A New Approach for Normalizing Continuous Data, Applicable in Parametric and Nonparametric Continuous Studies

M.M Saber[1], M. Taghipour[2*], M. Salehi[2], H.M Yousof[3]

[1] *Department of Statistics, Higher Education Center of Eghlid, Eghlid, Islamic Republic of Iran.*
[2] *Department of Statistics, Faculty of Sciences, University of Qom, Qom, Islamic Republic of Iran.*
[3] *Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University, Egypt*

Received: 5 November 2024 / Revised: 31 May 2025 / Accepted: 13 July 2025

Abstract

Satisfying the normality assumption is fundamental to many statistical inferences, as its violation can significantly affect the validity and reliability of conclusions drawn from the data. In this paper, we introduce a novel method for normalizing data that applies to both parametric and non-parametric cases. This method is grounded in a refined version of the empirical distribution function (EDF), which enhances its flexibility and accuracy compared to traditional normalization techniques. By leveraging this new EDF formulation, our approach effectively addresses common issues associated with existing methods, such as sensitivity to outliers and the inability to handle skewed distributions efficiently. A key advantage of our technique is its reversibility, which enables normalized data to be effortlessly transformed back into their original form, thereby preserving the integrity of the raw data for further analysis or interpretation. To demonstrate the efficacy of our method, we evaluate its performance using multiple real-world examples, including datasets related to the COVID-19 pandemic. These datasets, characterized by their complexity and variability, provide a rigorous test of the proposed normalization approach. The results confirm that our method successfully normalizes the data while maintaining their underlying structure and relationships, thus improving the robustness of subsequent statistical analyses. This innovation not only expands the toolkit available for data preprocessing but also enhances the applicability of standard statistical techniques to a broader range of real-life datasets.

Keywords: Normality assumption; Box-Cox transformation; Yeo-Johnson Transformation; Empirical distribution function.

Introduction

Statistical distributions are important in practical fields such as reliability engineering, medicine, and finance. The normal distribution is particularly important among these distributions, especially for classical statistical analysis, including confidence intervals, hypothesis testing, and regression analysis. These

[*] Corresponding Author: Tel: 09171177616; Email m.taghipour@qom.ac.ir

parametric methods rely heavily on the assumption of normality (1). However, it is important to acknowledge that in certain cases, the central limit theorem cannot be used to assume the normality of the data.

Many researchers often aim to develop new models that retain key features of the original model while also adhering to all necessary assumptions. This can involve techniques such as applying appropriate transformations to the data or filtering out questionable data points that may be regarded as outliers. This approach has been discussed in the works of Thöni (2) and Hoyle (3), among others. Nevertheless, there are techniques available to transform data and align it with the normal distribution. By utilizing such transformations, statistical analysis can be tailored to better suit the specific data and take advantage of the benefits associated with the normal distribution.

Two popular methods for data normalization, the Box-Cox (4, 5) and Yeo-Johnson transformations (6) exist. However, these methods have limitations that may restrict their applicability. For example, the Box-Cox transformation assumes positive and continuous data, with the existence of variance being a crucial condition. Applying this transformation to heavy-tailed data can be problematic and render the results invalid due to the absence of variance in such distributions. Additionally, both Box-Cox and Yeo-Johnson transformations can be substantially influenced by extreme outliers in the dataset, potentially distorting the shape and resulting in distorted normalized data.

Furthermore, while the goal of these transformations is data normalization, the transformed values obtained are not always easy to interpret. Consequently, their use may require additional explanation or conversion back to the original scale for meaningful inference. To overcome these challenges, continued exploration and refinement of normalization techniques are necessary. Identifying and developing new methods that mitigate the limitations of the Box-Cox and Yeo-Johnson transformations would contribute to advances in data normalization and enhance their overall applicability across diverse datasets.

This topic has garnered significant attention from researchers in recent years. For instance, in the paper (8), the Box–Cox transformation a fundamental tool in statistical modeling is comprehensively reviewed and further developed. The authors provide an in-depth exploration of its historical evolution and diverse applications. Additionally, they introduce an extended Yeo-Johnson transformation, which allows for separate power transformations for positive and negative response values. The necessity of this extension is demonstrated through robust data analysis, highlighting its utility in addressing asymmetry and heteroscedasticity in datasets.

Furthermore, Riani et al. (9) propose an automated approach for applying robust versions of the Box–Cox and extended Yeo-Johnson transformations to regression models. This method ensures that the response variable achieves approximate normality, even when it contains both positive and negative values. By incorporating robust statistical techniques, their approach mitigates the influence of outliers and enhances the reliability of model assumptions, thereby improving the overall validity of regression analyses.

In this study, we propose a new transformation for normalizing data that works accurately for all data (positive and negative) with a parametric continuous distribution. This transformation is based on a continuous version of the empirical distribution function (EDF) and can be applied to nonparametric data with the same efficiency. It is also applicable for distributions that do not have a cumulative distribution of closed form. The details of the proposed method, in two cases parametric and nonparametric, are presented in Sections 3 and 4 respectively.

Materials and Methods

In this section, we will examine some techniques for normalizing data, after which we will introduce our proposed method.

1. The Current methods

1.1 The Box-Cox transformation

Tukey (7) introduced a family of power transformations designed to ensure that the transformed values represent a monotonic function of the observations within a suitable range, typically indexed by

$$x_t^{(\lambda)} = \begin{cases} \log(x_t), & \lambda = 0 \\ x_t^{\lambda}, & \lambda \neq 0 \end{cases}$$

for $x_t > 0$.

However, this family was subsequently refined by Box & Cox (1) to address the discontinuity that arises at $\lambda = 0$.

The Box-Cox transformation is a widely used statistical technique for normalizing data. It involves transforming a target variable into a normalized variable using a power transformation. This transformation is controlled by a parameter, the lambda, which is chosen such that it achieves the best approximation to a normal distribution. The Box-Cox method can be applied to various types of data, except negative data, and is commonly used in fields such as finance, economics, and engineering.

To transform a target variable x into a normalized variable w, we use Equation (1), where t represents the period, and λ is a parameter that ranges from -5 to 5.

$$w_t = \begin{cases} \log(x_t), & \lambda = 0 \\ \frac{(x_t^{\lambda}-1)}{\lambda}, & \lambda \neq 0 \end{cases} \tag{1}$$

It is mentioned that the current transformation can be performed on non-time series data as well.

1.2 Yeo-Johnson transformation

Yeo-Johnson transformation is also a statistical technique that normalizes data similar to the Box-Cox transformation. It is an extension of the Box-Cox transformation and can be applied to both positive and negative values of the target variable. The transformation is controlled by a parameter estimated using maximum-likelihood methods to achieve the best approximation of a normal distribution. Yeo-Johnson transformation is commonly used in various fields, including finance, engineering, and social sciences.

The Yeo-Johnson transformation is defined by Equation (2).

$$\psi(\lambda, y) = \begin{cases} \frac{(y+1)^{\ \lambda}-1}{\lambda}, & \lambda \neq 0, y \geq 0 \\ \log\ (y+1), & \lambda = 0, y \geq 0 \\ -\frac{(-y+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2, y < 0 \\ -\log(-y+1), & \lambda = 2, y < 0 \end{cases} \tag{2}$$

If y is strictly positive, then the Yeo-Johnson transformation is the same as the Box-Cox transformation of $(y+1)$.

If y is strictly negative, then the Yeo-Johnson transformation is the Box-Cox transformation of $(-y+1)$ with power $2-\lambda$. For both negative and positive values, the transformation is a mixture of them, using different powers for each.

2. The proposed method for the parametric case

In this section, a link between two continuous random variables through the distribution function method is established. Applying the following theorem, we can derive a transformation for any arbitrary continuous random variable that results in a standard normal distribution.

Theorem 1. Let X and Y be continuous random variables with cumulative distribution functions (CDFs) F_X and F_Y, respectively. If the function h is defined as $h(y) = {F_X}^{-1}(F_Y(y))$; then $h(Y) \overset{D}{=} X$.

Proof. The CDF of $h(Y)$ is computed as follows:

$$F_{h(Y)}(x) = P\ (h(Y) \leq\ x\ ) = P\ (\ {F_X}^{-1}(\mathrm{F}_Y(\mathrm{Y}))\ \leq x\ )$$
$$= P\ (F_Y(Y)) \leq F_X(x)),$$

then

$$F_{h(Y)}(x) = P\left(Y \leq {F_Y}^{-1}(F_X(x))\right) = F_Y\left({F_Y}^{-1}(F_X(x))\right)$$
$$= F_X(x).$$

Since, $F_{h(Y)}(x) = F(x)$ for all x, the proof is complete.

The following corollaries resulted directly from Theorem 1.

Corollary 1. Let X and Y have exponential distribution with parameters λ_1 and λ_2, respectively. Let $F_X$and F_Y be their CDFs. then,

$${F_X}^{-1}(x) = -\frac{1}{\lambda_1}\ln(1-x)$$

and $F_Y(y) = 1 - e^{-\lambda_2 y}$. Therefore,

$${F_X}^{-1}(F_Y(y)) = -\frac{1}{\lambda_1} ln(1 - 1 + e^{-\lambda_2 y}) = \frac{\lambda_2}{\lambda_1} y.$$

Thus, $\frac{\lambda_2}{\lambda_1} Y \overset{D}{=} X$.

Corollary 2. Let $X \sim weibull(\alpha, \lambda)$ and $Y \sim logistic(0,1)$, with CDFs F_X and F_Y, respectively. Then,

$${F_X}^{-1}(x) = \left[-\frac{1}{\lambda}\ln(1-x)\right]^{\frac{1}{\alpha}},$$

and $F_Y(y) = \tan^{-1}(y)$.
Therefore,

$${F_X}^{-1}(F_Y(Y))\ \left[-\frac{1}{\lambda}\ln(1 - \tan^{-1}(Y))\right]^{\frac{1}{\alpha}} \overset{D}{=} X.$$

Now, the basic theorem of this section is stated.

Theorem 2. For any continuous random variable Y with CDF F_Y, the transformation

$$h(Y) = \boldsymbol{\Phi}^{-1}(\mathrm{F_Y}(Y)), \tag{3}$$

has a standard Normal distribution. Here, $\boldsymbol{\Phi}$ is the cdf of standard Normal distribution.

Proof: Substituting X as a standard Normal variable in Theorem 1, proves Theorem 2.

The equation (3) converts data with CDF F_Y to normalized data. For probable application, the inverse of (3) returns normalized data to the original distribution. In other words,

$$h^{-1}(Z) = F_Y{}^{-1}(\boldsymbol{\Phi}(Z)). \tag{4}$$

This converts the normal random variable Z to a random variable Y with CDF F_Y.

For instance, if $Y \sim \text{Exp}(\lambda)$ then

$$\boldsymbol{\Phi}^{-1}(1 - e^{-\lambda Y}) \sim N(0,1).$$

In practice, we can utilize the "$qnorm$" function directly from the R software as " $\boldsymbol{\Phi}^{-1}$". This means that any dataset $\boldsymbol{y}$ with a cumulative distribution function of F_Y can be transformed into a normally distributed dataset by using $qnorm(F_Y(y))$".

To evaluate the effectiveness of the proposed transformation method, a simulation study is conducted. Suppose that $Y \sim E(\lambda)$ then $\Phi^{-1}(1 - e^{-\lambda Y}) \sim N(0,1)$. Therefore, to transform a sample y drawn from an exponential distribution, we apply the transformation:

$$\text{" } qnorm(1 - exp\,(-\lambda \boldsymbol{y}))\text{"}$$

This transformation maps the exponential data into standard normal scores. To assess the performance of this method across different scenarios, various combinations of sample sizes and values of λ are considered. The resulting transformed datasets are then tested for normality using the Kolmogorov–Smirnov and Shapiro–Wilk tests.

The corresponding p-values obtained from these tests are summarized in Table 1. As shown in the table, at the significance level of α=0.05, the original data do not follow a normal distribution, whereas the transformed data satisfy the normality assumption based on both the K-S and S-W tests.

However, after applying a normalizing transformation, normality is satisfied according to both K-S and S-W tests under the level of significance $\alpha = 0.05$. Interestingly, for sample sizes of $n = 5$ and $n = 10$, the K-S test suggests that the original data also follows a normal distribution. However, these results are attributed to the fact that the K-S test is asymptotic and may not perform accurately enough for small sample sizes. Hence, for the two cases, the normality of the original data should only be assessed using the S-W test. After applying the transformation method, the data demonstrates substantial improvements in normality, with K-S values increasing significantly (e.g., from 0.027 to 0.998 for (40,1.5) and from 0.034 to 0.788 for (20,1.5)). Likewise, the S-W test values increase noticeably, with some cases (e.g., (40,1.5) and (5,0.1)) achieving p-values above 0.5, indicating much better adherence to normality. However, for cases with lower λ values and smaller sample sizes (e.g., (10,1.5) and (30,0.4)), normality is not fully achieved, suggesting that the transformation method is more effective for larger datasets and higher rate parameters.

2.1. Application to Skewed data

There are instances where the CDF of the original distribution lacks a closed form. In such cases, we can adapt the equation $F_Y(y) = \boldsymbol{\Phi}(y)$ to normalize the data. It is worth noting that in this context, F_Y (y) represents the CDF of the original distribution. Suppose that Y has a skew-normal distribution denoted by $Y \sim \text{SN}(\alpha)$ for $\alpha \neq 0$. Then,

$$f_Y(y) = 2\phi(y)\boldsymbol{\Phi}(\alpha y),$$

$$F_Y(y) = \int_{-\infty}^{y} 2\phi(w)\boldsymbol{\Phi}(\alpha w)\, dw,$$

where ϕ and $\boldsymbol{\Phi}$ are the standard normal density and distribution function, respectively.

Let $\alpha = 1$, $F_Y(y) = \boldsymbol{\Phi}^2(y)$. Then $\boldsymbol{\Phi}^{-1}(\boldsymbol{\Phi}^2(\text{Y})\,) \sim N(0,1)$. When $\alpha \neq 1$, $F_Y(y)$ does not a close form. In this case, the transformation is carried out through (4). Therefore,

$$\boldsymbol{\Phi}^{-1}\left(\int_{-\infty}^{y} 2\phi(w)\boldsymbol{\Phi}(\alpha w)\, dw\,\right) \sim N(0,1).$$

3. The proposed method for nonparametric case

Theorem 2 is a useful tool for normalizing data in various applications. If the distribution of the data is known, this corollary can be implemented directly. However, in cases where the data lacks a parametric

Table 1. Results of Normality Tests for Original and Transformed Exponential Data at α=0.05 Significance Level.

	$(\boldsymbol{n}, \boldsymbol{\lambda})$	$(\mathbf{40}, \mathbf{1.5})$	$(\mathbf{20}, \mathbf{1.5})$	$(\mathbf{10}, \mathbf{1.5})$	$(\mathbf{30}, \mathbf{0.4})$	$(\mathbf{15}, \mathbf{0.4})$	$(\mathbf{5}, \mathbf{0.1})$
	K-S for E(λ)	0.747	0.121	0.719	0.390	0.409	0.479
Original data	K-S	0.027	0.034	0.130	0.031	0.070	0.175
	S-W	0.000	0.000	0.001	0.000	0.000	0.011
Transformed data	K-S	0.998	0.788	0.694	0.540	0.606	0.773
	S-W	0.588	0.578	0.444	0.184	0.200	0.537

distribution, one can utilize the empirical distribution function (EDF) denoted by F_Y^* instead of F_Y in Theorem 2.

Let $Y_{(1)} \leq, Y_{(2)} \leq \cdots \leq Y_{(\mathrm{n})}$ denote the order statistics of the sample. Then,

$$F_Y^*(y) = \begin{cases} 0, & y < Y_{(1)} \\ \frac{i}{n}, & Y_{(i)} \leq y < Y_{(i+1)}; i = 1, \dots, n-1 \\ 1, & y \geq Y_{(n)} \end{cases} \quad (5)$$

The EDF presented in (5) is not continuous and one-to-one, two conditions that are necessary for Theorem 2. To address this issue, we propose a new version of the empirical distribution function called the Generalized Empirical Distribution Function (GEDF), denoted by $F_{new}^*(y)$. The GEDF is defined in the following form and is recommended for use instead.

$$F_{new}^*(y) = \begin{cases} 0, & y < Y_{(1)} - \delta \\ \frac{y - Y_{(1)} + \delta}{n\delta}, & Y_{(1)} - \delta \leq y < Y_{(1)} \\ \frac{1}{n}\left(i + \frac{y - Y_{(i)}}{Y_{(i+1)} - Y_{(i)}}\right), & Y_{(i)} \leq y < Y_{(i+1)}; i = 1, \dots, n-1 \\ 1, & y \geq Y_{(n)} \end{cases} \quad (6)$$

The value of δ can be determined based on the range of data and is typically a small value. Specifically, we use the formula

$$\delta = \min_{i = 1, \dots, n-1} \left[\frac{1}{n^2}\left(Y_{(i+1)} - Y_{(i)}\right)\right]$$

to calculate its value.

The GEDF in (6) is continuous everywhere and is one-to-one for all $y \leq Y_{(n)}$ as well. Its inverse function is given by Equation (7):

$$F_{new}^{*-1}(p) = \begin{cases} Y_{(1)} - \delta(1 - np) & p < \frac{1}{n} \\ Y_{(i)} + (Y_{(i+1)} - Y_{(i)})(np - i) & \frac{i}{n} \leq p \leq \frac{i+1}{n}; i = 1, \dots, n-1 \end{cases} \quad (7)$$

Equations (5) and (6) show that $F_Y{}^*\left(Y_{(i)}\right) = F_{new}^*\left(Y_{(i)}\right) = \frac{i}{n}$. However, in practical statistical applications, we deal with data values $Y_{(i)}$ for $i = 1, \dots, n-1$, and there is no data available in the intervals $(Y_{(i)}, Y_{(i+1)})$ for $i = 1, \dots, n-1$, or below $Y_{(1)}$. Therefore, there is no difference between using (5) or (6) in such cases. However, the key advantage of GEDF is its continuity property, which distinguishes it from EDF. Moreover, as mentioned earlier, GEDF can be used as per Theorem 2 to obtain the desired conversions. To address this issue, we recommend using half ($[\frac{n}{2}]$) for constructing random data using (6) and then normalizing the remaining data with this function. Subsequently, we can swap these two datasets in the next step.

Suppose $y_1, y_2, \dots, y_{\mathrm{n}}$ is a nonparametric data set and let $m_1 = [\frac{n}{2}]$ and $m_2 = n - m_1$. Choose $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_{m_1}^{(1)}$ randomly from $y_1, y_2, \dots, y_{\mathrm{n}}$ and the name remained data as $Y_1^{(2)}, Y_2^{(2)}, \dots, Y_{m_2}^{(2)}$. The two following transformations convert $y_1, y_2, \dots, y_n$ to normalized data.

$$\delta^{(1)} = \min_{i = 1, \dots, m_1 - 1} \left(\frac{Y_{(i+1)}^{(1)} - Y_{(i)}^{(1)}}{m_1^2}\right)$$

$$\delta^{(2)} = \min_{i = 1, \dots, m_2 - 1} \left(\frac{Y_{(i+1)}^{(2)} - Y_{(i)}^{(2)}}{m_2^2}\right)$$

For $Y_1^{(1)}, \dots, Y_{m_1}^{(1)}$, $h(y)$ is used as follows

$$h(y) = \begin{cases} -3.5, & y < Y_{(1)}^{(2)} - \delta^{(2)} \\ \mathbf{\Phi}^{-1}\left(\frac{y - Y_{(1)}^{(2)} + \delta^{(2)}}{m_2\, \delta^{(2)}}\right), & Y_{(1)}^{(2)} - \delta^{(2)} \leq y \leq Y_{(1)}^{(2)} \\ \mathbf{\Phi}^{-1}\left(\frac{i}{m_2} + \frac{y - Y_{(i)}^{(2)}}{m_2\left[Y_{(i+1)}^{(2)} - Y_{(i)}^{(2)}\right]}\right), & Y_{(i)}^{(2)} \leq y \leq Y_{(i+1)}^{(2)}; i = 1, \dots, m_2 - 1 \\ 3.5, & Y_{(m_2)}^{(2)} < y \end{cases} \quad (8)$$

and for $Y_1^{(2)}, \dots, Y_{m_2}^{(2)}$, $h(y)$ is applied in the following

$$h(y) = \begin{cases} -3.5, & y < Y_{(1)}^{(1)} - \delta^{(1)} \\ \mathbf{\Phi}^{-1}\left(\frac{y - Y_{(1)}^{(1)} + \delta^{(1)}}{m_1\, \delta^{(1)}}\right), & Y_{(1)}^{(1)} - \delta^{(1)} \leq y \leq Y_{(1)}^{(1)} \\ \mathbf{\Phi}^{-1}\left(\frac{i}{m_1} + \frac{y - Y_{(i)}^{(1)}}{m_1\left[Y_{(i+1)}^{(1)} - Y_{(i)}^{(1)}\right]}\right), & Y_{(i)}^{(1)} \leq y \leq Y_{(i+1)}^{(1)}; i = 1, \dots, m_1 - 1 \\ 3.5, & Y_{(m_1)}^{(1)} < y \end{cases} \quad (9)$$

The inverse conversion for returning normalized data to the original is

$$h^{-1}(z) = \begin{cases} Y_{(1)} - \delta\big(1 - n\,\mathbf{\Phi}(\mathrm{z})\big), & \mathbf{\Phi}(\mathrm{z}) < \frac{1}{n} \\ Y_{(i)} + \big(Y_{(i+1)} - Y_{(i)}\big)(n\mathbf{\Phi}(\mathrm{z}) - i), & \frac{i}{n} \leq \mathbf{\Phi}(\mathrm{z}) \leq \frac{i+1}{n}; i = 1, \dots, n-1 \end{cases} \tag{10}$$

where δ is defined as

$$\delta = \min_{i = 1, \dots, n-1} \left(\frac{Y_{(i+1)} - Y_{(i)}}{n^2}\right).$$

4. Application to read data sets

In this section, we demonstrate proficiency of the proposed method in two real data sets.

4.1. Confirmed case rate of the COVID-19 data

This data demonstrates the effectiveness of normalizing transformations for nonparametric data using several real-world datasets. Specifically, we analyze the confirmed case rate (CCR) of the COVID-19 virus across ten different countries: Canada, France, Germany, Iran, Iraq, Italy, Mexico, Netherlands, Turkey, and the UK. These datasets were sourced from publicly available data provided by the World Health Organization (WHO) at https://covid19.who.int/

The CCR data exhibit significant variability and skewness, making them ideal candidates for normalization techniques. To address this, we applied Equations (9) and (10), which represent mathematical transformations designed to reduce skewness and approximate a normal distribution. After applying these transformations, we evaluated the normalized data using two widely used statistical tests: The Kolmogorov-Smirnov (K-S) test and the Shapiro-Wilk (S-W) test. The K-S test assesses the goodness-of-fit between the empirical distribution of the data and a theoretical normal distribution, while the S-W test is particularly sensitive to deviations from normality in smaller sample sizes. The results of both tests are reported in Table 2, where it is evident that the transformations significantly improve the normality of the datasets. This improvement is reflected in the p-values of the tests, which indicate a closer adherence.

The table clearly shows that none of the original datasets are normally distributed, while all datasets normalized using nonparametric transformations have a normal distribution.

To have a comparison with previous methods of normalization, data also were transformed by Box-Cox transformation with the best value of λ. This work is done by function "boxcox" in package "MASS". The results demonstrate that none of these transformed data have a normal distribution regarding the K-S test. However, regarding the S-W test, the transformed data of 4 countries have a normal distribution. This certifies the better performance of our method concerning, for, to Box-Cox transformation.

The results in Table 2 indicate that the original COVID-19 CCR data strongly deviates from normality, as evidenced by extremely small p-values from both the K-S and S-W tests. This suggests that the data distribution is significantly non-normal across all countries. Upon applying the Box-Cox transformation, slight improvements are observed in normality, particularly in the S-W test for some countries (e.g., Canada: 0.2695, Mexico: 0.7005, and the UK: 0.833). However, many values remain well below conventional thresholds for normality ($p > 0.05$), indicating that the

Table 2. Normality tests for original and transformed CCR of COVID-19 data.

	Country	**Canada**	**France**	**Germany**	**Iran**	**Iraq**
Original data	K-S	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$
	S-W	0.0073	$< 2\times 10^{-16}$	2×10^{-15}	4×10^{-16}	3.9×10^{-7}
Transformed data by Box-Cox	K-S	$< 2.2\times 10^{-16}$	$< 2.2\times 10^{-16}$	$< 2.2\times 10^{-16}$	$< 2.2\times 10^{-16}$	$< 2.2\times 10^{-16}$
	S-W	0.2695	1.49×10^{-5}	8.20×10^{-11}	1.89×10^{-5}	8.59×10^{-7}
Transformed data by method of paper	K-S	0.4992	0.9991	0.3439	0.9642	0.786
	S-W	0.1385	0.8113	0.8621	0.2411	0.4
	Country	**Italy**	**Mexico**	**Netherlands**	**Turkey**	**UK**
Original data	K-S	$<7\times 10^{-16}$	$< 2\times 10^{-16}$	1.2×10^{-15}	$< 2\times 10^{-16}$	4.6×10^{-7}
	S-W	0.0047	1.6×10^{-6}	0.0433	$< 2\times 10^{-16}$	0.8188
Transformed data by Box-Cox	K-S	$<7\times 10^{-15}$	$< 2.2\times 10^{-16}$	1.221×10^{-15}	$< 2.2\times 10^{-16}$	2.49×10^{-13}
	S-W	0.0422	0.7005	0.8813	1.30×10^{-9}	0.833
Transformed data by method of paper	K-S	0.9635	0.8221	0.9409	0.6051	0.7644
	S-W	0.2861	0.5015	0.2793	0.2368	0.2571

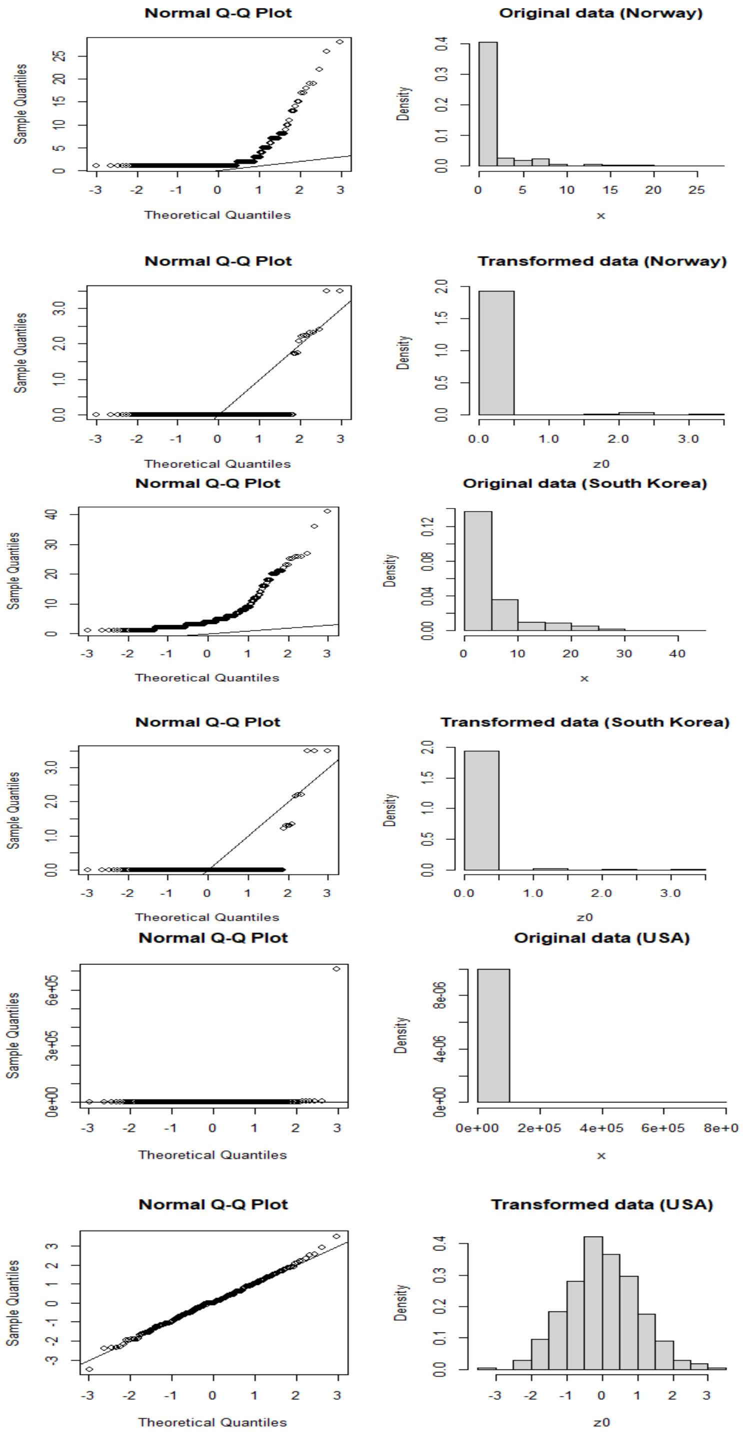


Figure 1. "Effect of Proposed Transformations on Normality: Q-Q Plots and Density Curves for Norway, South Korea, and the USA"

Box-Cox transformation alone is insufficient for achieving a normal distribution in many cases.

Conversely, the transformation method proposed in the paper demonstrates a substantial improvement in normality. K-S test results show much higher p-values (closer to 1), indicating that the transformed distributions better fit a normal distribution. Likewise, the S-W test results show a significant shift towards normality, with values such as 0.8113 (France), 0.8621 (Germany), and 0.5015 (Mexico) reflecting stronger normality characteristics than the Box-Cox transformation. This suggests that the proposed transformation method is more effective at normalizing the data, making it a preferable approach for statistical modeling.

Figure 1 illustrates the effects of the proposed transformation on datasets from Norway, South Korea, and the United States. The original data demonstrated significant non-normality, primarily characterized by right skewness. Following the transformation, the data aligned much more closely with a normal distribution, as evidenced by decreased skewness and the emergence of more symmetrical, bell-shaped density curves. Although some minor deviations from normality persist in certain cases, the transformation markedly enhances the data's suitability for parametric analyses. Each subplot includes a Normal Q-Q plot and a density curve, providing clear visual confirmation of the transformation's success in improving distributional symmetry and overall normality.

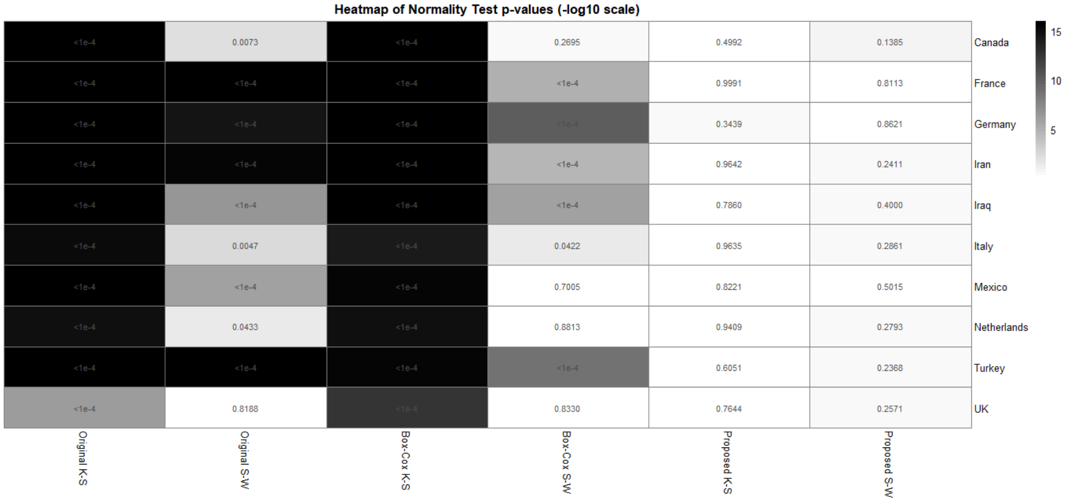
Figure 2 displays a grayscale heatmap of log10-transformed p-values from normality tests, illustrating how well COVID-19 CCR data conform to a normal distribution across different countries and transformation techniques. In this visualization, darker shades represent stronger evidence of deviation from normality (lower p-values), while lighter shades indicate data closer to a normal distribution (higher p-values). For example, the original data for Canada are depicted in very dark hues, reflecting extremely low p-values (<1e-16) and a significant departure from normality. In contrast, after applying the proposed transformation, the color intensity becomes noticeably lighter, with p-values exceeding 0.3, demonstrating a marked improvement in normality fit. This comparison underscores the effectiveness of the proposed method, which outperforms both the untransformed data and the Box-Cox transformation in achieving normality.

4.2. The mortality rate and confirmed case rate of COVID-19 data

It is essential to clarify that this article focuses exclusively on the normalization of continuous or non-discrete data. As such, the methodology proposed herein may not be suitable for normalizing discrete data, a point we will further explore in this section through a practical situation that highlights how aggregated data can sometimes be treated as discrete.

We examine the mortality rate (MR) and confirmed case rate (CCR) of COVID-19 across three countries: the USA, South Korea, and Norway. The relevant data can be downloaded from https://covid19.who.int/

In this case, the MR and CCR data for South Korea and Norway contain many zeros and are thus approximately discrete, while the same variables for the USA are continuous. As shown in Table 3, both the K-S and S-W tests confirm that none of the six datasets has a normal distribution. Although the transformed datasets for the USA are normally distributed, the data



	Original K-S	Original S-W	Box-Cox K-S	Box-Cox S-W	Proposed K-S	Proposed S-W
Canada	<1e-4	0.0073	<1e-4	0.2695	0.4992	0.1385
France	<1e-4	<1e-4	<1e-4	<1e-4	0.9991	0.8113
Germany	<1e-4	<1e-4	<1e-4	<1e-4	0.3439	0.8621
Iran	<1e-4	<1e-4	<1e-4	<1e-4	0.9642	0.2411
Iraq	<1e-4	<1e-4	<1e-4	<1e-4	0.7860	0.4000
Italy	<1e-4	0.0047	<1e-4	0.0422	0.9635	0.2861
Mexico	<1e-4	<1e-4	<1e-4	0.7005	0.8221	0.5015
Netherlands	<1e-4	0.0433	<1e-4	0.8813	0.9409	0.2793
Turkey	<1e-4	<1e-4	<1e-4	<1e-4	0.6051	0.2368
UK	<1e-4	0.8188	<1e-4	0.8330	0.7644	0.2571

Figure 2 "Comparison of Normality Fit for COVID-19 CCR Data Across Countries: Grayscale Heatmap of Log10-Transformed p-Values"

Table 3. Normality tests for original and transformed MR and CCR of COVID-19 data**.**

	Country	**USA**		**South Korea**		**Norway**	
	variable	**MR**	**CCR**	**MR**	**CCR**	**MR**	**CCR**
Original data	K-S	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$
	S-W	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	8×10^{-14}	$< 2\times 10^{-16}$	8×10^{-14}	$< 2\times 10^{-16}$
Transformed data by Box-Cox	K-S	$< 2.2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$
	S-W	3.23×10^{-6}	$< 2\times 10^{-16}$	4.87×10^{-7}	1.12×10^{-10}	$< 2\times 10^{-16}$	$< 2\times 10^{-16}$
Transformed data by method of paper	K-S	0.9912	0.431	8.69×10^{-8}	$< 2\times 10^{-16}$	3.5×10^{-5}	$< 2\times 10^{-16}$
	S-W	0.7046	0.5023	3.27×10^{-8}	$< 2\times 10^{-16}$	2.9×10^{-6}	$< 2\times 10^{-16}$

for South Korea and Norway do not have the normal distribution yet. The results of the Box-Cox transformation demonstrate that none of these transformed data have a normal distribution regarding the K-S and S-W tests. Therefore, our method has better performance of Box-Cox transformation.

Table 3 presents normality test results for the Mortality Rate (MR) and Case Fatality Rate (CCR) of COVID-19 data across the USA, South Korea, and Norway. The original data exhibits strong deviations from normality, as indicated by extremely small p-values in both the Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) tests, confirming significant non-normality. Applying the Box-Cox transformation yields minor improvements in a few cases (e.g., the S-W test for MR in the USA: 3.23×10^{-6}), but in general, normality is not achieved, as many values remain exceptionally low. This suggests that the Box-Cox transformation is insufficient for properly normalizing the data. On the other hand, the proposed transformation method in the paper significantly enhances normality for MR in the USA, where K-S and S-W tests return p-values of 0.9912 and 0.7046, respectively. However, for CCR in South Korea and Norway, the transformation does not substantially improve normality, with many p-values remaining near zero. This indicates that while the method is effective in certain cases, its performance varies across datasets and variables, necessitating further investigation.

Results

The normality of data and its distribution is crucial for many statistical analyses, as it simplifies the statistical modeling process. In this article, recognizing the importance of normalization conversions and their application in various scientific fields, we have introduced a new and straightforward normalization conversion method for both parametric and non-parametric continuous data. Our approach outperforms conventional methods such as the Yeo-Johnson transformation and Box-Cox transformation and overcomes their limitations. Additionally, we have introduced a new empirical distribution function that allows the proposed conversion to be used for non-parametric data. Unlike conventional empirical distribution functions, this new definition is continuous, enabling us to normalize non-parametric data.

Conclusion

To further strengthen the study, additional validation should be conducted by comparing the proposed transformation with other techniques such as Yeo-Johnson and log transformations, along with supplementary normality tests like the Anderson-Darling test. Visualizing the distributions through histograms, density plots, and Q-Q plots would provide intuitive confirmation of normality improvements. Moreover, evaluating the impact of normalization on downstream statistical analyses, such as regression modeling and hypothesis testing, would demonstrate its practical benefits. Expanding the application to other epidemiological datasets and periods would help assess its generalizability. Finally, a deeper theoretical discussion on why the proposed transformation outperforms Box-Cox, particularly in terms of skewness and kurtosis reduction, would provide a stronger mathematical justification for its effectiveness.

Moreover, to further validate the findings, the proposed transformation should be compared with alternative techniques such as Yeo-Johnson or log transformations, and additional normality tests like the Anderson-Darling test should be conducted. Visualizing the distributions through histograms, Q-Q plots, and density plots would provide clearer insights into the effectiveness of each transformation. Evaluating the impact of normality improvements on subsequent statistical analyses, such as predictive modeling and hypothesis testing, would help establish its practical benefits. Additionally, applying the method to a broader range of countries and datasets would assess its generalizability. Finally, a theoretical discussion on why the transformation is more effective for MR than CCR, possibly exploring the influence of skewness and kurtosis, would strengthen the study's conclusions.

References

1. Graybill FA. The Theory and Applications of the Linear Model (London, Duxbury Press). 1976.
2. Thöni H. A table for estimating the mean of a lognormal distribution. Journal of the American Statistical Association. 1969 Jun 1;64(326):632-6
3. Hoyle MH. Transformations: An introduction and a bibliography. The International Statistical Review. 1973; 41(203-223).
4. Box GEP and Cox, DR. An analysis of transformations. Journal of the Royal Statistical Society. 1964 B; 26(211-234).
5. Sakia RM. The Box-Cox transformation technique: A review. The Statistician. 1992; 41(169-178).
6. Yeo IK, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000 Dec 1;87(4):954-9.
7. Tukey JW. The comparative anatomy of transformations. Annals of Mathematical Statistics 1957; 28(602-632).
8. Atkinson AC, Riani M, Corbellini A. The box–cox transformation: Review and extensions. Statistical Science. 2021;36(2):239–55.
9. Riani M, Atkinson AC, Corbellini A. Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression. Statistical Methods & Applications. 2023;32(1):75-102.