ENHANCING PROCESS CAPABILITY ANALYSIS FOR LOGNORMAL DATA UTILIZING BOX COX TRANSFORMATION AND GOODNESS OF FIT TESTS

J. Krishnan[1] and R. Vijayaraghavan[2]

•

(1). Department of Mathematics, Sri Krishna Adithya College of Arts and Science
Coimbatore – 641042, Tamil Nadu, INDIA
(2). Department of Statistics, Bharathiar University, Coimbatore 641 046,
Tamil Nadu, INDIA
[1]krrishme92@gmail.com, [2]vijaystatbu@gmail.com

Abstract

Process capability analysis is a valuable tool in quality assurance, but deviations from normal distribution necessitate adjustments to basic process capability indices. Process control literature offers solutions for non-normality, with data transformation being a common approach. The Box-Cox transformation (BCT) is often used to normalize non-normal data, relying on maximum likelihood estimation (MLE) to determine the transformation parameter, lambda. Alternative methods exist for estimating the single transformation parameter lamda, employing goodness-of-fit tests instead of the MLE method. This study explores two expressions within the Box-Cox transformation (BCT), encompassing both optimal and rounded values of lambda. The primary goal is to identify an effective method for transforming non-normal data into a distribution closer to normality through goodness-of-fit tests, aiming to obtain accurate estimates for process capability analysis in alignment with six sigma standards. Furthermore, this study focuses on the influence of utilizing both optimal and rounded values of lambda when transforming non-normal data to normal, and how these lambda values impact the estimates of process capability analysis. The findings reveal that methods such as Shapiro-Wilk's (SW) and Artificial Covariate (AC) outperform the MLE method. Moreover, employing the optimal lambda value during data transformation leads to improved estimates of process capability. Data simulation and analysis were conducted using Minitab software and the R programming language.

Keywords: Goodness of fit tests, Box-Cox Transformation, Asymmetric, MLE, Lognormal distribution, Six sigma.

I. Introduction

Process capability indices (PCIs) are essential tools in quality control, commonly utilized across manufacturing industries to ensure processes meet required standards. Process capability analysis (PCA) evaluates how effectively a manufacturing process adheres to specified targets. However, traditional PCIs assume a normal distribution, which may inaccurately assess non-normal processes. Kane (1986) suggests that transforming data to preserve a somewhat normal distribution improves the accuracy of process capability analysis [5]. Empirical studies have shown that transformed data yields superior results compared to original data [4]. Based on many literature surveys transformation methods, especially for non-normal distributions like Lognormal and Weibull, consistently outperform Non-Transformation (NT) methods. NT methods are inadequate for assessing process capability when distributions deviate significantly from normal [15]. Hence, transformation methods are preferred, as they provide more reliable assessments,

even for distributions distanced from normality.

In process capability analysis (PCA), the variability of a process is measured using the standard deviation. This variability can be divided into short-term and long-term variations. Short-term variability is determined by the estimated standard deviation obtained from random sample observations, which is then used in calculating process capability indices. On the other hand, long-term variability is assessed for computing process performance indices. Consequently, capability indices are computed using short-term variation, while performance indices utilize all data points, considering long-term variation. The commonly used capability indices are denoted as Cp and Cpk, while the respective performance indices are represented as Pp and Ppk. Various methods for handling non-normality in calculating process capability indices are discussed in [13]. Among these, the most widely applied indices in the manufacturing industry are the process capability index Cp and process capability ratio Cpk, as shown in Table 1 below, along with their respective performance indices. Here, $\bar{x}$ denotes the sample mean, USL refers to the upper specification limit, and LSL indicates the lower specification limit.

Table 1: *Process Capability and Process Performance Indices*

Process capability indices	Process performance indices
$C_p = \frac{USL-LSL}{6\sigma_W}$	$P_p = \frac{USL-LSL}{6\sigma_{overall}}$
$C_{pk} = min\ (C_{PU}, C_{PL})$	$P_{pk} = min\ (C_{PU}, C_{PL})$
$C_{PU} = \frac{USL-\bar{\bar{x}}}{3\sigma_W}, \quad C_{PL} = \frac{\bar{\bar{x}}-LSL}{3\sigma_W}$	$P_{PU} = \frac{USL-\bar{\bar{x}}}{3\sigma_{overall}}, \ P_{PL} = \frac{\bar{\bar{x}}-LSL}{3\sigma_{overall}}$

In [2], researchers employed the method of maximum likelihood estimation (MLE) to determine the optimal parameter λ in the Box-Cox transformation. Other approaches to the MLE methods, which rely on goodness of fit tests (specifically normality tests), were developed in [1], [3], [9], [10] and [14]. Through an examination of the impact of transforming non-normal data into normal data using different goodness of fit tests, [3] illustrated that the MLE method for estimating the λ parameter in BCT could be biased and inefficient. Furthermore, as indicated in [18] employing various goodness of fit tests instead of the MLE method for estimating the BCT parameter λ leads to improved estimates of process capability and process performance for non-normal data. The effectiveness of different goodness of fit tests was also assessed in [3] using various error measures, estimates of process capability and process performance indices, and defective parts per million (PPM) products. The results of different goodness of fits tests are recorded and presented to help the practitioner to choose the method which will produce the improvised results in various asymmetric situations, *viz*., low, moderate and high. Thus, the objectives of this paper is to examine the effectiveness of the different goodness of fit tests involving transformation of non-normal data into normal data using BCT and to recommend a superior test that will produce higher values of process capability with minimum of error and PPM values particularly, for lognormal distribution. Additionally this paper focuses on the impact of optimal and rounded value of transforming parameter λ in BCT. It also verifies whether the proposed method produce the results within the standard of six sigma level.

II. Methodology

Converting non-normal data into a normal distribution is a common practice when observed data fail to meet normality assumptions. Several methods are employed for this purpose in practical applications, including Johnson's system of transformation (JST), Box-Cox transformation (BCT), and Rosenblatt transformation (RT). While both JST and BCT approaches are effective, BCT is generally preferred over JST, particularly in situations where computer-assisted analysis is

available, as it tends to outperform other methods [12]. Additionally, BCT is noted for its superior accuracy and precision compared to the JST method. BCT offers a range of power transformations designed to optimally normalize specific variables. According to [2], the BCT method transforms non-normal data into normal data for positive response variable x, as expressed below:

$$x^{\lambda} = \begin{matrix} \frac{x^{\lambda}-1}{\lambda}, for\ \lambda\ \neq 0 \\ \log x, for\ \lambda = 0 \end{matrix} \quad (1)$$

It may be noted that since an analysis of variance is unchanged by a linear transformation, the expressions given (1) is equivalent to

$$x^{\lambda} = \begin{matrix} x^{\lambda}, for\ \lambda \neq 0 \\ \log x, for\ \lambda = 0 \end{matrix} \quad (2)$$

The form (1) is slightly preferable for theoretical analysis because it is continuous at λ = 0, refer [2]. The major effort in Box-Cox transformation is connected to the transformation X to X^{λ}, with the parameter λ possibly a vector describing a specific transformation. A single transforming parameter λ is the main source of dependence for this family of transformations, and its value is determined using maximum likelihood estimation [2].

The results of the earlier studies presented in the literature, particularly in [1], [6], [9], [10], [13], [14], [15] and [17] would be useful to understand the significance of tests of goodness of fit while transforming non-normal data into normal data. The estimation of λ is done through various goodness of tests for normality, that are available in the literature, which includes tests, such as Shapiro - Wilk (SW), Anderson Darling (AD), Cramer Von Mises (CVM), Pearson Chi-square (PC), Shapiro - Francia (SF), Lillefors (Kolmogorov - Simirnov) (LT / KS), Jarque - Bera (JB), and artificial covariate method (AC). Additionally, Minitab Transformation (M_T) is included in the evaluation, which employs a rounded value of λ compared to the optimal value of λ used in goodness-of-fit tests. Since, the choice of the value for lambda (λ) in a Box-Cox transformation might have a significant impact on the result of process capability or process performance analysis. [9] Shows that the test based on SW statistic is a powerful test of normality for a variety of non-normal distributions, the SW statistic is reliable for small samples and in regression applications, the statistic would yield higher R2. It is asserted in [6] that the test based on SW statistic is the most powerful test for non-normal distributions. According to [18], the current MLE technique could be effectively substituted by using goodness of fits tests in Box-Cox transformation to get data close to normal as possible and achieving desired results in estimating process capability analysis.

III. Lognormal Distribution

The log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed. When the logarithm of a log-normal distributed variable is taken, it results in a normal distribution. However, when looking at the original data itself, it doesn't follow a normal distribution. It typically exhibits skewness and can have a long tail on one side. It's often used to model phenomena where the logarithm of the variable is normally distributed, such as stock prices, incomes, and certain biological measurements.

$$f(x\,|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2x}}\, e^{-(\ln(x)-\mu)^2/(2\sigma)^2} \quad (3)$$

Where,

x>0 is the value of the random variable.
μ is the mean of the natural logarithm of the variable.
σ is the standard deviation of the natural logarithm of the variable.

The mean and variance of the log-normal distribution is given by

$$E(x) = e^{\mu+\frac{\sigma^2}{2}} \quad (4)$$

$$V(x) = \left(e^{\sigma^2} - 1\right).e^{2\mu+\sigma^2} \quad (5)$$

The lognormal distribution was examined at various asymmetric levels, characterized by different mean and standard deviation pairs: (0, 0.25), (0, 0.50), and (0, 1). These parameter sets were grouped to evaluate the impact of low, moderate and high asymmetry in transforming non-normal data to normal data and conducting process capability analysis. Figure 1 illustrates the shape of the density function for each parameter set.

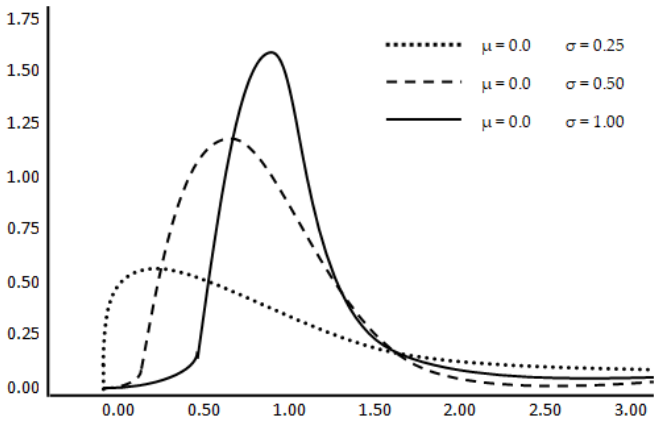

Figure 1: *The asymmetric behavior of lognormal distribution used for simulation study*

IV. Numerical Illustrations

The log normal distribution is applicable to a wide range of non-normal processes because it is capable of generating a variety of distinct curves based on its parameters. A log-normally distributed random variable only accepts positive real values. It is an easy-to-use model that can be applied to measurements in the exact sciences, engineering, medicine, economics, and other fields (such as energies, concentrations, lengths, prices of financial instruments, and other metrics). For simulation set-up, the data set of the size is taken as 100 and generated using different asymmetric levels of lognormal distribution. The lower and upper specification limits were taken as 0.01 and 10. The defined specification limit in this study of process capability analysis might be appropriate in certain situations where the process parameter being analyzed is bounded by very low values and 0.01 represents a meaningful lower limit for the process output. Here are some scenarios where these specification limits could be reasonable such as chemical concentrations, precision engineering, analytical instruments, environmental monitoring, biomedical applications and so on.

The study evaluates the effectiveness of the method using a combination of box plots, descriptive statistics, measures of errors (Bias, Percentage Bias, Median Absolute Error (MdAE), Root Mean Square Error (RMSE)), and radar charts. Due to space limitations, only error measures and radar plots are included. Bias, MdAE, and RMSE serve as error metrics for transforming non-normal data into normal data using various goodness-of-fit tests in Box Cox transformation. After transformation, the data are utilized to estimate process capability and performance index, aiding in the selection of the most effective approach among different goodness-of-fit tests.

A process is considered to be under six sigma controls if both process capability and performance indices such as Cp and Cpk and Pp and Ppk are greater than or equal to 2 and 1.5, respectively. In the automotive industry, a Cpk value of 1.33 is used as a benchmark for assessing process capability. According to Pearn W.L. and Chen K.S. (2002), a process is considered

inadequate if its process capability index (PCI) is less than 1.00, capable if PCI is between 1.00 and 1.33, satisfactory if PCI is between 1.33 and 1.50, excellent if PCI is between 1.50 and 2.00, and super if PCI is equal to or greater than 2.00 [7]. As outlined by Sibalija TV and Majstorovic VD (2010), the primary objective for quality and industry practitioners is to achieve 6σ limits, with a corresponding defect rate of 3.4 PPM associated with the process [11]. One may refer to [8] and [16] for the details on the concepts of six-sigma tools and process capability analysis for non-normal data, respectively. Table 2 displays the process fallout in defective parts per million products alongside the proportion of good items and PPM values for different sigma levels.

Table 2: P*rocess fallout in defective parts per million with respect to different sigma levels*

Sigma Level	Percentage	PPM Values
6	99.9997%	3.4
5	99.98%	233
4	99.4%	6,210
3	93.3%	66,807
2	69.1%	308,537
1	30.9%	691,462

I. Low Asymmetric Distribution

The simulation study focuses on utilizing a low asymmetric lognormal distribution with a skewness of 0.36 and 0.63, where the mean and standard deviation are 0 and 0.25, respectively. To assess the effectiveness of various methods in transforming non-normal data into a normal distribution, two sets of data are analyzed. One with a Skewness (Sk) of 0.36 and another with a skewness of 0.63. For the dataset with ln(0, 0.25)(1), methods like M_T, PT, LT, and JB transforms data more like a normal distribution with fewer errors. Likewise, for the dataset with ln(0, 0.25)(2), methods such as SF, JB, SW, AC, and MLE are transforms data getting closer to a normal with fewer errors. For further details, refer to Table 3 and Figure 2. Subsequently, the transformed data from different goodness-of-fit tests are used to estimate process capability/performance. This analysis helps identify the most effective method for handling non-normal, low asymmetric distributions.

Table 3: *Various measures of error values for low asymmetric data after the data transformation*

Goodness of fit tests	Low Asymmetry (Sk=0.36) Lognormal distribution (μ=0, σ =0.25)			Low Asymmetry (Sk=0.63) Lognormal distribution (μ=0, σ=0.25)		
	Bias	MdAE	RMSE	Bias	MdAE	RMSE
SW	1.0156	1.0068	1.0158	1.0231	1.0104	1.0236
AD	1.0156	1.0068	1.0158	1.0373	1.0168	1.0388
CVM	1.0180	1.0078	1.0182	1.0423	1.0191	1.0442
PT	0.9916	0.9962	0.9916	1.0484	1.0218	1.0510
SF	1.0159	1.0069	1.0161	1.0223	1.0100	1.0228
LT	1.0147	1.0065	1.0149	1.0355	1.0160	1.0368
JB	1.0152	1.0067	1.0153	1.0228	1.0102	1.0233
AC	1.0161	1.0071	1.0163	1.0245	1.0110	1.0251
MLE	1.0161	1.0070	1.0163	1.0246	1.0111	1.0252
M_T	-0.0008	0.0735	0.1049	1.0262	1.0118	1.0269

All the methods of data transformation used in this study results within the standard of six sigma but the only few methods produces better estimates of Pp/Cp and Ppk/Cpk, such methods are CVM, AC, MLE, SF, SW and AD for data set ln(0, 0.25)(1) and SF, JB, SW, AC and MLE for data set ln(0, 0.25)(2). Though only the tests SF, SW, AC, and MLE were considered appropriate

procedures to deal non-normal low asymmetric distributions in order to obtain desirable results with less errors, better estimates, and PPM values within the six sigma limits. For more information see the table 2, 4 and 5.

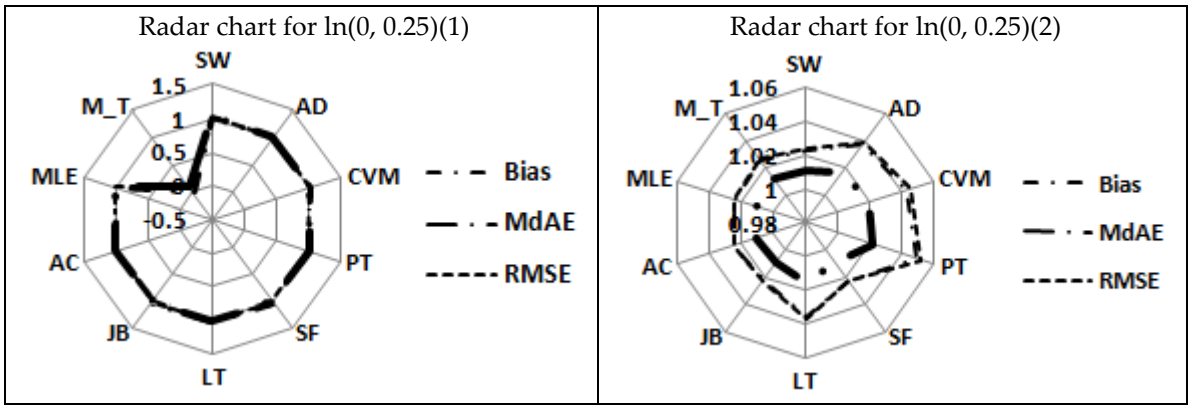

Figure 2: *Radar chart for various measures of errors after the normalization of low asymmetric distribution*

Table 4: *Estimates of process capability and process performance indices for ln(0, 0.25)(1) data after normalization via goodness of fit tests*

Goodness of fit tests	λ Value	LSL	USL	PCI (Within Capability)			PPI (Overall Capability)		
				Cp	Cpk	PPM	Pp	Ppk	PPM
ln(0, 0.25)(1)	-	0.01	10	7.62	1.49	3.87	7.893	1.544	1.81
SW	0.31	-2.45	3.36	4.38	3.65	0.00	4.54	3.79	0.00
AD	0.31	-2.45	3.36	4.38	3.65	0.00	4.54	3.79	0.00
CVM	0.21	-2.95	2.96	4.44	4.39	0.00	4.6	4.55	0.00
PT	1.38	-0.72	16.66	13.2	1.09	5.64	13.7	1.12	373
SF	0.3	-2.50	3.32	4.39	3.73	0.00	4.55	3.86	0.00
LT	0.35	-2.29	3.54	4.40	3.42	0.00	4.57	3.54	0.00
JB	0.33	-2.37	3.45	4.39	3.54	0.00	4.55	3.67	0.00
AC	0.29	-2.54	3.27	4.38	3.79	0.00	4.54	3.92	0.00
MLE	0.29	-2.54	3.28	4.39	3.79	0.00	4.55	3.92	0.00
M_T	0.5	0.100	3.16	4.65	2.69	0.00	4.82	2.79	0.00

Table 5: *Estimates of process capability and process performance indices for ln(0, 0.25)(2) data after normalization via goodness of fit tests*

Goodness of fit tests	λ Value	LSL	USL	PCI (Within Capability)			PPI (Overall Capability)		
				Cp	Cpk	PPM	Pp	Ppk	PPM
ln(0, 0.25)(2)	-	0.01	10	7.041	1.436	8.25	7.020	1.432	8.73
SW	0.12	-3.54	2.65	4.55	3.89	0.00	4.51	3.86	0.00
AD	-0.43	-14.5	1.46	11.68	2.15	0.00	11.55	2.12	0.00
CVM	-0.62	-26.4	1.23	20.03	1.80	0.00	19.78	1.78	0.00
PT	-0.85	-57.7	1.01	41.93	1.47	5.27	41.37	1.45	6.90
SF	0.15	-3.33	2.75	4.47	4.03	0.00	4.43	4	0.00
LT	-0.36	-11.8	1.57	9.80	2.31	0.00	9.69	2.28	0.00
JB	0.13	-3.47	2.68	4.52	3.92	0.00	4.48	3.9	0.00
AC	0.065	-3.98	2.48	4.75	3.64	0.00	4.71	3.61	0.00
MLE	0.06	-4.02	2.47	4.78	3.63	6.10	4.73	3.6	0.00
M_T	0.0	-4.61	2.31	5.086	3.387	0.00	5.039	3.356	0.00

II. Moderate Asymmetric Distribution

The lognormal distribution, characterized by parameters μ = 0 and σ = 0.50, offers a means to generate moderately asymmetric data with respective skewness values of 0.96 and 1.32. Through a simulation study, it is confirmed that for data set ln(0, 0.5)(1) the LT, AD, SW, SF, CVM and JB and for data set ln(0, 0.5)(2), the M_T, LT, AC, MLE, SF and SW methods of goodness of fit tests effectively transform the non-normal data into normal distributions with minimal errors. Consequently, the transformed datasets are subjected to further examination to evaluate their efficiency in estimating process capability/performance for moderately asymmetric distributions. See the table 6 and figure 3.

Table 6: *Various measures of error values for moderate asymmetric data after the data transformation*

Goodness of fit tests	Moderate Asymmetry (Sk=0.96) Lognormal distribution (μ=0, σ=0.5)			Moderate Asymmetry (Sk=1.32) Lognormal distribution (μ=0, σ=0.5)		
	Bias	MdAE	RMSE	Bias	MdAE	RMSE
SW	1.1050	1.0392	1.1141	1.1634	1.0599	1.1856
AD	1.1036	1.0386	1.1124	1.1703	1.0627	1.1938
CVM	1.1065	1.0397	1.1158	1.1648	1.0605	1.1873
PT	1.1079	1.0403	1.1174	1.1648	1.0605	1.1873
SF	1.1050	1.0392	1.1141	1.1621	1.0594	1.1840
LT	1.1007	1.0375	1.1091	1.1538	1.0561	1.1741
JB	1.1065	1.0397	1.1158	1.1675	1.0616	1.1905
AC	1.1086	1.0406	1.1182	1.1608	1.0589	1.1825
MLE	1.1079	1.0403	1.1174	1.1607	1.0588	1.1823
M_T	0.1071	0.1667	0.3422	1.1483	1.0539	1.1675

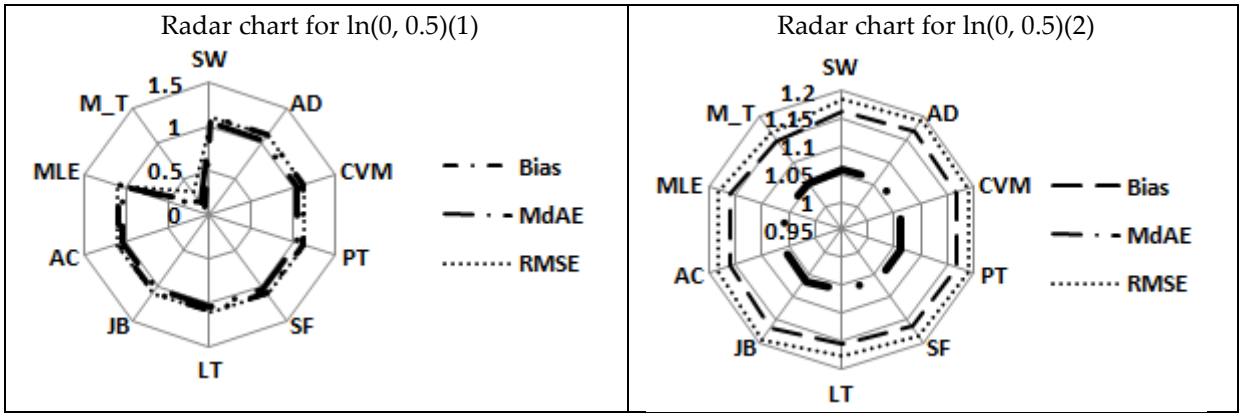

Figure 3: *Radar chart for various measures of errors after the normalization of moderate asymmetric distribution*

In the simulation study, transformed data yields improved estimates of Pp/Cp and Ppk/Cpk using methods such as AC, PT, MLE, JB, CVM, SW, and SF for dataset ln(0, 0.5)(1), and AC, MLE, SF, and SW for dataset ln(0, 0.5)(2). Furthermore, the PPM values indicate that for dataset ln(0, 0.5)(1), the results fall within the standard six sigma limits, while for dataset ln(0, 0.5)(2), PPM values range between 5σ and 6σ limits (with recorded values ranging from 17 to 128, against benchmark values of 233 for 5σ and 3.4 for 6σ, as detailed in Table 2). This close alignment with the six sigma standard is promising. Upon considering various measures of errors, process capability/performance indices, and associated PPM values, it becomes evident that the AC, MLE, SW, and SF approaches outshine other methods, as depicted in Tables 2, 7, and 8.

Table 7: *Estimates of process capability and process performance indices for ln(0, 0.5)(1) data after normalization via goodness of fit tests*

Goodness of fit tests	λ Value	LSL	USL	PCI (Within Capability)			PPI (Overall Capability)		
				Cp	Cpk	PPM	Pp	Ppk	PPM
ln(0, 0.5)(1)	-	0.01	10	2.85	0.64	22623	2.81	0.64	28271
SW	0.29	-2.54	3.28	1.80	1.59	0.88	1.80	1.59	0.95
AD	0.3	-2.50	3.32	1.80	1.57	1.24	1.80	1.56	1.34
CVM	0.28	-2.59	3.23	1.80	1.62	0.56	1.80	1.62	0.60
PT	0.27	-2.64	3.19	1.80	1.65	0.36	1.80	1.65	0.38
SF	0.29	-2.54	3.28	1.80	1.59	0.88	1.80	1.59	0.95
LT	0.32	-2.41	3.40	1.80	1.52	2.71	1.79	1.51	2.93
JB	0.28	-2.59	3.23	1.80	1.62	0.56	1.80	1.62	0.60
AC	0.27	-2.66	3.17	1.80	1.66	0.300	1.80	1.66	0.32
MLE	0.27	-2.64	3.19	1.80	1.65	0.36	1.80	1.65	0.38
M_T	0.5	0.10	3.16	1.88	1.15	285	1.87	1.14	310

Table 8: *Estimates of process capability and process performance indices for ln(0, 0.5)(2) data after normalization via goodness of fit tests*

Goodness of fit tests	λ Value	LSL	USL	PCI (Within Capability)			PPI (Overall Capability)		
				Cp	Cpk	PPM	Pp	Ppk	PPM
ln(0, 0.5)(2)	-	0.01	10	2.59	0.59	38718	2.60	0.59	37847
SW	-0.11	-6.00	2.03	2.51	1.28	62.67	2.54	1.30	49.46
AD	-0.16	-6.81	1.93	2.72	1.22	127.9	2.77	1.24	102.4
CVM	-0.12	-6.15	2.01	2.55	1.27	72.39	2.59	1.29	57.30
PT	-0.12	-6.15	2.01	2.55	1.27	72.39	2.59	1.29	57.30
SF	-0.1	-5.85	2.06	2.47	1.30	50.39	2.51	1.32	39.42
LT	-0.04	-5.06	2.20	2.27	1.38	17.58	2.30	1.40	13.52
JB	-0.14	-6.47	1.97	2.63	1.24	93.68	2.67	1.26	76.73
AC	-0.091	-5.72	2.08	2.44	1.31	43.33	2.47	1.33	33.92
MLE	-0.09	-5.71	2.08	2.43	1.31	43.33	2.47	1.33	33.96
M_T	0.0	-4.61	2.30	2.16	1.44	8.00	2.19	1.46	6.00

III. High Asymmetric Distribution

The lognormal distribution, with parameters $\mu = 0$ and $\sigma = 1$, can lead to highly skewed distributions. Through numerical examples, it's clear that methods like SF, AC, SW, MLE, and JB effectively transform non-normal data into normal distributions with minimal errors for dataset ln(0, 1)(1), and methods M_T, SF, AC, MLE, and SW do the same for dataset ln(0, 1)(2). Refer to Table 9 and Figure 4 for more details. In terms of producing accurate estimates of Pp/Cp and Ppk/Cpk, methods like SF, AC, JB, SW, and MLE perform well for dataset ln(0, 1)(1), and SW, SF, AC, and MLE perform well for dataset ln(0, 1)(2). Considering various error measures, estimates of process capability/performance indices, and corresponding PPM values, it's evident that methods like AC, MLE, SW, and SF perform better than others. The estimates and their PPM values in this study fall within the standard 4σ and 5σ limits. For dataset ln(0, 1)(1), values range from a minimum of 424 to a maximum of 495, while for dataset ln(0, 1)(2), values range from a minimum of 5664 to a maximum of 6230. The standard PPM value for 4σ limits is 6210, and for 5σ limits is

233. These values closely approach the standard of 5σ limits for highly asymmetric distributions. Refer to Tables 2, 10, and 11 for further details.

Table 9: *Various measures of error values for high symmetric data after the data transformation*

Goodness of fit tests	High Asymmetry (Sk=1.81) Lognormal distribution (μ=0, σ=1)			High Asymmetry (Sk=2.45) Lognormal distribution (μ=0, σ=1)		
	Bias	MdAE	RMSE	Bias	MdAE	RMSE
SW	1.4662	1.1958	1.6223	1.6526	1.3108	1.9611
AD	1.5072	1.2156	1.6798	1.6903	1.3319	2.0109
CVM	1.5335	1.2301	1.7174	1.7012	1.3379	2.0253
PT	1.5715	1.2509	1.7729	1.7514	1.3655	2.0912
SF	1.4612	1.1935	1.6153	1.6472	1.3072	1.9540
LT	1.5124	1.2185	1.6872	1.6848	1.3288	2.0038
JB	1.4662	1.1958	1.6223	1.6686	1.3198	1.9824
AC	1.4657	1.1956	1.6216	1.6506	1.3096	1.9585
MLE	1.4662	1.1958	1.6223	1.6526	1.3108	1.9611
M_T	1.5335	1.2301	1.7174	1.6102	1.2810	1.9043

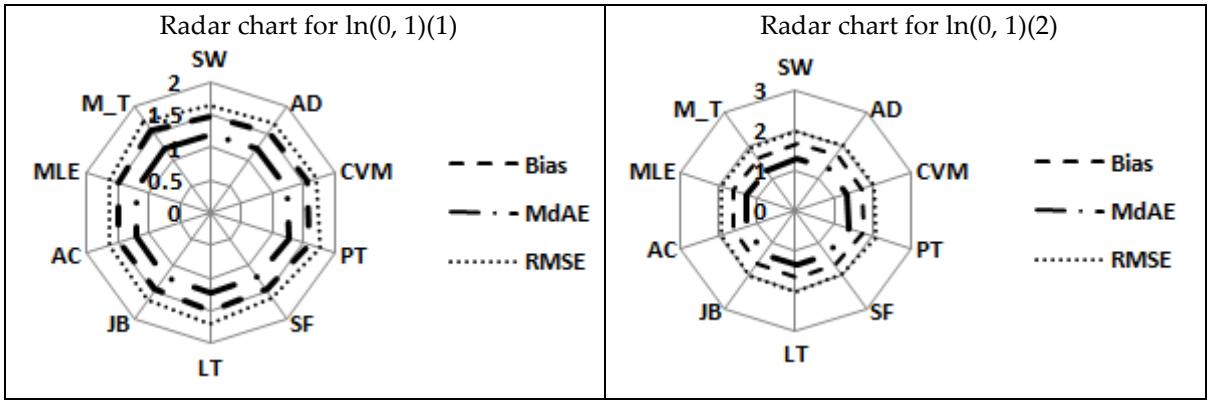

Figure 4: *Radar chart for various measures of errors after the normalization of high asymmetric distribution*

Table 10: *Estimates of process capability and process performance indices for ln(0, 1)(1) data after normalization via goodness of fit tests*

Goodness of fit tests	λ Value	LSL	USL	PCI (Within Capability)			PPI (Overall Capability)		
				Cp	Cpk	PPM	Pp	Ppk	PPM
ln(0, 1)(1)	-	0.01	10	1.143	0.324	165823	1.17	0.33	161042
SW	0.13	-3.47	2.68	1.25	1.10	492.37	1.28	1.13	355.35
AD	0.05	-4.11	2.44	1.31	1.00	1325.51	1.35	1.03	1016.34
CVM	0	-	-	-	-	-	-	-	-
PT	-0.07	-5.43	2.13	1.47	0.86	4947.72	1.50	0.88	4127.10
SF	0.14	-3.39	2.72	1.24	1.12	424.31	1.27	1.15	302.77
LT	0.04	-4.21	2.41	1.33	0.99	1548.60	1.36	1.01	1206.37
JB	0.13	-3.47	2.68	1.25	1.10	488.93	1.28	1.13	352.46
AC	0.13	-3.46	2.69	1.25	1.10	475.10	1.28	1.13	342.26
MLE	0.13	-3.47	2.68	1.25	1.10	494.90	1.28	1.13	357.47
Minitab	0.0	-4.61	2.30	1.096	0.766	10808	1.12	0.780	9660

Table 11: *Estimates of process capability and process performance indices for ln(0, 1)(2) data after normalization via goodness of fit tests*

Goodness of fit tests	λ Value	LSL	USL	PCI (Within Capability)			PPI (Overall Capability)		
				Cp	Cpk	PPM	Pp	Ppk	PPM
ln(0, 1)(2)	-	0.01	10	0.970	0.284	196712	0.953	0.83	200849
SW	-0.08	-5.57	2.10	1.47	0.83	6211.92	1.44	0.82	6896.12
AD	-0.15	-6.64	1.95	1.62	0.78	9864.76	1.60	0.77	10845.7
CVM	-0.17	-6.99	1.91	1.66	0.76	11394.4	1.64	0.75	12448.9
PT	-0.26	-8.89	1.73	1.94	0.68	19977.2	1.90	0.67	21670.8
SF	-0.07	-5.43	2.13	1.45	0.84	5663.84	1.43	0.83	6298.88
LT	-0.14	-6.47	1.97	1.59	0.78	9466.45	1.57	0.77	10412.8
JB	-0.11	-6.00	2.03	1.53	0.81	7697.25	1.51	0.79	8552.19
AC	-0.08	-5.52	2.11	1.46	0.84	6036.33	1.44	0.82	6704.05
MLE	-0.08	-5.57	2.10	1.47	0.83	6230.13	1.44	0.82	6918.05
Minitab	0.0	-4.61	2.30	1.189	0.839	5916	1.13	0.80	8528

V. Result and Discussion

This study investigates two main areas, focusing on data transformation and the estimation of process capability analysis. The effectiveness of different goodness-of-fit tests is evaluated based on various error measures, estimates of process capability/performance, and PPM values that closely adhere to the six sigma standard. It also explores the impact of optimal and rounded values of lambda when transforming non-normal data into normal data for estimating process capability analysis. To achieve desired outcomes, it is essential that the transformed data closely resemble a normal distribution with minimal errors. Additionally, consistency in producing standard estimates and lower PPM values from the extended transformed data serves as evidence that the methodology employed in this study yields the desired results.

In each of the three distinct asymmetric scenarios examined a range of goodness-of-fit tests, notably SW, SF, AC, and MLE, exhibit proficiency in converting non-normal datasets into normal distributions with minimal error values, the estimated values of Pp/Cp and Ppk/Cpk meet or exceed benchmark standards. The corresponding PPM values fall within or near the 6σ limits, only when low and moderate asymmetric distributions. For highly asymmetric distributions, the transformed dataset demonstrates reduced errors, yet the estimates of Pp/Cp and Ppk/Cpk deviate from standard results, and the corresponding PPM values do not align with the 6σ benchmark. It is noteworthy that across all asymmetric scenarios, error values are minimized for the JB, M_T, LT, and PT methods of goodness-of-fit tests. However, the associated estimates and PPM values do not correspond with the desired outcomes.

The primary objective of this paper is to obtain improved estimates of process capability or process performance indices using Box-Cox Transformation (BCT) through goodness-of-fit tests. When applying BCT to convert non-normal data into a normal distribution, selecting the transformation parameter λ becomes crucial. BCT provides an optimal and rounded value of lambda for data transformation. In this study, goodness-of-fit tests utilize the optimal value of λ, whereas M_T employs the rounded value of λ. Based on the numerical illustrations, it is observed that the PPM values estimated in this study using the optimal value of λ are higher than those estimated using other methods. Notably, for moderately asymmetric distributions, employing the M_T method results in higher PPM values compared to methods utilizing the optimal value of lambda for ln(0, 0.5)(1). However, for ln(0, 0.5)(2), PPM values of minimum compared to goodness of fit tests but results in lesser values of Pp/Cp and Ppk/Cpk compared to benchmark standards.

Similarly, for highly asymmetric distributions, the M_T method produces higher PPM values compared to methods utilizing the optimal value of λ for ln(0, 1)(1). Nonetheless, for ln(0, 1)(2) PPM values are minimum compared to goodness of fit tests but results in lesser values of Pp/Cp and Ppk/Cpk compared to benchmark standards.

Table 12: *Efficiency comparison over different goodness of fit tests in data transformation and estimation of process capability and process performance indices for lognormal distribution*

Goodness of fit tests	Efficiency in data transformation						Efficiency in estimation of PCI					
	Low Asymmetric		Moderate Asymmetric		High Asymmetric		Low Asymmetric		Moderate Asymmetric		High Asymmetric	
Skewness	0.36	0.63	0.96	1.32	1.81	2.45	0.36	0.63	0.96	1.32	1.81	2.45
SW	✓	✓	✓	✓	✓	✓	✓*	✓*	✓*	✓*	✓*	✓*
AD	✓		✓				✓					
CVM			✓				✓$		✓			
PT	✓								✓$			
SF		✓	✓	✓	✓	✓	✓	✓*	✓*	✓*	✓*	✓*
LT	✓		✓	✓						✓		
JB	✓	✓	✓		✓	✓		✓	✓		✓*	✓*
AC		✓		✓	✓	✓	✓$	✓*	✓$	✓*	✓*	✓*
MLE		✓		✓	✓	✓	✓$	✓*	✓$	✓*	✓*	✓*
M_T	✓		✓	✓		✓	✓	✓$				
DME												

DME – Direct Minitab Estimation | ✓- Less Error and/or Better Estimate | ✓* - Better Estimate with less error and lesser PPM values | ✓$ - Better Estimates with less PPM and higher error values.

This is clearly indicates that using rounded values of λ is somewhat less efficient in transforming non-normal data into normal data, resulting in corresponding estimates of process capability/performance that do not meet the benchmark standards compared to the results obtained from methods utilizing the optimal value of λ. This discrepancy arises because the use of rounded value of λ does not accurately reflect the transforming pattern needed to achieve a close approximation to a normal distribution, as opposed to the optimal value of λ. Therefore, it is evident that opting for the optimal value of λ to attain improved estimates in process capability analysis would be the superior choice. One may refer table 4, 5, 7, 8, 10 and 11. A table of data is formed for the better understanding of the efficiency of different normality tests under various asymmetric behaviors of lognormal distribution based on the numerical example, result and discussion. See table 12 for more information.

V. Conclusion

The core objective of this research work is to analyze the impact of the transformation parameter lambda on enhancing the process capability assessment for lognormal distribution. A test that satisfies all the requirements including data closely adhering to a normal distribution with less error, enhanced estimates of process capability/performance and reduced PPM values, to achieve the desired result is looked at utilizing low, moderate, and high asymmetric log normal distribution. Accordingly, based on the findings, result and discussion, SW, SF, AC, and MLE methodologies of goodness of fit tests have more intense power to estimate process capability/performance indices with smaller PPM values and also have higher accuracy in data

transformation. The SW test performs better than the other approaches in every way to produce enhanced estimates of process capability analysis. However, other test methods, like the SF, AC, and MLE methods, position themselves subsequent places for dealing with non-normal quality characteristics, particularly lognormal distribution and delivering remarkably good results.

M_T (Minitab Transformation) utilizes the rounded value of λ instead of the optimal value to ascertain the transforming parameter. Optimal λ is typically required for superior results as it accurately reflects the transforming pattern, unlike a rounded value. This approach ensures that all values are brought as close to normal as possible. Based on the numerical illustrations, this study produces large amount of error values while using rounded value of λ, except in low asymmetry situations, resulting in the transformed data not close enough to normal distribution and less efficient estimates when compared to methods that are utilizing an ideal value of λ during data transformation and estimation. Furthermore, it is concluded and recommended that when dealing with non-normal data specifically lognormal distribution, utilizing an optimum value of λ is typically required for better results and Shapiro Wilk's (SW) test is one such method among the different goodness of fit tests to transform non normal data into normal and estimating process capability/ performance in order to get enhanced results.

References

[1] Asar O, Ilk O and Dag O (2013). Estimating Box Cox power transformation parameter via goodness of fit tests. Communication in statistics – simulation and computation, 46(1), pp 91 – 105.

[2] Box.G.E.P., Cox. D.R., (1964). An analysis of Transformations. Journal of the royal statistical society. Series B(Methodological), 26(2), 211-252.

[3] Dag O, Asar O and Ilk O (2013). A Methodology to Implement Box-Cox Transformation When No Covariate is Available. Communication in statistics – simulation and computation, 43(7), pp 1740 – 1759.

[4] Gunter, B.H. (1989). The use and abuse of Cpk ,1-4, Quality Progress, 22(1), 72-73(3), 108-109(5), 79-80(7).

[5] Kane V E (1986), Process capability indices. Journal quality technology, 18(1), pp 41 – 52.

[6] Oztuna D, Elhan A and Tuccar E (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions, Turkish Journal of Medical Sciences, 36(3), pp 171-176.

[7] Pearn, W. L., & Chen, K. S. (2002). One-sided capability indices CPU and CPL: decision making with sample information. International Journal of Quality & Reliability Management, 19(3), 221–245.

[8] Pyzdek T, The Six Sigma Handbook, ISBN 0-07-141015-5, McGraw-Hill Companies, Inc, New York 2003.

[9] Rahman M (1999). Estimating the box cox transformation via Shapiro wilk W statistic. Communication in statistics – simulation and computation, 28(1), pp 223 – 241.

[10] Rahman M and Pearson LM (2008). Anderson-darling statistic in estimating the box-cox transformation parameter. Journal of applied probability & statistics, 3(1), pp 23 – 35.

[11] Sennaroglu B and Senvar O (2015). Performance comparison of Box-Cox transformation and weighted variance methods with weibull distribution. Journal of Aeronautics and Space Technologies, 8(2), pp 49-55.

[12] Sibalija TV and Majstorovic VD (2010). Process performance analysis for non normal data distribution. International Journal "Total Quality Management & Excellence", 38(3), pp 1-4.

[13] Tang LC and Than SU EE, (1999). Computing process capability indices for non normal data: A review and comparative study. Quality and Reliability Engineering International, 15, pp 339 – 353.

[14] Thadewald T and Buning H (2007), Jarque-Bera Test and its Competitors for Testing

Normality - A Power Comparison, Journal of Applied Statistics, 34(1), pp 87-105.

[15] Yang Y and Zhu H (2018). A study on process capability analysis based on Box-Cox transformation, IEEE, pp 240 – 243.

[16] Yap B W and Sim C H (2011), Comparisons of various types of normality tests. Journal statistical computation and simulation, 18(12), pp 2141 – 2155.

[17] Yoap T (2006). Process capability analysis for non normal data with Minitab, Six Sigma: Advances Tools for Black Belts and Master Black Belts, pp 131 – 149.

[18] Krishnan J and Vijayaraghavan R (2024), Process capability analysis for non normal data Based on box-cox transformation through Tests of goodness of fit, Reliability: Theory & Applications. March 1(77), 297-309.

[19] Swamy, D. R., Nagesh, P., and Wooluru, Y. (2016). Process Capability Indices for Non-normal Distribution – A Review, Proceedings of the International Conference on Operations Research and Management, January 21 – 22, Mysuru, India.