



## Original Article

**Short title: Nitin Chandurkar et al.: Data Distribution and Transformation in Medical Research**

### Beyond the Bell Curve: Understanding Data Distribution and Transformation in Medical Research

Nitin Chandurkar<sup>1</sup>, Kumar Bhaskar Naidu<sup>1</sup>

<sup>1</sup> Clinical Research & Development, Ipca Laboratories Ltd.

#### Abstract

The normal distribution is central to many statistical methods used in medical research, yet real-world clinical data frequently deviate from this idealised bell-shaped pattern. Such deviations may arise from biological variability, disease heterogeneity, natural limits of measurement, extreme observations, or multiplicative processes underlying certain clinical and pharmacokinetic variables. Importantly, statistical normality should not be confused with clinical normality, and non-normal data should not automatically be interpreted as abnormal or poor-quality data. Appropriate assessment of distribution therefore requires more than reliance on a formal normality test and should incorporate graphical evaluation, descriptive measures, sample size, and the assumptions of the intended statistical model. When substantial skewness is present, data transformation may be useful to improve symmetry, stabilise variance, or better represent the underlying mechanism of variability. Logarithmic transformation is particularly relevant for variables influenced by proportional or multiplicative effects, as illustrated by pharmacokinetic parameters such as maximum plasma concentration and area under the concentration–time curve. However, transformation should not be performed merely to force data into a normal distribution; alternative summaries, non-parametric methods, or distribution-specific statistical models may be more appropriate in some situations. This review provides a practical and intuitive overview of normal and non-normal data in medical research, with emphasis on understanding the origin of observed distributions and selecting statistical methods that provide valid and clinically meaningful interpretation.

#### INTRODUCTION

The normal distribution is one of the most familiar concepts in statistics and underlies many commonly used statistical methods in medical research. Characterised by a symmetric, bell-shaped pattern, it represents a situation in which most observations cluster around a central value, while progressively fewer observations occur towards the extremes.<sup>1</sup> In relatively homogeneous populations, several physiological measurements may therefore show an approximately normal distribution.

However, the term *normal* in statistics should not be confused with what is considered normal or healthy in clinical medicine. Statistical normality describes the shape of a distribution, not the clinical desirability of the observed values. For example, systolic blood pressure in a group of patients with hypertension may follow an approximately normal distribution even though the values themselves are clinically elevated. Conversely, a clinically healthy population may show a skewed distribution for certain biological variables. Thus, a statistically normal distribution does not necessarily imply clinical normality, and a non-normal distribution does not necessarily indicate abnormal or poor-quality data.

Medical data frequently depart from the idealised bell-shaped pattern. Clinical studies generally involve patients who may differ in disease severity, comorbidities, treatment exposure, physiological

characteristics, and underlying biological responses. Consequently, medical variables may show varying degrees of skewness or other departures from normality.<sup>2,3</sup> Such departures may themselves reflect the underlying biological or clinical process rather than a flaw in the dataset.

Understanding the distribution of data is important because it influences the choice of descriptive statistics, statistical models, and methods of inference. Assessment of normality should therefore not rely solely on a statistical test but should also consider graphical assessment, sample size, the magnitude of departure from normality, and the assumptions of the statistical method being applied.<sup>2,3</sup> Importantly, for many commonly used parametric models, the relevant assumption concerns the distribution of model residuals rather than requiring the original observed variable itself to be normally distributed.

Data transformation is another commonly used approach when observations are markedly skewed or when variability is more appropriately represented on a different scale. However, transformation should not be viewed merely as a procedure for

**Address for Correspondence:** Kumar Bhaskar Naidu, Clinical Research & Development, Ipca Laboratories Ltd.

E-mail: [kumarnaidu2@gmail.com](mailto:kumarnaidu2@gmail.com)

**How to cite this article:** Chandurkar N, Naidu KB. Beyond the bell curve: understanding data distribution and transformation in medical research. Indian J Health Allied Sci. 2026;1(3):84-87.

**Received:**25/08/2026 **Accepted:**16/09/2026 **Published:**30/09/2026

#### Access this article online

Quick Response Code



Website:

<https://ijhas.varnavpublishers.com/>

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

forcing data to resemble a normal distribution. The choice of transformation should reflect the nature of the variable, the mechanism generating its variability, and the statistical model being used.

This review provides a practical and intuitive overview of normal and non-normal data in medical research, the reasons why medical data may depart from normality, approaches for assessing distributional assumptions, and the appropriate use of data transformations. Particular emphasis is placed on interpreting these concepts in a clinically meaningful manner rather than treating normality as an obligatory property of medical data.

### Understanding the Normal Distribution in Medical Research

The normal distribution provides a simple way of understanding how measurements are spread around a central value. In a perfectly normal distribution, observations are distributed symmetrically around the mean, producing the familiar bell-shaped curve. The mean, median and mode coincide at the centre, while the frequency of observations gradually decreases as the values move farther away from this central point.<sup>1</sup> Approximately 68 per cent of observations lie within one standard deviation of the mean, about 95 per cent within two standard deviations, and nearly all observations within three standard deviations.

An intuitive way to visualise this pattern is to imagine a continuous flow of sand falling at approximately the same point. Most of the sand accumulates near the centre, while progressively less is found towards either side, creating a mound that resembles a bell-shaped curve when viewed in cross-section. This analogy is useful for understanding the concentration of observations around a central value, although a physical sand pile is not mathematically a normal distribution. In biological data, a similar pattern may arise when numerous relatively small influences contribute to variation around a typical value.

Several physiological measurements can show an approximately normal distribution in a sufficiently homogeneous population. Human height is a commonly used example: most individuals tend to cluster around an average height, while progressively fewer individuals are extremely short or extremely tall. However, the shape of a distribution depends not only on the variable itself but also on the characteristics of the population being studied. Combining individuals with substantially different biological or clinical characteristics can broaden the distribution, introduce skewness or, in some situations, produce more than one peak.<sup>2,3</sup>

The normal distribution is therefore best viewed as a useful statistical model for describing a particular pattern of variability rather than as an expected property of all medical measurements. Its importance lies in providing a framework for summarising data and for many commonly used statistical methods. In clinical research, however, departures from this idealised pattern are frequent and may themselves provide useful information about the underlying biological or clinical process.

### Why Medical Data Deviate from Normality

Departures from normality are common in medical research and should not necessarily be viewed as statistical irregularities.

Clinical data arise from complex biological processes and are often collected from populations that are inherently heterogeneous. Differences in disease severity, age, comorbidities, treatment exposure, lifestyle, genetic factors and physiological responses can influence the distribution of a measured variable.<sup>2,3</sup>

**Skewness** is one of the most frequently encountered departures from normality. In a right-skewed distribution, most observations occur at the lower or intermediate end of the range, while a smaller number of relatively high values form a long right tail. Many laboratory and clinical variables may show this pattern. For example, triglyceride concentrations may be relatively low or moderate in most individuals but markedly elevated in a smaller subgroup. Such high values may represent genuine biological or clinical differences rather than measurement error. Similarly, inflammatory biomarkers, liver enzymes, duration of hospital stay and healthcare utilisation may show right-skewed distributions.<sup>2,3</sup>

The characteristics of the study population can also influence the shape of a distribution. A measurement obtained in a relatively homogeneous group may appear approximately symmetric, whereas combining individuals with substantially different clinical or biological characteristics can produce skewness, increased variability or even more than one peak. Thus, the shape of a distribution may itself provide useful information about heterogeneity within the population rather than merely indicating a statistical problem.

Another important reason for non-normality is the **mechanism by which variability is generated**. Some variables are influenced mainly by additive effects, whereas others are affected by proportional or multiplicative changes. When several factors act multiplicatively, the resulting observations may show a long right tail and can often be better represented by a lognormal rather than a normal distribution.<sup>4</sup> This concept is particularly relevant to pharmacokinetic variables such as maximum plasma concentration (C<sub>max</sub>) and area under the concentration–time curve (AUC), for which variability is commonly expressed proportionally and logarithmic transformation is routinely used in bioequivalence analysis.<sup>5,6</sup>

Natural boundaries may further contribute to non-normality. Many biological measurements cannot take negative values, while clinical scales may have predefined lower or upper limits. Floor and ceiling effects can therefore result in asymmetric distributions. However, the mere fact that a variable is restricted to positive values does not imply that it must be lognormally distributed. The underlying mechanism of variability remains more important than positivity alone.

Extreme observations also require careful interpretation. An unusually high or low value may arise from a measurement or data-entry error, but it may equally represent a genuine patient with an uncommon clinical characteristic. Such observations should therefore be investigated rather than automatically excluded simply because they affect the apparent normality of the data.

Recognising **why** a variable is non-normally distributed is often more informative than merely classifying it as normal or non-normal. The observed distribution should be interpreted in the context of the biology of the measurement, characteristics of the study population and the mechanism by which variability is generated.

### Assessing Normality: Beyond Statistical Tests

Assessment of data distribution is an important step in statistical analysis, but normality should not be decided on the basis of a single statistical test. A combination of **graphical examination, descriptive measures and formal tests** generally provides a more informative assessment.<sup>2,3</sup> Histograms can reveal

asymmetry, long tails, multiple peaks and unusual observations, while quantile–quantile (Q–Q) plots allow the observed data to be compared visually with the pattern expected under a normal distribution. **Figure 1** illustrates the contrasting appearance of approximately normal and right-skewed data using histograms and corresponding Q–Q plots.

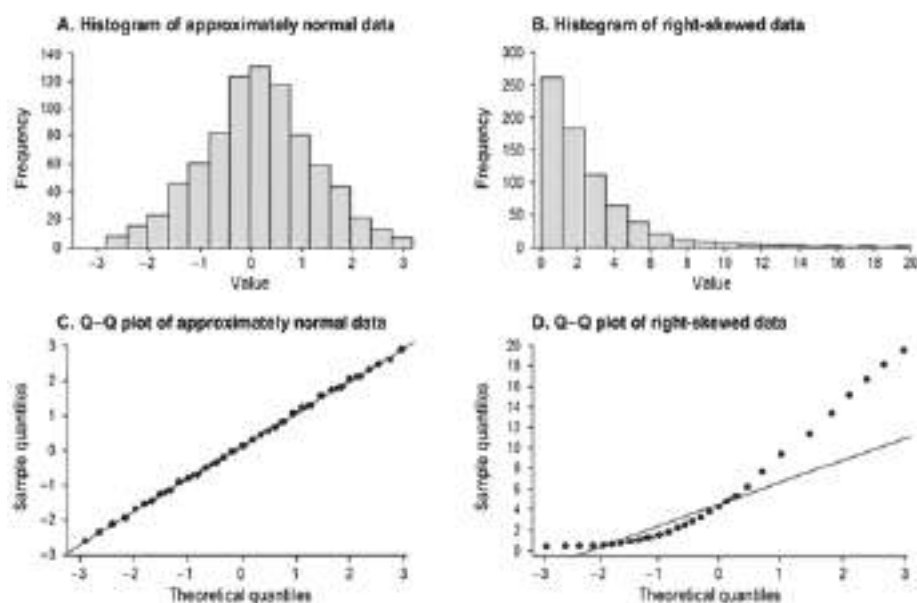


**Figure 1. Graphical assessment of data distribution**

Formal tests of normality, particularly the **Shapiro–Wilk test**, are widely used to assess whether the observed distribution shows evidence of departure from normality.<sup>7</sup> The resulting *P* value, however, should be interpreted cautiously. A non-significant test does not prove that the data are normally distributed, while a statistically significant result does not necessarily imply that the departure from normality is large enough to affect the planned analysis.<sup>2,3</sup>

Sample size has an important influence on this interpretation. With a small sample, a normality test may have limited ability to detect meaningful departures from normality. In contrast, with a large sample, even relatively minor deviations may produce a statistically significant result. Moreover, many commonly used parametric procedures can be reasonably robust to moderate departures from normality, particularly with adequate sample sizes.<sup>8</sup> Thus, automatically selecting a non-parametric test whenever a normality test gives  $P < 0.05$  may be unnecessarily restrictive.

It is also important to clarify **what is actually required to be normally distributed**. For procedures such as linear regression, analysis of variance and analysis of covariance, the normality assumption principally concerns the model errors or residuals rather than requiring the original outcome variable itself to follow a perfect normal distribution. Consequently, examining the distribution of residuals after fitting the statistical model may be more relevant than testing the raw observations alone.<sup>8</sup>

Normality assessment should therefore be treated as a process of statistical judgement rather than a binary decision of “normal” or “non-normal”. The magnitude and nature of the departure,

presence of influential observations, sample size, study design and robustness of the proposed statistical method should all be considered together. The objective is not to demonstrate that every dataset conforms perfectly to a bell-shaped curve, but to determine whether the distributional assumptions of the intended analysis are sufficiently reasonable for valid interpretation.

### Managing Non-Normal Data: Transformation and Appropriate Statistical Methods

When data show substantial departure from normality, the first response should not be to transform them automatically. The reason for the observed distribution should first be considered. Skewness may arise from genuine biological variability, extreme but valid observations, natural limits of the measurement, or the presence of clinically distinct subgroups. The statistical approach should therefore be chosen according to the characteristics of the variable and the objective of the analysis.

**Data transformation** changes the scale on which observations are analysed and may improve symmetry, stabilise variance and reduce the disproportionate influence of very large observations.<sup>5</sup> The logarithmic transformation is among the most commonly used approaches for positive, right-skewed biomedical data. Other transformations include the square-root transformation for moderately skewed or count-type data, reciprocal transformations for more pronounced right skewness, and the Box–Cox family of transformations, which provides a systematic approach to selecting an appropriate transformation. However, achieving a visually perfect normal distribution

should not itself be regarded as the objective of transformation.

The logarithmic transformation is particularly meaningful when variability is **multiplicative rather than additive**.<sup>4</sup> This can be illustrated by pharmacokinetic parameters such as maximum plasma concentration (C<sub>max</sub>) and area under the concentration–time curve (AUC). These parameters are affected by several biological and formulation-related processes whose effects are often proportional. Consequently, their distributions frequently show positive skewness and are commonly analysed on the logarithmic scale in bioequivalence studies.<sup>6</sup> Importantly, this behaviour is not simply because C<sub>max</sub> and AUC cannot take negative values. Other positive variables, such as body weight, may still show an approximately symmetric distribution. The more relevant consideration is the mechanism generating the variability.

Logarithmic transformation also changes the interpretation of the statistical comparison. A difference between two treatment means on the logarithmic scale becomes a **ratio after back-transformation**:

$$[\exp[\log(T) - \log(R)]] = \frac{T}{R}$$

This is why bioequivalence analyses of C<sub>max</sub> and AUC are naturally expressed as geometric mean Test-to-Reference ratios with confidence intervals rather than as simple arithmetic mean differences.<sup>6</sup>

Transformation, however, is not always necessary or desirable. When meaningful skewness persists, or when the transformed scale is difficult to interpret, investigators may consider reporting the **median and interquartile range**, using appropriate non-parametric procedures, or applying statistical models that directly accommodate the observed distribution. The choice should depend on the research question, study design and assumptions of the proposed analysis rather than on a rigid requirement that all medical data must be made normally distributed.

The broader principle is therefore that **the statistical method should adapt to the nature of the data, rather than forcing the data to conform to a preferred statistical method**.

### Practical Considerations and Conclusion

Assessment of data distribution should be incorporated into routine statistical planning rather than treated as a mechanical preliminary step. Before selecting a statistical method, investigators should consider the type of variable, its biological characteristics, the shape of its distribution, the presence of extreme observations, and the assumptions of the proposed statistical model. Graphical tools such as histograms and Q–Q plots, together with descriptive measures and formal tests, can help determine whether departures from normality are likely to be important for the intended analysis.<sup>2,3,8</sup>

A practical distinction should also be made between **statistical significance of non-normality and practical relevance of that departure**. In large datasets, minor deviations may be detected by formal normality tests without materially affecting the validity of a robust parametric analysis. Conversely, in small samples, important departures may remain undetected. Statistical judgement should therefore accompany formal testing.<sup>8</sup>

When substantial skewness is present, investigators should first ask whether it reflects the underlying biological process.

Transformation may be appropriate when it improves the representation of variability or better matches the statistical model, particularly when effects are proportional or multiplicative.<sup>4,5</sup> However, transformation is not obligatory. Depending on the research question, alternative summaries such as the median and interquartile range, non-parametric methods, or statistical models designed for non-normal outcomes may provide a more appropriate analysis.

The pharmacokinetic example illustrates this principle particularly well. C<sub>max</sub> and AUC are analysed on the logarithmic scale not merely because they are positive variables, but because their variability is commonly proportional and multiplicative.<sup>6</sup> The transformed scale therefore corresponds more naturally to the biological and statistical behaviour of these parameters and leads directly to clinically interpretable treatment ratios after back-transformation.

Ultimately, normality should be viewed as a useful statistical model rather than a requirement that medical data must fulfil. **Non-normal data are not necessarily abnormal data**, and departures from the bell-shaped distribution may themselves provide information about disease processes, patient heterogeneity or biological variability. The objective of statistical analysis should therefore not be to force every dataset towards normality, but to understand how the data were generated and to select a method that provides valid, transparent and clinically meaningful inference.

**FINANCIAL SUPPORT AND SPONSORSHIP:** Nil

**CONFLICT OF INTEREST:** None

### REFERENCES

- Altman DG, Bland JM. Statistics notes: The normal distribution. *BMJ*. 1995;310:298. doi:10.1136/bmj.310.6975.298.
- Ghasemi A, Zahediasl S. Normality tests for statistical analysis: A guide for non-statisticians. *Int J Endocrinol Metab*. 2012;10:486–489. doi:10.5812/ijem.3505.
- Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth*. 2019;22:67–72. doi:10.4103/aca.ACA\_157\_18.
- Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: Keys and clues. *BioScience*. 2001;51:341–352. doi:10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2.
- Bland JM, Altman DG. Transformations, means, and confidence intervals. *BMJ*. 1996;312:1079. doi:10.1136/bmj.312.7038.1079.
- U.S. Food and Drug Administration. *Statistical Approaches to Establishing Bioequivalence: Guidance for Industry*. Silver Spring, MD: U.S. Food and Drug Administration; May 2026.
- Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52:591–611. doi:10.1093/biomet/52.3-4.591.
- Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health*. 2002;23:151–169. doi:10.1146/annurev.publhealth.23.100901.140546.

