Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.6, No.2, 2016
www.iiste.org

Assessing Univariate and Multivariate Normality, A Guide For Non-Statisticians

Felix Boakye Oppong[1*] Senyo Yao Agbedra [2]

1. Hasselt University, Agoralaan Gebouw D, BE-3590 Diepenbeek, Belgium
2. Brong Ahafo Regional Hospital, P.O.Box 2090, Sunyani, Ghana.

Abstract

Most parametric methods rely on the assumption of normality. Results obtained from these methods are more powerful compared to their non-parametric counterparts. However for valid inference, the assumptions underlying the use of these methods should be satisfied. Many published statistical articles that make use of the assumption of normality fail to guarantee it. Hence, quite a number of published statistical results are presented with errors. As a way to reduce this, various approaches used in assessing the assumption of normality are presented and illustrated in this paper. In assessing both univariate and multivariate normality, several methods have been proposed. In the univariate setting, the Q-Q plot, *histogram, box plot, stem-and-leaf plot or dot plot* are some graphical methods that can be used. Also, the properties of the normal distribution provide an alternative approach to assess normality. The Kolmogorov-Smirnov (K-S) test, Lilliefors corrected K-S test, Shapiro-Wilk test, Anderson-Darling test, Cramer-von Mises test, D'Agostino skewness test, Anscombe-Glynn kurtosis test, D'Agostino-Pearson omnibus test, and the Jarque-Bera test are also used to test for normality. However, Kolmogorov-Smirnov (K-S) test, Shapiro-Wilk test, Anderson-Darling test, and Cramer-von Mises test are widely used in practice and implemented in many statistical applications.

For multivariate normal data, marginal distribution and linear combinations should also be normal. This provides a starting point for assessing normality in the multivariate setting. A scatter plot for each pair of variables together with a Gamma plot (Chi-squared Q-Q plot) is used in assessing bivariate normality. For more than two variables, a Gamma plot can still be used to check the assumption of multivariate normality. Among the many test proposed for testing multivariate normality, Royston's and Mardia's tests are used more often and are implemented in many statistical packages.

When the normality assumption is not justifiable, techniques for non-normal data can be used. Likewise, transformation to near normality is another alternative.

Keywords: Univariate normal, Multivariate normal, Q-Q plot, Gamma plot, Kolmogorov-Smirnov test, Shapiro-Wilk test, Mardia's test, Royston's test.

1. Introduction

In the statistics literature, the normality assumption is tagged as an omnipresent assumption. This is due to its wide use in both univariate and multivariate analysis (Rosner, 2006). Many of the statistical methods including correlation, linear regression, t-tests, Analysis of Variance (ANOVA), just to name a few, are based on the normal distribution (Neter *et al.,* 2005). Aside these, many multivariate statistical techniques such as, Multivariate Analysis of Variance (MANOVA), Principal Component Analysis (PCA), canonical correlation analysis, discriminants analysis, etc. rely on the multivariate normality assumption in order to make inferences (Johnson & Wichern, 2007). In essence, this assumption requires that a set of data upon which a statistical test of significance or statistical modelling is to be applied must either exactly or approximately be normally distributed. Essentially, this is due to the fact that almost all of these tests and models are developed with the normal distribution. By making this assumption about the data, parametric tests are more powerful than non-parametric test.

The validity of any statistical analysis depends on the assumptions underlying its use. This is no exception to statistical methods that make use of the assumption of normality. It is therefore important to ensure that the normality assumption is satisfied before use. However, for large sample, violation of the normality assumption should not be of major concern, due to the central limit theorem. The theorem states that, the mean of random sample from any distribution will have normal distribution (Altman & Bland, 1995). For this reason, if we have samples consisting of hundreds of observations, we can ignore the distribution of the data. There are several tools used in checking the validity of the normality assumptions. The purpose of this paper is to provide a concise overview of how to assess both univariate and multivariate normality, using both graphical techniques and formal test. In the absence of a more detailed and rigorous mathematical formulas, several references are provided. To achieve this, a random normal and non-normal data is generated and used for the purpose of illustration. Also, to illustrate how to check multivariate normality, the lumber stiffness data is used.

2. Assessing normality in univariate data

In the univariate setting, several tools (graphical and formal tests) are available for checking the normality assumption. The Q-Q plot is one of the most popular graphical methods used in testing univariate normality even though, there are many other graphical techniques as well (Stevens, 2001). Some of the other graphical tools include the use of *histogram, box plot, stem-and-leaf plot or dot plot*. These tests allow a quick and simple means of evaluating the shape of the distribution. However, they require large samples in order to provide reliable information (Neter *et al.,* 2005). In this paper, Q-Q plot together with histogram is used for illustrative purpose.

2.1 Q-Q plot, Histogram and correlation test

In a Q-Q plot, the sample quantiles are plotted against the quantiles that would be expected if the sample came from a normal distribution. If the data are normally distributed, the result would be a straight diagonal line. Also, for normally distributed data, the histogram should approximates the bell shape (symmetry) of a normal distribution. Two different set of data were generated: normal data and a random generated data with no known distribution. For the normal data, data of sizes 20 and 200 were generated to assess the effect of sample size on normality. Figure 1 presents both the Q-Q plot and the histogram for the two univariate data sampled from a normal distribution with size 20 and 200.

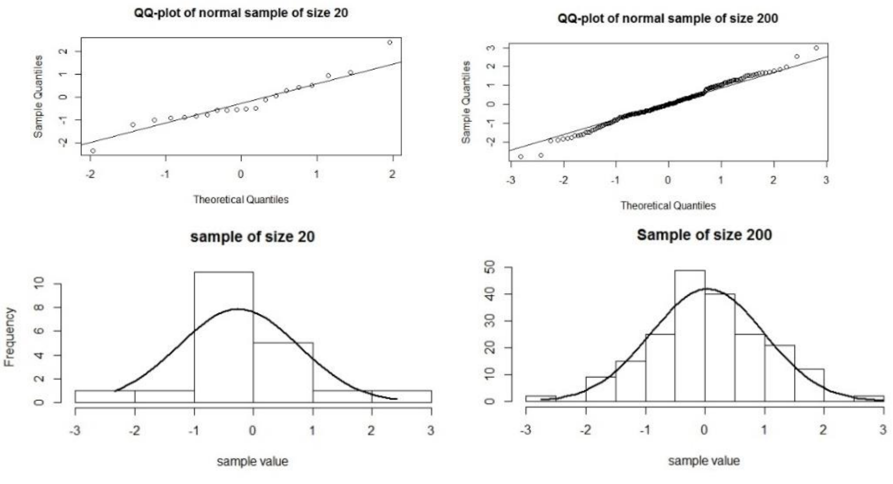

Figure 1. Q-Q plot and Histogram for generated normal data of sizes 20 and 200

From the plots, both samples are normally distributed. This is obvious since the data was generated from a normal distribution. Though normal, the histogram for the sample with fewer observations does not approximate the bell shape as expected. To use the histogram to check normality, there should be reasonably large number of observations (Neter *et al.,* 2005). This can be seen in the histogram of the normal sample of size 200. The histogram has the bell shape (symmetric shape) as expected for normal data. On the other hand, using the Q-Q plot, both sample are normally distributed. It is observed that in both samples, most of the points fall or lie close to the diagonal line.

In practice, a plot that is nearly linear suggests normality. Since linearity is subjective particularly with graphics, the correlation between the sample quantiles and the expected (theoretical) quantiles can be computed. This is called the correlation test for normality (Neter *et al.,* 2005). A high coefficient of correlation is an indication of normality. As an alternative, some authors have develop a rule for making conclusions using the correlation test. This rule depends on the sample size as well as the level of significance (Johnson & Wichern, 2007). For large samples and large α, a large correlation coefficient is require to achieve normality.

In addition, as a simple guide, Figure 2 shows Q-Q plots where the distributions depart from a normal distribution.

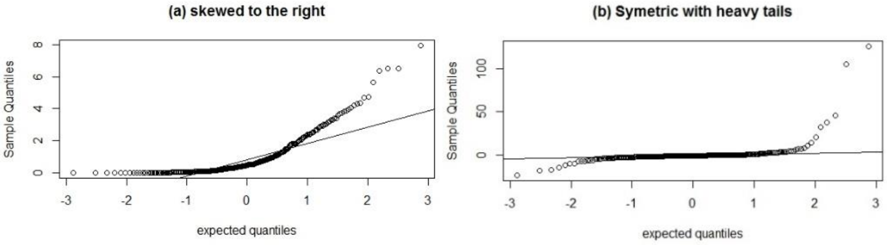

Figure 2. Q-Q plot for non-normal data

Figure 2(a) shows a Q-Q plot where the distribution is highly skewed to the right. This is indicated in the concave-upward shape in the plot. Plots that show concave-downwards shape are also not normally distributed: such plots are indication of left skewed distributions. Similarly, Figure 2(b) presents a Q-Q plot of a non-normal data which is symmetric but have heavy tails. These can be used as a simple guide to assess whether a univariate data is normally distributed or otherwise.

The Q-Q plot and histogram for the random generated data with no known distribution is also presented in Figure 3.

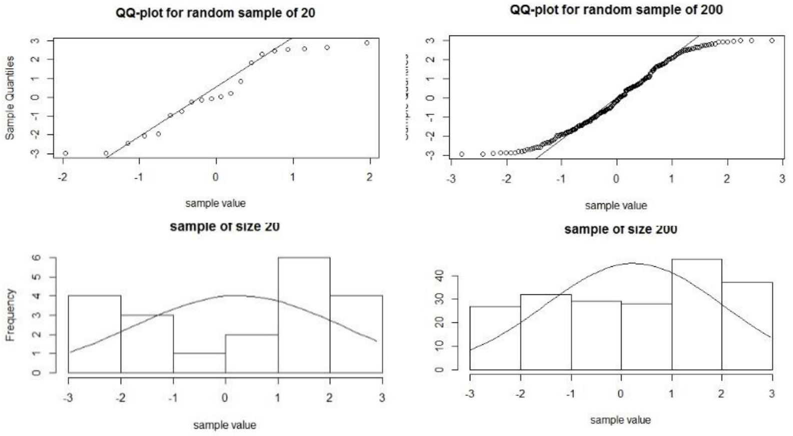

Figure 3. Q-Q plot for random generated data with no known distribution

From the plot (both Q-Q plot and histogram), it can clearly be observed that the samples are not normally distributed. The Q-Q plot for the sample of size 20 seem to exhibit some pattern. Also for the larger sample (sample size 200), the Q-Q plot looks symmetric with heavy tails. Likewise, the shape of the histogram in both samples do not approximate that of the bell shape for normally distributed data

2.2 Properties of normal distribution

From the empirical rule of normal distribution, it has been established that if a random sample is normally distributed then, 68.3%, 95% and 95.4% of the observed values must lie within 1, 1.96 and 2 standard deviation (*sd*) of the mean respectively. Hence for large sample, it is expected that the observed proportion of the sample lying within $\bar{x} \pm \sqrt{s}$, $\bar{x} \pm 1.96\sqrt{s}$ and $\bar{x} \pm 2\sqrt{s}$ should be about 68.3%, 95% and 95.4% respectively (Johnson and Wichern 2007). Consequently, it is observed that, for departure from normality,

$|\hat{p}_{j1} - 0.683|$ will be $> 3\sqrt{((0.683\times0.317)/200)}$. $\bar{x}$ is sample mean, s is the sample variance, $\sqrt{s}$ is sample standard deviation (*sd*) and $\hat{p}_{j1}$ is the proportion of observation within $\bar{x} \pm \sqrt{s}$ (Hogg *et al.*, 2004). Using the two generated samples each of size 200, the percentage of observations within 1 and 1.96 *sd* of the mean is calculated and presented in Table 1. From the results, the percentage of observations within 1 *sd* of the mean is at least 68% for the normal sample and 58% for the non-normal sample. The proportion for the non-normal sample is far below 68.3%.

Table 1. Percentage of observations within 1 and 1.96 *sd* of the mean

	% within 1 *sd* of mean	% within 1.96 *sd* of mean	$\|\hat{\boldsymbol{p}}_{j1} - \mathbf{0.683}\|$
Normal sample	68	96	0.003
Non-normal sample	58	100	0.103

Likewise, due to the heavy tails in the non-normal sample as shown in Figure 3, all the observations in that sample are within 1.96 *sd* of the mean. Comparing $|\hat{p}_{j1} - 0.683|$ with 3√ ((0.683×0.317)/200) = 0.098, it can be concluded that the random sample from the unknown distribution is not normally distributed.

2.3 Formal tests for univariate normality

The normality tests are complement to the graphical techniques. The main tests for the assessment of normality are Kolmogorov-Smirnov (K-S) test, Lilliefors corrected K-S test, Shapiro-Wilk test, Anderson-Darling test, Cramer-von Mises test, D'Agostino skewness test, Anscombe-Glynn kurtosis test, D'Agostino-Pearson omnibus test, and the Jarque-Bera test (Ghasemi & Zahediasl, 2012). Among these, Kolmogorov-Smirnov (K-S) test, Shapiro-Wilk test, Anderson-Darling test, and Cramer-von Mises test are widely used in practice and implemented in many statistical applications (SAS Guide, 2004). These tests have their own pros and cons hence, in practice, one can perform more than a single test to be sure of the results obtained. In this paper, the Shapiro-Wilk test together with the Kolmogorov-Smirnov test is used. Detailed formulas can be consulted in the literature and are not presented here.

The results of the test for normality for the two generated samples is presented in Table 2. The hypotheses for these tests is the following: H_0: Sample is normally distributed against H_1: Sample is not normally distributed. At a 5% level of significance, we fail to reject the null hypothesis of normality for the normal sample. However, for the non-normal data, we reject the null hypothesis at a 5% level of significance and conclude that, the random sample is not from a normal distribution. The p-value for both test is < 0.05 for the non-normal sample and otherwise in the normal samples.

Table 2. Shapiro-Wilk and Kolmogorov tests for normality

	Normal sample				Non-normal sample	
	Sample size = 20		Sample size =200		Sample size =200	
Test	Test statistics	p-value	Test statistics	p-value	Test statistics	p-value
Shapiro - Wilk	0.938	0.224	0.994	0.58	0.945	<0.0001
Kolmogorov -Smirnov	0.284	0.065	0.07136	0.26	0.2774	<0.0001

3. Assessing normality in multivariate data

Just like with univariate normality, several methods exist for assessing normality for multivariate data. However, unlike in uinvariate analysis, we would want to check the assumption of normality for all distributions of $2,3,\ldots,p$ dimensions. If $\boldsymbol{X}$ is multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$**,** we write $\boldsymbol{X} \sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$. Many statistical techniques such as multivariate analysis of variance (MANOVA), principal component analysis (PCA), canonical correlation, discriminant analysis and many others make use of the multivariate normal assumption in making inference. Furthermore, most of the theories in multivariate data analysis have been developed assuming multivariate normality (Johnson & Wichern, 2007). This is because, procedures based on normal populations are simple and more efficient hence, are very common in statistical applications.

As noted earlier, for large samples, it is assumed that the data is approximately normal regardless of underlying distribution. Even though this is so, the approximation will be better if the distribution is closer to normality. Therefore, the detection of departures from normality is crucial. The properties of multivariate normal data will be used as a starting point in assessing the assumption of multivariate normality. The marginal distributions and linear combinations of a multivariate normal distribution are normal. Hence, one can begin the assessment of multivariate normality by checking univariate normality and bivariate normality.

Several authors suggest restricting the investigation to the univariate and the bivariate margins (Johnson & Wichern, 2007, Rencher, 2002), however, multivariate distribution as a whole should be investigated. To check univariate normality, the procedure outlined earlier can be used.

A scatter plot of the pairs of variables as well as a bivariate Gamma plot (chi-squared Q-Q plot) can be used in checking bivariate normality (Johnson & Wichern, 2007). With the scatter plot, if the plot points do not look ellipsoidal, then the normality assumption is questionable. On the other hand, if the plot is ellipsoidal, there is no guarantee of bivariate normality since many other distributions have this shape. So, for an ellipsoidal shape, we are only sure that the random samples do not come from a non-normal distribution.

Assume $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots, \boldsymbol{x_n}$ is a random sample from a p dimensional multivariate normal population. Then, the generalized square distance, $d_j^2 = (\boldsymbol{x_j} - \overline{\boldsymbol{x}})'\boldsymbol{S^{-1}}(\boldsymbol{x_j} - \overline{\boldsymbol{x}}) \sim \chi_p^2$ (Hogg *et al.,* 2004). A Gamma plot is used to check if d_j^2 has a χ_p^2 **.** A Gamma plot is a plot of the ordered generalized squared distances $d_j^2,$ against the corresponding percentile of the chi-squared distribution. The points on the plot should be linear if the data is bivariate/multivariate normal.

To illustrate the use of scatter plot and Gamma plot to check bivariate/multivariate normality, the data for lumber stiffness is used. The data set contains four different measures of stiffness X_1, X_2, X_3 and X_4 for $n = 30$ boards. This data has been studied in detail by Johnson and Wichern (2007) hence, more detail can be obtained from their analysis. In this paper, we use the data set for illustration, in checking whether it is multivariate normal.

3.1 Scatter plot and Gamma plot

To begin with, a scatter plot of all pair of variables is provided. Since there are four variables in total, we will require 6 bivariate scatter plots. Again, for illustration, only two of the plots are presented. It should be emphasized that, the remaining four plots are similar to the ones presented in this paper. From the plot as presented in Figure 4, most of the points are found inside the ellipse. As confidence curve, the ellipse shows where 95% of the data should lie, assuming a bivariate normal distribution. Since many other distributions are within the class of elliptically contoured distributions e.g. multivariate *t*, multivariate Cauchy, etc. (Kibria, 2008), bivariate normality cannot be guaranteed. Likewise, for the Gamma plot, most of the point fall on or closer to the straight line. However, for both plots, some points are distance away from the ellipse (for scatter plot) and straight line (Gamma plot). In practical application, observations like this should prompt the analyst to check the data for outlying observations.

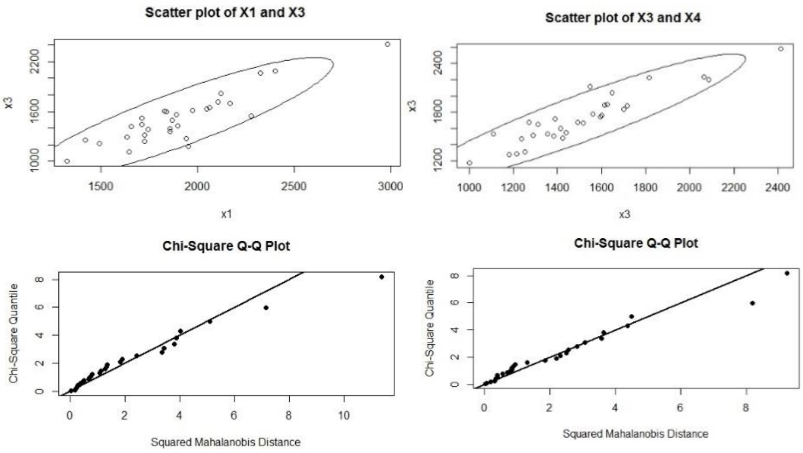

Figure 4. Bivariate scatter plot and Gamma plot for X_1 against X_3 and X_3 against X_4**.**

As mentioned earlier, restricting the investigation to the univariate and the bivariate margins is sometimes sufficient, however, a gamma plot can also be constructed for more than two variables. A Gamma plot for the four variables used in the illustration is presented in Figure 5. If the data is multivariate normal, it is expected that most of the points should fall on the diagonal line and should not show any pattern as shown in Figure 2. From the Gamma plot (Figure 5), a substantial number of points lie away from the line particularly, in the right tail. Hence, the assumption of multivariate normality is questionable. However, in order to confirm this, other methods for testing multivariate normality will also be used.

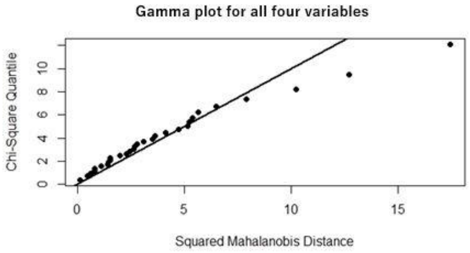

Figure 5. Gama plot for all variables

3.2 *Formal tests for multivariate normality*

Several tests have been proposed for testing multivariate normality (Selcuk *et al.,* 2015). Unfortunately, there is no known uniformly most powerful test and it is recommended to perform several tests before coming up with a conclusion on normality (Thomas & Jon, 2003). It is recommended that the choice of the method should be based on the kind of departures from normality that one wishes to investigate depending on the particular problem at hand (Cox & Small, 1978). All the methods proposed can be grouped into four main categories namely: goodness of fit techniques, procedures based on skewness and kurtosis, consistent and invariant tests, and graphical and correlational approaches (Patrick *et al.,* 2006).

For illustration, we discuss two commonly used methods namely; Royston's test and Mardia's test. These methods are also implemented in many statistical software. Royston's test is an extension of the Shapiro and Wilk goodness of fit test for univariate normality while Mardia’s test is a multivariate extension of measures of skewness and Kurtosis. In the absence of extensive mathematical formulations and exhaustive detail on Royston's and Mardia's tests, Selcuk *et al.,* 2015, Thomas & Jon, 2003, Patrick *et al.,* 2006, Mardia *et al.,* 2003, De Carlo, 1997 and Domanski, 2009 are recommended.

Royston's and Mardia's test is applied to the lumber stiffness data to check if it is multivariate normal. The hypotheses for these tests is the following: H_0 : The measurements have a multivariate normal distributed, against H_1: The measurements do not have a multivariate normal distributed. The results obtained from both tests are presented in Table 3. At a 5% level of significance, the null hypothesis of multivariate normality is rejected. The results from both test reveal that the data is not multivariate normal. In most situations, outliers and small sample size are found to have a substantial influence in the results of the test for normality (Johnson & Wichern, 2007).

Table 3. Test for normality Using Royston's test and Mardia's test

Test	Royston’s	Mardia’s	
		Skewness	Kurtosis
Test Statistics	9.8238	37.68	0.58
p-value	0.0095	0.00967	0.01114

To illustrate what to look for if data is multivariate normal, a random sample of size 250 bivariate normal variates with $\mu = \begin{pmatrix}0.0\\2.0\end{pmatrix}$, $\Sigma = \begin{pmatrix}1.0 & 0.5\\0.5 & 1.0\end{pmatrix}$ is used. A scatter plot of the pair of variables, a bivariate Gamma plot as well as Mardia's and Royston's test are used in checking multivariate normality. Both plots show that the random sample has a multivariate normal distribution. With the scatter plot, almost all the points fall inside the ellipse. The ellipse shows where 95% of the observations should lie if the data is bivariate/multivariate normal. However, if the scatter plot for any pair of variables have an ellipsoidal shape, there is no guarantee of bivariate normality. Aside the normal distribution, several distributions (the multivariate normal, matric-t, multivariate Student's t, and multivariate Cauchy) belong to the class of elliptically contoured distributions (Kibria, 2008). The Gamma plot in Figure 6 also provides evidence of a bivariate normal sample. For bivariate/multivariate normal data, the points on the Gamma plot should be linear. The plot in Figure 6 is spectacular, almost all the points fall on the diagonal line. This should not be a surprise since we have a bivariate normal sample. However, in practical applications, one should not always have a plot similar to this to conclude multivariate normality. Plots like Figures 2, 3 and 5 do not guarantee bivariate / multivariate normality.

Again, the results obtained from both Mardia's and Royston's tests (Table 4) confirm bivariate normality. For Mardia's test, the p-values for both skewness and kurtosis are far greater than the significance level of 0.05.

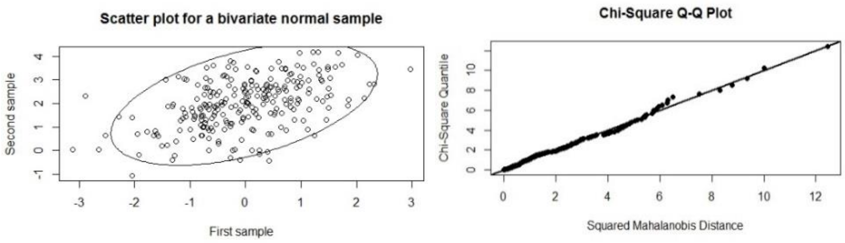

Figure 6. Scatter plot and Gamma plot for a bivariate normal data

Table 3. Test for normality Using Royston's test and Mardia's test

Test	Royston's	Mardia's	
		Skewness	Kurtosis
Test Statistics	4.1797	3.3447	0.040
p-value	0.1248	0.5019	0.9679

3. Conclusion

The assumption of normally distributed data is widely used in practice. Many statistical techniques rely on this assumption for making inference. Methods that make use of this assumptions have higher power compared to their non-parametric counterparts. However, the validity of these methods depend on the validity of the assumption of normality. Several methods have been outlined to help assess this assumptions. In the univariate case, the Q-Q plot, *histogram, box plot* and *dot plot* can be used as graphical techniques for checking normality. In the case of the Q-Q plot, the correlation test for normality can also be used. Using the properties of normal data, another alternative for assessing normality can be applied. Likewise, Kolmogorov-Smirnov (K-S) test, Lilliefors corrected K-S test, Shapiro-Wilk test, Anderson-Darling test, Cramer-von Mises test, D'Agostino skewness test, Anscombe-Glynn kurtosis test, D'Agostino-Pearson omnibus test, and the Jarque-Bera test can also be used (Ghasemi & Zahediasl, 2012). However, Kolmogorov-Smirnov test, Shapiro-Wilk test, Anderson-Darling test, and Cramer-von Mises test are widely used in practice and implemented in many statistical applications (SAS Guide, 2004). To be sure of the results obtained from the assessment of normality, it is recommended to use several methods and not just one method.

For data to be multivariate normal, the marginal distributions and linear combinations should be normally distributed as well. Hence, the univariate methods should first be used for checking normality of each variable. Then, all the pair of variables should be bivariate normal. For each pair, a scatter plot together with a gamma plot can be used to test bivariate normality. Though in many applications, restricting attention to the univariate and the bivariate margins is sufficient (Johnson & wichern, 2007, Rencher, 2002), a Gamma plot, and other methods can be used in testing multivariate normality. Royston's and Mardia's tests can be used as alternative methods in checking multivariate normality.

Many alternative methods exits when the normality assumption is not tenable. Techniques for non-normal data can be used (Christensen, 1997). Also, the data can be transformed to make it normal or near normal (Johnson & wichern, 2007, Altman & Bland, 1995).

A natural question to ask is whether all forms of departure from normality can be captured by the methods presented in this paper. Likewise, the statistical power associated with using these methods deserves further investigation and creates a path for further research.

References

Altman, D. G., & Bland, J.M. (1995). Statistics notes: the normal distribution. *BMJ*, 310, 298.

Christensen, L. A. (1997). Introduction to building a linear regression model. *Proceedings of the Twenty-Second Annual SAS Users Group International Conference.*

Cox, D. R., & Small, N. J. H. (1978). Testing multivariate normality. *Biometrika,* 65(2), 263-272.

De Carlo, L. T. (1997). On the Meaning and Use of Kurtosis. *Psychological Methods,* 2(3), 292-307.

Domański, C. (2009). Attempt to Assess Multivariate Normality Tests. *Acta Universitatis Lodziensis. Folia Oeconomica* 225, 75-90.

Ghasemi, A., & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *Int J Endocrinol Metab.* 10(2), 486-489.

Hogg, R. V., Craig, A. T. & Mckean, J. W. (2004). *Introduction to Mathematical Statistics*. (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Kibria, B. M. (2008). Robust Predictive Inference for Multivariate Linear Models with Elliptically Contoured Distribution Using Bayesian, Classical and Structural Approaches. *Journal of Modern Applied Statistical Methods*, 7(2), 19.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (2003). *Multivariate Analysis*. London: Academic Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (2005). *Applied Linear Statistical Models*. (5th ed.). New York: McGraw-Hill Education.

Patrick J., Matias S., & Katarzyna N. (2006). On tests for multivariate normality and associated simulation studies. *Journal of Statistical Computation & Simulation,* 77, 1065-1080.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. (2nd ed.). Wiley series in probability and mathematical statistics. New York: John Wiley & Sons.

Rosner, B. (2006). *Fundamentals of Biostatistics.* (6th ed.). Belmont, CA: Thomson-Brooks/Cole.

SAS Institute. (2004), “SAS 9.1.3 Procedures Guide”, Volume 4. Cary, NC: SAS Institute.

Selcuk K., Dincer G., & Gokmen Z. (2015). MVN: An R Package for Assessing Multivariate Normality. Available: https://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf (November 13, 2015).

Stevens, J. (2001). *Applied multivariate statistics for the social sciences.*(4th ed.).

Thomas S. & Jon W. (2003). Tests For Assessing Multivariate Normality And The Covariance Structure Of Mimo Data. *Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '03), IEEE International Conference*, 4.