

## An Update on Data Distribution and Techniques of Data Transformation

Ahmad Najmi<sup>1\*</sup>, Avik Ray<sup>2</sup>

<sup>1</sup>Assistant Professor, Dept. of Pharmacology, AIIMS Bhopal

<sup>2</sup>Avik Ray, Masters student, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, USA.

### ABSTRACT

The distribution in biostatistics can be defined as distribution of frequencies of values of a given variable in a sample. Distribution can be broadly classified into normal and skewed distribution. Normal distribution is a symmetrical bell shaped curve.  $\pm 1$  standard deviation covers 65% of values around median value and  $\pm 2$  S.D. covers 95% of values around median value. Mean, median & mode are equal for normal distribution curve. Parametric test like t test and ANOVA are based on the assumption that the data follows normal distribution. In skewed or asymmetrical distribution, there is clustering of cases in either right side or left side of the curve. In right sided skewness, the tail of curve is on the right side. In left skewed distribution, the tail is on the left side. Non-parametric test can be used in case of skewed data. Parametric test are more robust as compare to non-parametric test. The alternative is to transform the numerical variable into another scale where the values do satisfy the assumptions needed for the desired parametric or "normal" statistical methods. These technique include logarithm transformation, generalized linear modelling, and bootstrapping.

**Keywords:** data distribution, data transformation

### Article History

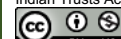
Received: 31.10.2022

Accepted: 21.11.2022

\*Corresponding Author

Dr. Ahmad Najmi,  
Assistant Professor, Dept. of  
Pharmacology, AIIMS Bhopal  
E mail: ahmad.pharm@aiimsbhopal.edu.in

**Copyright:** © the author(s). IABCR is an official publication of Ibn Sina Academy of Medieval Medicine & Sciences, registered in 2001 under Indian Trusts Act, 1882.



This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial

## INTRODUCTION


The distribution in biostatistics can be defined as distribution of frequencies of values of a given variable in a sample. Distribution can be broadly classified into normal and skewed distribution. Normal distribution or Gaussian distribution<sup>1,2</sup> is a type of distribution which comes under probability distribution. The person who first derived the formula, Abraham De Moivre (1667- 1754)<sup>3,4</sup>, was solving a gambling problem whose solution depended on finding the sum of the terms of a binomial distribution. Later work, especially by Gauss<sup>5</sup> about 1800, used the normal distribution to describe the pattern of random measurement error in observational data. The term "Normal" is derived from Latin word "Normalis", which means perpendicular. As Gauss draw perpendicular lines on both sides of median to describe distribution, so it is known as normal distribution. The term normal distribution has no

relationship with clinical normality. Even data which does not follow normal distribution may be absolutely clinically normal.

### Characteristics of normal distribution

Normal distribution is a symmetrical bell shaped curve<sup>6</sup>.  $\pm 1$  S.D. covers 65% of values around median value &  $\pm 2$  S.D. covers 95% of values around median value. Mean, median & mode are equal for normal distribution curve. In normal distribution or symmetrical distribution, there is clustering of most of the values in the centre. Most of the physiological parameters like heart rate, blood pressure, blood sugar, serum cholesterol, height, weight follow normal distribution. There are some exceptions like antibody titre. Although antibody titre is bodily parameter, but it does not follow normal distribution.

### Access this article online

| Website:   | Quick Response code   |
|--|---|
| <a href="http://www.iabcr.org">www.iabcr.org</a> |  |
| DOI: 10.21276/iabcr.2022.8.4.1                   |   |

**How to cite this article:** Najmi A., Ray A. An update on data distribution and techniques of data transformation. Int Arch BioMed Clin Res. 2022;8(4):PH1-PH3.

**Source of Support:** Nil, **Conflict of Interest:** None

Distribution also depends on sample size. If we measure blood pressure of 5-10 individuals & draw a graph, we may not get normal distribution. But if we take blood pressure of 500 individuals & draw a graph, we will get normal distribution. Variables which are measured as ranks or scores do not follow normal distribution. Example is Apgar score, Glasgow coma score, visual analogue score for measurement of pain, severity of any disease in term of mild, moderate & severe etc. Variables measured in the form of counts do not follow normal distribution. For an example number of people attended OPD, number of people experiencing headache, number of people encountered road traffic accidents etc. They follow binomial distribution. If we know about value of mean & S.D. then we may draw assumption about distribution. If the mean value is less than twice of standard deviation, then it does not follow normal distribution.

### Implications

Many statistical tests like t test & ANOVA & are based on the assumption<sup>7</sup> that the data follows normal distribution. We can predict about population parameters based on our sample parameter, if the data follows normal distribution. There is an air of suspicion around the word 'normal'. It depicts some sort of blandness and strictness about the data. However, multiple statistical tools like the Student t-tests, Analysis of Variance, linear regression, all assume that the underlying is distributed normally. If we try to explain in a non-statistical manner, the Gaussian distribution is stated "normal" from the perspective that it is typical. As per the central limit theorem, on combining multiple independent random variables, the cumulative distribution is thought of as Gaussian. There are multiple tests to verify, but of course failure to show a significant deviation may only mean that the sample was too small for an adequately powered test. There is also a graphical assessment known as the Q-Q (for quartile-quartile) plot which can be used to visually compare the sample distribution against a Gaussian ideal. And of course, a simple histogram of values may be enough to assess normality of the distribution and whether the deviation is due to outliers or erroneous data. We are considering continuous outcomes here, such as cost, blood loss, or days in hospital. Event counts such as complication or mortality rate do not follow continuous distributions. However, they may be assumed to follow  $\chi^2$  distributions, and the  $\chi^2$  distribution is actually derived from the Gaussian: it is the sum of squared Gaussian terms.

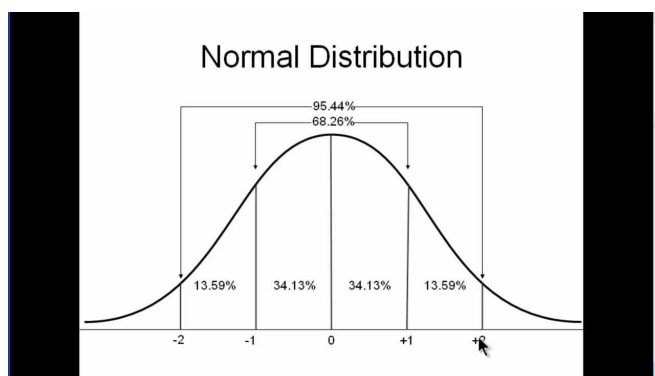


Figure 1: showing normal distribution

### Skewed Distribution

A common deviation from the symmetric Gaussian is skewed data<sup>8,9</sup>, which can happen often with "count data" such as days in hospital or accumulated costs during a hospital stay. In skewed or asymmetrical distribution, there is clustering of cases in either right side or left side of the curve. In right sided skewness, the tail of curve is on the right side. In left skewed distribution, the tail is on the left side. If most of the students are obese in a classroom, then it will give left skewed curve. If most of the students are underweight, then it will provide right skewed data. In these particular cases, we expect that relatively fewer patients will be seeing large stays or large bills, and smaller counts will be more common. For skewed data, mean and SD may not be informative. Skewed data are better described by the median and either the interquartile range (middle 50% of points) or the full range of the data.

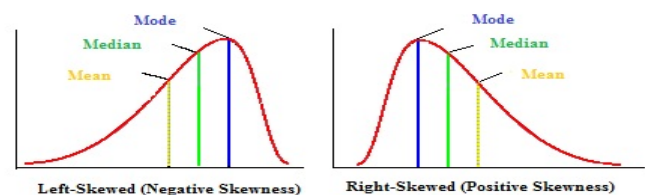


Figure 2: Skewed distribution

If examination of the data indicates that the outcome variable is highly skewed or otherwise non-Gaussian, there are a number of approaches<sup>10</sup>. Each one has advantages and disadvantages. The overall advantage, of course, is that each one can be applied separately on the same data set. The most straightforward strategy is to use different statistical strategies that do not require normal data. Statistical tests such as the Mann-Whitney, Kruskal-Wallis, or Friedman test are done on ranked data. A new variable is created in which the smallest number is assigned 1, the next smallest is 2, and so on, with ties given the average of ranks if they had not been tied. There is also a comparable correlation coefficient, called Spearman rho (or  $\rho$ ), which is like a correlation coefficient but between 2 rank-transformed variables, and is useful for assessing relationships which may follow a curved line. These tests are called nonparametric because they make no assumptions or predictions about the distribution of outcomes. A major drawback of these tests is that they do not give any information about how much the outcome actually varies (in physical units) with the predictor making the results of these tests less practically informative.

To undertake statistical analysis on a nonnormally distributed sample, there are two options. The simplest is to use statistical tests that do not depend on any assumptions about the distribution of the data (e.g., an assumption of normality). These are termed "non-parametric" tests and include Kaplan-Meier estimation, Kruskal-Wallis one-way analysis of variance, Mann-Whitney U test, Mc Nemar's test, Spearman's rank correlation coefficient, and Wilcoxon signed-rank test.

### Data transformation techniques

The alternative is to transform the numerical variable into another scale where the values do satisfy the assumptions needed for the desired parametric or "normal" statistical methods. The logarithmic (log) transformation<sup>11</sup> is, perhaps, the most popular transformation, probably because it is particularly useful for positively skewed data (skewed to the right). In a log transformation, the values of the variable are

replaced by their logarithm. Other transformations include square ( $X^2$ ), cube ( $X^3$ ), reciprocal ( $1/X$ ), and square root ( $X^{1/2}$ ) or cubed root ( $X^{1/3}$ ), although there are many other alternatives. Data with nonnormal distributions other than positively skewed may be normalized by these other transformations (e.g., square or cube transformation for negatively skewed data). Why are some medical variables not normally distributed? If we go into the population at random and test haemoglobin, we would expect to get a normally distributed set of results. The reference range we use to decide if a patient is anaemic is based exactly on this type of process, and represents the 95% confidence intervals of normal distribution. However, when we sample from a population of patients with a disease, it is not all that surprising to find that they have a different distribution of results, as these patients are more likely to have lower haemoglobin values. The distribution will also be influenced by the way we treat patients that have abnormal values. In the case of haemoglobin, we are unlikely to find patients with very low values as they are likely to have exhibited symptoms and been treated. These factors combine to give a distribution that has patients bunched around a lower median value than the healthy population, with a larger tail to the right and a short left tail; i.e. positively skewed dataset.

The type of data distribution determines which central tendency characteristic best describes the middle of the data set. If the data are normally distributed, the mean and median will be the same or similar. Mean is often the preferred central tendency measure, but if the data are significantly skewed the mean is less representative and median is more appropriate and useful.

Non-normal data sets can also be modified to facilitate analysis. Techniques include data transformation, generalized linear modelling, and bootstrapping.<sup>12</sup> It is possible to mathematically transform data to approximate a normal distribution. There are various rules of thumb for picking specific transformations, such as the logarithm, the square root, or the inverse to modify the data. Many statistical packages will have options for Box-Cox power transformations, which help determine the best transformation to normalize data or whether it is feasible to work with the raw data, without transforming. The caveat here is that the only mathematical manipulation that will return intuitively useful descriptions of effect is the logarithm, which maps multiplication onto addition; apply the inverse transform; and you can describe the effects on outcome as percentage increase or decrease.

A relatively new tool is the Generalized Linear Model or GLM. It allows the user to select a "link function" to transform the data and a probability distribution to describe measurement error. Logistic regression may be a familiar example of a GLM: probability estimates from 0 to 1 are transformed by a logistic link function, and a binomial error distribution is assigned. A particular advantage of a GLM is that parameters are returned in the same units as the outcome, so the effect of predictors on the outcome of interest could be described directly as days, dollars, etc., in an intuitive manner. In contrast, the link function might be problematic if its use results in something nonsensical, such as a negative length of stay. Which probability distribution to pick may depend on the existing knowledge of the system being studied, or it may be a matter of trying several and seeing which one gives the best predictive performance in an empiric manner.

Finally, bootstrapping combines useful features of all of the above. New synthetic data sets are created by randomly resampling the existing data with replacement: each of these synthetic sets will contain about two-thirds of the original data, with some records appearing multiple times in the place of missing data. Confidence intervals for mean, mean difference, etc. can be generated by bootstrapping hundreds or thousands of times and taking the values at the 2.5% and 97.5% percentiles to define the 95% confidence interval. One drawback is that the returned confidence interval will be slightly different for each run because of the randomization. Another consideration when using this technique, which may be an advantage or disadvantage depending on how you feel about P-values, is that bootstrapping methods do not return P-values at all.

## CONCLUSION

Distribution can be broadly classified into normal and skewed distribution. Normal distribution is a symmetrical bell shaped curve.  $\pm 1$  standard deviation covers 65% of values around median value and  $\pm 2$  standard deviation covers 95% of values around median value. In skewed or asymmetrical distribution, there is clustering of cases in either right side or left side of the curve. Assessment of the frequency distribution and comparison of the median and mean is the simplest way to assess normality of data. A variety of other statistical tests such as the Kolmogorov-Smirnov test, the Shapiro-Wilk test, the Shapiro-Francia test, or a test of skewness and kurtosis are available for assessment of normality of data. Parametric test can be used, if data is distributed normally. Non-parametric test can be used in case of skewed data. Parametric test are more robust as compare to non-parametric test. The alternative is to transform the numerical variable into another scale where the values do satisfy the assumptions needed for the desired parametric or "normal" statistical methods. These techniques include logarithm transformation, generalized linear modelling, and bootstrapping.

## REFERENCES

1. Krithikadatta J. Normal distribution. J Conserv Dent. 2014 Jan;17(1):96-7.
2. Limpert E, Stahel WA. Problems with using the normal distribution--and ways to improve quality and efficiency of data analysis. PLoS One. 2011;6(7):e21403
3. Peters, W.S. (1987). Normal Distribution. In: Counting for Something. Springer Texts in Statistics. Springer, New York, NY. [https://doi.org/10.1007/978-1-4612-4638-1\\_8](https://doi.org/10.1007/978-1-4612-4638-1_8)
4. DEMING, W. De Moivre's "Miscellanea Analytica", and the Origin of the Normal Curve. *Nature* **132**, 713 (1933). <https://doi.org/10.1038/132713a0>
5. Bennett MR. The origin of Gaussian distributions of synaptic potentials. *Prog Neurobiol.* 1995 Jul;46(4):331-50.
6. Sartori, R. The Bell Curve in Psychological Research and Practice: Myth or Reality?. *Qual Quant* **40**, 407-418 (2006)
7. Bennett MR. The origin of Gaussian distributions of synaptic potentials. *Prog Neurobiol.* 1995 Jul;46(4):331-50.
8. Delucchi KL, Bostrom A. Methods for analysis of skewed data distributions in psychiatric clinical studies: working with many zero values. *Am J Psychiatry.* 2004 Jul;161(7):1159-68.
9. Higgins JP, White IR, Anzueto-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. *Stat Med.* 2008 Dec 20;27(29):6072-92
10. Manikandan S. Data transformation. *J Pharmacol Pharmacother.* 2010 Jul;1(2):126-7
11. Feng C, Wang H, Lu N, Chen T, He H, Lu Y, Tu XM. Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry.* 2014 Apr;26(2):105-9
12. Henderson AR. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clin Chim Acta.* 2005 Sep;359(1-2):1-26