

Descriptive Statistics and Normality Tests for Statistical Data

Abstract

Descriptive statistics are an important part of biomedical research which is used to describe the basic features of the data in the study. They provide simple summaries about the sample and the measures. Measures of the central tendency and dispersion are used to describe the quantitative data. For the continuous data, test of the normality is an important step for deciding the measures of central tendency and statistical methods for data analysis. When our data follow normal distribution, parametric tests otherwise nonparametric methods are used to compare the groups. There are different methods used to test the normality of data, including numerical and visual methods, and each method has its own advantages and disadvantages. In the present study, we have discussed the summary measures and methods used to test the normality of the data.

Keywords: Biomedical research, descriptive statistics, numerical and visual methods, test of normality

Introduction

A data set is a collection of the data of individual cases or subjects. Usually, it is meaningless to present such data individually because that will not produce any important conclusions. In place of individual case presentation, we present summary statistics of our data set with or without analytical form which can be easily absorbable for the audience. Statistics which is a science of collection, analysis, presentation, and interpretation of the data, have two main branches, are descriptive statistics and inferential statistics.^[1]

Summary measures or summary statistics or descriptive statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible. Descriptive statistics are the kind of information presented in just a few words to describe the basic features of the data in a study such as the mean and standard deviation (SD).^[2,3] The another is inferential statistics, which draw conclusions from data that are subject to random variation (e.g., observational errors and sampling variation). In inferential statistics, most predictions are for the future and generalizations about a population by studying a smaller sample.^[2,4] To draw the inference from the study participants in

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 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.

For reprints contact: reprints@medknow.com

terms of different groups, etc., statistical methods are used. These statistical methods have some assumptions including normality of the continuous data. There are different methods used to test the normality of data, including numerical and visual methods, and each method has its own advantages and disadvantages.^[5] Descriptive statistics and inferential statistics both are employed in scientific analysis of data and are equally important in the statistics. In the present study, we have discussed the summary measures to describe the data and methods used to test the normality of the data. To understand the descriptive statistics and test of the normality of the data, an example [Table 1] with a data set of 15 patients whose mean arterial pressure (MAP) was measured are given below. Further examples related to the measures of central tendency, dispersion, and tests of normality are discussed based on the above data.

Descriptive Statistics

There are three major types of descriptive statistics: Measures of frequency (frequency, percent), measures of central tendency (mean, median and mode), and measures of dispersion or variation (variance, SD, standard error, quartile, interquartile range, percentile, range, and coefficient of variation [CV]) provide simple summaries about the sample

How to cite this article: 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.

**Prabhaker Mishra,
Chandra M Pandey,
Uttam Singh,
Anshul Gupta¹,
Chinmoy Sahu²,
Amit Keshri³**

*Departments of Biostatistics and Health Informatics,
¹Haematology, ²Microbiology and ³Neuro-Otology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, India*

Address for correspondence:
Dr. Anshul Gupta,
Department of Haematology,
Sanjay Gandhi Postgraduate
Institute of Medical Sciences,
Lucknow - 226 014,
Uttar Pradesh, India.
E-mail: anshulhaemat@gmail.com

Access this article online

Website: www.annals.in

DOI: 10.4103/aca.ACA_157_18

Quick Response Code:



and the measures. A measure of frequency is usually used for the categorical data while others are used for quantitative data.

Measures of Frequency

Frequency statistics simply count the number of times that in each variable occurs, such as the number of males and females within the sample or population. Frequency analysis is an important area of statistics that deals with the number of occurrences (frequency) and percentage. For example, according to Table 1, out of the 15 patients, frequency of the males and females were 8 (53.3%) and 7 (46.7%), respectively.

Measures of Central Tendency

Data are commonly describe the observations in a measure of central tendency, which is also called measures of central location, is used to find out the representative value of a data set. The mean, median, and mode are three types of measures of central tendency. Measures of central tendency give us one value (mean or median) for the distribution and this value represents the entire distribution. To make comparisons between two or more groups, representative values of these distributions are compared. It helps in further statistical analysis because many techniques of statistical analysis such as measures of dispersion, skewness, correlation, *t*-test, and ANOVA test are calculated using value of measures of central tendency. That is why measures of central tendency are also called as measures of the first order. A representative value (measures of central tendency) is considered good when it was calculated using all observations and not affected by extreme values because these values are used to calculate for further measures.

Computation of Measures of Central Tendency

Mean

Mean is the mathematical average value of a set of data. Mean can be calculated using summation of the observations divided by number of observations. It is the most popular measure and very easy to calculate. It is a unique value for one group, that is, there is only one answer, which is useful when comparing between the groups. In the computation of mean, all the observations are used.^[2,5] One disadvantage with mean is that it is affected by extreme values (outliers). For example, according to Table 2, mean MAP of the patients was 97.47 indicated that average MAP of the patients was 97.47 mmHg.

Median

The median is defined as the middle most observation if data are arranged either in increasing or decreasing order of magnitude. Thus, it is one of the observations, which occupies the central place in the distribution (data). This is also called positional average. Extreme values (outliers) do not affect the median. It is unique, that is, there is only

Table 1: Distribution of mean arterial pressure (mmHg) as per sex

	Patient number														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MAP	82	84	85	88	92	93	94	95	98	100	102	107	110	116	116
Sex	M	F	F	M	M	F	F	M	M	F	M	F	M	F	M

MAP: Mean arterial pressure, M: Male, F: Female

Table 2: Descriptive statistics of the mean arterial pressure (mmHg)

Mean	SD	SE	Q1	Q2	Q3	Minimum	Maximum	Mode
97.47	11.01	2.84	88	95	107	82	116	116

SD: Standard deviation, SE: Standard error, Q1: First quartile, Q2: Second quartile, Q3: Third quartile

one median of one data set which is useful when comparing between the groups. There is one disadvantage of median over mean that it is not as popular as mean.^[6] For example, according to Table 2, median MAP of the patients was 95 mmHg indicated that 50% observations of the data are either less than or equal to the 95 mmHg and rest of the 50% observations are either equal or greater than 95 mmHg.

Mode

Mode is a value that occurs most frequently in a set of observation, that is, the observation, which has maximum frequency is called mode. In a data set, it is possible to have multiple modes or no mode exists. Due to the possibility of the multiple modes for one data set, it is not used to compare between the groups. For example, according to Table 2, maximum repeated value is 116 mmHg (2 times) rest are repeated one time only, mode of the data is 116 mmHg.

Measures of Dispersion

Measures of dispersion is another measure used to show how spread out (variation) in a data set also called measures of variation. It is quantitatively degree of variation or dispersion of values in a population or in a sample. More specifically, it is showing lack of representation of measures of central tendency usually for mean/median. These are indices that give us an idea about homogeneity or heterogeneity of the data.^[2,6]

Common measures

Variance, SD, standard error, quartile, interquartile range, percentile, range, and CV.

Computation of Measures of Dispersion

Standard deviation and variance

The SD is a measure of how spread out values is from its mean value. Its symbol is σ (the Greek letter sigma) or *s*. It is called SD because we have taken a standard value (mean) to measures the dispersion. Where x_i is individual value, \bar{x} is mean value. If sample size is <30 , we use " $n-1$ " in denominator, for sample size ≥ 30 , use " n " in denominator. The variance (s^2) is defined as the average

of the squared difference from the mean. It is equal to the square of the SD (s).

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad s^2 = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

For example, in the above, SD is 11.01 mmHg. When $n < 30$ which showed that approximate average deviation between mean value and individual values is 11.01. Similarly, variance is 121.22 [i.e., $(11.01)^2$], which showed that average square deviation between mean value and individual values is 121.22 [Table 2].

Standard error

Standard error is the approximate difference between sample mean and population mean. When we draw the many samples from same population with same sample size through random sampling technique, then SD among the sample means is called standard error. If sample SD and sample size are given, we can calculate standard error for this sample, by using the formula.

Standard error = sample SD / $\sqrt{\text{sample size}}$.

For example, according to Table 2, standard error is 2.84 mmHg, which showed that average mean difference between sample means and population mean is 2.84 mmHg [Table 2].

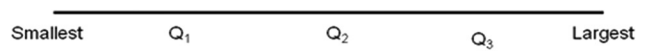
Quartiles and interquartile range

The quartiles are the three points that divide the data set into four equal groups, each group comprising a quarter of the data, for a set of data values which are arranged in either ascending or descending order. Q1, Q2, and Q3 are represent the first, second, and third quartile's value.^[7]

For *i*th Quartile = $[i * (n + 1)/4]^{\text{th}}$ observation, where *i* = 1, 2, 3.

For example, in the above, first quartile (Q1) = $(n + 1)/4 = (15 + 1)/4 = 4^{\text{th}}$ observation from initial = 88 mmHg (i.e., first 25% number of observations of the data are either ≤ 88 and rest 75% observations are either ≥ 88), Q2 (also called median) = $[2 * (n + 1)/4] = 8^{\text{th}}$ observation from initial = 95 mmHg, that is, first 50% number of observations of the data are either less or equal to the 95 and rest 50% observations are either ≥ 95 , and similarly Q3 = $[3 * (n + 1)/4] = 12^{\text{th}}$ observation from initial = 107 mmHg, i.e., indicated that first 75% number of observations of the data are either ≤ 107 and rest 25% observations are either ≥ 107 . The interquartile range (IQR) is a measure of variability, also called the midspread or middle 50%, which is a measure of statistical dispersion, being equal to the difference between 75th (Q3 or third quartile) and 25th (Q1 or first quartile) percentiles. For example, in the above example, three quartiles, that is, Q1, Q2, and Q3 are 88, 95, and 107, respectively. As the first and third

quartile in the data is 88 and 107. Hence, IQR of the data is 19 mmHg (also can write like: 88–107) [Table 2].

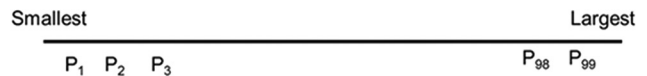


Percentile

The percentiles are the 99 points that divide the data set into 100 equal groups, each group comprising a 1% of the data, for a set of data values which are arranged in either ascending or descending order. About 25% percentile is the first quartile, 50% percentile is the second quartile also called median value, while 75% percentile is the third quartile of the data.

For *i*th percentile = $[i * (n + 1)/100]^{\text{th}}$ observation, where *i* = 1, 2, 3, ..., 99.

Example: In the above, 10th percentile = $[10 * (n + 1)/100] = 1.6^{\text{th}}$ observation from initial which is fall between the first and second observation from the initial = 1st observation + $0.6 * (\text{difference between the second and first observation}) = 83.20$ mmHg, which indicated that 10% of the data are either ≤ 83.20 and rest 90% observations are either ≥ 83.20 .



Coefficient of Variation

Interpretation of SD without considering the magnitude of mean of the sample or population may be misleading. To overcome this problem, CV gives an idea. CV gives the result in terms of ratio of SD with respect to its mean value, which expressed in %. $CV = 100 * (SD/\text{mean})$. For example, in the above, coefficient of the variation is 11.3% which indicated that SD is 11.3% of its mean value [i.e., $100 * (11.01/97.47)$] [Table 2].

Range

Difference between largest and smallest observation is called range. If A and B are smallest and largest observations in a data set, then the range (R) is equal to the difference of largest and smallest observation, that is, $R = A - B$.

For example, in the above, minimum and maximum observation in the data is 82 mmHg and 116 mmHg. Hence, the range of the data is 34 mmHg (also can write like: 82–116) [Table 2].

Descriptive statistics can be calculated in the statistical software “SPSS” (analyze → descriptive statistics → frequencies or descriptives).

Normality of data and testing

The standard normal distribution is the most important continuous probability distribution has a bell-shaped density curve described by its mean and SD and extreme values in the data set have no significant impact on

the mean value. If a continuous data is follow normal distribution then 68.2%, 95.4%, and 99.7% observations are lie between mean ± 1 SD, mean ± 2 SD, and mean ± 3 SD, respectively.^[2,4]

Why to test the normality of data

Various statistical methods used for data analysis make assumptions about normality, including correlation, regression, *t*-tests, and analysis of variance. Central limit theorem states that when sample size has 100 or more observations, violation of the normality is not a major issue.^[5,8] Although for meaningful conclusions, assumption of the normality should be followed irrespective of the sample size. If a continuous data follow normal distribution, then we present this data in mean value. Further, this mean value is used to compare between/among the groups to calculate the significance level (*P* value). If our data are not normally distributed, resultant mean is not a representative value of our data. A wrong selection of the representative value of a data set and further calculated significance level using this representative value might give wrong interpretation.^[9] That is why, first we test the normality of the data, then we decide whether mean is applicable as representative value of the data or not. If applicable, then means are compared using parametric test otherwise medians are used to compare the groups, using nonparametric methods.

Methods used for test of normality of data

An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing. There are two main methods of assessing normality: Graphical and numerical (including statistical tests).^[3,4] Statistical tests have the advantage of making an objective judgment of normality but have the disadvantage of sometimes not being sensitive enough at low sample sizes or overly sensitive to large sample sizes. Graphical interpretation has the advantage of allowing good judgment to assess normality in situations when numerical tests might be over or undersensitive. Although normality assessment using graphical methods need a great deal of the experience to avoid the wrong interpretations. If we do not have a good experience, it is the best to rely on the numerical methods.^[10] There are various methods available to test the normality of the continuous data, out of them, most popular methods are Shapiro–Wilk test, Kolmogorov–Smirnov test, skewness, kurtosis, histogram, box plot, P–P Plot, Q–Q Plot, and mean with SD. The two well-known tests of normality, namely, the Kolmogorov–Smirnov test and the Shapiro–Wilk test are most widely used methods to test the normality of the data. Normality tests can be conducted in the statistical software “SPSS” (analyze \rightarrow descriptive statistics \rightarrow explore \rightarrow plots \rightarrow normality plots with tests).

The Shapiro–Wilk test is more appropriate method for small sample sizes (<50 samples) although it can also

be handling on larger sample size while Kolmogorov–Smirnov test is used for $n \geq 50$. For both of the above tests, null hypothesis states that data are taken from normal distributed population. When $P > 0.05$, null hypothesis accepted and data are called as normally distributed. Skewness is a measure of symmetry, or more precisely, the lack of symmetry of the normal distribution. Kurtosis is a measure of the peakedness of a distribution. The original kurtosis value is sometimes called kurtosis (proper). Most of the statistical packages such as SPSS provide “excess” kurtosis (also called kurtosis [excess]) obtained by subtracting 3 from the kurtosis (proper). A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. If mean, median, and mode of a distribution coincide, then it is called a symmetric distribution, that is, skewness = 0, kurtosis (excess) = 0. A distribution is called approximate normal if skewness or kurtosis (excess) of the data are between -1 and $+1$. Although this is a less reliable method in the small-to-moderate sample size (i.e., $n < 300$) because it can not adjust the standard error (as the sample size increases, the standard error decreases). To overcome this problem, a *z*-test is applied for normality test using skewness and kurtosis. A *Z* score could be obtained by dividing the skewness values or excess kurtosis value by their standard errors. For small sample size ($n < 50$), *z* value ± 1.96 are sufficient to establish normality of the data.^[8] However, medium-sized samples ($50 \leq n < 300$), at absolute *z*-value ± 3.29 , conclude the distribution of the sample is normal.^[11] For sample size > 300 , normality of the data is depend on the histograms and the absolute values of skewness and kurtosis. Either an absolute skewness value ≤ 2 or an absolute kurtosis (excess) ≤ 4 may be used as reference values for determining considerable normality.^[11] A histogram is an estimate of the probability distribution of a continuous variable. If the graph is approximately bell-shaped and symmetric about the mean, we can assume normally distributed data^[12,13] [Figure 1]. In statistics, a Q–Q plot is a scatterplot created by plotting two sets of quantiles (observed and expected) against one another. For normally distributed data, observed data are approximate to the expected data, that is, they are statistically equal [Figure 2]. A P–P plot (probability–probability plot or percent–percent plot) is a graphical technique for assessing how closely two data sets (observed and expected) agree. It forms an approximate straight line when data are normally distributed. Departures from this straight line indicate departures from normality [Figure 3]. Box plot is another way to assess the normality of the data. It shows the median as a horizontal line inside the box and the IQR (range between the first and third quartile) as the length of the box. The whiskers (line extending from the top and bottom of the box) represent the minimum and maximum values when they are within 1.5 times the IQR from either end of the box (i.e., $Q1 - 1.5 * IQR$ and

$Q3 + 1.5 * IQR$). Scores >1.5 times and 3 times the IQR are out of the box plot and are considered as outliers and extreme outliers, respectively. A box plot that is symmetric with the median line at approximately the center of the box and with symmetric whiskers indicate that the data may have come from a normal distribution. In case many outliers are present in our data set, either outliers are need to remove or data should treat as nonnormally distributed^[8,13,14] [Figure 4]. Another method of normality of the data is relative value of the SD with respect to mean. If SD is less than half mean (i.e., $CV < 50\%$), data are considered normal.^[15] This is the quick method to test the normality. However this method should only be used when our sample size is at least 50.

For example in Table 1, data of MAP of the 15 patients are given. Normality of the above data was assessed. Result showed that data were normally distributed as skewness (0.398) and kurtosis (-0.825) individually were within ± 1 . Critical ratio (Z value) of the skewness (0.686) and kurtosis (-0.737) were within ± 1.96 , also evident to normally distributed. Similarly, Shapiro-Wilk test ($P = 0.454$) and Kolmogorov-Smirnov test ($P = 0.200$)

were statistically insignificant, that is, data were considered normally distributed. As sample size is <50 , we have to take Shapiro-Wilk test result and Kolmogorov-Smirnov test result must be avoided, although both methods indicated that data were normally distributed. As SD of the MAP was less than half mean value ($11.01 < 48.73$), data were considered normally distributed, although due to sample size <50 , we should avoid this method because it should use when our sample size is at least 50 [Tables 2 and 3].

Conclusions

Descriptive statistics are a statistical method to summarizing data in a valid and meaningful way. A good and appropriate measure is important not only for data but also for statistical methods used for hypothesis testing. For continuous data, testing of normality is very important because based on the normality status, measures of central tendency, dispersion, and selection of parametric/nonparametric test are decided. Although there are various methods for normality testing but for small sample size ($n < 50$), Shapiro-Wilk test should be used as it has more power to detect the nonnormality

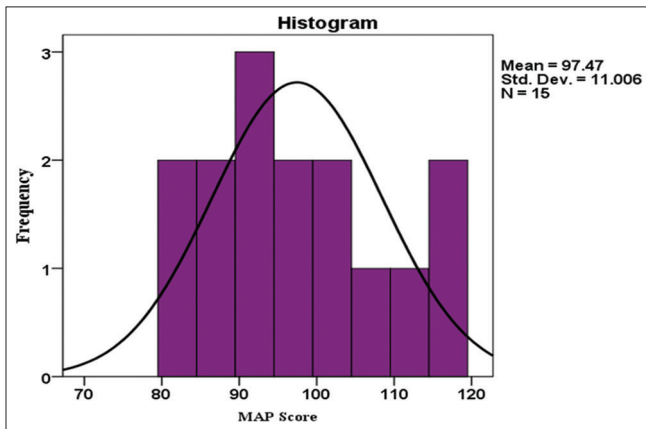


Figure 1: Histogram showing the distribution of the mean arterial pressure

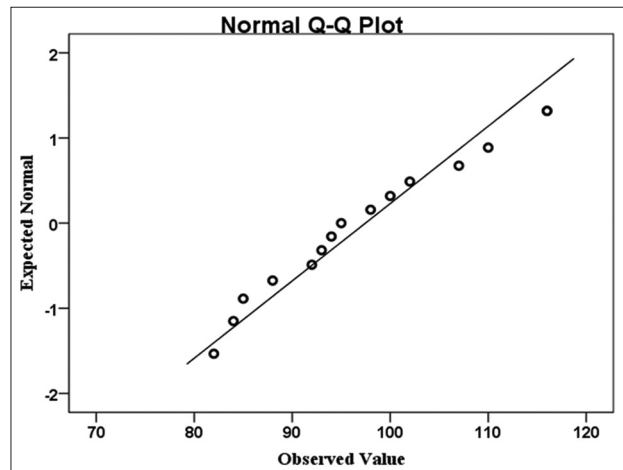


Figure 2: Normal Q-Q Plot showing correlation between observed and expected values of the mean arterial pressure

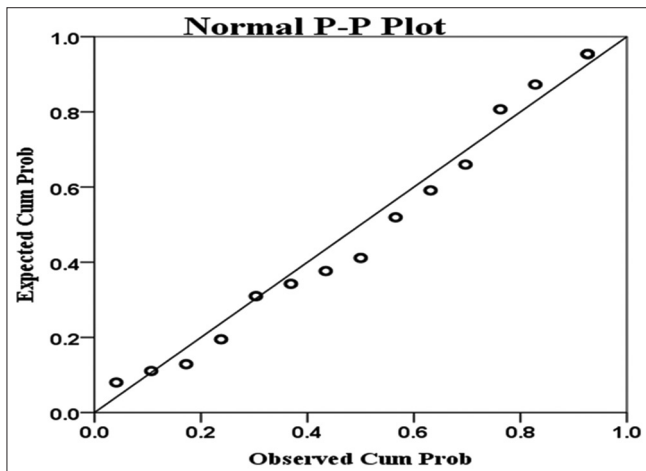


Figure 3: Normal P-P Plot showing correlation between observed and expected cumulative probability of the mean arterial pressure

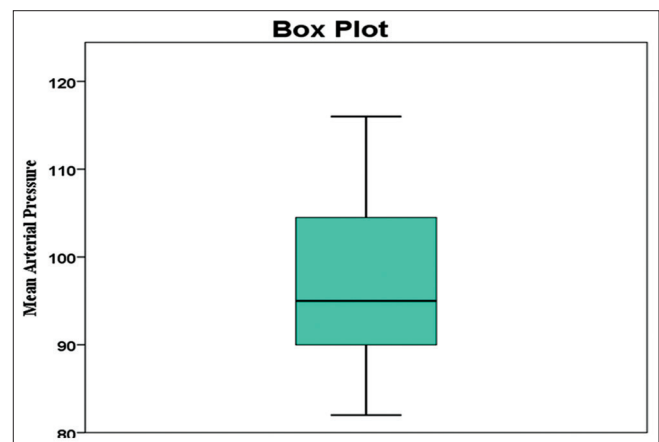


Figure 4: Boxplot showing distribution of the mean arterial pressure

Table 3: Skewness, kurtosis, and normality tests for mean arterial pressure (mmHg)

Variable	Skewness			Kurtosis			P	
	Value	SE	Z	Value	SE	Z	K-S test with Lilliefors correction	Shapiro-Wilk test
MAP score	0.398	0.580	0.686	-0.825	1.12	-0.737	0.200	0.454

K-S: Kolmogorov–Smirnov, SD: Standard deviation, SE: Standard error

and this is the most popular and widely used method. When our sample size (n) is at least 50, any other methods (Kolmogorov–Smirnov test, skewness, kurtosis, z value of the skewness and kurtosis, histogram, box plot, P–P Plot, Q–Q Plot, and SD with respect to mean) can be used to test of the normality of continuous data.

Acknowledgment

The authors would like to express their deep and sincere gratitude to Dr. Prabhat Tiwari, Professor, Department of Anaesthesiology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, for his critical comments and useful suggestions that was very much useful to improve the quality of this manuscript.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

- Lund Research Ltd. Descriptive and Inferential Statistics. Available from: <http://www.statistics.laerd.com>. [Last accessed on 2018 Aug 02].
- Sundaram KR, Dwivedi SN, Sreenivas V. Medical Statistics: Principles and Methods. 2nd ed. New Delhi: Wolters Kluwer India; 2014.
- Bland M. An Introduction to Medical Statistics. 4th ed. Oxford: Oxford University Press; 2015.
- Campbell MJ, Machin D, Walters SJ. Medical Statistics: A text book for the health sciences, 4th ed. Chichester: John Wiley & Sons, Ltd.; 2007.
- Altman DG, Bland JM. Statistics notes: The normal distribution. *BMJ* 1995;310:298.
- Altman DG. Practical Statistics for Medical Research Chapman and Hall/CRC Texts in Statistical Science. London: CRC Press; 1999.
- Indrayan A, Sarmukaddam SB. Medical Bio-Statistics. New York: Marcel Dekker Inc.; 2000.
- Ghasemi A, Zahediasl S. Normality tests for statistical analysis: A guide for non-statisticians. *Int J Endocrinol Metab* 2012;10:486-9.
- Indrayan A, Satyanarayana L. Essentials of biostatistics. *Indian Pediatr* 1999;36:1127-34.
- Lund Research Ltd. Testing for Normality using SPSS Statistics. Available from: <http://www.statistics.laerd.com>. [Last accessed 2018 Aug 02].
- Kim HY. Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod* 2013;38:52-4.
- Armitage P, Berry G. Statistical Methods in Medical Research. 2nd ed. London: Blackwell Scientific Publications; 1987.
- Barton B, Peat J. Medical Statistics: A Guide to SPSS, Data Analysis and Clinical Appraisal. 2nd ed. Sydney: Wiley Blackwell, BMJ Books; 2014.
- Baghban AA, Younespour S, Jambarsang S, Yousefi M, Zayeri F, Jalilian FA. How to test normality distribution for a variable: A real example and a simulation study. *J Paramed Sci* 2013;4:73-7.
- Jeyaseelan L. Short Training Course Materials on Fundamentals of Biostatistics, Principles of Epidemiology and SPSS. CMC Vellore: Biostatistics Resource and Training Center (BRTC); 2007.