Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 6** : 1–16.
ISSN-0037-8712
doi:10.5269/bspm.79145

Evaluating Normality in Continuous Data: Evidence from LASI Wave-1

Kanchan Yadav* and Dechenla Tshering Bhutia

ABSTRACT: Descriptive statistics are indispensable in empirical research as they furnish concise summaries that clarify the fundamental attributes of study variables. These summaries generally include measures of central tendency and variability, which are essential for precise data interpretation. Nonetheless, whether to use parametric or non-parametric statistical tests depends mostly on how well the dataset satisfies the normality assumption. Consequently, the evaluation of the normality of continuous variables emerges as a pivotal preliminary phase in statistical analysis. The present study examines the assessment of normality utilizing both quantitative and visual methodologies on continuous variables sourced from the Longitudinal Ageing Study in India (LASI) Wave-1 dataset. Quantitative approaches incorporate the Anderson-Darling test, while visual evaluations are performed through histograms, Q-Q plots, and P-P plots. The data were subjected to analysis utilizing SPSS and Minitab software to compare the efficacy and interpretability of each technique. The findings of the study suggest that visual methods offer intuitive understanding of data distribution, especially when used alongside statistical tests. The Anderson-Darling test is particularly effective for medium to large sample sizes, providing a reliable assessment of deviations from normality. Descriptive and visual analyses indicate that the data largely follow a normal distribution. However, no single method is universally best, as the choice of method depends on factors such as sample size, data characteristics and the research context. This study underscores the necessity of performing normality assessments prior to hypothesis testing to ensure the validity and reliability of research findings. The utilization of an integrated approach combining both quantitative and visual methods facilitate a more comprehensive understanding of data distribution, especially when engaging with extensive secondary datasets such as LASI.

Keywords: LASI, normality testing, descriptive statistics, visual and numerical methods, statistical tests.

Contents

* Corresponding author.
2020 *Mathematics Subject Classification:* 62F03, 62G10, 62P25.
Submitted September 24, 2025. Published April 29, 2026.

1. Introduction

Statistical analysis frequently presupposes that datasets adhere to a normal distribution, particularly within the realm of parametric methodologies, including t-tests, ANOVA, and regression models [1, 2]. The assumption of normality is critical for ensuring the validity of inferential statistics as well as the robustness of the resultant findings [3]. Secondary data, which is gathered for objectives that diverge from the current research aims, often exhibits variability in quality and structure, thus raising issues regarding the relevance of normality assessments. In contrast, primary data enables researchers to mitigate biases and guarantee the implementation of suitable data collection techniques. However, secondary data presents challenges such as the presence of missing values, undetermined measurement inaccuracies and heterogeneous sample characteristics [4, 5]. It is vital to ascertain whether normality testing holds significance for secondary data, as this determination impacts data preprocessing, the selection of statistical methods, and the interpretation of findings [6, 7].

Several previous studies [8, 9] have scrutinized the process of normality testing within primary datasets, frequently advocating for the implementation of transformations or alternative statistical methodologies when the underlying assumption is violated [1]. Nevertheless, there exists a lack of research concentrated on secondary data, despite its rapid increase in application in domains such as epidemiology, social sciences and economic inquiry [10]. Investigations by previous studies [8, 11], have assessed normality tests, yet they did not specifically consider their applicability to secondary datasets.

Although the increasing availability of robust and non-parametric statistical methods has reduced strict dependence on the normality assumption, still normality testing remains an important component of exploratory data analysis [12]. Assessing normality provides valuable insights into the underlying structure and distributional characteristics of data, informs the selection of appropriate summary measures, and enhances transparency in analytical decision-making [13]. Even when robust or non-parametric methods are employed, an understanding of distributional behaviour aids in the interpretation of results and the identification of outliers or data irregularities [7]. This is particularly relevant in large-scale secondary datasets, where heterogeneity and measurement variability are common [4, 8]. Consequently, normality assessment continues to play a meaningful role in contemporary statistical practice, complementing rather than competing with modern analytical approaches.

The extent of literature reveals an absence of consensus regarding the necessity of normality testing in the analysis of secondary data [14, 15]. Although some researchers maintain that the need for normality test is no longer necessary due to modern robust statistical methods, while others maintain that the verification of distributional assumptions enhances the reliability of study findings [15]. This disparity underscores the imperative for a methodical investigation into the circumstances and rationale for the application of normality testing to secondary data. This research study explains the function of normality testing within secondary data analysis, thereby guiding the researchers in the selection of suitable statistical methodologies. Through the examination of empirical datasets and the execution of simulation studies, we provide practical guidance on the conditions under which normality tests should be employed, as well as instances in which alternative strategies may be preferable. The objective of this study is to provide a comprehensive examination of the methodologies employed for assessing normality within the context of statistical analysis on secondary data, specifically the Longitudinal Ageing Study in India (LASI), utilizing both Minitab (version 22) and SPSS (version 26) software applications.

2. Materials and Methods

Data Source, Sampling Procedure and Sample Size

The dataset from the Longitudinal Ageing Study in India (LASI) Wave 1 (2017–2018) were used in this study, which constitutes a nationally and state-representative longitudinal large-scale inquiry into the phenomena of ageing and health, particularly focusing on older adults aged 45 years and above, along with their spouses irrespective of age. LASI delivers valid, reliable and continuous scientific insights regarding the health, social, mental and economic well-being of the specified population. The study sample includes non-institutionalized individuals from across India, selected using a multistage, stratified, area-based cluster sampling method that ensures representation from all states and Union Territories. [16].

LASI employed a three-stage sampling design in rural contexts and a four-stage sampling design within urban settings. At the initial stage, based on the 2011 Indian census, the enumeration of sub-districts (Tehsils/Talukas) was utilized as Primary Sampling Units (PSUs) for each state or Union Territory. Within each region, the PSUs were selected employing Probability Proportional to Size (PPS) sampling, where the number of households within a PSU served as the size metric. The subsequent stage involved the selection of a predetermined number of Secondary Sampling Units (SSUs), which comprised villages in rural settings and wards in urban contexts of the chosen PSUs. In rural areas, the third stage involved a systematic selection of a predetermined number of households from the identified villages. Conversely, the urban sampling process necessitated an additional stage. During the third stage, one Census Enumeration Block (CEB) was randomly chosen from each selected urban ward. In the fourth stage, a predetermined number of households from this CEB were systematically selected. The overarching objective of this comprehensive sampling framework was to ensure the acquisition of a representative sample at each stage of the selection process [16].

The analysis was conducted using a combined dataset that included information from both individual level and biomarker data. Subsequent to the integration of these datasets, with a particular emphasis on middle-aged and older adults, the conclusive analysis preserved data on 73,396 individuals aged 45 years and older. Additional details concerning the sampling methodologies and sample sizes may be obtained from the LASI India Report [16, 17].

Data Accessibility and Ethical Considerations

Data for the study were collected through field surveys conducted by survey agencies, with prior informed consent obtained from all participants. The study is based on publicly available data. The ethical approval was taken from the Indian Council of Medical Research (ICMR) during the data collection of the LASI project [17].

Description of Study Variables

For this study, there are four (quantitative) study variables (age, height, weight and BMI) selected for normality assessment. All respondents aged 45 years and above along with height (in cms), weight (in kgs) and BMI (kg/m2) in India covered under LASI (2017- 2018) data.

Research Hypothesis

Null hypothesis (H_0): Data does not follow a normal distribution

Alternative hypothesis (H_1): Data follows a normal distribution

3. Statistical Analysis: Methods for Normality Assessment

The most important distribution in the field of statistical data analysis is said to be the normal distribution. Various statistical tests are formulated under the presumption that the dependent variable or model residuals conform to a normal distribution. Methodologies for assessing data normality encompass both graphical representations and quantitative techniques.

Descriptive Statistics

There exist three principle classifications of descriptive statistics i.e, measures of frequency (including frequency and percentage), measures of central tendency (comprising mean, median and mode), and measures of dispersion or variation (which encompass variance, standard deviation, standard error, quartiles, interquartile range, percentiles, range, and coefficient of variation) along with skewness and kurtosis [12]. These measures yield straight-forward summaries regarding the sample and the associated metrics. Measures of frequency are predominantly employed for categorical data, whereas the latter measures are typically utilized for quantitative datasets.

4. Measures of Central Tendency

The representative value of a dataset is determined by describing the observations in a measure of central tendency, also known as a measure of central location. There are three different ways to measure central tendency i.e, mean, median and mode. One value (the mean or median) for the distribution is provided by measures of central tendency, and this value is representative of the entire distribution. Representative values of these additional statistical distributions are compared in order to compare two or more groups. It aids in a variety of statistical analysis methods, including the calculation of measures of central tendency for the t-test, ANOVA test, correlation, skewness, and dispersion [12]. For this reason, measures of central tendency are also referred to as first-order measurements. Since measures of central tendency are used to calculate other measures, a representative value is deemed good if it was computed using all observations and unaffected by extreme values.

Mean

The mean is a set of data's mathematical average value. The mean can be computed by dividing the total number of observations by their sum. The most widely used metric is also the most straightforward to compute. When comparing groups, it is helpful to know that there is only one answer, meaning that it is a unique value for one group. Every observation is used in the mean calculation [18, 19]. The mean of age, height, weight and BMI of the respondents were shown in Table 1.

Median

When data are arranged in either ascending or descending order of magnitude, the median is the middlemost observation. As a result, it is among the observations that hold the most central position in the distribution (data). Another name for this is positional average. The median is unaffected by extreme values, or outliers. It is distinct in that there is only one dataset's median, which is helpful for group comparisons. The fact that the median is less popular than the mean is one drawback [20]. Table 1 shows that the respondents' median age was 57, meaning that 50 percent of the data observations were either younger than or equal to 57 years old, while the remaining 50 percent were either older or equal to 57.

Mode

In a collection of observations, the value that appears the most frequently is called the mode. In other words, the observation with the highest frequency is the mode. A dataset may contain more than one mode, or it may not have any modes at all. Table 1 shows, for instance, that the mode of the age, height, weight and BMI in the dataset, while the remaining values are repeated only once in the data.

5. Measures of Dispersion

Measures of dispersion, also known as measures of variation, are another way to illustrate how dispersed (variable) a dataset is. It is a quantitative measure of how values vary or disperse within a population or sample. More precisely, it is demonstrating the under-representation of central tendency measures, typically mean/median. These indices help us determine whether the data is homogeneous or heterogeneous [18, 20].

Similar measures are coefficient of variation (CV), quartile, interquartile range, percentile, range, variance, standard deviation (SD) and standard error (SE).

Standard deviation and Variance

The standard deviation (SD) serves as an index for quantifying the extent to which values deviate from their arithmetic mean. Its representation is denoted by the symbol "σ"(the Greek letter sigma) or "s". The term "standard deviation" is employed because it utilizes a reference point (the mean) to assess the extent of variability. In this context, x_i represents an individual observation, while $\bar{x}$ signifies the mean value.

Standard Error

The estimated difference between the population mean and the sample mean is known as the standard error. Standard error is the difference between the sample means when we use a random sampling strategy to draw many samples from the same population with the same sample size. We may use the formula to determine the standard error for this sample if the sample size and sample SD are provided.

$$\text{Standard Error (S.E.)} = \frac{\text{Sample SD}}{\sqrt{\text{Sample Size}}}$$

Quartiles and Interquartile range

The quartiles are the three points, i.e., for a collection of data values that are sorted in either ascending or descending order, which split the dataset into four equal groups, each of which contains a quarter of the data. Q_1, Q_2, and Q_3 stand for the values of the first, second, and third quartiles, respectively [21]. As a measure of statistical dispersion, the interquartile range (IQR) is sometimes referred to as the middle 50% or mid-spread. It is equivalent to the difference between the 25th (Q_1 or first quartile) and 75th (Q_3 or third quartile) percentiles [$\text{IQR} = Q_3 - Q_1$].

Coefficient of Variation (CV)

Without taking into account the sample or population mean's magnitude, interpreting SD could be deceptive. CV provides an idea to solve this issue. The result is presented by CV as the SD ratio in relation to the mean value, which is given as a percentage i.e,

$$\text{CV} = 100 \times \left(\frac{\text{SD}}{\text{mean}}\right)$$

Range

Range is described as the difference between the largest and smallest observation in the dataset. In a data collection, if A and B are the smallest and largest observations, then the range (R) is equal to the difference between the two, meaning that

$$R = A - B.$$

6. Histogram

A histogram is a rough depiction of the distribution of numerical data. Before doing a formal statistical test, it is standard procedure to generate a histogram in order to visually examine the data distribution and possible outliers. If the sample size is sufficiently large, a bell-shaped histogram frequently indicates that the data distribution is roughly normal, if not, outliers and severe skewness i.e, a measure of symmetry in a distribution, as well as higher kurtosis i.e, a measure of the "tailedness" of the distribution of random variables indicate a violation of normalcy. It should be used in conjunction with other statistical tests, though, as it is a subjective approach [2]. For instance, the histogram (Figure 1) is roughly bell-shaped and shows a normal distribution.

7. Normal Q-Q (Quantile-Quantile) Plot

A scatter plot known as a Q-Q plot contrasts the quantiles of the actual data with those of a theoretical distribution, most often a normal distribution. If the data is regularly distributed, the points will roughly fall on the reference line at 45 degrees. Deviations from the line signify skewness or kurtosis, which are deviations from normalcy. For the purpose of identifying minute deviations from normalcy, Q-Q plots are more accurate than histograms [22]. A Q-Q plot is quite similar to a P-P plot, except instead of plotting each individual score in the data as it depicts the quantiles-values that divide a dataset into equal sections. Additionally, when there are huge sample numbers, the Q-Q graphs are simpler to understand [2].

8. Normal P-P (Probability-Probability or Percent-Percent) Plot

A P-P plot constitutes a graphical methodology employed to evaluate whether a dataset adheres to a particular theoretical distribution, such as the Gaussian distribution. The cumulative distribution function (CDF) of the actual data and CDF of the selected theoretical distribution are contrasted in this plot. In the process of constructing a P-P plot, the data are initially ranked and organized, followed by the computation of the cumulative probabilities for each individual data point. These cumulative probabilities are subsequently transformed into expected z-scores based on the theoretical distribution (for eg, the Gaussian distribution). Concurrently, the actual empirical data values are also converted into z-scores. The observed z-scores are then depicted graphically in relation to the expected z-scores. In instances where the data exhibit a normal distribution, the points represented on the plot will closely approximate a 45-degree diagonal line, thereby signifying a favourable fit. P-P plots exhibit heightened sensitivity to discrepancies occurring in the central region of the distribution [8]. They might, however, be less successful than Q-Q plots at identifying tail variations [2].

9. Anderson Darling Test

The Anderson-Darling test represents a statistical methodology employed to evaluate whether a particular sample of data originates from a certain distribution, most commonly the normal distribution. By giving more weight to the distribution's extremes, it improves upon the Kolmogorov-Smirnov test and makes it more sensitive to differences at the data's tails [23]. This test's technique compares the sample data's cumulative distribution function (CDF) with the suggested theoretical distribution's CDF. It computes the squared differences between these two distributions, applying a weighting factor that is the inverse of the variance of the theoretical distribution. This methodology guarantees that deviations occurring at the tails have a more pronounced influence on the ultimate test statistic compared to those occurring in the central region of the distribution [24]. The Anderson-Darling test is frequently favoured over alternative normality assessments, such as the Kolmogorov-Smirnov or Shapiro-Wilk tests, particularly when the objective is to identify subtle deviations from normality, especially in larger datasets or in contexts where tail behaviour is critical (for instance, in risk assessment or detection of outliers) [8].

Assuming that the theoretical distribution is completely described (e.g., a normal distribution with known mean and variance), this test falls within the category of parametric tests. In fact, it is necessary to modify the crucial values when the distribution's parameters are estimated from the data (as is frequently done), or software programs may offer approximations of p-values for inferential analysis [13].

10. Normality of Data and Testing

The most significant continuous probability distribution is the standard normal distribution, which has a bell-shaped density curve that is defined by its mean and standard deviation (SD). The mean value of the dataset is not significantly affected by extreme values. Around 68.2% of observations that fall within the range of (mean $\pm$ 1 SD) and 95.4% of observations fall within the range of (mean $\pm$ 2 SD), and similarly 99.7% of observations fall within the range of (mean $\pm$ 3 SD), respectively, if the continuous data follows a normal distribution [18,25].

Why normality test is required?

Many of the statistical procedures used in data analysis, including correlation, regression, t-tests, and analysis of variance, assume that the data may be normal. The central limit theorem states that when the sample size encompasses 100 or more observations, the violation of the assumption of normality does not constitute a significant concern [1, 19].

However, regardless of sample size, the assumptions of normality should be maintained in order to draw valid conclusions. In instances where continuous data follows a normal distribution, this data is typically expressed in terms of its mean value. Furthermore, this mean value serves as a basis for comparative analyses among groups, facilitating the calculation of the significance level (p-value). In situations where the data deviates from a normal distribution, the resultant mean may not accurately represent the underlying dataset. An erroneous selection of the representative value of a dataset, coupled with the subsequent calculation of significance levels derived from this value, may lead to misinterpretations [26].

Consequently, it is most important to initially assess the normality of the data prior to determining the appropriateness of the mean as a representative value. If deemed applicable, means are compared utilizing parametric tests, otherwise medians are employed for group comparisons, utilizing nonparametric methodologies.

Results

Various techniques used to assess normality

In Table 1, the demographic information includes age, weight, height, and body mass index (BMI) of the respondents are presented. The normality of the aforementioned data was evaluated. The findings indicated that the data had a normal distribution, as evidenced by the skewness values (0.43, 0.1, and 0.58) and kurtosis values (-0.10, 0.22, and 0.56) for age, weight and height, respectively, which were all within the range of ± 1. Also, this study shows through its findings that if Mean $>$ 2 SD, then data follows Normal distribution or in other words, the standard deviations of age, height, weight, and BMI were less than half of the mean values, the data were considered to be normally distributed [27]. However, it is imperative to employ this methodology since it is recommended that the sample size should be at least 50. An additional method for evaluating the normality of the data involves the relative value of the standard deviation in relation to the mean. If the standard deviation is less than half of the mean (i.e., coefficient of variation, CV $< 50\%$), the data can be classified as normal [27]. This represents a rapid approach to assessing normality. Nevertheless, this approach should only be used when the sample size is greater than 50.

Table 1: Descriptive Statistics for age, height, weight and body mass index (BMI).

Variable	**Age**	**Height**	**Weight**	**BMI**
Mean	57.91	155.36	55.57	22.94
SE Mean	0.04	0.03	0.05	0.01
Standard Deviation (S.D.)	11.69	8.87	13.03	4.77
Variance	136.78	78.68	170.01	22.81
Coefficient of Variation (C.V.)	20.19	5.71	23.46	20.81
Minimum	18	57.3	20.05	9.33
Q1	49	149	46.1	19.47
Median	57	154.8	54.4	22.46
Q3	66	161.6	63.7	25.80
Maximum	116	195.8	149.1	145.58
Range	98	138.5	129.05	136.25
IQR	17	12.6	17.6	6.33
Mode	45	150.20	50	21.86
Skewness	0.43	0.1	0.58	1.04
Kurtosis	-0.10	0.22	0.56	9.17

A histogram (shown in Figure 1) serves as an approximation of the probability distribution pertaining to a continuous variable. If the bell-shaped curve in the graphical depiction is symmetrical around the mean, it can be assumed that the data is distributed normally [28, 29].

In the realm of statistics, a Q–Q plot (shown in Figure 2) is a scatterplot that is generated by plotting two sets of quantiles (observed versus expected) against each other. In instances of normally distributed data, the observed values closely align with the expected values, signifying statistical equivalence.

A P–P plot (shown in Figure 3) functions as a visual methodology for evaluating the degree of concordance between two datasets (observed and expected). It typically manifests as an approximate linear trajectory when the data adhere to a normal distribution. Deviations from this linearity signify a deviation from normality.

The prominent tests for normality, specifically the Anderson–Darling test, represent the most widely used approaches for evaluating normality in extensive datasets. The results of the Anderson–Darling test demonstrated statistical significance, thereby allowing the data to be classified as normally distributed.

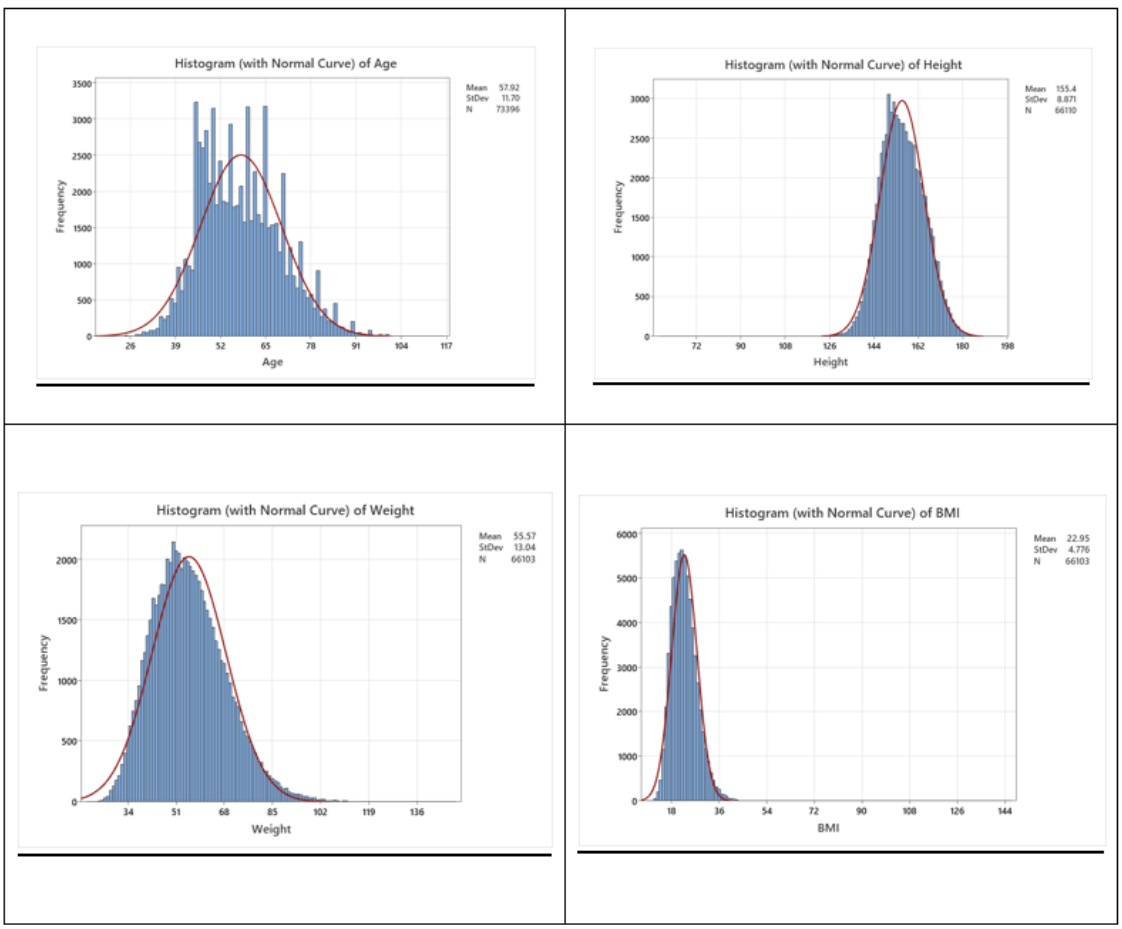

Figure 1: Histogram accompanied by normal distribution curves showing the statistical distribution of age, height, weight and body mass index (BMI).

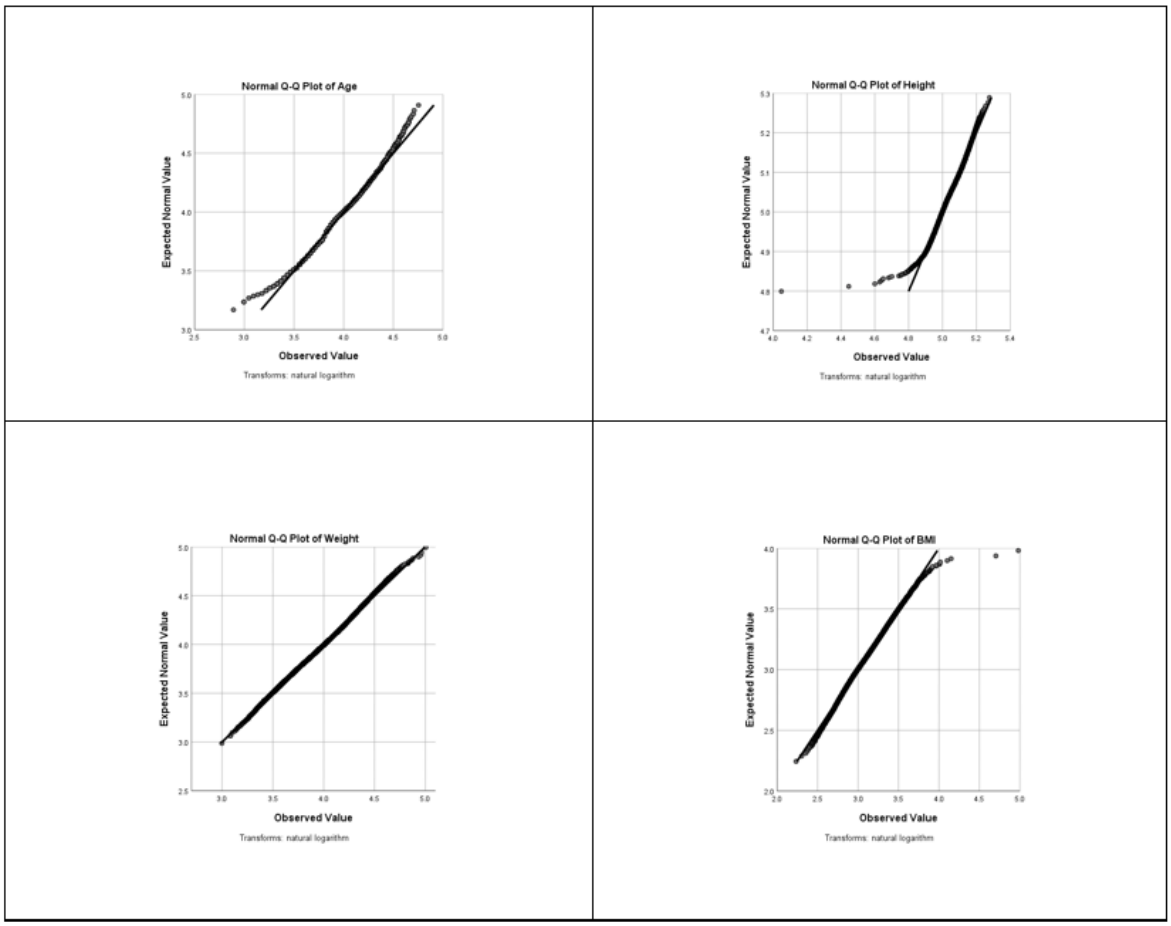

Figure 2: Normal Q–Q Plot showing the correlation between the observed and expected values of age, height, weight and body mass index (BMI).

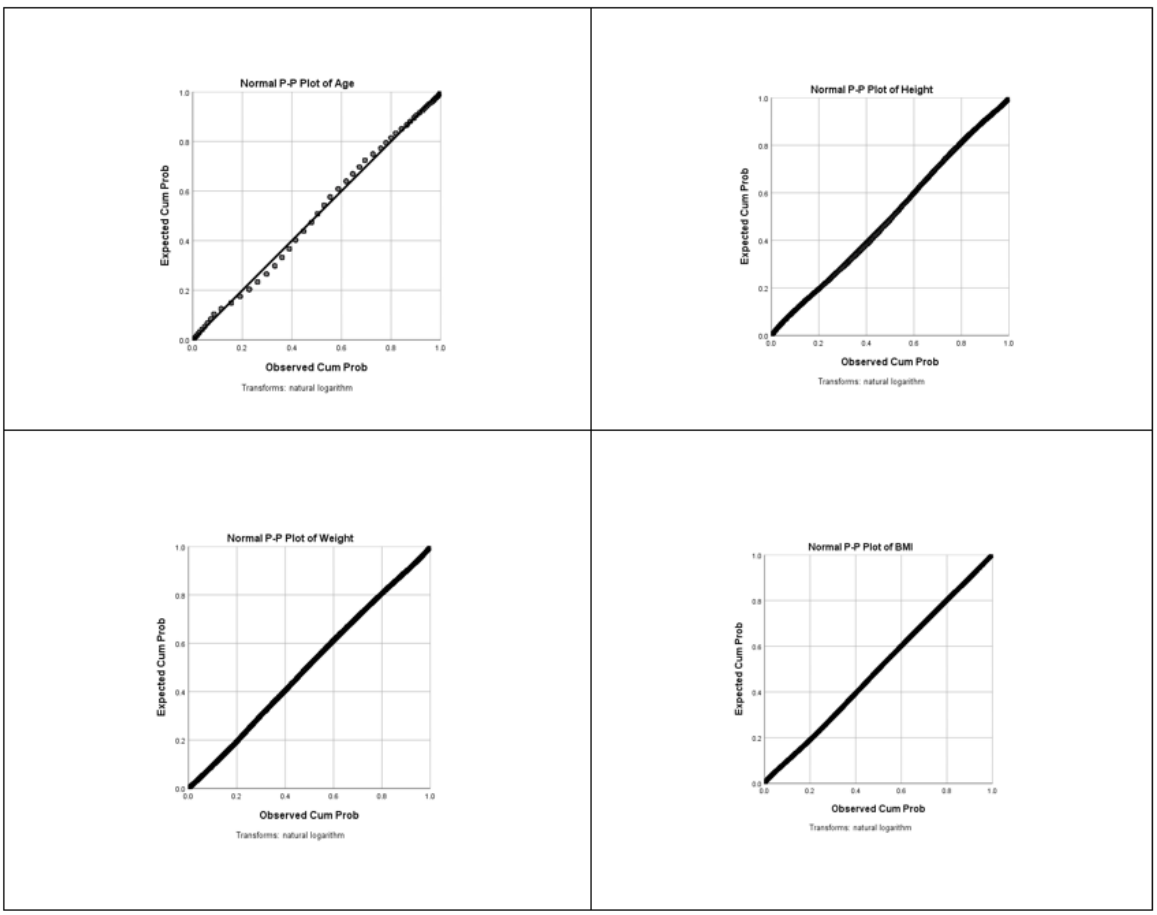

Figure 3: Normal P–P Plot showing the correlation between the observed and expected cumulative probabilities of age, height, weight and body mass index (BMI).

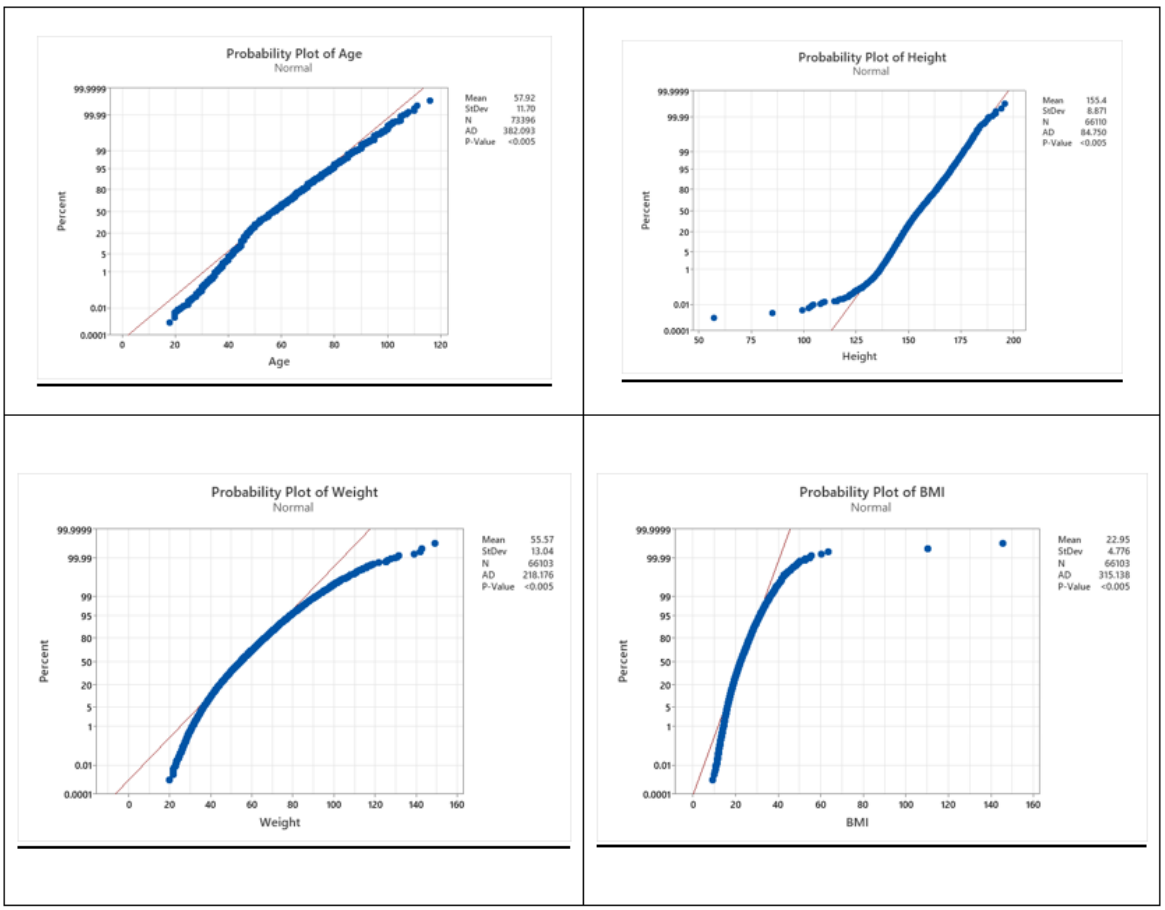

Figure 4: Normality test through Anderson–Darling Test for age, height, weight and body mass index (BMI).

The findings of the Anderson–Darling test (shown in Figure 4) should take precedence over those of the Shapiro–Wilk and Kolmogorov–Smirnov tests, as the sample size exceeds 50. The Anderson–Darling (AD) test enhances the Kolmogorov–Smirnov (KS) test by assigning greater significance to the distribution tails [30]. Although there is a greater computational demand for this test, it has a better degree of statistical power than the KS test. The null hypothesis in this test states clearly that the data come from a population that is not normally distributed. According to Figure 4, the p-value is less than 0.005, hence the null hypothesis is rejected, leading to the conclusion that the data follows a normal distribution.

Discussion

Assessing the normality of continuous data constitutes a crucial step for the implementation of numerous parametric statistical tests, which assume that the data follow a normal distribution [1]. Within the realms of biomedical and public health research, particularly in the analysis of extensive datasets such as LASI Wave-1, confirming this assumption helps improve the reliability of statistical conclusions and supports the appropriate selection of analytical methods.

Two principle categories of techniques are utilized for the evaluation of normality i.e, graphical methods and numerical (statistical) tests [8, 31]. Numerical tests such as the Shapiro-Wilk, Kolmogorov-Smirnov and Anderson-Darling tests provide the advantage of delivering an objective measure of deviation from normality [1, 23]. Nonetheless, their efficacy depends upon sample size. In small samples, these tests may lack power and fail to detect true non-normality. In contrast, in large samples, they may be too sensitive, identifying even minor, practically insignificant deviations as statistically significant [9].

Graphical methods, such as histograms, box plots, P-P plots, and Q-Q plots, furnish an intuitive visualization of data distribution and can augment numerical tests, particularly in instances where those tests yield borderline or ambiguous outcomes [2, 22]. However, interpreting these visualizations correctly require a basic understanding of statistics. Researchers lacking experience may overlook important patterns or misinterpret normal variation as non-normality [32]. Consequently, it is advisable that graphical methods be employed in conjunction with statistical tests to strengthen the accuracy of normality assessment [13].

Among the descriptive techniques, skewness and kurtosis furnish quantitative insights into the distribution's shape. Skewness measures the asymmetry of data relative to the mean, where a value of zero signifies perfect symmetry [33]. A positive skew denotes a longer tail on the right side, whereas a negative skew signifies a longer tail on the left. Conversely, kurtosis assesses the peakedness or tailedness of the distribution. A normal distribution exhibits a kurtosis of 3 (mesokurtic), while values exceeding 3 indicate heavy tails (leptokurtic), and values below 3 suggest light tails (platykurtic) [34].

Therefore, it is important to note that in large samples, formal normality tests such as the Anderson-Darling, Shapiro-Wilk and Kolmogorov-Smirnov tests are highly sensitive and may identify statistically significant deviations from normality even when these deviations are negligible in practical terms. As sample size increases, minor departures from a perfect normal distribution can result in very small p-values, potentially leading to an over-rejection of the null hypothesis [14]. Therefore, in large-scale datasets like LASI Wave-1, the interpretation of normality should not rely solely on p-values from statistical tests but should be supported by graphical methods (histograms, Q-Q plots and P-P plots) and descriptive indicators such as skewness and kurtosis. This integrated approach allows researchers to distinguish between statistically detectable but practically unimportant deviations and meaningful violations of the normality assumption [4].

In this study, which employs LASI Wave-1 data, both numerical and graphical techniques were utilized to evaluate the normality of key continuous variables. Given the substantial sample size of LASI, reliance exclusively on statistical tests would likely produce significant results for even minimal deviations from normality. Therefore, additional graphical analyses such as Q-Q plots and histograms provided valuable context. These methods or analyses helped determine whether the deviations were meaningful enough to require non-parametric tests, or if parametric tests could still be used.

From an applied standpoint, the findings of this study provide actionable guidance for researchers analysing large-scale secondary health datasets. The results highlight the importance of supplementing formal normality tests with visual inspection and descriptive summaries to better understand data behaviour in practice. Such an approach facilitates appropriate analytical planning, supports consistency in

reporting and helps prevent overinterpretation of statistically significant results that may lack substantive relevance [4,7]. In the context of population-based health research, where data heterogeneity and large sample sizes are common, this pragmatic strategy can aid researchers in producing more reliable, interpretable and contextually meaningful empirical evidence [8]. In a nutshell, this study highlights the importance of using multiple methods when testing for normality in large-scale secondary datasets like LASI. It also reflects the real-world challenges where researchers face in balancing statistical rigor with data limitations. A careful and informed approach to normality testing improves the validity and interpretation of findings in large-scale, population-based health research.

Limitations

As no study is complete without its limitations, hence, this study has certain limitations that should be acknowledged. First, the analysis is based on secondary data from LASI Wave-1, which restricts control over data collection procedures, measurement precision and missing data mechanisms. Second, the normality assessment was limited to selected continuous variables (age, height, weight and BMI), and the findings may not be directly generalizable to other types of variables or distributions commonly encountered in health research. Third, given the large sample size, formal normality tests may exhibit heightened sensitivity, potentially identifying statistically significant deviations that are of limited practical relevance. Despite these limitations, the study provides useful methodological insights by demonstrating the value of integrating numerical and graphical approaches when assessing normality in large-scale population-based datasets.

Conclusion

Evaluating the normality of continuous data constitutes a critical procedure in the selection of suitable statistical methodologies, especially when using parametric tests that presuppose a normal distribution. A proper assessment of normality not only ensure the accuracy of descriptive statistics, including the mean and standard deviation, but also informs the decision-making process regarding the utilization of parametric versus non-parametric tests in hypothesis testing. Although, many statistical and graphical methods exist for the assessment of normality, it is generally advised that researchers adopt a combination of approaches, incorporating visual methods (such as histograms, P-P plots, and Q-Q plots) in conjunction with formal statistical tests. Among statistical tests, the Shapiro–Wilk test is best suited for small sample sizes ($n < 50$) as a result of its enhanced power in identifying deviations from the normal distribution. For larger datasets ($n \geq 50$), additional methods such as Kolmogorov–Smirnov test and Anderson–Darling test, as well as testing of skewness and kurtosis, alongside visual representations, can be effectively utilized. It is crucial to acknowledge that relying on the Kolmogorov–Smirnov test is not recommended due to its low statistical power and relative ineffectiveness in assessing normality. Therefore, the assessment of normality should be considered as an integral aspect of data analysis in both primary and secondary data, to ensure the use of appropriate statistical techniques and increase the reliability of study findings. Researchers are encouraged to carefully examine the characteristics of their data prior to determining their statistical strategies. In short, from a methodological perspective, this study reinforces that normality assessment should be viewed as a diagnostic tool rather than a strict decision rule, particularly when working with large-scale secondary datasets. Formal statistical tests, while informative, may be overly sensitive in large samples and should therefore be interpreted alongside graphical methods and descriptive indicators such as skewness and kurtosis. An integrated approach to normality evaluation supports more informed choices between parametric and non-parametric methods, enhances the transparency of analytical decisions and improves the robustness and interpretability of statistical findings in population-based health research.

Abbreviations

AD test – Anderson–Darling test
ANOVA – Analysis of Variance
BMI – Body Mass Index
CDF – Cumulative Distribution Function

CEB – Census Enumeration Block
CV – Coefficient of Variation
H_0 – Null Hypothesis
H_1 – Alternative Hypothesis
IIPS – International Institute for Population Sciences
IQR – Interquartile Range
KS test – Kolmogorov–Smirnov test
LASI – Longitudinal Ageing Study in India
P-P Plot – Probability–Probability Plot
PPS – Probability Proportional to Size
PSU – Primary Sampling Unit
Q-Q Plot – Quantile–Quantile Plot
SD – Standard Deviation
SE – Standard Error

Acknowledgments

The authors would like to thank the anonymous reviewers for their suggestions and remarks that contributed to improving this research article.

Author contributions

KY and DTB conceptualized the study. KY and DTB did the analysis and wrote the main manuscript text. KY and DTB prepared all the tables and figures. DTB guided and supervised in the entire process. Both authors reviewed the final version of the manuscript. The author(s) read and approved the final manuscript.

Funding

No funding was received for this study by any organization or individual.

Data availability

The study utilises a secondary source of data that is freely available in the public domain through https://g2aging.org/

Data declarations

Ethical approval and consent to participate

The study is based on the freely available data source, and survey agencies conducted the field survey for the data collection have also collected a prior consent from the respondent. The Indian Council of Medical Research (ICMR) extended the necessary guidance and ethical approval for conducting the LASI, ruled that no formal ethical approval was required to carry out research from this data source.

Consent for publication

Not applicable.

Competing interests

All authors declare no potential competing interests.

Author details

[1]Corresponding author, Research Scholar, Department of Community Medicine, Sikkim Manipal Institute of Medical Sciences (SMIMS), Sikkim Manipal University (SMU), Tadong, Gangtok, Sikkim, 737102, India.
[2]Professor and Head, Department of Community Medicine, Sikkim Manipal Institute of Medical Sciences (SMIMS), Sikkim Manipal University (SMU), Tadong, Gangtok, Sikkim, 737102, India.

References

1. Ghasemi, A. and Zahediasl, S., Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 2012. (1726-913X (Print)).
2. Field, A., *Discovering statistics using IBM SPSS statistics*. 2024: Sage publications limited.
3. Norman, G., Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 2010. 15: p. 625–632.
4. Johnston, M.P., Secondary data analysis: A method of which the time has come. *Qualitative and quantitative methods in libraries*, 2014. 3(3): p. 619–626.
5. Smith, E., Weinberger, M., Katz, B., and Moore, P., Predictors of quality of secondary data in health research. *Health Services Research*, 2011. 46(3): p. 868–887.
6. Blanca Mena, M.J., et al., Non-normal data: Is ANOVA still a valid option? *Psicothema*, 2017. 29(4): p. 552–557.
7. Osborne, J., Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, 2010. 15(1).
8. Razali, N.M. and Wah, Y.B., Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2011. 2(1): p. 21–33.
9. Yap, B.W. and Sim, C.H., Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 2011. 81(12): p. 2141–2155.
10. Dominici, F. and Cefalu, M., Confounding adjustment and exposure prediction in environmental epidemiology: Additional insights. *Epidemiology*, 2015. 26(2): p. e28.
11. Kim, H.-Y., Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics*, 2013. 38(1): p. 52.
12. Mishra, P., et al., Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, 2019. (0974-5181 (Electronic)).
13. Thode, H.C., *Testing for normality*. 2002: CRC press.
14. Campbell MJ, Machin D, Walters SJ. Medical Statistics: A text book for the health sciences, 4th ed. Chichester: John Wiley & Sons, Ltd.; 2007..
15. Tsagris, M. and Pandis, N., Normality test: Is it really necessary? *American journal of orthodontics and dentofacial orthopedics*, 2021. 159(4): p. 548–549.
16. International Institute for Population Sciences (IIPS). Longitudinal Ageing Study in India (LASI). 2020.
17. International Institute for Population Sciences (IIPS); National Programme for Health Care of Elderly (NPHCE), Ministry of Health & Family Welfare (MoHFW); Harvard T.H. Chan School of Public Health; University of Southern California. *Longitudinal Ageing Study in India (LASI) Wave 1, 2017-18: India Report*. Mumbai: 2020.
18. Sundaram, K.R., Dwivedi, S.N., and Sreenivas, V. *Medical Statistics: Principles and Methods*. 2nd ed. New Delhi: Wolters Kluwer India; 2014.
19. Altman, D.G. and Bland, J.M., Statistics notes: the normal distribution. *BMJ*, 1995. 310(6975): p. 298.
20. Altman, D.G., *Practical statistics for medical research*. 1990: Chapman and Hall/CRC.
21. Indrayan, A. and Sarmukaddam, S.B. *Medical Bio-Statistics*. New York: Marcel Dekker Inc.; 2000.
22. Wilk, M.B. and Gnanadesikan, R., Probability plotting methods for the analysis of data. *Biometrika*, 1968. 55(1): p. 1–17.
23. Stephens, M.A., EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 1974. 69(347): p. 730–737.
24. D'Agostino, R.B., *Goodness-of-fit-techniques*. Vol. 68. 1986: CRC press.
25. Campbell, M.J., Machin, D., and Walters, S.J., *Medical statistics: a textbook for the health sciences*. 2010: John Wiley & Sons.
26. Indrayan, A. and Satyanarayana, L., Essentials of biostatistics. *Indian Pediatrics*, 1999. 36: p. 1127–34.
27. Jeyaseelan, L., *Short Training Course Materials on Fundamentals of Biostatistics, Principles of Epidemiology and SPSS*. CMC Vellore: Biostatistics Resource and Training Center (BRTC), 2007.
28. Armitage, P., Berry, G., and Matthews, J.N.S., *Statistical methods in medical research*. 2013: John Wiley & Sons.
29. Barton, B. and Peat, J., *Medical statistics: A guide to SPSS, data analysis and critical appraisal*. 2014: John Wiley & Sons.
30. Anderson, T.W. and Darling, D.A., Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, 1952: p. 193–212.

31. Shapiro, S.S. and Wilk, M.B., An analysis of variance test for normality (complete samples). *Biometrika*, 1965. 52(3–4): p. 591–611.

32. Delwiche, L.D. and Slaughter, S.J., *The little SAS book: a primer*. 2019: SAS institute.

33. Doane, D.P. and Seward, L.E., Measuring skewness: a forgotten statistic? *Journal of statistics education*, 2011. 19(2).

34. DeCarlo, L.T., On the meaning and use of kurtosis. *Psychological methods*, 1997. 2(3): p. 292.

Kanchan Yadav,
Research Scholar, Department of Community Medicine,
Sikkim Manipal Institute of Medical Sciences (SMIMS),
Sikkim Manipal University (SMU), Tadong, Gangtok, Sikkim, 737102, India.
E-mail address: yadavkanchan73@gmail.com

and

Dechenla Tshering Bhutia,
Professor and Head, Department of Community Medicine,
Sikkim Manipal Institute of Medical Sciences (SMIMS),
Sikkim Manipal University (SMU), Tadong, Gangtok, Sikkim, 737102, India.
E-mail address: dtsering16@gmail.com