CONDITIONAL PROBABILITY INTEGRAL TRANSFORMATIONS FOR MULTIVARIATE NORMAL DISTRIBUTIONS

Santiago Rincon Gallardo, C. P. Quesenberry, F. J. O'Reilly

Institute of Statistics
Mimeograph Series No. 1148
Raleigh, November 1977

CONDITIONAL PROBABILITY INTEGRAL TRANSFORMATIONS FOR MULTIVARIATE NORMAL DISTRIBUTIONS

Santiago Rincon-Gallardo[1] and C. P. Quesenberry[2]

North Carolina State University

and

Federico J. O'Reilly

IIMAS, Universidad Nacional Autónoma de México

Let $X_1, \cdots, X_n$ be a random sample from a full-rank multivariate normal distribution $N(\mu,\Sigma)$. The two cases (i) μ unknown and $\Sigma = \sigma^2\Sigma_0, \Sigma_0$ known, and (ii) μ and Σ completely unknown are considered here. Transformations are given that transform the observation vectors to a (smaller) set of i.i.d. uniform rv's. These transformations can be used to construct goodness-of-fit tests for these multivariate normal distributions.

1. *Introduction and Summary*. There is a large literature that considers the multivariate normal distribution. However, there is very little available in the way of goodness-of-fit tests for multivariate normality, and nothing whatever that is based upon exact distribution theory and therefore is applicable for small and moderate size samples. Recently, Moore (1976) has commented upon the need for goodness-of-fit tests for multivariate normality. In this work we give transformations which can be used to construct exact level goodness-of-fit tests for multivariate normality.

O'Reilly and Quesenberry (1973), O-Q, introduced the conditional probability integral transformations, CPIT's. Transformations were given in that paper for a multivariate normal parent $N(\mu,\Sigma)$ for the case when μ is unknown and

[1]Supported by IIMAS, UNAM.

[2]Research supported by National Science Foundation Grant MCS76-82652.

$\Sigma = \Sigma_0$ is known. Here we give transformations for the two cases: (i) μ unknown, and $\Sigma = \sigma^2\Sigma_0$ with Σ_0 known, and (ii) μ and Σ unknown. In both of these cases the components of the observation vectors are transformed using certain Student-t distribution functions. The transformations and model testing techniques considered here can be carried out on high-speed computers.

2. Notation and Preliminaries. Let $\mathcal{P}$ denote the class of k-variate full-rank normal distributions with mean vector μ and variance-covariance matrix Σ. For $X_1, \cdots, X_n$ i.i.d. (column) vector rv's from $P \in \mathcal{P}$, with corresponding probability density function f and distribution function F, both defined on R_k, a complete sufficient statistic for $\mathcal{P}$ is $T_n = (\bar{X}_n, S_n)$, where $\bar{X}_n = (1/n) \sum_{i=1}^{n} X_i$ and $S_n = \sum_{i=1}^{n} X_i X_i' - n\bar{X}_n\bar{X}_n'$. It is readily verified that T_n is doubly transitive (cf. O-Q), i.e., $\sigma(T_n, X_n) = \sigma(T_{n-1}, X_n)$, where $\sigma(W)$ denotes the σ - algebra induced by a statistic W.

Consider the conditional distribution function $\tilde{F}$ of a single observation given the statistic T_n. For $n > k+1$ $\tilde{F}$ is absolutely continuous and possesses a density function $\tilde{f}$ which is the minimum variance unbiased, MVU, estimator of the parent density function. These functions were obtained by Ghurye and Olkin (1969), p. 1265, cases 3.2 and 3.4, for the cases (i) and (ii) above. Case (ii) will be developed in detail, but we will only summarize the results for case (i). The next lemma gives the density $\tilde{f}$ in a form that will be convenient in this work. The indicator function of the set satisfying condition $[\cdot]$ is denoted by $I[\cdot]$.

LEMMA 2.1. If $X_1, \cdots, X_n$ are i.i.d. rv's with a common multivariate normal distribution $P \in \mathcal{P}$, the MVU estimator $\tilde{f}$ of the corresponding normal probability density function is

$$(2.1)\qquad \tilde{f}_n(x) = \frac{[n/(n-1)]^{(n-3)/2}\Gamma[(n-1)/2]}{\pi^{k/2}\,\Gamma[(n-k-1)/2]}|S_n|^{-\frac{1}{2}}$$

$$\cdot\{[(n-1)/n] - (x-\bar{X}_n)'S_n^{-1}(x-\bar{X}_n)\}^{\frac{1}{2}(n-k-3)}$$

$$\cdot I[(x-\bar{X}_n)'\,S_n^{-1}(x-\bar{X}_n) \le (n-1)/n],\ n > k+1\ .$$

PROOF. This result is immediate from Ghurye and Olkin (1969) and the two facts:

(a) For $B(k \times k)$ nonsingular and $x(k \times 1)$, $|B-xx'| = |B|(1-x'B^{-1}x)$.

(b) If B is p.d. then $B-xx'$ is p.d. iff $x'Bx < 1$.

LEMMA 2.2. Suppose Y is a rv which has for fixed T_n the conditional density function $\tilde{f}_n$ of (2.1). Then for

$$Z_n = A_n(Y-\bar{X}_n)/\{((n-1)/n) - (Y-X_n)'S_n^{-1}(Y-\bar{X}_n)\}^{\frac{1}{2}}\ ,$$

where $A_n'A_n = S_n^{-1}$, the conditional density function of Z_n given T_n is

$$(2.2)\quad \tilde{g}(z) = \Gamma(\tfrac{1}{2}(n-1))\{\pi^{\frac{1}{2}k}\,\Gamma(\tfrac{1}{2}(n-k-1))\}^{-1}\,\{1+z'z\}^{-\frac{1}{2}(n-1)},\ n > k+1\ .$$

PROOF. If $z = A_n(y-\bar{X}_n)/\{((n-1)/n) - (y-\bar{X}_n)'\,S_n^{-1}(y-\bar{X}_n)\}^{\frac{1}{2}}$,

then $y = A_n^{-1}z((n-1)/n)^{\frac{1}{2}}\,(1+z'z)^{-\frac{1}{2}} + \bar{X}_n$. The Jacobian is

$$\{((n-1)/n)/(1+z'z)\}^{+\frac{k}{2}}\,|A_n|^{-1}|I - zz'/(1+z'z)|\ ,$$

and using the relations

$$|I - zz'/(1+z'z)^{-1}| = (1+z'z)^{-1} \text{ and } (y-\bar{X}_n)'S_n^{-1}(y-\bar{X}_n) = ((n-1)/n)z'z/(1+z'z),$$

the result follows from Lemma 2.1.

The density function $\tilde{g}$ of (2.2) has the form of a generalized multivariate t distribution. Dickey (1967), Theorems 3.2 and 3.3, gives conditional and marginal distributions for generalized multivariate t distributions from which the conditional and marginal distributions of $\tilde{g}$ of (2.2) can be obtained. Let G_ν denote the distribution function of a univariate Student-t distribution with ν degrees of freedom. Then the following can be obtained from results given by Dickey.

LEMMA 2.3. Let $Z' = (Z_1, \cdots, Z_k)$ denote a vector rv with (conditional) probability density function $\tilde{g}(z)$ of equation (2.2). Then

$$(2.3) \qquad \tilde{P}(Z_i \le z_i | Z_1 = z_1, \cdots, Z_{i-1} = z_{i-1})$$

$$= G_{n-k+i-2}\{z_i[(n-k+i-2)/(1 + \sum_{j=1}^{i-1} z_j^2)]^{\frac{1}{2}}\} ,$$

for $i = 1, \cdots, k$.

Consider again the original sample $X_1, \cdots, X_n$, and put

$$Z_j = A_j(X_j - \bar{X}_j)/\{[(j-1)/j] - (X_j - \bar{X}_j)'S_j^{-1}(X_j - \bar{X}_j)\}^{\frac{1}{2}} ,$$

and denote $Z_j' = (Z_{1,j}, \cdots, Z_{k,j})$ for $j = k+2, \cdots, n$. Then the next theorem follows from Lemma 2.3 and a slight extension of Theorem 5.1 of O-Q.

THEOREM 2.1. The $k(n-k-1)$ random variables given by

$$(2.4) \quad U_{i,j} = G_{j-k+i-2}\{Z_{i,j}[(j-k+i-2)/(1 + Z_{1,j}^2 + \cdots + Z_{i-1,j}^2)]^{\frac{1}{2}}\} ,$$

for $j = k+2, \cdots, n$ and $i = 1, \cdots, k$; are i.i.d. $U(0,1)$ rv's.

We now summarize the results for case (i) when μ is unknown and $\Sigma = \sigma^2\Sigma_0$ for Σ_0 known. For $\bar{X}_n$ and S_n defined above put here

$$A'A = \Sigma_0^{-1}, \ s_n = \operatorname{tr} \Sigma_0^{-1}S_n ,$$

and

$$Z_j = A(X_j - \bar{X}_j)/\{[(j-1)s_j/j] - (X_j - \bar{X}_j)'\Sigma_0^{-1}(X_j - \bar{X}_j)\}^{\frac{1}{2}} ,$$

and denote $Z_j' = (Z_{1,j}, \cdots, Z_{k,j})$ for $j = 3, \cdots, n$.

THEOREM 2.2. For $X_1, \cdots, X_n$ i.i.d. from $N(\mu, \sigma^2 \Sigma_0)$, Σ_0 known, the $(n-2)k$ random variables given by

$$(2.5) \quad U_{i,j} = G_{[(j-2)k+i]}\{Z_{i,j}[((j-2)k+i)/(1 + Z_{1,j}^2 + \cdots + Z_{i-1,j}^2)]^{\frac{1}{2}}\} ,$$

for $j = 3, \cdots, n$ and $i = 1, \cdots, k$; are i.i.d. $U(0,1)$ rv's.

3. Discussion. After the multivariate sample $X_1, \cdots, X_n$ has been transformed by using either (2.4) or (2.5), then a level α goodness-of-fit test for the corresponding composite multivariate normal null hypothesis class (case (i) or case (ii)) can be made by testing the surrogate simple null hypothesis that the transformed values are i.i.d. $U(0,1)$. Quesenberry and Miller (1977) and Miller and Quesenberry (1977) have studied power properties of omnibus tests for uniformity and recommend either the Watson U^2 test (Watson (1962)) or the Neyman smooth test (Neyman (1937)) for testing simple uniformity.

REFERENCES

Dickey, J. M. (1967). Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. Ann. Math. Statist. 38, 511-18.

Ghurye, S. G. and Olkin, I. (1969). Unbiased estimation of some multivariate probability densities and related functions. Ann. Math. Statist. 40, 1261-71.

Miller, F. L., Jr. and Quesenberry, C. P. (1977). Power studies of tests for uniformity, II. Commun. Statist. submitted.

Moore, D. D. (1976). Recent developments in chi-square tests for goodness-of-fit. Department of Statistics, Purdue University, Mimeograph Series #459.

Neyman, Jerzy (1937). "Smooth" test for goodness-of-fit. Skandinavisk Aktuarietidskrift **20**, 149-99.

O'Reilly, F. J. and Quesenberry, C. P. (1973). The conditional probability integral transformation and applications to obtain composite chi-square goodness-of-fit tests. Ann. Statist. **1**, 74-83.

Quesenberry, C. P. and Miller, F. L., Jr. (1977). Power studies of some tests for uniformity. J. Statist. Comput. Simul. **5**, 169-91.

Watson, G. S. (1962). Goodness-of-fit tests on a circle. II. Biometrika **49**, 57-63.