Advances and Applications in Statistics
Volume 24, Number 2, 2011, Pages 157-162
Published Online: November 2011
Available online at http://pphmj.com/journals/adas.htm
Published by Pushpa Publishing House, Allahabad, INDIA

A CORRELATION TEST FOR BIVARIATE NORMALITY

H. F. Coronel-Brizio, A. R. Hernández-Montoya and M. E. Rodríguez-Achach

Facultad de Fisica e Inteligencia Artificial
Universidad Veracruzana
Mexico
e-mail: hcoronel@uv.mx

Abstract

Roy's union-intersection principle can be applied to construct a test of multivariate normality based on any test statistic used for the univariate case. In this paper, the maximum correlation test statistic is shown to be useful for testing multivariate normality. The technique is illustrated by considering a test for bivariate normality.

1. Introduction

Let $\mathbf{X_1}, ..., \mathbf{X_n}$ be a p-variate random sample from a normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Also, define $Z_i = \mathbf{c}^\top \boldsymbol{X_i}$, where $\mathbf{c}$ is a p-dimensional non-random vector.

Then, if $Z_{(1)}, ..., Z_{(n)}$ are the ordered transformed observations, we have for all non-null vectors $\mathbf{c}$,

$$E(Z_{(i)}) = \mathbf{c}^\top \mu + \{\mathbf{c}^\top \boldsymbol{\Sigma} \mathbf{c}\}^{\frac{1}{2}} m_i,$$

2010 Mathematics Subject Classification: 62-07, 62F10, 62F12, 62N01, 62N02.

Keywords and phrases: goodness-of-fit, bivariate normality, correlation statistics.

Received August 1, 2011

where $\boldsymbol{m}^{\top} = (m_1, \ldots, m_n)$ is the vector of expected values of the order statistics from a standard normal distribution. Note that the mean value of the m_i is zero.

Then, define $R^2(\mathbf{c})$ to be the squared correlation coefficient between **Z**, the vector of order statistics corresponding to the transformed observations, and **m**. Thus

$$R^2(\mathbf{c}) = \frac{\left(\sum_{i=1}^{n} Z_{(i)} m_i\right)^2}{\sum_{i=1}^{n} m_i^2 \sum_{i=1}^{n} (Z_i - \bar{Z})^2}. \tag{1}$$

The squared correlation $R^2(\mathbf{c})$ is then independent of the parameters of the distribution, and Roy's [1] union-intersection principle can be applied to this situation by finding the vector $\mathbf{c}^*$ so that $R^2(\mathbf{c}^*)$ is a minimum.

The null hypothesis of multivariate normality will then be rejected for small values of the correlation or, equivalently, large values of the test statistic

$$T_n = n[1 - R^2(\mathbf{c}^*)]. \tag{2}$$

The asymptotic normality of these correlation-type statistics have already been studied by Lockhart [3] and McLaren and Lockhart [5]. In their findings they report an extremely slow rate of convergence to the normal distribution which results useless from a practical point of view. For the exponential case, Lockhart [3] found that the percentage of cases in which the squared correlation exceeded 1, was about 2.3% with a sample size of $n = e^{16}$. The interested reader is also referred to the work by D'Agostino and Stephens [2].

2. The Bivariate Case

We illustrate the use of $R^2(\mathbf{c})$ for dimension two. For this case, the minimization process reduces to a simple unidimensional search of the normalized vector $\mathbf{c}^*$. Using polar coordinates we can write the normalized vector **c** as

$$\mathbf{c}^\top = (\sin\theta, \cos\theta).$$

Hence the problem reduces to find the value θ^*, of θ, which minimizes

$$T_n(\theta) = n[1 - R^2(\theta)].$$

Empirical percentage points of the distribution of the test statistic $T_n(\theta^*)$ were obtained empirically from ten thousand simulations and are shown in Table 1.

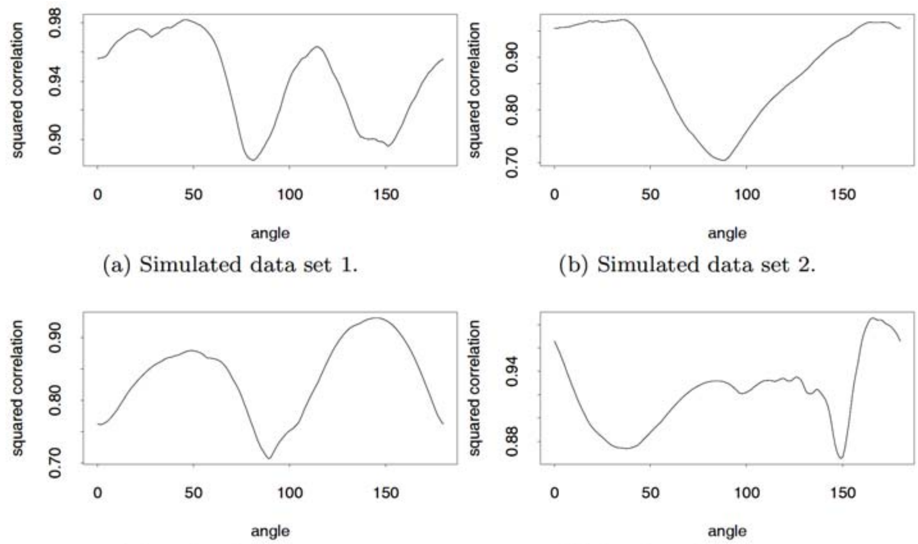
Table 1. Empirical percentage points of $T(\theta^*)$ for bivariate normality

Significance level

n	0.500	0.250	0.150	0.100	0.050	0.025	0.010
10	1.34	1.73	2.00	2.20	2.54	2.82	3.24
20	1.57	2.04	2.38	2.60	3.03	3.44	4.00
30	1.72	2.23	2.58	2.87	3.34	3.86	4.53
40	1.82	2.33	2.67	2.96	3.46	3.92	4.55
50	1.88	2.41	2.79	3.05	3.51	4.00	4.68
60	1.94	2.48	2.82	3.13	3.63	4.15	4.72
70	1.98	2.52	2.89	3.17	3.66	4.20	4.78
80	2.00	2.55	2.90	3.19	3.71	4.24	5.00
90	2.02	2.57	2.96	3.26	3.74	4.22	4.90
100	2.01	2.63	3.02	3.29	3.82	4.38	4.99

3. Examples

In the following four examples, simulated data sets are analyzed. In each case, a search in the interval $[0^\circ, 180^\circ)$ was conducted using increments of one degree. This magnitude for the increment produces a precision greater than 0.01 in the calculated value of θ^*.



(a) Simulated data set 1.

(b) Simulated data set 2.

(c) Simulated data set 3.

(d) Simulated data set 4.

Figure 1. Plot of $R^2(\theta)$ vs θ.

Example 1. The first data set consists of 20 observations from a normal distribution with zero mean vector and identity covariance matrix. A plot of $R^2(\theta)$ vs θ is shown in Figure 1(a). In this case, the minimum value of the correlation is 0.8856.

Example 2. In the second simulated data set, the first component of the vector was sampled from a chi-squared distribution with one degree of freedom. The corresponding plot is shown in Figure 1(b). The minimum value of the correlation for this sample is 0.7048.

Example 3. Figure 1(c) shows the plot for the third simulated data set consisting of 20 vectors, in which both components of the vector were sampled from a chi-squared distribution with one degree of freedom. Here, the minimum value of the correlation was 0.7070.

Example 4. The fourth data set was constructed to illustrate the situation in which the components of the random vector are marginally normal but not jointly normal. Two independent standard normal values u and v were first generated. The second component of the vector was set equal to v whereas the first component was set equal to u, if $uv > 0$ or $-u$ if $uv < 0$. A sample of twenty vectors was simulated. The corresponding plot is shown in Figure 1(d). The minimum correlation for this data set was found to be 0.8653.

4. Conclusions

The type of plots illustrated in Figures 1(a) to 1(d) is a useful tool to summarize relevant information about the type of departure from bivariate normality exhibited by the data. For instance, non-normality in the first component translates into a significant drop of the value of $R^2(\theta^*)$ in the neighbourhood of 90°; non-normality of both components will produce significant drops in the neighbourhoods of 0° and 90°. The proposed test based on the maximum correlation method combines simultaneously the properties of graphical and formal statistical methods and can be easily extended for larger dimensions, in particular for dimension three.

Acknowledgements

The authors acknowledge the valuable help of Selene Jiménez in typesetting the final version of this paper. They also acknowledge support from the Sistema Nacional de Investigadores (S.N.I.), National Council for Science and Technology (CONACyT) under Science Research Project No. 155492.

References

[1] S. N. Roy, On a heuristic method of test construction and its use in multivariate analysis, Ann. Math. Statist. 24 (1953), 220-238.

[2] R. B. D'Agostino and M. A. Stephens, Goodness-of-fit Techniques, R. B. D'Agostino and M. A. Stephens, eds., Marcel Dekker, New York, 1986.

[3] R. A. Lockhart, The asymptotic distribution of the correlation coefficient in testing fit to the exponential distribution, Canad. J. Satist. 13(3) (1985), 253-256.

[4] C. G. McLaren, Some contributions to goodness-of-fit, Ph.D. Thesis, Department of Mathematics and Statistics, Simon Fraser University, 1985.

[5] C. G. McLaren and R. Lockhart, On the asymptotic efficiency of certain correlation tests of fit, Canad. J. Statist. 15(2) (1987), 159-167.