BAU Journal - Science and Technology

Volume 3 | Issue 2
ISSN: 2706-784X

Article 7

June 2022

# NORMALITY TESTING METHODS AND THE IMPORTANCE OF SKEWNESS AND KURTOSIS IN STATISTICAL ANALYSIS

Georges Hatem
*University of Porto, Portugal, Lebanese University, Lebanon*, georges.r.hatem@gmail.com

Joe Zeidan
*University of Limoges, France*, joe.zeidan@xlim.fr

Mathijs Goossens
*Centre for cancer detection*, mathieu.goossens@bevolkingsonderzoek.be

Carla Moreira
*University of Porto, Portugal*, carlamgmm@gmail.com

# NORMALITY TESTING METHODS AND THE IMPORTANCE OF SKEWNESS AND KURTOSIS IN STATISTICAL ANALYSIS


## Abstract

The normal distribution (bell curve or Gaussian distribution) is a distribution that happens commonly in many circumstances. Real-life data rarely, if ever, follow a perfect normal distribution. Many tests are useful to test normality and more particularly skewness and kurtosis tests assess the comparability of a given distribution from a normal distribution. These tests are widely used in statistics, business, and epidemiological data including blood pressure, heights, IQ scores and measurement errors. This report provides a review assessing the essential methods employed for testing normality and highlighting the importance of skewness and Kurtosis in statistics. Moreover, it gives some examples of the importance of normality in epidemiological health studies analysis.



## Keywords

Normality, Skewness, Kurtosis, test, normal distribution.

## 1. INTRODUCTION

The normal distribution, alias Gaussian distribution or bell-curved distribution, includes a family of distributions with continuous probability. They all share the same general shape, but differ in their position (the mean or average) and scale parameters (the standard deviation) (Ahsanullah et al, 2014).

Normal distribution also designated as "standard normal distribution" is obtained by taking $\mu = 0$ and $\sigma^2 = 1$ in a general normal distribution. Whereas, normal distribution could be adapted to a standard normal distribution by changing variables to $Z \equiv (X - \mu)/\sigma$, so $dz = dx/\sigma$ (Feller et al, 1971).

Many proprieties undergo with normal distributions where random variants with unknown distributions are sometimes assumed to be normal in some sciences (physics and astronomy) (Metcalf, 2021). This approximation is based on the central limit theorem. The aforementioned states that the mean of any set of variants with any distribution having a defined mean and variance has a propensity to the normal distribution. Many common attributes such as test scores, height and weight follow roughly normal distributions, with many observations in the middle and few members at the high and low ends (Weisstein et al, 2021).

To draw the inference from the observations made on the study sample in terms of different groups, statistical methods are used. These methods require many assumptions including normality of the continuous data (Mishra et al, 2019). Various methods exist to test the normality of data such as numerical and visual methods, and each method undergoes its own advantages and limitations. Furthermore, normality tests are based on many characteristics including the empirical distribution, moments, correlation, and regression (Altman et al, 1995).

## 2. BACKGROUND OF THIS REPORT

One of the common practices in statistic and data analysis is to find out if a data sample originated from a normally-distributed population (Hernandez, 2021). Currently, many methods have been suggested in the scientific literature for evaluating normality. However, knowing the best method for testing normality could have substantial benefits on the analysis of the sample data. More particularly, normality can be tested through two other measures of shape: skewness and excess kurtosis (Garren et al, 2021). This report provides a review assessing the essential methods employed for testing normality and highlights the importance of skewness and kurtosis in statistics. Moreover, it scrutinizes some examples that emphasize the importance of implementing such methodology in epidemiological health studies analysis.

## 3. OVERVIEW OF NORMALITY TESTS

Normality testing starts by drawing hypothesis where a null statement, assuming that the studied data is not statistically different than normal versus an alternate statement that the normal distribution and the data are not compatible.

Given the importance of normality in statistical analysis, many tests exist. Two of the most frequent methods to test normality are Kolmogorov-Smirnov (KS) test and Shapiro-Wilk (SW) test (Razali et al, 2011).

Apart than KS and SW tests, additional methods are applied for checking the normality of a designated data set. Among those methods, graphical methods are applicable: Histogram, boxplot or probability-probability (P-P) plots (Demir et al, 2016). For example, the shape of the histogram for a data set is explored to see if it flows into the normal distribution shape or not. Despite its frequent use, one of the main disadvantages is the subjectivity of decisions. Nevertheless, computing the results of graphical methods with different approaches to check the shape of the distribution could be informative. Therefore, it will be useful to use graphical methods along with other practices (Orcan et al, 2020). Graphical tests are mainly one of the noteworthy conducted methods (i.e. histograms where the bars should feature a symmetric bell, steam-and-leaf plot and the box-and-whisker plot) (Das et al, 2016). Other tests are based on the Empirical Cumulative Distribution Function (i.e. the Crámer-von Mises, the Kolmogorov-Smirnov test, the Anderson-Darling improvement of the CVM test, the Kuiper) versus the tests based on Regression and Correlation (Order Statistics) and tests based on moments (Bakshaev, 2009).

While the importance of determining the best method to test normality cannot be overstated, its identification is highly challenging. The most common approach is by determining their power using a Monte Carlo simulation approach, where different samples of different size are obtained from different non-normal distributions. In general, regression-based normality tests are considered more accurate when compared to other types of tests (Hernandez, 2021).

## 4. SKEWNESS AND KURTOSIS

The normality of data can be checked also according to skewness and kurtosis. Even though skewness and kurtosis are widely used in practice, there is no general rule defining the values that indicate normality. Some papers reported that up to an absolute value of 1 for skewness and kurtosis might be translated to normality, while other reports suggested much larger values of skewness and kurtosis to test the normality of the data (Şirin et al, 2018). In this part, this report will start by defining each of the terms, then evaluate some of the applications of skewness and Kurtosis in public health studies.

**Skewness** assesses the extent to which a variable’s distribution is symmetrical. If the distribution of responses for a variable stretches toward the right or left tail of the distribution, then the distribution is referred to as skewed (Hair, 2017). More precisely, it is a measure of the lack of symmetry and the relative size of the two tails. Ideally, a skewness equal to 0 is noted in a symmetrical dataset or a normal distribution. Figure 1 illustrates the different type of skewness (Weinstein, 2021).

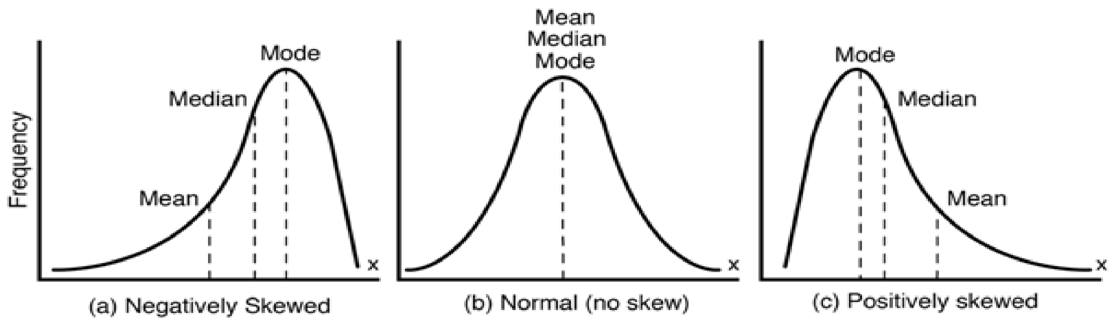


Fig.1: Geometrical presentation of the different type of skewness

In normal distribution (b), the mean, median and mode values are close to each other and converge to the expected value of the studied parameter. Thus, skewness of the data can be evaluated according to the relationship between these values whereas in (a) Left Skewed or Negatively Skewed: the mean and the median are less than the mode and the mean-median difference in (a) is relatively minor than (c) Right Skewed or Positively Skewed: the mean and median are greater than the mode and the mean is greater than the median.

On the other side, **Kurtosis** compares the dataset distribution to the normal distribution. Data with high or positive kurtosis are those who tend to have heavy tails or outliers versus low or negative kurtosis (light tails) with lack of outliers (Figure2).

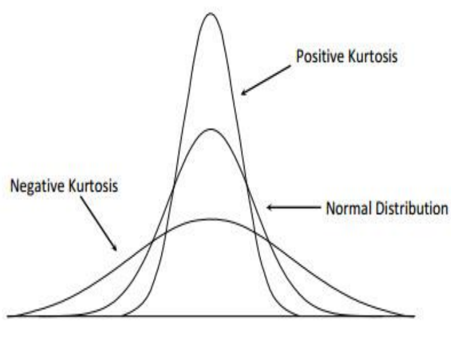


Fig.2: Presentation of the different types of kurtosis

The classical kurtosis was thought to evaluate the peakedness of the distribution which was later on considered as main misconception (Kevin et al, 1988). One of the interpretations of the kurtosis defines it as a measure of dispersion around the two values $\mu \pm \sigma$ (Moors et al, 1986). More particularly, the probability in the tails is measured by Kurtosis and then the value is compared to the one of the normal distribution which takes the value of three. However, sometimes, kurtosis is reported as "excess kurtosis" by subtracting three and therefore making it comparable to 0 (SPC, 2018). Table 1 below reports the statistical formulas to get skewness and kurtosis and an overview of their interpretation.

**Table 1. Skewness and kurtosis formulas and interpretation**

| Skewness | Kurtosis |
|---|---|
| $a_3 = \sum \frac{(X_i - \bar{X})^3}{ns^3}$ | $a_4 = \sum \frac{(X_i - \bar{X})^4}{ns^4}$ |
| $Skewness = \frac{n}{(n-1)(n-2)} \sum \frac{(X_i - \bar{X})^3}{s^3} = \frac{n}{s^3(n-1)(n-2)} (S_{above} - S_{below})$ | $Kurtosis = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \frac{(X_i - \bar{X})^4}{s^4} \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$ |
| -If the skewness is between **-0.5 and 0.5**, the data are fairly symmetrical<br>-If the skewness is between **-1 and – 0.5** or between 0.5 and 1, the data are moderately skewed<br>-If the skewness is less than **-1 or greater than 1**, the data are highly skewed | -If the kurtosis is close to **0**, then a normal distribution is often assumed (mesokurtic).<br>-If the kurtosis is **less** than zero, then the distribution has light tails (platykurtic).<br>-If the kurtosis is **greater** than zero, then the distribution has heavier tails (leptokurtic). |

Additionally, most of the classical tests and intervals assume that the data are of a normal distribution. However, normality assumptions cannot be made when obtaining significant skewness and kurtosis. In this particular case, many approaches exists to try to reach the normality of data (Thrun et al, 2020). These approaches include applying some transformations to make the data normal like the Box-Cox transformation or more particularly in the case of a moderate right skewness taking the square root or the log of the data set (Osborne et al, 2010). Moreover, in most cases, quantitative traits are not normally distributed. In this case, it is suggested to perform a parametric transformation to approximate the normality. For instance, in meta-analysis, transformation is an important strategy by combining two or more populations to enhance the statistical power (Goh et al, 2009).

To understand more the importance of the applications of such transformations, R software (4.0.3) was used to simulate 10000 samples in three different scenarios from left (negative) skewed data to right (positive) skewed data, light and heavy tailed data.

## 4.1 Scenario I

The data was simulated from a beta distribution with parameters α=10 and β= 2. A skewness of -1.0788 and a kurtosis of 3.6821 were reported. These values describe a highly skewed and light tailed distribution (Figure 3 right histogram). Therefore, a transformation y=$x^4$ was performed and the acquired results were -0.11177 and 2.2844 for skewness and kurtosis respectively. Due to this transformation, the data follows a normal distribution (Figure 3 left histogram).

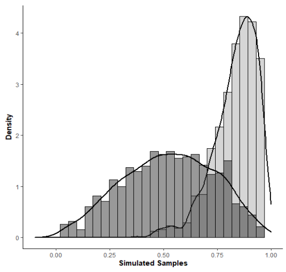


Fig.3: Transformation of a highly skewed light tailed distribution (right) to a normal distribution (left)

## 4.2 Scenario II

The data was simulated from a beta distribution with parameters α=1 and β= 5. In this scenario, a skewness of 1.2194 and a kurtosis of 4.3740 were obtained. These findings illustrate a highly skewed and light tailed distribution (Figure 4 left histogram). Therefore, a transformation y=$\sqrt[3]{x}$ was computed resulting in a skewness of -0.13251 and a kurtosis of 2.5114. Consequently, the data follows a normal distribution (Figure 4 right histogram).

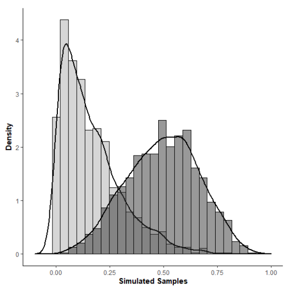


Fig.4:.Transformation of a highly skewed light tailed distribution (left) to a normal distribution (right)

### 4.3 Scenario III

The data was simulated from a lognormal distribution with parameters μ=1 and σ= 5. The results obtained were 1.6674 for the skewness and 7.7328 for the kurtosis. In this scenario also an initial highly skewed and light tailed distribution was observed (Figure 5 right histogram). Therefore, a transformation y=$log(x)$ was performed deriving a skewness of 0.027948 and a kurtosis of 2.9485. As a result, the data follow a normal distribution (Figure 5 left histogram).

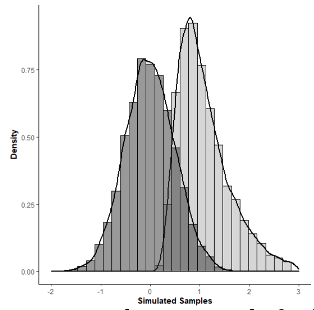

Fig.5: Transformation of a highly skewed light tailed distribution (right) to a normal distribution (left)

## 5. IMPORTANCE OF NORMALITY IN HEALTH EPIDEMIOLOGY

The main strength to the continuum view of the observations is that it explains uncertainties in clinical practice, where sometimes abnormality shades with normality especially in the case of COVID-19 epidemiological research (Lee et al., 2021). Based on the values of skewness and kurtosis, the statistical analysis and tests used will change where Pearson correlation and stepwise regression analyses were conducted based on evidence of data normality (Uğraş et al., 2021; Hatem et al., 2021).

In public health researches, the data sets involve different types of variables (Demirtas, 2012). For example in the mental and psychiatric diseases, the deluded intersects with the eccentric, and schizophrenia combines with affective disorder and personality subtypes. This is consistent with the expression “the medical model”; where the symptoms and signs of an organ failure (kidney, liver, and heart) can only be understood from a model of normal organ functioning (David, 2010). This same model can be applied in any epidemiological field where to get a statistical reliable results, normality should be considered and results should be compared to what is expected to happen if there were no exposure or outcome.

Additionally, normality assumption can be achieved in public health data given the impact of the big sample size which discards the effect of extreme and moderate skewness and kurtosis (Lumley, 2002). Accordingly, they are efficient methods to compare distinctive epidemiological studies throughout different periods and groups more particularly for weight and height measurements (Werner, 2006). However, a recent study elucidated that low skewness is not efficient when determining whether or not the t-test should be costumed, but low kurtosis is a provocative reason to avoid using this test (Garren, 2021).

## 6. CONCLUSION

The normal distribution fits into many natural phenomena (heights, blood pressure, measurement error, and IQ scores) (Brandt, 2017). It is symmetrical around its mean and has two main parameters (the mean and the standard deviation). Normality tests are crucial for deciding the measures of central tendency and assessing the best statistical methods for data analysis. In many studies related to the health evaluation, one of the primary steps is to test the normal distribution of the data to ensure an efficient analysis.

## REFERENCES


- Ahsanullah M., Kibria B.M.G., Shakil M. (2014) Normal Distribution. In: Normal and Student´s t Distributions and Their Applications. Atlantis Studies in Probability and Statistics, vol 4. Atlantis Press, Paris.
- Feller, W (1971). An Introduction to Probability Theory and Its Applications, Vol. 2, 3rd ed. New York: Wiley, p. 45, 1971.
- Metcalf, R. B. (2021). Practical Statistics for Physics & Astronomy.
- Weisstein, Eric W. (2021) "Normal Distribution." From MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/NormalDistribution.html
- Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. Annals of cardiac anaesthesia, 22(1), 67–72.

- Altman DG, Bland JM. Statistics notes: The normal distribution. BMJ. (1995); 310:298.
- Hugo Hernandez (2021). Testing for Normality: What is the Best Method? ForsChem Research, Vol. 6, 2021-05
- Garren, S. T., & Osborne, K. M. (2021). Robustness of T-test Based on Skewness and Kurtosis. Journal of Advances in Mathematics and Computer Science, 36(2), 102-110.
- Razali N.M. & Wah, Y.B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, Journal of Statistical Modeling and Analytics, 2(1), 21-33.
- Demir, E., Saatcioğlu, Ö. & İmrol, F. (2016). Uluslararası dergilerde yayımlanan eğitim araştırmalarının normallik varsayımları açısından incelenmesi, Current Research in Education, 2(3), 130-148.
- Orcan, F. (2020). Parametric or Non-parametric: Skewness to Test Normality for Mean Comparison. International Journal of Assessment Tools in Education, 7 (2) , 255-265 .
- Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. American Journal of Theoretical and Applied Statistics, 5(1), 5-12
- Bakshaev, A. (2009). Goodness of fit and homogeneity tests on the basis of N-distances. Journal of statistical planning and inference, 139(11), 3750-3758
- Şirin, Y.E., Aydın, Ö. & Bilir, F.P. (2018). Transformational-transactional leadership and organizational cynicism perception: physical education and sport teachers sample. Universal Journal of Educational Research, 6(9), 2008-2018
- Hair, J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. (2017). A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). 2nd Ed. Thousand Oaks, CA: Sage.
- Kevin P. Balanda & H. L. Macgillivray (1988) Kurtosis: A Critical Review, the American Statistician, 42:2, 111-119.
- J. J. A. Moors (1986). The Meaning of Kurtosis: Darlington Reexamined, The American Statistician, 40:4, 283-284
- Are the Skewness and Kurtosis useful statistics? SPC for excel (2008), Revised in 2016.
- Thrun MC, Gehlert T, Ultsch A (2020) Analyzing the fine structure of distributions. PLoS ONE 15(10): e0238835.
- Osborne, Jason (2010) "Improving your data transformations: Applying the Box-Cox transformation," Practical Assessment, Research, and Evaluation: Vol. 15, Article 12.
- Goh, L., Yap, V.B. (2009). Effects of normalization on quantitative traits in association test. BMC Bioinformatics 10, 415.
- Lee, S. M., So, W. Y., & Youn, H. S. (2021). Importance-Performance Analysis of Health Perception among Korean Adolescents during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, *18*(3), 1280.
- Uğraş, S., Sağın, A. E., Karabulut, Ö., & Özen, G. (2021). The relationship between college athletes' health literacy and COVID-19 anxiety levels. *Physical education of students*, *25*(1), 36-42.
- Hatem, G., Ghanem, D., Kellen, E., AlZaim, I., & Goossens, M. (2021). Knowledge and Beliefs of Cancer Risk Factors and Early Cancer Symptoms in Lebanon: A Cross-sectional Survey among Adults in the Community. *Cancer Control*, *28*, 10732748211053149.
- Demirtas, H., Hedeker, D., & Mermelstein, R. J. (2012). Simulation of massive public health data by power polynomials. *Statistics in Medicine*, *31*(27), 3337-3346.
- David, A. S. (2010). Why we need more debate on whether psychotic symptoms lie on a continuum with normality. *Psychological medicine*, *40*(12), 1935-1942.
- Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health*, *23*(1), 151-169.
- Werner, B. O., Bodin, L., & Bremberg, S. (2006). Data on height and weight from school health records as a national public health surveillance tool: the case of Sweden. *Scandinavian journal of public health*, *34*(4), 406-413.
- Garren, S. T., & Osborne, K. M. (2021). Robustness of T-test Based on Skewness and Kurtosis. *Journal of Advances in Mathematics and Computer Science*, 102-110.
- Brandt, L., & Addison, W. E. (2017). A low-anxiety introduction to the standard normal distribution and measures of relative standing.