

Evaluating Multivariate Normality: A Graphical Approach

¹Shahla Ramzan, ¹Faisal Maqbool Zahid and ²Shumila Ramzan

¹Department of Statistics, Government College University, Faisalabad, Pakistan

²University of Agriculture Faisalabad, Pakistan

Abstract: The statistical graphics play an important role in providing the insights about data in the process of data analysis. The main objective of this paper is to provide a comprehensive review of the methods for checking the normality assumption. Multivariate normality is one of the basic assumptions in multivariate data analysis. Univariate normality is essential for the data to be multivariate normal. This paper reviews graphical methods for evaluating univariate and multivariate normality. These methods are applied on a real life data set and the normality is investigated.

Key words: Bootstrapping . chi-squared plot . mahalanobis distance . normality . outlier . Q-Q plot . simulation

INTRODUCTION

The use of visual analysis of the data in research is strongly backed by the advancement in other associated fields as: graphical methodologies for graphical representation of complex data sets, psychology of graphical perception and advancement in computer technology alongwith development and dissemination of appropriate software [1].

In statistical modeling, it is often crucial to verify if the data at hand satisfy the underlying distributional assumptions. Many times such an examination may be needed for the residuals after fitting various models. For most multivariate analyses, it is very important that the data indeed follow the multivariate normal or if not exactly at least approximately. If the answer to such a query is affirmative, it can often reduce the burden of searching for procedures which are robust to departure from multivariate normality.

Normality of a data refers to the situation where the data are drawn from a population that has a normal distribution. This distribution is inarguably the most important and the most frequently used distribution in both the theory and application of Statistics. A random variable X is said to be distributed normally with mean μ and variance σ^2 if it assumes the probability density function

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

on the domain $x \in (-\infty, \infty)$. While statisticians and mathematicians uniformly use the term "normal distribution" for this distribution, physicists sometimes call it a Gaussian distribution. Because of its curved flaring shape, social scientists refer to it as the "bell curve." Feller [2] uses the symbol $\phi(x)$ for $P(x)$ in the above equation, but then switched to $\eta(x)$ in Feller [3]. De Moivre developed the normal distribution as an approximation to the binomial distribution and it was subsequently used by Laplace in 1783 to study measurement errors and by Gauss in 1809 in the analysis of astronomical data [4].

There are several things that can cause the data to appear non-normal, such as:

- The data come from two or more different sources. This type of data will often have a multi-modal distribution. This can be solved by identifying the reason for the multiple sets of data and analyzing the data separately.
- The data come from an unstable process. This type of data is nearly impossible to analyze because the results of the analysis will have no credibility due to changing nature of the process.

Correspondence Author: Shahla Ramzan, Department of Statistics, Government College University Faisalabad, Pakistan

- The data were generated by a stable, yet fundamentally non-normal mechanism. For example, particle counts are non-normal by the varying nature of the particle generation process. Data of this type can be handled using transformations.

Statistical methods are based on various assumptions that uphold the methods. One of them is normality, which is commonly assumed. Thus, statistical models often require checking the normality of variables. Otherwise, interpretations and inferences based on the models are not reliable. This paper illustrates some visual methods for testing the assumption of normality in the univariate and multivariate data.

The rest of this paper is organized as follows. A brief detail of the graphical methods for checking normality is presented in Section 2. In Section 3 all the methods are applied on real data and the results are discussed and compared with each other. Conclusions reached in Section 2 and Section 3, are presented in Section 4. Section 5 concludes the discussion with throwing some light on merits and demerits of the use of graphical techniques discussed in the text for assessing multivariate normality.

METHODOLOGY

The statistical procedures and parametric tests are based on certain assumptions. These procedures are valid only if the assumptions hold. If a parametric test is applied on a nonparametric data, the results are likely to be inaccurate. Most statistical procedures are based on the assumption of normality. The assumption of normality implies that the population from which data are drawn follows normal distribution. Here is a brief introduction of some graphical methods for assessing the assumption of normality.

Assessing univariate normality: The assumption of normality underlies many statistical techniques. By univariate normality (UVN) or simply normality means the data at hand is drawn from a normal distribution. This assumption plays a significant role in multivariate analysis, e.g., discriminant analysis, although the awareness of multivariate tests is limited.

The assumption of univariate normality can be investigated graphically in several ways like Histogram with a normal curve overlay, Box-Whisker plot, Stem & leaf plot, Matrix plot, Dot plot, Q-Q plot and Normal probability plot etc.

Histogram: Histogram is a very simple and important graph of the frequency distribution. It was introduced by Pearson [5]. The data is presented in the form of adjacent rectangles with height of rectangle proportional to the frequency. A normal curve is drawn over the histogram to examine if the data follows normal pattern.

Stem & leaf plot: Another technique used to present and visualize quantitative data is stem and leaf plot. The most attractive feature of this display is that the original data or information is not lost after the formation of the graph. The data values are divided into two portions; a stem and a leaf. Then the leaves for each stem are shown separately in a display.

Box-whisker plot: The box-whisker plot or box plot introduced by Chambers *et al.* [6] is another summarized picture of the data. This plot uses the quartiles and extreme values of the data as a summary measure. The five-number summary is used to prepare the box plot, that is, smallest value, lower quartile Q_1 , median Q_2 , upper quartile Q_3 and the largest value. The plot consists of a rectangle (the box) in the central part of the observed data and whiskers are drawn to the lowest and highest values from the rectangle. The limits of the box are lower and upper quartiles and the middle line is the median.

Dot plot: The main purpose of this plot is the detection of any outliers or extreme values in the data. The observations are plotted simply on a real line. If there is any value that is far away from the rest of data, it appears on the graph significantly far away from other data values.

Normal probability plot: Probability plots are most commonly used for examining whether the data follows a specific distribution or not. The normal distribution is the most desirable property of certain statistical procedures. So

the most widely used probability plot is normal probability plot Chambers [6]. The data is plotted against the corresponding expected values from the normal distribution. The resulting plot should look like a straight line at 45°, if the data is drawn from a normal distribution. If the graph deviates from a straight line, it indicates departure from normality. An interesting feature of normal probability plot is that it screens out outliers or extreme values in the data.

Quantile-quantile (Q-Q) plot: This plot is used for the same purpose as the probability plot. The quantiles of the data are plotted against the expected values of desired distribution. This plot should look like a straight line. A quantile plot is a visual display that provides a lot of information about a univariate distribution (Chambers *et al.*, [6], Gnanadesikan [7]). The quantiles of a distribution are a set of summary statistics that locate at relative positions within the complete ordered array of data values. Specifically, the p^{th} quantile of a distribution, X , is defined as the value x_p such that approximately $p\%$ of the empirical observations have values lower than x_p .

Assessing multivariate normality: To assess multivariate normality, several visual procedures have been suggested in literature. These procedures use the properties of multivariate normal distribution for their application. Thode [8] has categorized these multivariate plotting procedures as Scatter plots of the component data, Probability plots of the marginal data and Probability plots of reduced data. As a first approach to assessing multivariate normality, univariate probability plots are used to independently assess each of the marginal variables. Healy [9] also proposed using scatter plots of all variables taken two at a time; although this is a more effective way of identifying outliers, it also allows identification of other nonlinear relationships between variables. Another approach includes ordering the marginal observations independently and plotting the ordered observations against each other taking the variates two at a time. Under the hypothesis of normality, these plots are equivalent to normal probability plots and should follow a linear pattern.

Chi-square plot of squared Mahalanobis distance: A widely used graphical procedure is based on the distribution of the ordered squared Mahalanobis distances of the individual sample points from their mean. A plot of the ordered squared distances d_i^2 and $100\left(\frac{i-0.5}{n}\right)$ quantiles of the chi-square distribution with p degrees of freedom is called a chi-square plot. Where $d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$, $i = 1, \dots, n$ and X_1, \dots, X_n are the sample observations each measured on p variables. When the population is multivariate normal and both n and $n-p$ are greater than 30, each of the squared distances should behave like a chi-square random variable. The following procedure illustrates the method to construct a chi-square plot.

- Order the squared distances d_i^2 from smallest to largest as $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$.
- Calculate the quantiles $q_{c,p}\left(\frac{i-0.5}{n}\right)$ related to the upper percentiles of a chi-square distribution.
- Plot the pairs $\left(q_{c,p}\left(\frac{i-0.5}{n}\right), d_{(i)}^2\right)$ to obtain a chi-square plot.

For p variables and a large sample size, the squared mahalanobis distances of the observations to the mean vector are distributed as chi-square with p degrees of freedom. However, the sample size must be quite large to have a chi-square distribution other than the situation where p is very small. This plot should resemble a straight line through the origin. A systematic curved pattern suggests lack of normality. This plot is sensitive to the presence of outliers and should be cautiously used as a rough indicator of multivariate normality.

Beta probability plot of squared mahalanobis distances: Another graphical approach to test multivariate normality is a QQ plot of the ordered squared Mahalanobis distances d_i^2 statistics against the quantiles of the beta distribution with parameters $p/2$ and $(n-p-1)/2$ as suggested by Gnanadesikan and Kettenring [10]. If the data have been sampled from p -variate multivariate normal population, then $\frac{n}{(n-1)^2} d_i^2 \sim \text{Beta}\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$. Small [11]

suggested to use the probability plots of the d_i^2 with beta order statistics using general plotting position $\frac{i-\alpha}{n-\alpha-\beta+1}$ for $\alpha = (p-2)/2p$ and $\beta = 0.5-(n-p-1)^{-1}$.

Detecting multivariate normality using characteristic function: Holgerson [12] suggested a different criterion than the previously discussed methods for detecting multivariate normality. The suggested method can be described as follows.

Let X_1, \dots, X_n be n i.i.d. random variables in \mathcal{R}^p such that $E(X_j) = \mu$ and $\text{cov}(X_j) = \Sigma$. If

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j \text{ and } S = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})^T$$

be the sample version of μ and Σ respectively, we define the characteristic function of X as $\phi_X(T) = E(e^{iT^T X})$. The normal distribution then can be characterized by $\phi_{\bar{X}, S}(T, U) = \phi_{\bar{X}}(T) \phi_S(U) \Leftrightarrow \phi_X(L) = e^{iL^T \bar{X} - L^T S L / 2}$, where T, U and L are fixed vectors in \mathcal{R}^p with finite and non-null elements. The characteristic function given in (1) relates to the normal distribution if and only if $L^T \bar{X}$ and $L^T S L$ are independent.

To detect normality for a multivariate data with this approach, B independent bootstrap samples X_1^*, \dots, X_B^* of size n with replacement are drawn from the original sample of size n which is denoted by $X = \{X_1, \dots, X_n\}$. For a general discussion of the nonparametric bootstrap see Efron and Tibshirani [13]. Each of B pairs of statistics $\{L^T \bar{X}_b^*, L^T S_b^* L\}$; $b=1, \dots, B$, are plotted in two-dimensional space. If graph displays a correlation pattern, the data will violate the assumption of normality of X . There are several possible choices for the constant vector L . To detect the normality for the full multivariate data set, all elements of L need to be non-zero. To exclude some variable in the normality detection process, the corresponding element in L is set to zero.

DATA AND APPLICATION

The data used by Johnson [14] is being used in this section for the application of all methods discussed in Section 2 and comparing the conclusions reached with these different methods. The data is of a firm who is attempting to evaluate the quality of its sales staff by selecting a random sample of 50 employees. Each individual is evaluated on two measures of performance: sales growth and sales profitability. Sales performance by taking a series of tests for each selected employee. Each employee receives four exams designed to measure their creativity, mechanical reasoning, abstract reasoning and mathematical ability.

To check if the sales person data follows multivariate normal distribution, the first step is to check the data for univariate normality. Different graphical approaches discussed in Section 2 are applied to this data. Figure 1

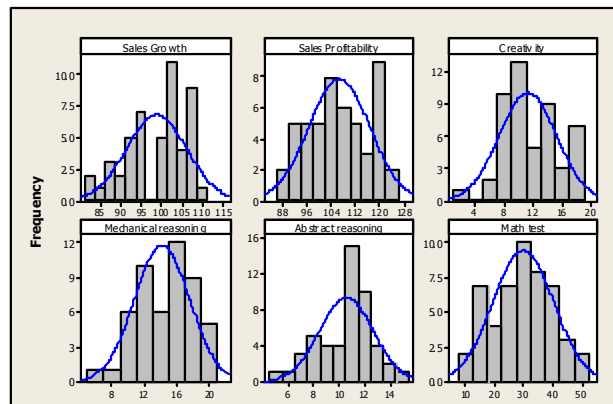


Fig. 1: Histogram

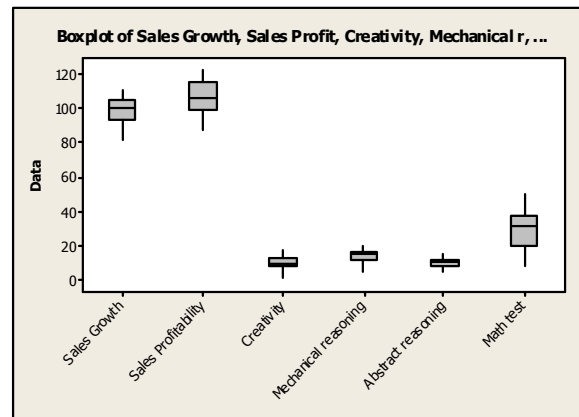


Fig. 2: Box plot

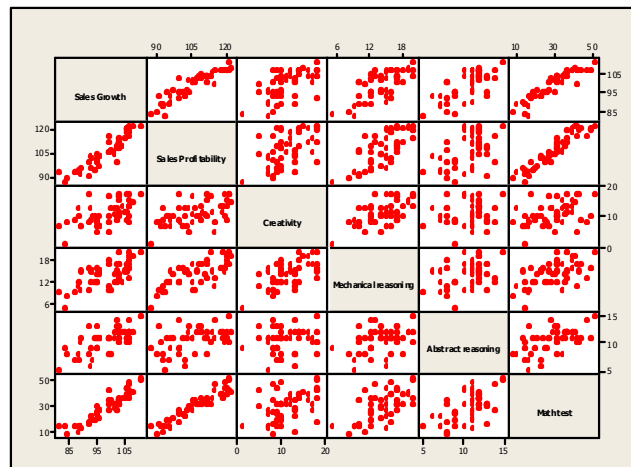


Fig. 3: Matrix plot

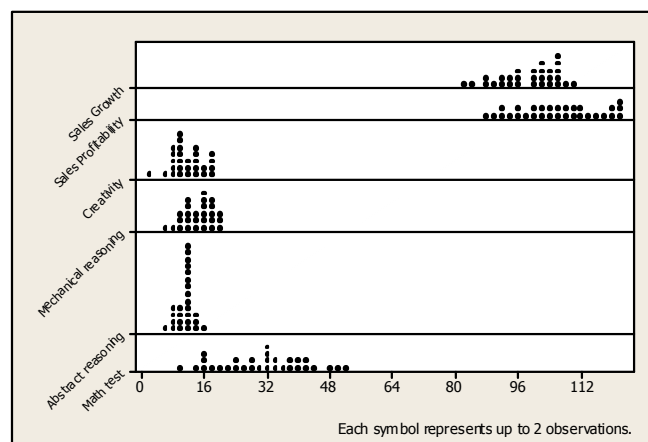


Fig. 4: Dot plot

shows the histogram of all the six variables with a normal curve overlay. None of the variables show a clear symmetric pattern. Therefore the box-whisker plots for all the variables are constructed and are shown in Fig. 2. The box plot for sales profitability and math test seem to be symmetrical. While creativity, mechanical reasoning and abstract reasoning have smaller variation as compared to others. The similar conclusion can be made from the stem and leaf plot.

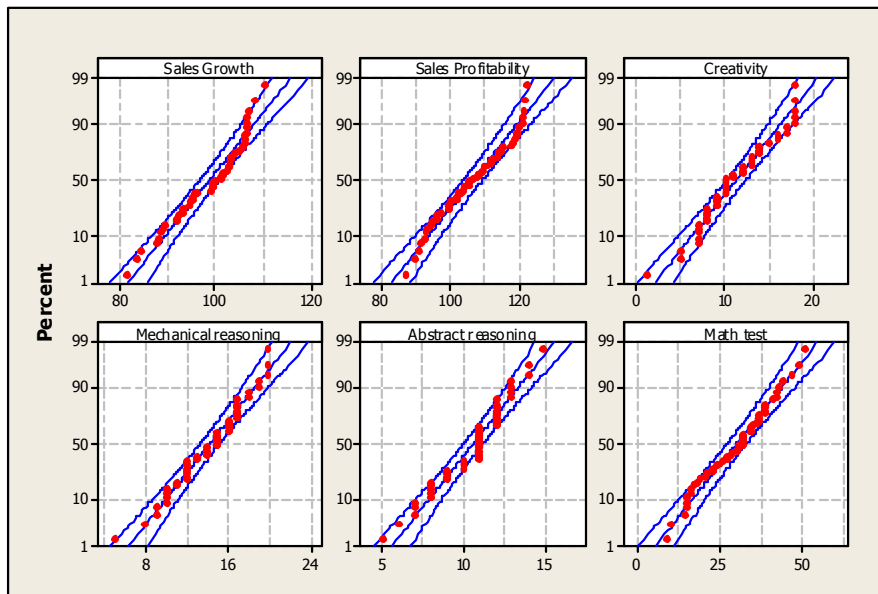


Fig. 5: Normal probability plots

The matrix plot is shown in Fig. 3. The variables are plotted in the form of pairs with each other on a scatter plot. If the underlying distribution is normal then this plot must show ellipses. But the graph shows that the sales growth and sales profitability have a linear relationship. Therefore this data cannot be regarded as drawn from the multivariate normal distribution. Another use of this graph is the detection of outliers. No outliers are present in the data as shown by Fig. 4 since there are no points far away from rest of the data.

For detection of outliers in multivariate data, another graphical display is dot plot. The dot plot for our data is shown in Fig. 4. This graph shows that no outliers are likely to present in the data. This is further confirmed by making the normal probability plots for all the variables shown in Fig. 5.

The QQ-plot of the data are shown in Fig. 6. Although a considerable amount of the data in the QQ-plots for Mathematical ability and Sales Profitability appears to fall on a straight line, it is obvious that taken as a whole, the data does not appear to be normally distributed. Therefore we must assess the hypothesis of normality by calculating the straightness of these QQ-plots using the correlation coefficient for each plot.

The straightness of the QQ-plots can be measured by calculating the correlation coefficient of the points in the plot. The correlation coefficient for QQ-plot is defined by

$$r_{Q,p} = \frac{\sum_{i=1}^n (x_{(i)} - \bar{x})(q_{(i)} - \bar{q})}{\sqrt{\sum_{i=1}^n (x_{(i)} - \bar{x})^2} \sqrt{\sum_{i=1}^n (q_{(i)} - \bar{q})^2}} = \frac{\sum_{i=1}^n (x_{(i)} - \bar{x})(q_{(i)})}{\sqrt{\sum_{i=1}^n (x_{(i)} - \bar{x})^2} \sqrt{\sum_{i=1}^n (q_{(i)})^2}} \text{ since } \bar{q} = 0$$

where $x_{(i)}$ are ordered observations and $q_{(i)}$ are the quantiles of the standard normal distribution. To calculate the values of $r_{Q,p}$, the standard normal quantiles are given in Table 2. The correlation coefficient for the first variable that is sales growth is calculated as

$$r_{Q,1} = \frac{\sum_{i=1}^n (x_{(i)} - \bar{x})(q_{(i)})}{\sqrt{\sum_{i=1}^n (x_{(i)} - \bar{x})^2} \sqrt{\sum_{i=1}^n (q_{(i)})^2}} = \frac{8.1282}{\sqrt{2638} \sqrt{48.7684}} = 0.0227$$

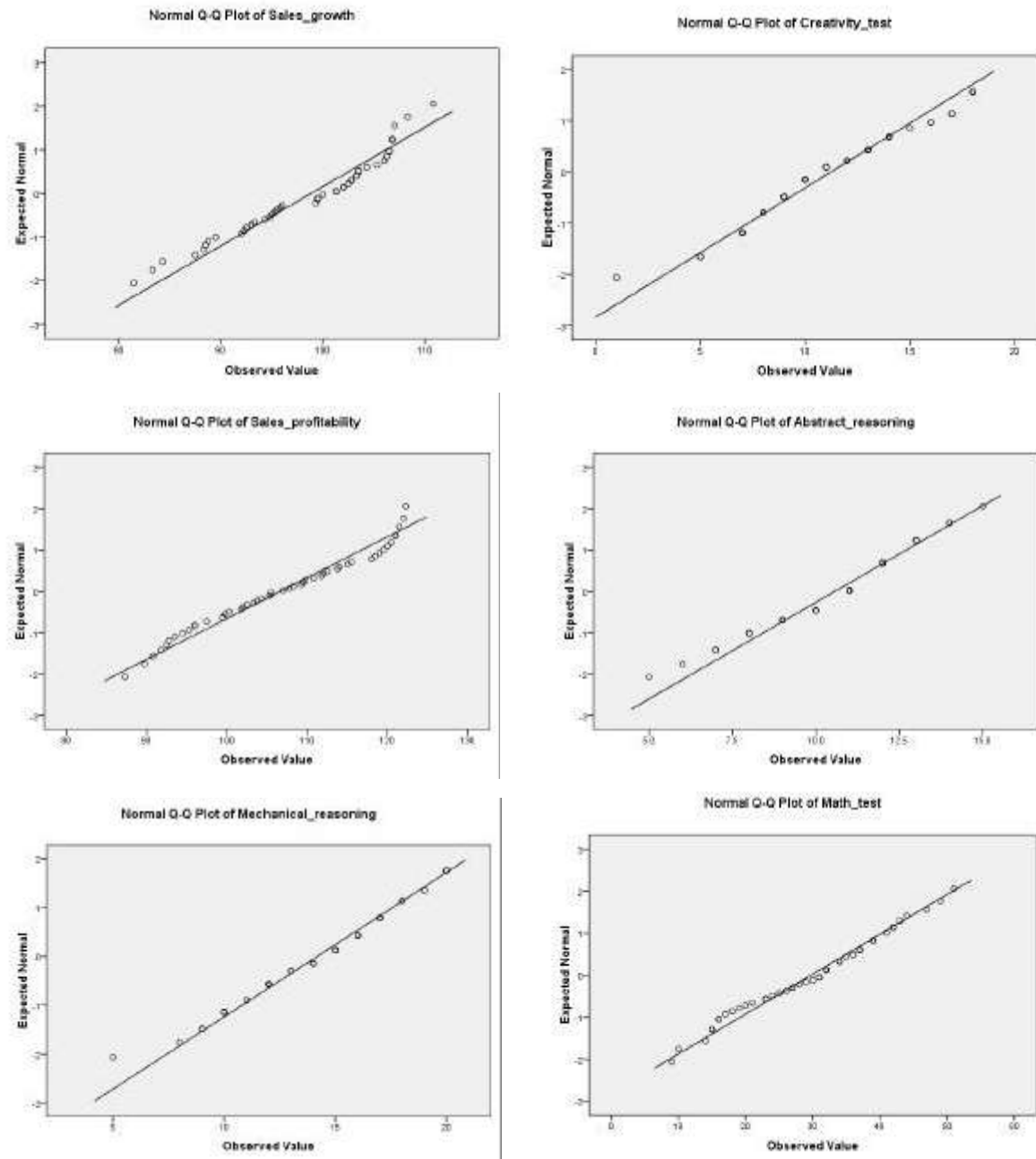


Fig. 6: Quantile quantile plot

Similarly, the values of $r_{Q,p}$ for all six variables are as under:

p	1	2	3	4	5	6
$r_{Q,p}$	0.0227	0.0998	0.1054	0.0417	0.0372	0.0239

We examine the normality of the data by referring to the table of the critical points of QQ-plot correlation coefficient for Normality. At the 10% level of significance, $r_{tab} = 0.9809$, corresponding to $n = 50$, $\alpha = 0.10$. Since $r_{Q,p} < 0.9809$, we reject the hypothesis of normality. Similarly we reject the hypothesis of normality at 5% and 1% levels of significance since $r_{Q,p} < 0.9768$ and $r_{Q,p} < 0.9671$ respectively.

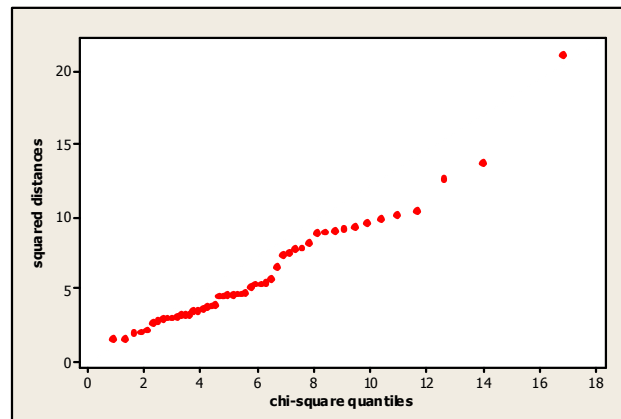


Fig. 7: Chi-square probability plot

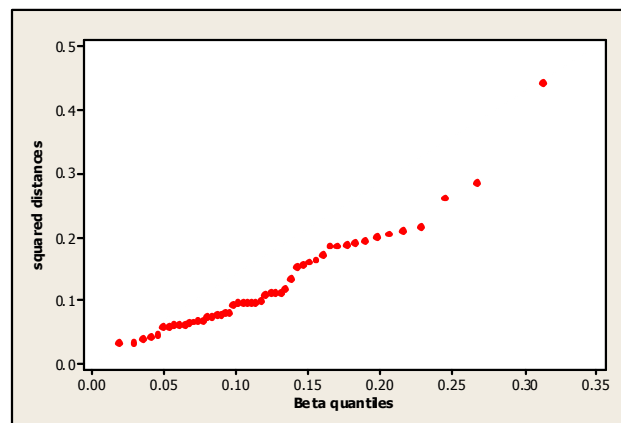


Fig. 8: Beta probability plot

We now test the normality of this data using multivariate techniques. First step in checking whether the data follows a multivariate normal distribution is that all the variables should be normally distributed. But converse is not generally true, that is, univariate normality does not necessarily imply multivariate normality. As the variables are not distributed normally at individual levels, multivariate normality cannot be assumed for this data. To confirm whether the data is chosen from a multivariate normal distribution, the chi-square and beta probability plots are constructed.

The mahalanobis distances for the data are calculated and are shown in Table 1. The corresponding chi-square percentiles, that is, $(i-0.5)/50$ are required to construct the chi-square plot. These percentiles are also reported in Table 1. The chi-square percentiles are plotted against the ordered mahalanobis distances to obtain the required chi-square plot. The resulting plot is shown in Fig. 7. This plot shows that the points do not follow a straight line pattern. Therefore the data do not follow a multivariate normal distribution. Another useful information which can be obtained from this plot is the presence of a multivariate outlier. The Beta probability plot is constructed for salesperson data and plotted against $\beta(3, 21.5)$ with $\alpha = 0.33$ and $\beta = 0.48$ shown in Fig. 8. Clearly, all the points do not fall on a straight line. Thus the data do not support the assumption of multivariate normality. However, the point that lies far above the line indicates a large distance, or an outlying observation and may require further attention.

To detect normality for the data under consideration using the characteristic function approach suggested by Holgerson [12], we have drawn $B=5000$ independent bootstrap samples X_1^*, \dots, X_{5000}^* of size $n=50$ with replacement from the original sample. We plotted each of $B=5000$ pairs of statistics $\{L^T \bar{X}_b^*, L^T S_b^{*2} L\}$; $b=1, \dots, B$, in two-dimensional space. All elements of the constant vector $L = (L, L, \dots, L)$ are set as $L = 1/p$. The graph of the said paired statistics is shown in Fig. 9(a), which shows a correlation pattern suggesting non-normality of the data. The same result is also achieved with the quantile-quantile (Q-Q) plot, when ordered Mahalanobis distances

Table 1: Ordered squared distances and Chi-square quantiles and Beta quantiles for the salesperson data

i	$d_{(i)}^2$	$q_{c,6}\left(\frac{i-1/2}{50}\right)$	Beta quantiles	i	$d_{(i)}^2$	$q_{c,6}\left(\frac{i-1/2}{50}\right)$	Beta quantiles
1	1.4732	0.8721	0.0192	26	4.6419	5.4296	0.1138
2	1.5500	1.3296	0.0291	27	4.7384	5.5954	0.1170
3	1.9307	1.6354	0.0357	28	5.1658	5.7652	0.1204
4	1.9821	1.8846	0.0410	29	5.3078	5.9395	0.1238
5	2.1328	2.1029	0.0457	30	5.3103	6.1189	0.1273
6	2.6985	2.3014	0.0499	31	5.3840	6.3041	0.1309
7	2.7677	2.4863	0.0538	32	5.6100	6.4958	0.1346
8	2.8756	2.6613	0.0575	33	6.4176	6.6948	0.1384
9	2.9901	2.8289	0.0610	34	7.3386	6.9021	0.1424
10	3.0133	2.9908	0.0644	35	7.4561	7.1188	0.1465
11	3.0547	3.1484	0.0676	36	7.6792	7.3464	0.1508
12	3.165	3.3028	0.0708	37	7.7913	7.5864	0.1553
13	3.2106	3.4546	0.0740	38	8.1381	7.8408	0.1601
14	3.2264	3.6046	0.0771	39	8.8763	8.1122	0.1652
15	3.4826	3.7535	0.0801	40	8.9253	8.4036	0.1706
16	3.5295	3.9015	0.0831	41	9.0035	8.7191	0.1764
17	3.6013	4.0493	0.0861	42	9.1458	9.0642	0.1827
18	3.6983	4.1973	0.0891	43	9.3033	9.4461	0.1896
19	3.8106	4.3457	0.0921	44	9.5216	9.8754	0.1973
20	3.8539	4.4951	0.0952	45	9.8024	10.3676	0.2061
21	4.4629	4.6456	0.0982	46	10.0361	10.9479	0.2163
22	4.5232	4.7978	0.1012	47	10.3209	11.6599	0.2286
23	4.5577	4.9519	0.1043	48	12.4992	12.5916	0.2444
24	4.568	5.1083	0.1074	49	13.6269	13.9676	0.2673
25	4.6301	5.2674	0.1106	50	21.1708	16.8119	0.3123

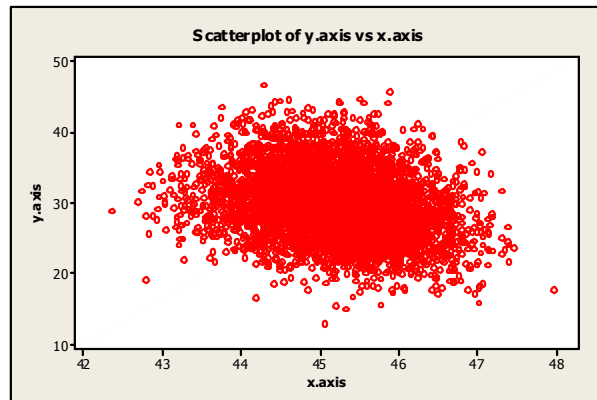


Fig. 9(a): Correlation plot

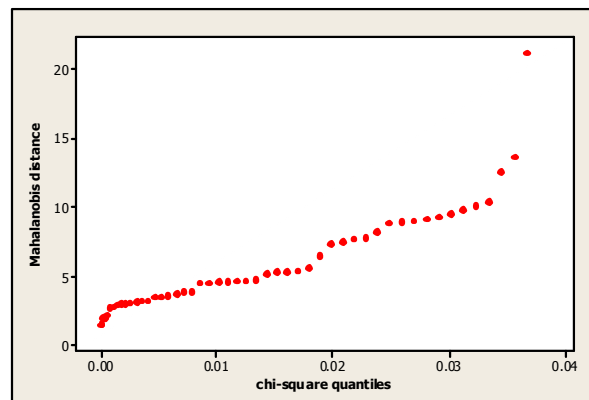


Fig. 9(b): Multivariate QQ plot

$d_j^2 = (\mathbf{X}_j - \bar{\mathbf{X}})^T \mathbf{S}^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})$ for $j = 1, \dots, n$ are plotted against the chi-square distribution quantiles $Q_p\left(\frac{j}{n+1}\right)$. The graph of $\left\{Q_p\left(\frac{j}{n+1}\right), d_{(j)}^2\right\}_{j=1}^n$ shown in Fig. 9(b) is not a fairly straight line indicating non-normality of the data.

CONCLUSION

The normality of the data, which is a key assumption for making valid inferences, can be tested using various statistical tests or visual inspection. According to Chambers *et al.* [6], there is no single statistical tool to assess the normality that is as powerful as a well-chosen graph. Assessing the normality using graphical methods do lack objectivity (the analysts make use of their experience in visualizing the graph to make a subjective judgement about the data) which is not the case when dealing with statistical tests. However, the assessment of normality using statistical tests is sensitive to the sample size. In case of small samples, the null hypothesis of normality is often not rejected and conversely for large samples where the inferences are relatively robust to the large samples, hypothesis of normality is rejected even for small violations. So, the graphical methods should be used to analyze the violation of normality in the light of sample size. In sum, combining graphical methods and test statistics will definitely improve our judgement on the normality of the data.

REFERENCES

1. Jacoby, W.G., 1997. Statistical Graphics for Univariate and Bivariate Data Sage University Papers Series. No. 07-117
2. Feller, W., 1968. An Introduction to Probability Theory and Its Applications, 3rd Ed. New York: Wiley, Vol: 1.
3. Feller, W., 1971. An Introduction to Probability Theory and Its Applications, 3rd Ed. New York: Wiley, Vol. 2.
4. Havil, J., 2003. Gamma: Exploring Euler's Constant. Princeton, NJ: Princeton University Press.
5. Pearson, K., 1895. Contributions to the Mathematical Theory of Evolution-II. Skew Variation in Homogeneous Material. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 186: 343-326.
6. Chambers, J.M., W.S. Cleveland, B. Kleiner and P.A. Tukey, 1983. Graphical Methods for Data Analysis. Pacific Grove, CA: Wadsworth and Brooks/Cole.
7. Gnanesikan, R., 1977. Methods for Statistical Analysis of Multivariate Data. New York: John Wiley.
8. Thode, H.C., 2002. Testing for Normality: Statistics, a Series of Textbooks and Monographs. Stony Brook, New York.
9. Healy, M.J.R., 1968. Multivariate normal plotting. Applied Statistics, 17: 157-161.
10. Gnanadesikan, R. and J.R. Kettenring, 1972. Robust estimates, residuals and outlier detection with multiresponse data. Biometrics, 28: 81-124.
11. Small, N.J.H., 1985. Testing for Multivariate normality. In Kotz, S. and N.L. Johnson (Eds.). Encyclopedia of Statistical Sciences, John Wiley and Sons, New York, Vol: 6.
12. Holgersson, H., 2006. A graphical method for assessing multivariate normality. Computational Statistics, 21 (1): 141-149.
13. Efron, B. and R.J. Tibshirani, 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC.
14. Johnson, R.A. and D.W. Wichern, 2007. Applied Multivariate Statistical Analysis. 6th Edn. Prentice Hall.