The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 385-392

A Rao-Robson Chi-Square Test for Multivariate Normality Based on the Mahalanobis Distances

Cheolyong Park[1)]

Abstract

Many tests for multivariate normality are based on the spherical coordinates of the scaled residuals of multivariate observations. Moore and Stubblebine's (1981) Pearson chi-square test is based on the radii of the scaled residuals, or equivalently the sample Mahalanobis distances of the observations from the sample mean vector. The chi-square statistic does not have a limiting chi-square distribution since the unknown parameters are estimated from ungrouped data. We will derive a simple closed form of the Rao-Robson chi-square test statistic and provide a self-contained proof that it has a limiting chi-square distribution. We then provide an illustrative example of application to a real data with a simulation study to show the accuracy in finite sample of the limiting distribution.

Keywords : Scaled residuals, Spherical coordiantes, Wald's method

1. Introduction

Let $X_1, X_2, \cdots, X_n$ be a random sample from a k-dimensional distribution. Many tests for multivariate normality are based on the scaled residuals

$$Z_i = S^{-1/2}(X_i - \overline{X}),$$

where $\overline{X}$ and S is the sample mean vector and sample covariance matrix. The scaled residuals remove the sample means and sample covariances and thus have a null sample mean vector and an identity sample covariance matrix. Those tests based on the scaled residuals usually utilizes their spherical coordinates R_i and U_i given by

$$R_i = \sqrt{Z_i^t Z_i} = \sqrt{(X_i - \overline{X})^t S^{-1}(X_i - \overline{X})}, \quad U_i = Z_i / R_i, \tag{1}$$

where 't' is a notation for transpose. For example, Moore and Stubblebine's (1981) chi-square test and Mardia's (1970) multivariate skewness and kurtosis tests are based on the radii R_i, Rayleigh statistic (Koziol, 1983) is based on the unit residuals U_i, and Quiroz and Dudley's

1) Associate Professor, Department of Statistics, Keimyung University, Taegu 704-701
E-mail : cypark1@kmu.ac.kr

(1991) chi-square test is based on both R_i and U_i.

We can note that the radii R_i are the sample Mahalanobis distances of X_i's from $\overline{X}$ and that R_i^2 have an asymptotic chi-square distribution with k degrees of freedom for large n when the sample is from a multivariate normal distribution. Thus the chi-square probability plot of R_i^2 can be an informal tool for checking the multivariate normality of the sample. Moore and Stubblebine's procedure is a formal chi-square test and is based on the cell counts of R_i's falling into a fixed partition of the real line. They derived the limiting distribution of the statistic only when the real line is partitioned such that the probability of R_i falling in each interval is equal. Park (1999) considered the general case where equiprobable intervals are not employed and derived the limiting distribution of the statistic which can be applied to unequally probable intervals as well as equiprobable intervals. However, the test statistic does not have a limiting chi-square distribution and we have some practical difficulty in calculating the asymptotic p-value of the test.

It is well known that the Pearson chi-square statistic does not have a limiting chi-square distribution when unkown parameters are estimated from ungrouped data since the work of Chernoff and Lehmann (1954). We will employ the generalized Wald's method so as to make the resulting chi-square test statistic have a limiting chi-square distribution. Such chi-square statistic is discovered by Rao and Robson (1974) and indepedently by Nikulin (1973) and will be denoted by the Rao-Robson statistic in this paper. We will provide a simple closed form of the Rao-Robson chi-square statistic corresponding to the Pearson statistic suggested by Moore and Stubblebine (1981). In order to show that the Rao-Robson chi-square statistic has a limiting chi-square distribution, we can apply a standard theorem on the quadratic forms of normal variates but will provide a self-contained direct proof instead.

In Section 2, we provide a simple closed form of the Rao-Robson chi-square statistic and show directly that it has a limiting chi-square distribution. In Section 3, we provide an illustrative example of application to a real data set with a simulation study to check the accuracy in finite samples of the limiting distribution.

2. Main Result

Before presenting main result for the Rao-Ronson chi-square test statistic, we will define some notations. Unless otherwise stated, vectors will be column vectors, but for convenience they will be written in text as row vectors.

For a given vector $y=(y_1, y_2, \cdots, y_m)$, we define the diagonal matrix $D(y)$ and the vector of square root values $\sqrt{y}$ to be

$$D(y) = \text{diag}(y_1, y_2, \cdots, y_m), \quad \sqrt{y} = (\sqrt{y_1}, \sqrt{y_2}, \cdots, \sqrt{y_m}). \tag{2}$$

Let $N_k(\mu, \Sigma)$ denote the k-variate normal distribution with mean vector μ and covariance

matrix Σ and let $\chi^2(k)$ denote the chi-square distribution with k degrees of freedom.

Let $X_1, X_2, \cdots, X_n$ be a random sample from $N_k(\mu, \Sigma)$ where Σ is nonsingular. Let $\theta=(\mu, \Sigma)$ be the parameter of the distribution and let the maximum likelihood estimator (MLE) of θ be denoted by $\theta_n=(\overline{X}, S)$, where n is used for the denominator of the sample covariance matrix S. Let $0=c_0<c_1<\cdots<c_M=\infty$ be a sequence of nonnegative real numbers which forms a partiton $I_1, I_2, \cdots, I_M$ of the range of the radii R_i, defined in (1), such that $I_i=[c_i, c_{i+1})$, $i=1,2,\cdots,M$. Let N_{in} denote the number of R_j^2's falling into the interval I_i, then the estimated probability p_{in} of R_j^2's falling into I_i is given by

$$p_{in}=P_{\theta_n}(R_1^2 \in I_i)=F_k(c_{i+1})-F_k(c_i),$$

where $F_k(\cdot)$ is the cumulative distribution function of $\chi^2(k)$. Note that the vector $p_n=(p_{1n}, \cdots, p_{Mn})$ of estimated probabilities does not depend on n. Let V_n be the M vector of standardized cell counts having i-th component

$$(N_{in}-np_{in})/\sqrt{np_{in}},$$

then the Pearson chi-square test statistic for multivariate normality suggested by Moore and Stubblebine (1981) is given by

$$\sum_{i=1}^{M}\frac{(N_{in}-np_{in})^2}{np_{in}}=V_n^t V_n. \tag{3}$$

The limiting distribution of this statistic is that of a weighted sum of chi-square variates but is not an exact chi-square distribution since θ is estimated by the MLE θ_n from ungrouped data. This is a well known fact since the work of Chernoff and Lehmann (1954). Although there have been much work on obtaining numerical approximations to the distibution of a weighed sum of chi-square variables (see Imhof (1961), Solomon and Stephens (1977), Farebrother (1990) among others), the asymptotic p-value of the statistic is not easy to calculate in practice.

The Rao-Robson chi-square statistic has a limiting chi-square distribution and thus is easy to calculate its asymptotic p-value. Since the statistic is known to be powerful (see Rao and Robson (1974) and Spruill (1976) among others), and is easy to compute in our case, it will be much suited for testing multivariate normality. Since we need to calculate the limiting variance of the standardized cell counts V_n in calculating the statistic, we first provide its limiting distribution in the following lemma:

Lemma 1. (Park, 1999) Under the assumptions described in this section,

$$V_n \xrightarrow{d} N_M(0, A) \quad \text{as} \quad n \to \infty$$

where

$$A=I-\sqrt{p_n}\sqrt{p_n}^t-2BB^t,\quad B=\{D(p_n)\}^{-1/2}(d_1 1_k,\cdots,d_M 1_k)^t$$

with 1_k the k-vector of ones, the square root vector $\sqrt{p_n}$ and the diagonal matrix $D(p_n)$ defined in (2), and

$$d_i=\left(c_{i-1}^{k/2}e^{-c_{i-1}/2}-c_i^{k/2}e^{-c_i/2}\right)b_k/2$$

$$b_k=\begin{cases}[k(k-2)\cdots 2]^{-1} & k \text{ even}\\ (2/\pi)^{1/2}[k(k-2)\cdots 1]^{-1} & k \text{ odd.}\end{cases}$$

Note that the limiting variance A of V_n does not depend on the unknown parameter θ. Thus the Rao-Robson chi-square statistic is given by the form $V_n^t A^- V_n$, where A^- is a generalized inverse of A. It is known by the general Wald's method (see p. 173 of Rao and Mitra (1971) for example) that if $x\sim N_p(0,\Sigma)$ and $\text{rank}(\Sigma)=k$ then $x^t\Sigma^- x\sim\chi^2(k)$, and thus we can easily show that

$$V_n^t A^- V_n \xrightarrow{d} \chi^2(M-1)\quad \text{as } n\to\infty.$$

We will not use this result since we can easily provide a self-contained direct proof that it holds for the Moore-Penrose inverse. We use the Moore-Penrose inverse A^+ in computing the closed form of the statistic since it is easy to calculate. Here is the main result on the Rao-Robson chi-square test statistic:

Thoerem 1. Under the preceding assumptions, the Rao-Robson chi-square test statistic is given by

$$T_n=V_n^t A^+ V_n=\sum_{i=1}^{M}\frac{(N_{in}-np_{in})^2}{np_{in}}+\frac{2k}{n(1-2kd^*)}\left[\sum_{i=1}^{M}d_i\frac{(N_{in}-np_{in})}{p_{in}}\right]^2$$

and it has the limiting $\chi^2(M-1)$ distribution, where d_i's are defined in Lemma 1 and $d^*=\sum_{i=1}^{M}(d_i^2/p_{in})$.

Proof: Define $D=B/\sqrt{kd^*}$, then the limiting variance of V_n is given by

$$A=I-\sqrt{p_n}\sqrt{p_n}^t-2kd^*DD^t.$$

Since $D^tD=1_k1_k^t/k$, it is easy to show that DD^t is an idempotent matrix of rank 1 and so is $\sqrt{p_n}\sqrt{p_n}^t$. Furthermore, since $\sum_{i=1}^{M}d_i=0$, two idempotent matrices DD^t and $\sqrt{p_n}\sqrt{p_n}^t$ are orthogonal to each other. Therefore A can be expressed as

$$(I-\sqrt{p_n}\sqrt{p_n}^t)(I-2kd^*DD^t),$$

whose Moore-Penrose inverse is given by

$$(I-2kd^*DD^t)^{-1}(I-\sqrt{p_n}\sqrt{p_n}^t)^+ = \left(I+\frac{2kd^*}{1-2kd^*}DD^t\right)(I-\sqrt{p_n}\sqrt{p_n}^t)$$
$$= I-\sqrt{p_n}\sqrt{p_n}^t+\frac{2kd^*}{1-2kd^*}DD^t.$$

This shows that

$$T_n = V_n^tA^+V_n = V_n^tV_n+\frac{2kd^*}{1-2kd^*}V_n^tDD^tV_n$$
$$= \sum_{i=1}^{M}\frac{(N_{in}-np_{in})^2}{np_{in}}+\frac{2k}{n(1-2kd^*)}\left[\sum_{i=1}^{M}d_i\frac{(N_{in}-np_{in})}{p_{in}}\right]^2,$$

where the second equality holds since $V_n^t\sqrt{p_n}=0$.

Now it remains to show that T_n has the limiting $\chi^2(M-1)$ distribution. Let $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_M\geq0$ be the eigenvalues of A and let $e_1, e_2, \cdots, e_M$ be the corresponding eigenvectors such that $e_i^te_j=\delta_{ij}$, where δ_{ij} is the Kronecker delta. By the orthogonality between $\sqrt{p_n}\sqrt{p_n}^t$ and DD^t, it is easy to show that $\lambda_1=\cdots=\lambda_{M-2}=1$, $\lambda_{M-1}=1-2kd^*$ and $\lambda_M=0$. Let $E=(e_1,\cdots,e_M)$ and $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_M)$. Then, by the spectral decomposition, $A=E\Lambda E^t$ and its Moore-Penrose inverse is given by $A^+=E\Lambda^+E^t$ where Λ^+ is a matrix obtained by replacing the nonzero elements of Λ by their reciprocals. Define $U_n=(\Lambda^+)^{1/2}E^tV_n$. Since V_n has the limiting $N_M(0,A)$ distribution by Lemma 1, U_n has the limiting $N_M(0, \mathrm{diag}(I_{M-1},0))$ distribution by continuous mapping theorem. This shows that $T_n=U_n^tU_n$ has the limiting $\chi^2(M-1)$ distribution, which completes proof. □

3. Application and Simulation

In this section, we provide an illustrative example of application to 'bone data' presented in p.34 of Johnson and Wichern (1992). The first two variables, the mineral content of the dominant and nondominant sides of radius, are examined for multivariate normality. We compare the asymptotic p-values of the Rao-Robson chi-square test statistic with those of the Pearson statistic. We then provide a simulation study to check the accuracy in finite samples of the limiting distribution of the Rao-Robson statistic.

Since 'bone data' contain 25 cases, the number M of intervals we consider are 3, 4, 5 and both equiprobable and unequally probable ways of forming intervals are considered. We consider only the equiprobable intervals for $M=5$ since some of unequally probable intervals have expected cell counts less than 5. Chi-square values and their asymptotic p-values of Pearson and Rao-Robson statistics are summarized in Table 1.

Here, the asymptotic p-values of the Pearson chi-square statistic are reported as

(the p-value of $\chi^2(M-2)$, the p-value of $\chi^2(M-1)$)

Table 1. Chi-square values and their asymptotic p-values

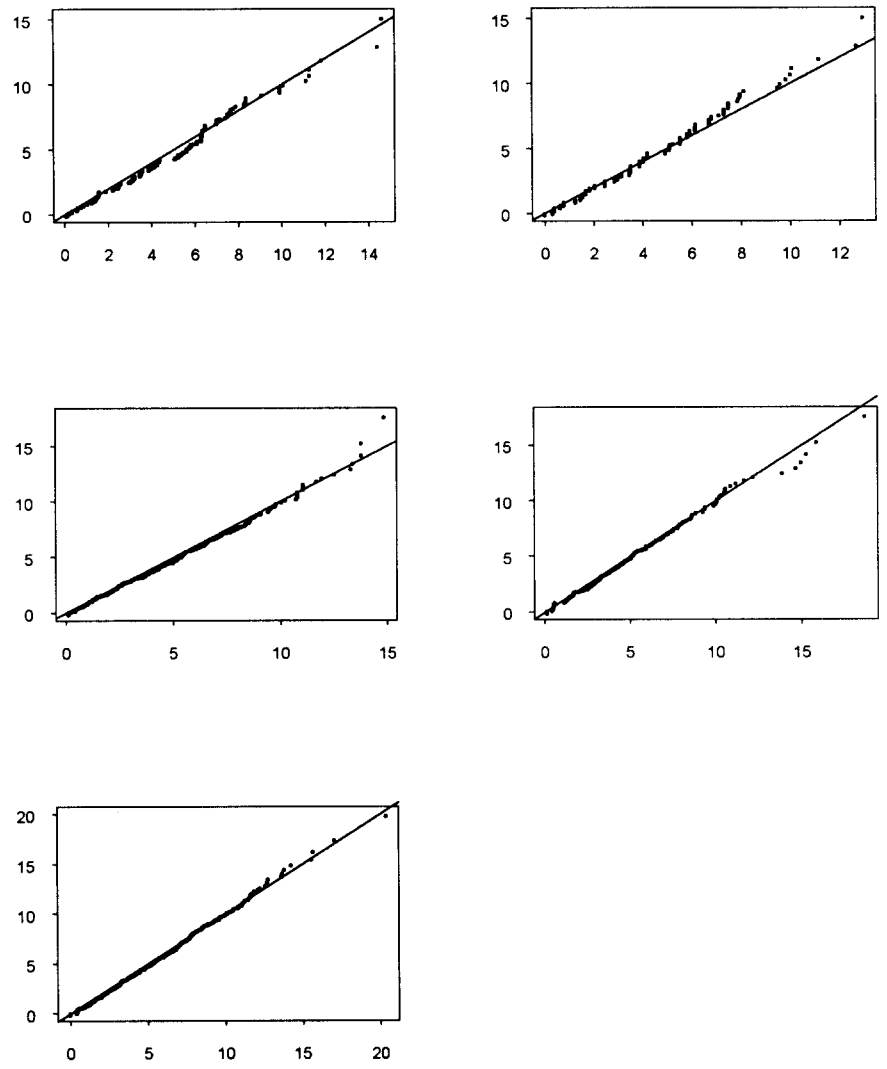
M	p_n	chi-square type	chi-square value	p-value
3	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	Pearson	2.96	(.0853, .2276)
		Rao-Robson	7.84	.0199
	$(\frac{1}{5}, \frac{2}{5}, \frac{2}{5})$	Pearson	10.7	(.0011, .0047)
		Rao-Robson	13.78	.0010
4	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	Pearson	5.24	(.0728, .1550)
		Rao-Robson	9.63	.0220
	$(\frac{1}{5}, \frac{1}{5}, \frac{3}{10}, \frac{3}{10})$	Pearson	16.53	(.0003, .0009)
		Rao-Robson	21.72	.0001
5	$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	Pearson	17.2	(.0006, .0018)
		Rao-Robson	27.73	.0001

since the limiting distribution of the statistic is between $\chi^2(M-2)$ and $\chi^2(M-1)$. From these tables, we can find that the unequally probable cases have smaller p-values than the corresponding equiprobable cases. This phenomenon is observed since those cases we choose are the same as in Park (1999) where we need to show that unequally probable intervals might have more power than equiprobable intervals. We can conclude that the Rao-Robson chi-square statistic is more powerful since it reports smaller p-value than the Pearson statistic.

To study the accuracy in finite samples of the limiting distribution of the Rao-Robson statistic for those cases we consider, we perform a small simulation study. Our simulation scheme is simple: we generate 1000 samples of size 25 from $N_2(0, I)$ and then calculate their Rao-Robson statistic values for those cases we consider. We do not need to consider other populations with $\mu \neq 0$ and $\Sigma \neq I_2$ since R_i and hence the test statistic are ancillary (see Lemma 3.1 of Park (1999) for details). We use the informal chi-square probability plot to check whether the statistic values are from the chi-square distribution with appropriate degrees of freedom. The chi-square probability plots are given in Figure 1.

Each row of the plots corresponds to the number of intervals: the first row is for $M=3$ and the second and third are for $M=4$ and $M=5$, respectively. Each column of the plots represent the way of forming intervals: the first column is for the equiprobable intervals and the second is for the unequally probable. We provide the reference line with intercept 0 and slope 1. This line represent 'ideal' case where the chi-square values coincide with those expected from their limiting distribution. From the figures, we can find that the points of plots for the equiprobable intervals are a little bit closer to the reference line than those for corresponding unequally probable intervals. We can also find that the chi-square approximation is fairly good except for those plots with $M=3$ where we can discover some discreteness of values.

Figure 1. Chi-square probability plots



References

[1] Chernoff, H., and Lehmann, E.L. (1954). The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit. *Annals of Mathematical Statistics* 25, 579-586.

[2] Farebrother, R.W. (1990). The Distribution of a Quadratic Form in Normal Variables. *Applied Statistics* 39, 294-309.

[3] Imhof, J.P. (1961). Computing the Distribution of Quadratic Forms in Normal Variables.

Biometrika 48, 419-426.

[4] Johnson, R.A., and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, Third Edition. Prentice Hall, New Jersey.

[5] Koziol, J.A. (1983). On Assessing Multivariate Normality. *Journal of the Royal Statistical Society - Series B* 45, 358-361.

[6] Mardia, K.V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika* 57, 519-530.

[7] Moore, D.D, and Stubblebine, J.B. (1981). Chi-Square Tests for Multivariate Normality with Application to Common Stock Prices. *Communications in Statistics - Theory and Methods* 10, 713-738.

[8] Nikulin, M.S. (1973). Chi-Square Test for Continuous Distributions with Shift and Scale Parameters. *Theory of Probability and Its Applications* 18, 559-568.

[9] Park, C. (1999). A Note on the Chi-Square Test for Multivariate Normality Based on the Sample Mahalanobis Distances. *Journal of the Korean Statistical Society* 28, 479-488.

[10] Quiroz, A.J., and Dudley, R.M. (1991). Some New Tests for Multivariate Normality. *Probability Theory and Related Fields* 87, 521-546.

[11] Rao, C.R., and Mitra, S.K. (1971). *Generalized Inverse of Matrices and Its Applications*. John Wiley & Sons, New York.

[12] Rao, C.R., and Robson, D.S. (1974). A Chi-Square Statistic for Goodness-of-Fit Tests within the Exponential Family. *Communications in Statistics* 3, 1139-1153.

[13] Solomon, H., and Stephens, M.A. (1977). Distribution of a Sum of Weighted Chi-Square Variables. *Journal of the American Statistical Association* 72, 881-885.

[14] Spruill, M.C. (1976). A Comparison of Chi-Square Goodness-of-Fit Tests Based on Approximate Bahadur Slope. *Annals of Statistics* 4, 409-412.