

Review Article
Editing, Writing & Publishing



Received: Nov 26, 2023
Accepted: Dec 21, 2023
Published online: Jan 15, 2024

Address for Correspondence:

Farrokh Habibzadeh, MD

Research and Development Unit, Petroleum
Industry Health Organization (PIHO) Polyclinic,
Eram Blvd., Shiraz 7143837877, Iran.

Email: Farrokh.Habibzadeh@gmail.com

© 2024 The Korean Academy of Medical
Sciences.

This is an Open Access article distributed
under the terms of the Creative Commons
Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>)
which permits unrestricted non-commercial
use, distribution, and reproduction in any
medium, provided the original work is properly
cited.

ORCID iD

Farrokh Habibzadeh
<https://orcid.org/0000-0001-5360-2900>

Disclosure

The author has no potential conflicts of
interest to disclose.

Data Distribution: Normal or Abnormal?

Farrokh Habibzadeh

Past President, *World Association of Medical Editors (WAME)*

Editorial Consultant, *The Lancet*

Associate Editor, *Frontiers in Epidemiology*

ABSTRACT

Determining if the frequency distribution of a given data set follows a normal distribution or not is among the first steps of data analysis. Visual examination of the data, commonly by Q-Q plot, although is acceptable by many scientists, is considered subjective and not acceptable by other researchers. One-sample Kolmogorov-Smirnov test with Lilliefors correction (for a sample size ≥ 50) and Shapiro-Wilk test (for a sample size < 50) are common statistical tests for checking the normality of a data set quantitatively. As parametric tests, which assume that the data distribution is normal (Gaussian, bell-shaped), are more robust compared to their non-parametric counterparts, we commonly use transformations (e.g., log-transformation, Box-Cox transformation, etc.) to make the frequency distribution of non-normally distributed data close to a normal distribution. Herein, I wish to reflect on presenting how to practically work with these statistical methods through examining of real data sets.

Keywords: Biostatistics; Statistical Distributions; Data Analysis; Normal Distribution; Epidemiologic Methods

INTRODUCTION

Appropriate data analysis is one of the cornerstones of every research study. An important part of data analysis is taking into account their distribution. Distribution of a data set affects both its report and analysis. For instance, normally distributed data should be reported as mean and the standard deviation (SD) and analyzed with parametric tests (e.g., Student's *t* test for independent samples, Pearson's correlation, linear regression, and one-way analysis of variance). Data that do not follow a normal distribution should be presented as median and the interquartile range (IQR) and analyzed with non-parametric tests (e.g., Mann-Whitney *U* test, Spearman's correlation, and Kruskal-Wallis test).¹⁻⁵ As parametric tests are generally more robust than their counterpart non-parametric tests, we prefer to use parametric tests for data analysis, if possible. Nonetheless, the parametric tests assume that the data to be analyzed follow a normal distribution. Violation of this assumption may result in unreliable results and biased estimates with trivial to critical consequences.⁶⁻⁹ Lack of awareness of these assumptions (e.g., normality of the data distribution) contributes to inappropriate use of statistical methods and reporting of results in scientific articles.¹⁰⁻¹² In fact, all researchers are expected to plan and report the results of their assessments of

the underlying assumptions in their study protocols and manuscripts.¹⁰ There are many assumptions. In this paper, the focus is on one of the assumptions made by all parametric statistical tests — the normality assumption.

One of the first steps in data analysis, after data cleaning, is to check whether the distribution of the data to be examined follows a normal (Gaussian, bell-shaped) distribution or not, and to try to transform the non-normally distributed data to a data set the distribution of which is closer to the normal. Herein, I wish to reflect on the common ways we can employ to determine whether a data set is normally distributed or not and describe two frequently used transformations to make the distribution of a non-normally distributed data closer to a normal distribution. The discussion is mainly based on data analysis of subsets of real data sets taken from my previous studies. Taking into consideration those journal editors and researchers with limited knowledge of statistics, throughout the article, I will try to emphasize the pragmatic issues of data analysis and avoid statistical details as much as possible.

NORMAL/GAUSSIAN DISTRIBUTION

A normal distribution, also called Gaussian distribution, has a symmetrical shape with the highest frequency at the center of the distribution. It has certain characteristics that help researchers to make predictions based on only the mean and the SD of the data. For example, about 95% of the data values are within the interval $\text{mean} \pm 2 \times \text{SD}$. As an example, **Fig. 1** shows the frequency distribution of hepatitis B surface antigen (HBs Ag) measured in 150 study participants (a subset of data from one of our previous studies).¹³ The data has a normal distribution, visually — the data distribution (gray curve) has an “acceptable” overlap over the hypothetical normal distribution having the same mean and SD of 0.38 and 0.09, respectively. There is another commonly used graphical method to determine whether a distribution follows a normal distribution or not. The graph is called Q-Q plot, which stands for “quantile-quantile plot.” The ordinate of the graph represents the quantile of the sample data; the abscissa, the quantile of a hypothetical data set should the data follow a normal distribution. As a rule of thumb, if the points are “close enough” to a straight line, it can be construed that the data distribution is normal (**Fig. 1B**); otherwise, it is not (**Fig. 2B**). These techniques although easy to do, are subjective.¹⁴ For example, the word “acceptable” is quite ambiguous and unscientific, and one might ask how close points should be to the line to be considered “close enough?” These terms become more meaningful with experience, but there are also quantitative measures to check the normality of a data distribution. One of the commonly used statistical tests for doing so is the one-sample Kolmogorov-Smirnov (K-S) test. It is a non-parametric test which tests the null hypothesis that the distribution of the data set follows a normal distribution. A significant *P* value ($P < 0.05$) implies that we should reject the null hypothesis and that the data distribution does not follow a normal distribution; with a non-significant *P* value ($P \geq 0.05$), the null hypothesis can be retained, and the data distribution can be assumed normal. The one-sample K-S test can technically only be used when the parameters of the distribution of interest (mean and the SD of the normal distribution) are known; otherwise, the results would be extremely conservative, and the test rejects the normality. We therefore need to correct the results when we examine a data sample. Lilliefors correction is what generally is used for this reason.¹⁵ Furthermore, the one-sample K-S test is usually recommended when the sample size is 50 or more. When the sample size is less than 50, another test for normality, the so-called Shapiro-Wilk (S-W) test, is better to be used.¹⁶

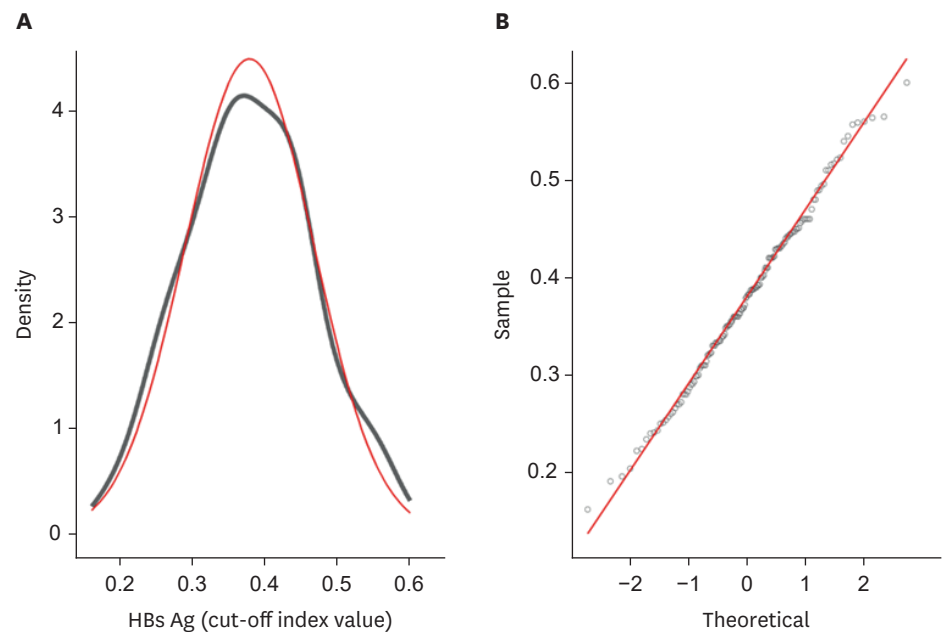


Fig. 1. Frequency distribution and Q-Q plot. **(A)** Frequency distribution of HBs Ag measured in 150 study participants taken from a previous study¹³ (bell-shaped gray curve) along with the fitted normal distribution (having the same mean and the standard deviation). **(B)** The Q-Q plot of the data implies that the distribution can be assumed to be normal.
HBs Ag = hepatitis B surface antigen.

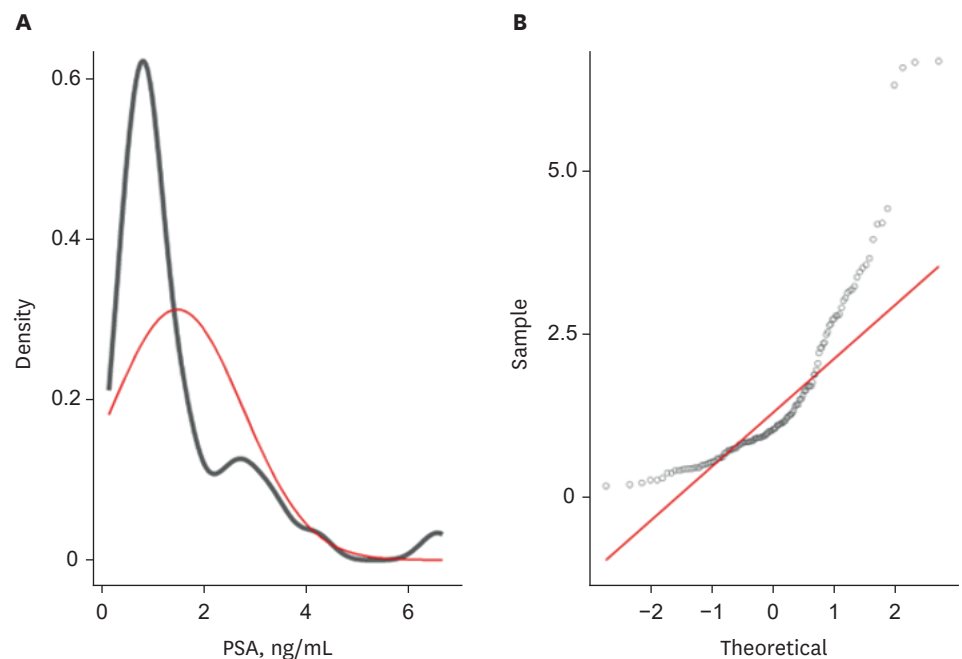


Fig. 2. Frequency distribution and Q-Q plot. **(A)** The frequency distribution of the PSA measured in 150 study participants taken from a previous study²⁰ (the highly positively skewed gray curve) along with the fitted normal distribution (having the same mean and the standard deviation). **(B)** The Q-Q plot of the data also implies that the data does not have a normal distribution.
PSA = prostate-specific antigen.

HOW TO CONSTRUCT THE GRAPHS AND PERFORM THE TESTS

There are numerous ways to examine the normality of the frequency distribution of a data set. Statistical Packages for Social Sciences (SPSS®) is a commonly used software for data analysis. In SPSS®, you can have both the Q-Q plot and the results of one-sample K-S test with Lilliefors correction and S-W test for a variable named HBS by running the following commands in the *Syntax Editor*:

```
EXAMINE VARIABLES=HBS
      /PLOT NPLOT
      /MISSING LISTWISE
      /NOTOTAL.
```

Or through the following path from the menu bar: Analyze → Descriptive Statistics → Explore (under the Plots, tick the box for “Normality plots with tests”). Alternatively, one-sample K-S test with Lilliefors correction can be done from the following path: Analyze → Nonparametric Tests → One Sample. The result of one-sample K-S test was not significant ($P = 0.200$; Fig. 3). Therefore, we could retain the assumption that the HBs Ag had a normal distribution, and predict that 95% of the 150 data (143 participants) had an HBs Ag level between 0.20 and 0.56; in fact, 141 (94%) did. We can also use parametric tests to analyze the data.

Another common method to perform the one-sample K-S test with Lilliefors correction is using the function *lillie.test* available from the R package *nortest*.¹⁷ The function *shapiro.test* from the same package can be used for performing S-W test. Q-Q plot can also be easily drawn using the *geom_qq* and *stat_qq_line* from the R package *ggplot2*.¹⁸

LOG-NORMAL DISTRIBUTION

Many researchers fallaciously believe that most biological variables have a normal distribution. However, non-normal skewed distributions are common in biomedicine. For example, the length of the latent period of many infectious diseases has a non-normal positively skewed distribution. This is because the period cannot be negative, the mean is usually short, and the variance [and the SD] is comparably large, usually more than half of the mean value.^{2,5} The frequency distribution of a variable is said to be log-normal if the distribution of logarithm of the variable values follows a normal distribution.¹⁹

	Tests of normality					
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HBS	0.041	150	0.200*	0.994	150	0.820

^aLilliefors significance correction

*This is a lower bound of the true significance.

Fig. 3. Output of the one-sample Kolmogorov-Smirnov test with Lilliefors correction and Shapiro-Wilk test from IBM® SPSS® Statistics ver. 26. Because the sample size was 150, the result of the first test is used. df = degrees of freedom.

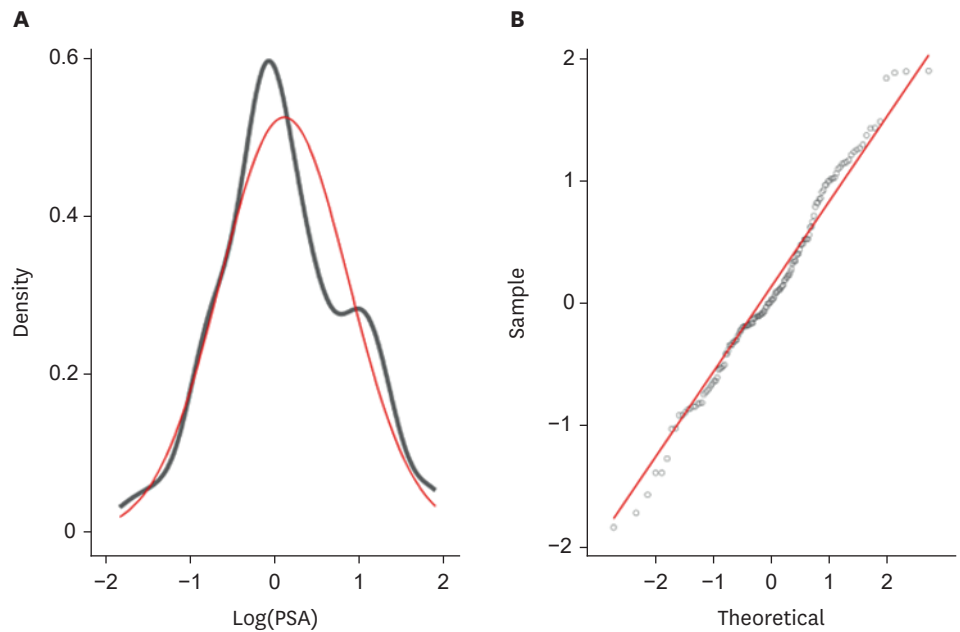


Fig. 4. The same graphs as those in Fig. 2 after log-transformation of the PSA, when log(PSA) is used instead of PSA. **(A)** The frequency distribution (gray curve) is now much closer to a normal distribution and **(B)** the point in the Q-Q plot lie close enough to the straight line to retain the assumption that the data distribution is normal. PSA = prostate-specific antigen.

As an example, Fig. 2 shows the frequency distribution of the prostate-specific antigen (PSA) measured in 150 study participants (a subset of data from one of our previous studies).²⁰ The data distribution (gray curve, Fig. 2A) is highly positively skewed and does clearly not fit the hypothetical normal distribution (red curve, Fig. 2A) having the same mean and SD of 1.50 and 1.28 ng/mL, respectively. The points in the Q-Q plot do not lie on the straight line too (Fig. 2B). One-sample K-S test with Lilliefors correction also resulted in a significant P value ($P < 0.001$), which is consistent with our visual methods. All these results imply that the frequency distribution of the PSA data set does not follow a normal distribution. That could be predictable from the mean and SD of the data set; the SD (1.28 ng/mL) exceeded half of the mean of 1.50 ng/mL.^{2,4,5} The logarithm of PSA, on the other hand, has a normal distribution (Fig. 4; one-sample K-S test with Lilliefors correction P value = 0.089).

Given that the logarithm of PSA does follow a normal distribution, we may construe that PSA follows a log-normal distribution. We may work with log(PSA) throughout the analysis and use parametric tests, but it is important to bear in mind that the final results should be reported in their original units (not the log-transformed values); we can easily back-transform the log-transformed values by exponentiating (inverse of logarithm) the results.

Log-transformation is a commonly used transformation to make the distribution of positively skewed distributions (e.g., the length of the incubation period of many infectious diseases, serum cholesterol level, and the distribution of minerals in the Earth's crust) closer to a normal distribution.¹⁹ However, it does not always work; other transformations may work better.

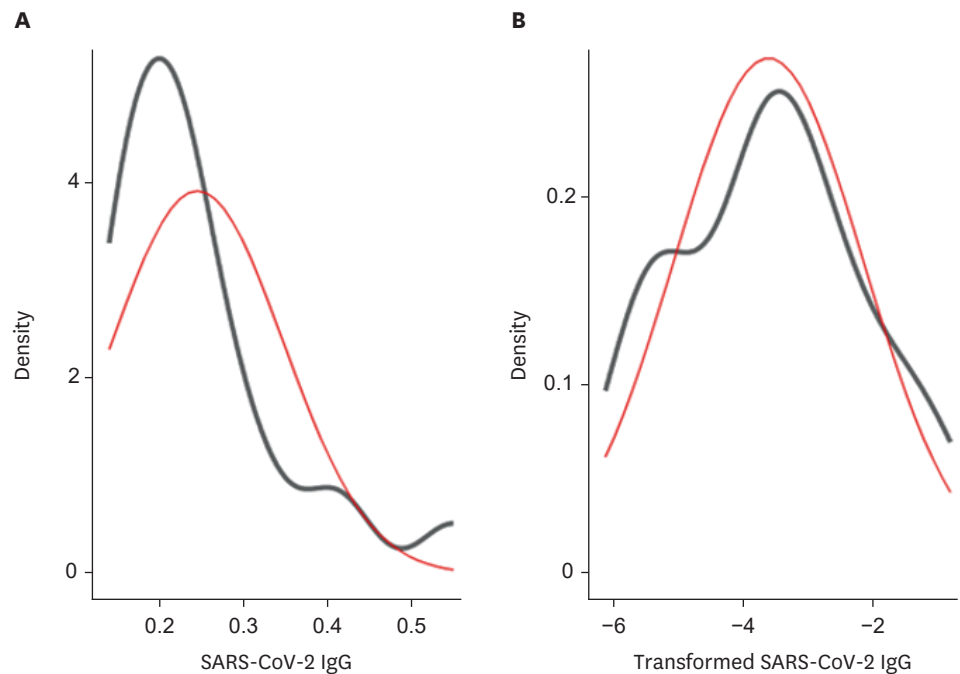


Fig. 5. Frequency distributions before and after transformation. **(A)** The frequency distribution of SARS-CoV-2 IgG level measured in 40 study participants taken from a previous study²¹ (gray curve). **(B)** Frequency distribution of the same data set after a Box-Cox transformation given a $\lambda = -1$ (Eq. 2). SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2, IgG = immunoglobulin G.

BOX-COX TRANSFORMATION

Fig. 5 shows the frequency distribution of the immunoglobulin (Ig) G against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) measured in 40 study participants (a subset of data from one of our previous studies).²¹ The SARS-CoV-2 IgG level had a mean of 0.25 (SD, 0.10). The frequency distribution of SARS-CoV-2 does not closely fit a normal distribution having the same mean and SD (**Fig. 5A**). Because the sample size was less than 50, S-W test was performed to test the normality of the distribution, which was found to be significant ($P < 0.001$), implying that the distribution was not normal. Log-transformation of the data could make the distribution closer to a normal distribution, but not enough to make the results of S-W test ($P = 0.012$) non-significant. The data were thus transformed using the Box-Cox transformation.

The Box-Cox transformation (transforming the value x to the new value y , given a parameter commonly designated by λ) is defined as follows²²:

$$y(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases} \quad (\text{Eq. 1})$$

The transformation acts differently depending on the value of λ . For certain values of λ , the transformation is equivalent to other well-known transformations (**Table 1**); for instance, if $\lambda = 0$, it turns to a log-transformation. But, which value of λ is better and makes the distribution of the transformed data closer to a normal distribution? To determine the most appropriate value for λ , we may use the function *boxcox* from the *R* package *EnvStats*.²³ The most appropriate value for λ for our data set was -1 . The transformation is then:

Table 1. Equivalent transformations for certain values of λ in a Box-Cox transformation (Eq. 1)

λ	Equivalent transformation
-2	$1/x^2$
-1	$1/x$
-0.5	$1/\sqrt{x}$
0	Log-transformation
0.5	\sqrt{x}
1	x
2	x^2

$$y = 1 - \frac{1}{x} \quad (\text{Eq. 2})$$

where x represents the SARS-CoV-2 IgG level and y the transformed value. Note that the transformation $y = 1/x$ works as well (**Table 1**), but for the time being, let us use Eq. 2. The frequency distribution of y (transformed SARS-CoV-2 IgG level) is close enough to a normal distribution (**Fig. 5B**) and the S-W test is non-significant ($P = 0.193$). Please note that here again the transformed variable y should be used in statistical analyses, but we should report the final values after they are back-transformed to the original scale using the following equation (solving Eq. 2 for x):

$$x = \frac{1}{1-y} \quad (\text{Eq. 3})$$

LIMITATIONS OF DATA TRANSFORMATION

Although the above-mentioned transformations may make the frequency distribution of our data set close to a normal distribution, which is favorable from the statistical point of view, the results obtained may not be meaningful and interpretable without back-transformation of the results.²⁴ Sometimes, finding an appropriate transformation is challenging; at times, despite all efforts made, no suitable transformation can be found. Under such circumstances, it would be better to use non-parametric statistical tests, although they are less robust than their parametric counterparts.

CONCLUSION

Examining the frequency distribution of a given data set is important in determining how to report and analyze the data. As parametric statistical tests, which assume normal distribution of the data, are more robust than their non-parametric counterparts, transforming a non-normally distributed data set to a set close to normal distribution would be very helpful. Graphical methods (e.g., Q-Q plot) are subjective and may not be reliable and replicable. The use of statistical tests (e.g., one-sample K-S and S-W tests) also have their own limitations. For example, for large samples, the test may show that the data distribution is not normal, even when the departure from normality is trivial and inconsequential. On the other hand, for small samples, serious departures from normality may not be detected by these tests.¹⁴ A combination of visual assessment and statistical tests may thus be necessary to come to a reasonable conclusion. Correct interpretation of results and choosing the appropriate transformation, if necessary, are skills that come with experience.

REFERENCES

1. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A, editors. *Science Editors' Handbook*. Exeter, UK: European Association of Science Editors; 2013.
2. Habibzadeh F. Statistical data editing in scientific articles. *J Korean Med Sci* 2017;32(7):1072-6. [PUBMED](#) | [CROSSREF](#)
3. Misra DP, Zimba O, Gasparyan AY. Statistical data presentation: a primer for rheumatology researchers. *Rheumatol Int* 2021;41(1):43-55. [PUBMED](#) | [CROSSREF](#)
4. Habibzadeh F. Common statistical mistakes in manuscripts submitted to biomedical journals. *Eur Sci Ed* 2013;39(4):92-4.
5. Habibzadeh F. How to report the results of public health research. *J Public Health Emerg* 2017;1:90. [CROSSREF](#)
6. Altman DG, Bland JM. Statistics notes: the normal distribution. *BMJ* 1995;310(6975):298. [PUBMED](#) | [CROSSREF](#)
7. Shatz I. Assumption-checking rather than (just) testing: the importance of visualization and effect size in statistical diagnostics. *Behav Res Methods*. Forthcoming 2023. DOI: 10.3758/s13428-023-02072-x. [PUBMED](#) | [CROSSREF](#)
8. Barker LE, Shaw KM. Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. *Am J Clin Nutr* 2015;102(3):533-9. [PUBMED](#) | [CROSSREF](#)
9. Casson RJ, Farmer LD. Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clin Exp Ophthalmol* 2014;42(6):590-6. [PUBMED](#) | [CROSSREF](#)
10. Nielsen EE, Nørskov AK, Lange T, Thabane L, Wetterslev J, Beyersmann J, et al. Assessing assumptions for statistical analyses in randomised clinical trials. *BMJ Evid Based Med* 2019;24(5):185-9. [PUBMED](#) | [CROSSREF](#)
11. Hu Y, Plonsky L. Statistical assumptions in L2 research: a systematic review. *Second Lang Res* 2019;37(1):171-84. [CROSSREF](#)
12. Hoekstra R, Kiers HA, Johnson A. Are assumptions of well-known statistical techniques checked, and why (not)? *Front Psychol* 2012;3:137. [PUBMED](#) | [CROSSREF](#)
13. Habibzadeh F, Roozbehi H. No need for a gold-standard test: on the mining of diagnostic test performance indices merely based on the distribution of the test value. *BMC Med Res Methodol* 2023;23(1):30. [PUBMED](#) | [CROSSREF](#)
14. Sawilowsky SS. Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney test for shift in location parameter. *J Mod Appl Stat Methods* 2005;4(2):598-600. [CROSSREF](#)
15. Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J Am Stat Assoc* 1967;62(318):399-402. [CROSSREF](#)
16. 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(1):67-72. [PUBMED](#) | [CROSSREF](#)
17. Gross J, Ligges U. *nortest: Tests for Normality*. R package version 1.0-4. The Comprehensive R Archive Network; 2015.
18. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY, USA: Springer-Verlag; 2016.
19. Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: keys and clues. *Bioscience* 2001;51(5):341-52. [CROSSREF](#)
20. Habibzadeh F, Habibzadeh P, Yadollahie M, Roozbehi H. On the information hidden in a classifier distribution. *Sci Rep* 2021;11(1):917. [PUBMED](#) | [CROSSREF](#)
21. Habibzadeh F, Habibzadeh P, Yadollahie M, Sajadi MM. Determining the SARS-CoV-2 serological immunoassay test performance indices based on the test results frequency distribution. *Biochem Med (Zagreb)* 2022;32(2):020705. [PUBMED](#) | [CROSSREF](#)
22. Box GE, Cox DR. An analysis of transformations. *J R Stat Soc B* 1964;26(2):211-52. [CROSSREF](#)
23. Millard SP. *EnvStats: An R Package for Environmental Statistics*. New York, NY, USA: Springer; 2013.
24. Lee DK. Data transformation: a focus on the interpretation. *Korean J Anesthesiol* 2020;73(6):503-8. [PUBMED](#) | [CROSSREF](#)