Ann. Inst. Statist. Math.
Vol. 42, No. 1, 149-155 (1990)

NULL DISTRIBUTION OF THE SUM OF SQUARED z-TRANSFORMS IN TESTING COMPLETE INDEPENDENCE

SHANDE CHEN AND GOVIND S. MUDHOLKAR

Department of Statistics, University of Rochester, Rochester, NY 14627, U.S.A.

(Received December 16, 1988; revised May 9, 1989)

Abstract. Brien *et al.* (1984, *Biometrika*, **71**, 545–554; 1988, *Biometrika*, **75**, 469–476) have proposed, illustrated and discussed advantages of using Fisher's z-transforms for analyzing correlation structures of multinormal data. Chen and Mudholkar (1988, *Austral. J. Statist.*, **31**, 105–110) have studied the sum of squared z-transforms of sample correlations as a test statistic for complete independence. In this paper Brown's (1987, *Ann. Probab.*, **15**, 416–422) graph-theoretic characterization of the dependence structure of sample correlations is used to evaluate moments of the test statistic. These moments are then used to approximate its null distribution accurately over a broad range of parameters, including the case where the population dimension exceeds the sample size.

Key words and phrases: Approximation, correlation analysis, dependence among sample correlations.

1. Introduction

It is obvious that the use of prior information regarding the covariance structure can improve the quality of multivariate data analysis. If the components are independent, then they can be analyzed separately using univariate methods and the results combined. In the presence of intermediate structures such as compound symmetry or sphericity, the analyses based upon specialized methods (see e.g., Arnold (1973)) are preferable to using general purpose multivariate methods. Problems involving hypotheses about covariance structures are therefore important. Using likelihood ratios is a common approach for testing hypotheses regarding covariance patterns. But as Brien *et al.* ((1984), see also (1988)) observe, these procedures often obscure intuitive detail and can be computationally costly. Moreover, they are meaningful only if the sample size exceeds the population dimension. Thus, for the simplest of these problems, testing complete independence in a p-dimensional normal population, the likelihood ratio is a function of the determinant of the sample correlation

matrix $\boldsymbol{R} = (r_{ij})$, which equals zero if the sample size n is no greater than p. Alternative test statistics such as $n \sum r_{ij}^2$ due to Nagao (1973) and

$$T = (n-3) \sum_{i<j} z_{ij}^2 , \tag{1.1}$$

where $z_{ij} = \tanh^{-1}(r_{ij})$, due to Chen and Mudholkar (1989), are reasonable for arbitrary n, but their null distributions are known only asymptotically.

In this paper we investigate the null distribution of T using a mix of analytical and numerical techniques so as to render T usable for practically relevant p and n. Specifically, we employ Brown's (1987) graph-theoretic characterization of the dependence structure of r_{ij}'s together with numerical integration techniques to obtain the first three moments of T. These moments are then used to construct some approximations for the null distribution of T, which are evaluated using Monte Carlo experiments. On the basis of these evaluations and consideration of simplicity, two of these approximations are seen to be practical and reasonably accurate for use over a broad range of p and n.

2. The moments of T

Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be a random sample from a p-dimensional normal population with dispersion matrix $\boldsymbol{V}$. Let r_{ij}'s be the sample correlation coefficients, z_{ij}'s their Fisher transforms, and consider the null hypothesis H_0: $\boldsymbol{V}$ is diagonal.

The likelihood ratio test which rejects H_0 for small values of the determinant $|\boldsymbol{R}|$ of the correlation matrix is impractical when $n \leq p$. Nagao (1973) has proposed $n \sum r_{ij}^2$ as a statistic for testing H_0, obtained an asymptotic expansion for its null distribution, and has shown it to be useful for $p = 3$ and $n \geq 100$. Appealing to the famous near-normality of the Fisher transforms of correlation coefficients, Chen and Mudholkar (1989) proposed and examined T as defined in (1.1) for testing H_0, and showed that its asymptotic $\chi^2_{p(p-1)/2}$ distribution (see Brien *et al.* (1984)) is adequate for moderate values of n, e.g., $n = 20$, provided p is small, e.g., ≤ 5. In order to improve upon this large sample approximation, we need at least the first three moments of T.

It is well known that if $\boldsymbol{V}$ is diagonal, then r_{ij}'s and consequently z_{ij}'s are pairwise independent (see e.g., Anderson (1984), p. 282). Hence we can use the marginal distribution of z to obtain the mean and variance of T. However, pairwise independence is not enough to obtain an expression for the third moment of T. Brown's (1987) analysis of the dependence structure of the r_{ij}'s is useful in this context. His result, which is in terms of the cycles in a graph, with edges corresponding to r_{ij}'s, is now outlined.

Let the vertices $1, 2,\ldots,p$ in a graph correspond to the components of a

random p-vector, and let its edges correspond to the $p(p-1)/2$ correlation coefficients of a random sample from the p-dimensional population. The subgraph corresponding to a subset of $\{r_{ij}, 1 \le i < j \le p\}$ or of z_{ij}'s, consists of corresponding $\{i_1, j_1, \ldots, i_k, j_k\}$ as its vertices, ignoring the repetitions, and $\{\overline{i_1 j_1}, \ldots, \overline{i_k j_k}\}$ as its edges. Then we have the following:

LEMMA 2.1. (Brown (1987)) *If the p components of a multivariate normal population are independent, then a subset of $\{z_{ij}, 1 \le i < j \le p\}$ is mutually independent if and only if the corresponding subgraph has no cycle.*

Now notice that if the dispersion matrix V is diagonal, the joint distribution of the $p(p-1)/2$ Fisher transforms z_{ij}'s of the correlation coefficients are invariant under permutations. Let $\alpha_1 = (n-3)E(z_{12}^2)$, $\alpha_2 = (n-3)^2 \operatorname{var}(z_{12}^2)$, $\alpha_3 = (n-3)^3 E\{(z_{12}^2 - E(z_{12}^2))^3\}$ and $\alpha_{123} = (n-3)^3 E(z_{12}^2 z_{13}^2 z_{23}^2) - \alpha_1^3$. Then we have the following:

THEOREM 2.2. *Under the null hypothesis that V is diagonal,*

$$\mu_1 \equiv E(T) = \binom{p}{2} \alpha_1 , \tag{2.1}$$

$$\mu_2 \equiv \operatorname{var}(T) = \binom{p}{2} \alpha_2 , \tag{2.2}$$

$$\mu_3 \equiv E\{(T-\mu_1)^3\} = \binom{p}{2} \alpha_3 + 6 \binom{p}{3} \alpha_{123} . \tag{2.3}$$

PROOF. (2.1) and (2.2) follow from the fact that z_{ij}'s are identically distributed and pairwise independent.

Now, let $U_{ij} = (n-3)z_{ij}^2 - \alpha_1$ and organize them lexicographically as $V_1, V_2, \ldots, V_{p(p-1)/2}$. Then V_i's are identically distributed, pairwise independent and $E(V_i) = 0$. Hence

$$\begin{aligned} E\{(T-\mu_1)^3\} &= E\left\{\left(\sum_i V_i\right)^3\right\} \\ &= E\left(\sum_i V_i^3 + 3\sum_{i \ne j} V_i^2 V_j + 6 \sum_{i<j<k} V_i V_j V_k\right). \end{aligned} \tag{2.4}$$

But $E(V_i^2 V_j) = 0$, $i \ne j$, as V_i's are pairwise independent with zero expectation. Also, $E(V_i V_j V_k) = 0$ if V_i, V_j and V_k are independent. When they are dependent, by Lemma 2.1, $E(V_i V_j V_k) = \alpha_{123}$, $i < j < k$. The number of such dependent triples V_i, V_j and V_k are $p(p-1)(p-2)/6$ since they must

correspond to a cycle U_{hl}, U_{lm} and U_{mh}. Hence we have (2.3), which completes the proof.

It is important to note that α_1, α_2, α_3 and α_{123} as defined above depend upon n only, and not upon p. α_1 and α_2 may be obtained from Gayen's (1951) expressions (see also Mudholkar (1983)) for the moments of z. However, for small values of n, e.g., $n = 10$, 15, these are not accurate enough for the present purpose. Hence the values of α_1, α_2, α_3 and α_{123} were obtained by numerical integration for $n = 8(1)50$. The following expressions for these quantities are constructed by using regression methods and Occam's razor in the light of asymptotic theory.

$$\alpha_1 \approx 1 - \frac{1}{3(n-1)^2} - \frac{7}{6(n-1)^3} - \frac{6}{(n-1)^4}, \tag{2.5}$$

$$\alpha_2 \approx 2 + \frac{2}{n-1} + \frac{8}{3(n-1)^2} - \frac{45}{(n-1)^4}, \tag{2.6}$$

$$\alpha_3 \approx 8 + \frac{24}{n-1} + \frac{187}{3(n-1)^2} + \frac{491}{3(n-1)^3} - \frac{644}{(n-1)^4}, \tag{2.7}$$

$$\alpha_{123} \approx \frac{4}{n-1} + \frac{4}{(n-1)^2} - \frac{165}{8(n-1)^3} - \frac{132}{7(n-1)^4}. \tag{2.8}$$

The errors in (2.5)–(2.8) are, respectively, within ± 0.00005, ± 0.00015, ± 0.0007 and ± 0.00005, except the errors in α_3 for $n = 8, 9, 10$ are approximately 0.00072, 0.0015 and 0.00084, respectively.

The first three moments of T can be obtained by substituting (2.5)–(2.8) in (2.1)–(2.3). These are used in the following section for constructing some simple approximations for the null distribution of T.

3. Approximations

Under the null hypothesis of complete independence, T is asymptotically distributed as a $\chi^2_{p(p-1)/2}$ variable. Hence in moderate-size samples, T may be approximated by three parameter random variables such as (i) $a\chi^2_\nu + b$, (ii) $cF(\nu_1, \nu_2)$ or (iii) $k(\chi^2_\nu)^h$, which agree with it in the first three moments. These, and some other, approximations were developed and studied in the present context, and the first two together were found to be appropriate to the purpose. They are now discussed:

Approximation 1. The idea of using $a\chi^2_\nu + b$, a translate of a gamma variable was proposed by Pearson (1959) for approximating the non-central χ^2 distribution. For approximating the distribution of T in this way,

the constants a, v and b, determined by equating the first three moments, are given by:

$$
\begin{aligned}
v &= 8\mu_2^3/\mu_3^2 , \\
a &= \mu_3/4\mu_2 , \\
b &= \mu_1 - 2\mu_2^2/\mu_3 ,
\end{aligned}
\tag{3.1}
$$

where μ_1, μ_2 and μ_3 are as in (2.1)–(2.3).

Approximation 2. Since T is an asymptotic χ^2-variable, in moderate size samples it may be approximated by a multiple $cF(\nu_1, \nu_2)$ of a variance ratio F, where the parameters c, ν_1 and ν_2 are given by

$$
\begin{aligned}
c &= \{2(1 + \theta - \phi)/(2 + 3\theta - 4\phi)\}\mu_1 , \\
\nu_1 &= 4(1 + \theta - \phi)/(\theta\phi - \theta + 4\phi) , \\
\nu_2 &= 4 + 2(\theta + 2)/(\theta - 2\phi) ,
\end{aligned}
\tag{3.2}
$$

where $\theta = \mu_3/(\mu_1\mu_2)$, $\phi = \mu_2/\mu_1^2$ and μ_1, μ_2, μ_3 are given by (2.1)–(2.3).

Evaluation. Evidently, the first approximation is simpler to use than the second if the tables of percentiles are used. When computers are available, it is very easy to write micros for either of the approximations in packages such as MINITAB and SAS. However, the parameter b in the first approximation is generally positive, implying that the approximation assigns zero probability to an interval of positive values of T. On the other hand, for $p > n$, the values of ν_2 in the second approximation become negative. The purpose of studying these two approximations is therefore to evaluate their quality in the presence of these anomalies. This was done using a Monte Carlo study. For certain combinations (p, n), for p in the range from 3 to 30 and $n = 10$, 15, 20, $N = 50{,}000$ samples, each of size n from a p-variate normal population with an identity dispersion matrix, were generated by use of NAG subroutine *g05ddf*, and the statistic T was computed for each sample. The empirical distribution functions of 50,000 simulations of T for each combination (p, n) were then used to estimate the missing probabilities pr $(T \leq b)$ and to evaluate the accuracy of the tail probabilities given by the two approximations. A selection of the results is given in Tables 1 and 2. From the results of the Monte Carlo study and other considerations, we recommend that Approximation 2, $T \approx cF(\nu_1, \nu_2)$ should be employed when $p \leq \min\{n, 10\}$. Otherwise, Approximation 1, $T \approx a\chi_v^2 + b$ is appropriate, especially if $p > n$.

Table 1. Empirical tail probabilities and missing probabilities pr $(T \leq b)$ using quantiles from the approximation $T \approx a\chi_\nu^2 + b$.

n	p	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$	b	pr $(T < b)$
10	3	0.1800	0.0490	0.0099	0.0048	0.0012	0.551	0.0940
	5	0.1026	0.0515	0.0097	0.0047	0.0010	2.883	0.0149
	10	0.0957	0.0480	0.0098	0.0049	0.0010	20.713	0.0002
	15	0.0993	0.0493	0.0101	0.0047	0.0011	59.306	0.0000
	20	0.0987	0.0492	0.0101	0.0053	0.0011	120.666	0.0000
	25	0.0980	0.0494	0.0103	0.0054	0.0012	205.666	0.0000
	30	0.1000	0.0505	0.0105	0.0049	0.0012	314.749	0.0000
15	3	0.1473	0.0507	0.0103	0.0054	0.0011	0.381	0.0559
	5	0.0988	0.0490	0.0104	0.0056	0.0013	2.146	0.0040
	10	0.0994	0.0480	0.0103	0.0056	0.0014	16.735	0.0000
	15	0.0990	0.0503	0.0107	0.0056	0.0012	50.037	0.0000
	20	0.0999	0.0511	0.0108	0.0055	0.0012	104.730	0.0000
	25	0.1005	0.0509	0.0102	0.0052	0.0010	182.149	0.0000
	30	0.0998	0.0503	0.0104	0.0052	0.0010	283.035	0.0000
20	3	0.1303	0.0482	0.0100	0.0055	0.0014	0.290	0.0380
	5	0.0978	0.0494	0.0099	0.0054	0.0013	1.700	0.0016
	10	0.0985	0.0493	0.0106	0.0055	0.0012	13.943	0.0000
	15	0.0962	0.0481	0.0096	0.0048	0.0010	42.979	0.0000
	20	0.0979	0.0475	0.0095	0.0049	0.0012	91.902	0.0000
	25	0.0974	0.0496	0.0100	0.0053	0.0011	162.423	0.0000
	30	0.0987	0.0485	0.0098	0.0054	0.0014	255.560	0.0000
SE		0.0013	0.0010	0.0004	0.0003	0.0001	—	—

The empirical probabilities are based upon 50,000 simulations.
SE = $\sqrt{\alpha(1-\alpha)/50000}$, the standard error of the estimated tail probabilities.

Table 2. Empirical tail probabilities using quantiles from the approximation $T \approx cF(\nu_1, \nu_2)$.

n	p	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
10	3	0.0954	0.0494	0.0100	0.0053	0.0013
	5	0.0977	0.0484	0.0103	0.0056	0.0011
	7	0.0973	0.0481	0.0099	0.0050	0.0009
	10	0.0999	0.0500	0.0103	0.0054	0.0012
	11	0.1003	0.0515	0.0105	0.0054	0.0011
15	3	0.0983	0.0493	0.0104	0.0052	0.0011
	5	0.1000	0.0488	0.0100	0.0053	0.0012
	7	0.0996	0.0502	0.0110	0.0054	0.0012
	10	0.1002	0.0501	0.0107	0.0051	0.0012
	15	0.1004	0.0507	0.0099	0.0049	0.0011
	16	0.0997	0.0504	0.0106	0.0056	0.0012
20	3	0.0987	0.0493	0.0103	0.0046	0.0009
	6	0.1007	0.0496	0.0096	0.0053	0.0013
	10	0.0996	0.0502	0.0107	0.0054	0.0011
	13	0.0994	0.0496	0.0094	0.0046	0.0009
	20	0.1014	0.0503	0.0096	0.0047	0.0011
	21	0.1004	0.0499	0.0096	0.0043	0.0011
SE		0.0013	0.0010	0.0004	0.0003	0.0001

The empirical probabilities are based upon 50,000 simulations.
SE = $\sqrt{\alpha(1-\alpha)/50000}$, the standard error of the estimated tail probabilities.

Acknowledgements

The authors are thankful to the referee for his comments, and to Professor M. S. Srivastava for some discussions.

REFERENCES

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Arnold, S. F. (1973). Application of the theory of products of problems to certain patterned covariance matrices, *Ann. Statist.*, **1**, 682–699.

Brien, C. J., Venables, W. N., James, A. T. and Mayo, O. (1984). An analysis of correlation matrices: Equal correlations, *Biometrika*, **71**, 545–554.

Brien, C. J., James, A. T. and Venables, W. N. (1988). An analysis of correlation matrices: Variables cross-classified by two factors, *Biometrika*, **75**, 469–476.

Brown, T. C. (1987). Independent subsets of correlation and other matrices, *Ann. Probab.*, **15**, 416–422.

Chen, S. and Mudholkar, G. S. (1989). A remark on testing significance of an observed correlation matrix, *Austral. J. Statist.*, **31**, 105–110.

Gayen, A. K. (1951). The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes, *Biometrika*, **38**, 219–247.

Mudholkar, G. S. (1983). Fisher's z-transformation, *Encyclopedia of Statistical Sciences*, (eds. S. Kotz, N. L. Johnson and C. B. Read), Vol. 3, 130–135, Wiley, New York.

Nagao, H. (1973). On some test criteria for covariance matrix, *Ann. Statist.*, **1**, 700–709.

Pearson, E. S. (1959). Note on an approximation to the distribution of non-central χ^2, *Biometrika*, **46**, 364.