Population*. https://hdc.moph.go.th/center/public/

Phureejarurot, P. P., Worawan., Tresutumas, N., Raksakom, S., & Sinsomboonthong, J. (2020). Efficiency Comparison of Data Transformation Methodsfor the Gamma Distributed Data. *Thai Science and Technology Journal*, *28*(8), 1321-1333.

Slifker, J. F., & Shapiro, S. S. (1980). The Johnson system: selection and parameter estimation. *Technometrics*, *22*(2), 239-246.

Sumranhun, P., Rungcharoen, S., Damrongwatthanayothin, B., Somtobburana, P., Patsaphat, B., & Nuchpum, T. (2023). The Reducing of the Error in Forecasting Demand for Industrial Products with High Variation by Data Transformation Techniques. *Journal of Management and Marketing, Rajamangala University of Technology Thanyaburi*, *10*(1), 36-54. https://so05.tci-thaijo.org/index.php/mmr/article/view/261443/178489

Watthanacheewakul, L. (2021, July 7-9). Transformations for left skewed data. *Proceedings of the World Congress on Engineering 2010* (pp. 1-6). London, U.K.

Yeo, I.-K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, *87*(4), 954-959. https://doi.org/10.1093/biomet/87.4.954