Society for Multidisciplinary & Advanced Research Techniques (SMART)
Trinity University, Lagos, Nigeria
SMART Scientific Projects & Research Consortium (SMART SPaRC)
Sekinah-Hope Foundation for Female STEM Education
Harmarth Global Educational Services
ICT University Foundations USA
IEEE Computer Society Nigeria Chapter

Proceedings of the 36th iSTEAMS Accra Bespoke Multidisciplinary Innovations Conference

Comparison Analysis of Normalization of Non-Normalized Uniform Distributed Data

[1]Omisore, Adedotun Olurin & [2]Adegbite, Ismaila Olawale
[1&2]Department of Statistics, Osun State Polytechnic, Iree, Nigeria
E-mail: [1]omisoreadedotunolurin@gmail.com [2]adegbiteonline@gmail.com

ABSTRACT

The problem of using non-normal data set as a normal data set has caused a lot of inadequacy and inefficient result for researcher's analysis. To solve this, we look in to this problem by exploring different options. This prompted us to use Shapiro-Wilk test and normal Q-Q plots to test whether the data sets violated the normality assumptions or not. The uniform distributed data was generated randomly via Microsoft Excel package and was used for the analysis. Four different methods of data normalization; Winzorising, Detection and Deletion of Outlier, Trimming and Transformation methods were performed on the non-normal data set generated in order to normalize them. The power test values show that the appropriate method to normalize the continuous data was inferred to be the Square root transformation method. It is thereby recommended that when left with options, Transformation method should be deployed in normalizing Uniform data for inferential purposes.

Keywords: Data set, Transformation, Continuous data, Normalization, Uniform distribution.

Proceedings Citation Format
Omisore, A.O. & Adegbite, I. O. (2023): Comparison Analysis of Normalization of Non-Normalized Uniform Distributed Data. Proceedings of the 36th iSTEAMS Accra Bespoke Multidisciplinary Innovations Conference. University of Ghana/Academic City University College, Accra, Ghana. 31st May – 2nd June, 2023. Pp 297-307
https://www.isteams.net/ghanabespoke2023. dx.doi.org/10.22624/AIMS/ACCRABESPOKE2023P27

1. INTRODUCTION

Normalization of data was carried out by different researchers and they came up with their findings. In this paper, we will be concentrating on, Transforming method, Winzorising method, Trimming method and Deletion of outlier Method. Normal Distribution is a distribution that has most of the data in the center with decreasing amounts evenly distributed to the left and the right (Hansen, Irizarry & Wu, 2012). Non-normal Distributions or Skewed Distribution is distribution with data clumped up on one side or the other with decreasing amounts trailing off to the left or the right. (Lumle, Diehr, Emerson & Chen, 2002; Kristin, 2002). Examples of this type of data are water rate or electricity consumption. On the other hand, normal data are data that are drawn or come from a population that has a normal distribution.

There are reasons why data sets area not normal, they are;

i. *Extreme values:* Too many extreme values in a data set will result in a skewed distribution (Helmes & Jackson, 1982). Normality of data can be achieved by cleaning the data. This involves determining measurement errors, data-entry errors and outliers, and removing them from the data for valid reasons. (Yannick, Johannes & Li-Xuan, 2022).
ii. *Overlap of two or more processes:* If two or more data sets that would be normally distributed on their own are overlapped, data may look bimodal or multimodal - it will have two or more most-frequent values. The remedial action for these situations is to determine which X's cause bimodal or multimodal distribution and then stratify the data (Boulund, Pereira, Jonsson & Kristiansson, 2018).
iii. *Insufficient data discrimination*: Round-off errors or measurement devices with poor resolution can make truly continuous and normally distributed data look discrete and not normal (Pereira, Wallroth & Jonsson, 2018). Insufficient data discrimination – and therefore an insufficient number of different values — can be overcome by using more accurate measurement systems or by collecting more data.
iv. *Sorted data:* Collected data might not be normally distributed if it represents simply a subset of the total output a process produced. This can happen if data is collected and analyzed after sorting.
v. *Values close to zero or a natural limit:* If a process has many values close to zero or a natural limit, the data distribution will skew to the right or left. In this case, a transformation, such as the Box-Cox power transformation, may help make data normal.
vi. *Data follows a different distribution:* There are many data types that follow a non-normal distribution by nature. Examples include: Weibull distribution, Log-normal distribution, Largest-extreme-value distribution, Exponential distribution, Poisson distribution, Binomial distribution among others.

If data are not normally distributed they cannot be used as a normal distributed data likewise they cannot assumed normality and use any of the following statistical tools; ANOVA t test, regression analysis, paired t- test, F-test, Bartlet's test, individual control chart, est. If a data sets are not normally distributed it is often possible to normalize them by Winsorization method, Trimming method, Outlier deletion method and Transformation method.

Winsorization Method

This is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers (Sufahani & Ahmad, 2012). A typical strategy is to set all outliers to a specified percentile of the data; for example, a 90% Winsorisation would see all data below the 5th percentile set to the 5th percentile, and data above the 95th percentile set to the 95th percentile. In a simpler form, usually, but necessarily, performed in a symmetrical fashion.

Rank data, and then give extremes the same value as adjacent rank. Re-compute statistics and test for normality. The approach is to shrink the smallest and largest observations to the next remaining observations, and count them as if they had those values. The α – Winsorized mean is computed as

$$\propto = \text{Winsorized mean} = \frac{1}{n}\left[(k+1)(X(k+1)+X(n-k)) + \sum_{i=k+2}^{n-k+1} X(i)\right] \qquad (1)$$

k is the smallest integer ≥ αn and X(i) is the order statistics.

Trimming Method

This is a method of averaging that removes a small percentage of the largest and smallest values before calculating the mean. After removing the specified observations, the trimmed mean is found using an arithmetic averaging formula. Investopedia explains that Trimmed Mean reduces the effects of outliers on the calculated average. This method is best suited for data with large, erratic deviations or extremely skewed distributions. A trimmed mean is stated as a mean trimmed by X%, where X is the sum of the percentage of observations removed from both the upper and lower bounds. The trimmed mean helps to reduce the effects of outliers on the calculated average (Luciano da Fontoura, 2022). This method is best suited for data with large, erratic deviations or extremely skewed distributions. A trimmed mean is stated as a mean trimmed by X%, where X is the sum of the percentage of observations removed from both the upper and lower boundary.

$$\propto = \text{trimmed mean} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X(i) \qquad (2)$$

k is the smallest integer ≥ αn and X(i) is the order statistics.

Outlier Detection And Deletion

Presence of outliers can easily skew normal-distributed data. Outlier can be caused by error in measurement or data entry, if they can be identified and removed, we might be able to obtain distributed data from the skewed data set. Outliers should only be removed if a specific cause of their extreme value is identified (Hansen, Irizarry & Wu, 2012). We have different ways of detecting outlier, such as; Z-score method via SPSS, Grubb test via Excels e.t.c.

Transformation Method

Data transformations are the application of a mathematical modification to the values of a variable. These are important options for analysts, but they do fundamentally transform the nature of the variable, making the interpretation of the results a time more complex. Transformation of data allows the use of parametric statistics and is completely legitimate. Transformation can be in any of these forms:

Logarithms: Growth rates are often exponential and log transforms will often normalize them. Log transforms are particularly appropriate if the variance increases with the mean. Log transformation can take different form as shown below:-

$$Y' = \text{Log}_{10}(X) \qquad (3)$$
$$Y' = \text{Log}_{10}(X+1) \qquad (4)$$
$$Y' = \text{Ln}(X) \qquad (5)$$
$$Y' = \text{Ln}(X+1) \qquad (6)$$

Reciprocal: If a log transform does not normalize your data you could try a reciprocal (1/x) transformation. It enables very small numbers to be very large, and very large numbers to be very small. This transformation has the effect of reversing the order of your scores. Thus, one must be careful to reflect, or reverse the distribution prior to applying an inverse transformation.

Square root: This transform is often of value when the data are counts, e.g. blood cells on a haemocytometer or woodlice in a garden. Carrying out a square root transform will convert data with a Poisson distribution to a normal distribution.

Various forms are;

$$x' = \sqrt{X} \quad (7)$$

$$x = \sqrt{X} + \sqrt{X+1} \quad (8)$$

$$x' = \sqrt{X + {}^{3}/_{8}} \quad (9)$$

$$x' = \sqrt[3]{X} \quad (10)$$

Arcsine: This transformation is also known as the angular transformation and is especially useful for percentages and proportions and involves in taking the arcsine of the square root of a number, with the resulting transformed data reported in radians. Because of the mathematical properties of this transformation, the variable must be transformed to the range -1.00 to 1.00. Whenever the data are proportions or percentages, one should consider an angular Transformation Percentages tend to usually follow a binomial distribution. Typical transforms: θ = arcsine $\sqrt{p}$

where θ ranges from 0 to 1.

$$\theta = \text{Arcsine}\sqrt{\frac{X + \frac{3}{8}}{N + \frac{3}{4}}} \quad (11)$$

Statistical power

The power of a statistical test is the probability of correctly rejecting a false null hypothesis; i.e. the probability of not committing a Type II error (R Core Team, 2017). As the Power increases, the chances of Type II error occurring decreases. The probability of a Type II error occurring is referred to as (β) and the power is equal to 1-β.

Components of statistical power analysis are;

a) Standardized effect size; (1). Effect size and (2). Variation (variability).
b) Sample size (N).
c) Test size (significant level).
d) Power of test (1-β).

The misuses of non-normal distributed data to be normal data has become rampant to the extent that the data are misused for normal data. But normal distribution does not happen as often as people think, yet the reality is that almost all analyses benefit from improving the normality of variables, particularly where substantial non-normality is present.

The focus of this paper is on transformations that improve normality of data and in arriving at efficient result and decision making by considering a data set that follows continuous Uniform distribution. In order to determine the appropriate method to use to transform data, the power test is required, i.e. by finding the power test value under each method of normalization and determines the method with highest power test value as the appropriate method.

2. AIM AND OBJECTIVES

This paper aims at comparing different methods of normalization of non-normalized uniform distributed data.

While the objectives are to:

1. use Zscore to check for the presence of outliers in the original generated continuous data.
2. use Trimming method, Winsorising, Detection & deletion of outlier and Transformation method to normalize a generated Uniform distributed data.
3. to determine the power values for each of the methods of data normalization used.

Hypothesis

H_0: The data is normally distributed.
H_1: The data is not normally distributed.

3. METHODOLOGY

A random number generator (RNG) which is a computational or physical device designed to generate a sequence of numbers or symbols that lack any pattern was used to generate random numbers that follows Uniform distribution using Microsoft Excel. Zsocre tool was used to detect outliers from the generated data while Winsorising method, Trimmed data method, Outlier detection method and Transformation methods were deployed to normalized the data generated data and make comparison to bring out the best method of normalization of discrete data among the methods with the aid of Power test statistic and Statistical Package for Social Scientists (SPSS V. 23).

4. ANALYSIS

a. Test for Normality of the generated Uniform distributed data.

The Shapiro-Wilk test is a specific test for normality, whereas the method used by Kolmogorov-Smirnov test is more general, but less powerful. Here we focus on Shapiro-Wilk test because is more appropriate for small sample sizes (< 50 samples), but it can also handle sample sizes as large as 2000. If the value of significant (i.e. P value) from Shapiro-Wilk is greater than α value (0.05), then we can conclude that the data is normalized.

Table 1: Non- normal Uniform distributed generated data and the Z-Scores values.

S/n	Uniform data	Zscore Uniform	S/n	Uniform data	Zscore Uniform	S/n	Uniform data	Zscore Uniform
1	59.0076	-0.82472	16	84.4687	-0.16353	31	30.6839	-1.56026
2	70.7485	-0.51983	17	102.4379	0.30311	32	83.2255	-0.19581
3	128.6392	0.98353	18	80.1134	-0.27663	33	65.2277	-0.66319
4	106.8914	0.41877	19	94.7497	0.10346	34	52.7630	-0.98689
5	142.9933	1.35630	20	102.1721	0.29621	35	90.3126	-0.01177
6	86.9756	-0.09843	21	142.7152	1.34904	36	135.7549	1.16832
7	44.9877	-1.18899	22	112.9479	0.57605	37	119.7037	0.75149
8	61.4613	-0.76100	23	28.6719	-1.61251	38	39.7176	-1.32567
9	133.5425	1.11087	24	29.5552	-1.58957	39	76.6865	-0.36562

S/n	Uniform data	Zscore Uniform	S/n	Uniform data	Zscore Uniform	S/n	Uniform data	Zscore Uniform
10	75.9176	-0.38559	25	112.9356	0.57573	40	117.9289	0.70540
11	27.7476	-1.63651	26	53.1474	-0.97691	41	148.4405	1.49775
12	48.0028	-1.11051	27	29.8824	-1.58108	42	90.6316	-0.00348
13	126.6722	0.93245	28	143.9298	1.38061	43	152.1537	1.59418
14	130.2423	1.02516	29	147.8434	1.48225	44	128.0544	0.96835
15	71.0062	-0.51313	30	115.8719	0.65198	45	56.9015	-0.87942

Table 1 shows the test of normality of the original generated Uniform distributed data.

Detrended Normal Q-Q Plot of original Uniform data

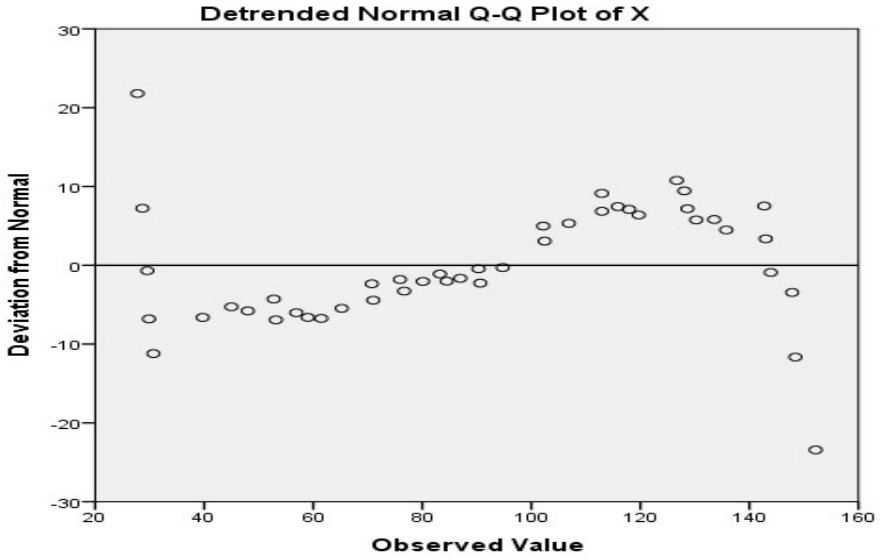

Fig.1: The graph that shows the detrended Normal Q-Q plot of the original generated Uniform data.

From Fig.1, since the value points are not close to the diagonal line on the Q-Q plot then it is concluded that the data is not normal.

Table 2: Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Original Uniform Data	.095	45	.200	.947	45	.041

Hypothesis:
H_0: The data is normally distributed.
H_1: The data is not normally distributed.

Conclusion: From Table 2, Since the Shapiro-Wilk value of significant is less than α, (p- value (0.041< 0.05), H_0 is thereby rejected and then conclude that the data is not normal.

b. Detection of outlier using Zscore method

Zscore was used to determine whether there is outlier in the uniform distributed data. The result obtained was shown in Table 1**.** The table shows that outlier is not present in the Uniform distributed data because the data falls in between the lower and upper limit values or boundaries; i.e. -3 to +3.

c. Normalising the original generated Uniform distributed data

The original generated Uniform distributed data sets was confirmed not to be a normalized data and was manipulated to be normal by the Winzorising, Trimmed data, Outlier detection and Transformation methods. This was presented in Table 3 below;

Table 3: The Table showing the original generated distributed data that was not normal and the normalized data with 3 different methods of data normalization.

S/N	Uniform Data	Winzorising	Trimming	Transformation (log transformation)
1	152.1537	148.4405		7.68
2	148.4405	148.4405	148.4405	8.41
3	147.8434	147.8434	147.8434	11.34
4	143.9298	143.9298	143.9298	10.34
5	142.9933	142.9933	142.9933	11.96
6	142.7152	142.7152	142.7152	9.33
7	135.7549	135.7549	135.7549	6.71
8	133.5425	133.5425	133.5425	7.84
9	130.2423	130.2423	130.2423	11.56
10	128.6392	128.6392	128.6392	8.71
11	128.0544	128.0544	128.0544	5.27
12	126.6722	126.6722	126.6722	6.93
13	119.7037	119.7037	119.7037	11.25
14	117.9289	117.9289	117.9289	11.41
15	115.8719	115.8719	115.8719	8.43
16	112.9479	112.9479	112.9479	9.19
17	112.9356	112.9356	112.9356	10.12
18	106.8914	106.8914	106.8914	8.95
19	102.4379	102.4379	102.4379	9.73
20	102.1721	102.1721	102.1721	10.11
21	94.7497	94.7497	94.7497	11.95
22	90.6316	90.6316	90.6316	10.63
23	90.3126	90.3126	90.3126	5.35
24	86.9756	86.9756	86.9756	5.44
25	84.4687	84.4687	84.4687	10.63
26	83.2255	83.2255	83.2255	7.29
27	80.1134	80.1134	80.1134	5.47
28	76.6865	76.6865	76.6865	12.00
29	75.9176	75.9176	75.9176	12.16
30	71.0062	71.0062	71.0062	10.76
31	70.7485	70.7485	70.7485	5.54
32	65.2277	65.2277	65.2277	9.12
33	61.4613	61.4613	61.4613	8.08
34	59.0076	59.0076	59.0076	7.26
35	56.9015	56.9015	56.9015	9.50
36	53.1474	53.1474	53.1474	11.65
37	52.7630	52.7630	52.7630	10.94
38	48.0028	48.0028	48.0028	6.30

S/N	Uniform Data	Winzorising	Trimming	Transformation (log transformation)
39	44.9807	44.9807	44.9807	8.76
40	39.7176	39.7176	39.7176	10.86
41	30.6839	30.6839	30.6839	12.18
42	29.8824	29.8824	29.8824	9.52
43	29.5552	29.5552	29.5552	12.34
44	28.6719	28.6719	28.6719	11.32
45	27.7476	28.6719		7.54

The data was tested whether it is normal after using the 4 different methods to normalize the original generated Uniform data. Shapiro-Wilk test & Normal Q-Q plot were used to test the normality of the data. The outputs of each method are shown below.

Condition: If P-value of Shapiro-Wilk value is greater than α=0.05, H_0 is accepted.

Table 4: The table showing the test of normality of Winzorising method of normalization using Shapiro-Wilk test of normality.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
Winsorized data	.097	45	.200	.943	45	.127

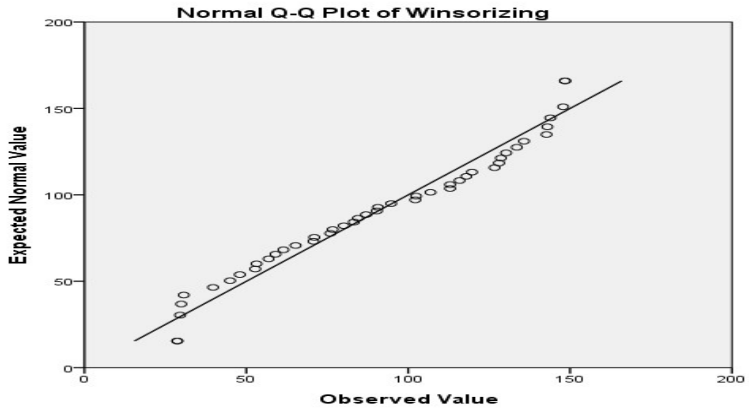

Fig. 2: Q-Q plots of Winzorising normalizing method.

Table 5: The table showing the test of normality of Trimming method of normalization using Shapiro-Wilk test of normality.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
Trimmed data	.097	43	.200	.950	43	.060

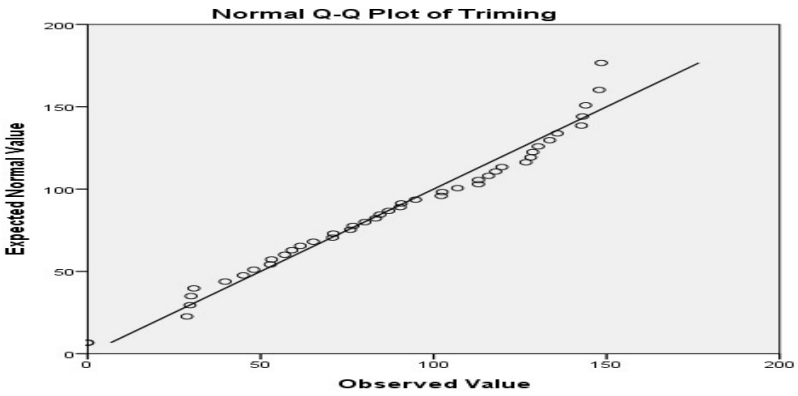

Fig. 3: Q-Q plots of Trimming normalizing method.

Table 6: The table showing the test of normality of Square root transformation method of normalization using Shapiro-Wilk test of normality.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
Transformed data	.092	45	.120	.950	43	.131

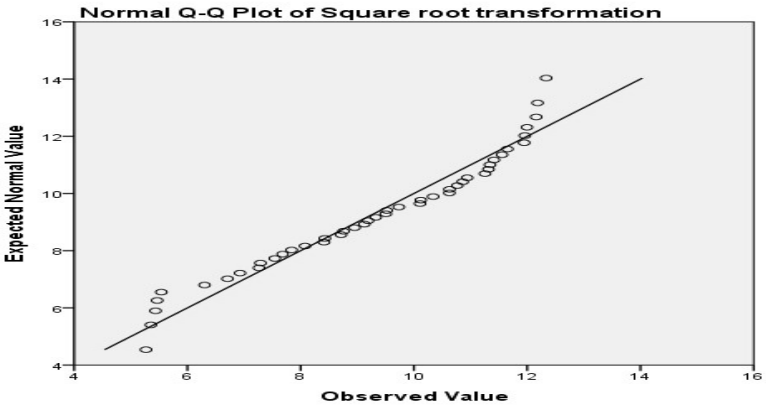

Fig. 4: Q-Q plots of Square root transformation normalizing methods.

Conclusion: From the above output (Table 4-6), taken α =0.05, the Shapiro-Wilk significant value (p value) is greater than α value. i.e., all the original generated Uniform data were normalized after using the three methods, Also the Normal Q-Q plot shows that the value points are close to and they are on the diagonal line. Therefore, all the null hypotheses were accepted and concluded that the data is normally distributed.

d. Comparison Analysis

The Power test value and different sample sizes of each method was found using G-Power programming package.

Table 7: Power test values of the four methods.

S/N	SAMPLE SIZE	TRIMMED POWER 1-βrr	WINSORIZED POWER 1-βrr	TRANSFORMATION 1-βrr
1	40	0.0500046	0.0500114	1.0000
2	41	0.0500047	0.0500117	1.0000
3	42	0.0500048	0.0500120	1.0000
4	43	0.0500049	0.0500123	1.0000
5	44	0.0500051	0.0500126	1.0000
6	45	0.0500052	0.0500129	1.0000
7	46	0.0500053	0.0500132	1.0000
8	47	0.0500054	0.0500135	1.0000
9	48	0.0500055	0.0500138	1.0000
10	49	0.0500057	0.0500141	1.0000
11	50	0.0500058	0.050014	1.0000

Decision Rule: The method with the highest power is the most appropriate method.

Remark: From Table 7, it is concluded that Transformation method using square root is the most appropriate and efficient method to be used to normalize the data because it has the highest power even across different sample sizes.

5. CONCLUSION

We compare the best method to normalize a Continuous Distribution, from the analysis it is observed that Transformation method is appropriate to normalize continuous distribution. It was also inferred that since the lower sample size and higher sample size do not affect the power test value of Square root transformation method therefore it is more efficient to use Square root transformation method to normalize a continuous distribution and uniform distribution in particular.

6. RECOMMENDATIONS

i. In order to normalize a uniform distributed data, Square root transformation method being the most appropriate is hereby suggested to employ.
ii. Researchers and statisticians should take pain to determine the powers of all methods that are being used to normalize a particular data in order to know the appropriate method for normalization before embarking on analysis so that results will be efficient and reliable to make a reliable decision.

REFERENCES

Boulund, F., Pereira, M.B., Jonsson, V., Kristiansson, E. (2018). Computational and statistical considerations in the analysis of metagenomic data In: Nagarajan, M.: Metagenomics: Perspectives, Methods and Applications. Academic Press, Cambridge.

Hansen, K.D., Irizarry, R.A, Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 13(2): 204–216. https://doi.org/ 10.1093/biostatistics/kxr054.

Helmes, E., & Jackson, D. N. (1982). A comparison of methods of normalizing a discrete distribution. **Journal of Clinical Psychology, 38**(3), 581–587. https://doi.org/ 10.1002/1097-4679(198207)38:3<581::AID-JCLP2270380318>3.0.CO;2-5.

Kristin, L. S. (2012). Dealing with Non-normal data. The American Academy of Physical Medicine and Rehabilitation. 4, 1001-1005. http://dx.doi.org/10.1016/j.pmrj.2012.10.013.

Luciano da Fontoura, C. (2022). Data Normalization in Signal and Pattern Analysis and Recognition: A Modelling Approach. ffhal-03688208v2f.

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annu Rev Public Health; 23:151-169.

Pereira, M., Wallroth, M., Jonsson, V. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC Genomics 19, 274-290. https://doi.org /10.1186/s12864-018-4637-6.

R Core Team. R: (2017). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.r-project.org.

Sufahani, S. & Ahmad, A. (2012). A Comparison between Normal and Non-Normal. Applied Mathematical Sciences, 6(92), 4547 – 4560.

Yannick, D., Johannes, L. & Li-Xuan, Q. (2022). Depth normalization of small RNA sequencing: using data and biology to select a suitable method Nucleic Acids Research, 50(10), 56-63, https://doi.org/10.1093/nar/gkac064.