

**EXACT SMALL SAMPLE SIGNIFICANCE POINTS
OF A DISTRIBUTION-FREE TEST FOR SYMMETRY
IN HIERARCHICAL DATA**

by

**Dr Tan Khye Chong
Division of Actuarial Science & Insurance
School of Accountancy & Business
Nanyang Technological University
Nanyang Avenue
Singapore 639798**

ABSTRACT

In Bhapkar and Gore (1973), a distribution-free test for the hypothesis of permutation invariance in hierarchical data was proposed. They showed that the test statistic S has an asymptotic chi-square distribution. In this note, the exact significance points of S for small samples are given for some selected cases. The results show that even for these small samples, the chi-square approximation is rather impressive.

1. Introduction

Suppose in an experiment a certain response to a particular treatment is measured on each subject at fixed intervals over a certain period of time say r periods. Clearly the set of measurements for each subject will be correlated. Here invariance of the joint distribution of all measurements on a subject under permutations among themselves will imply that the treatment is not effective to cause any change in the mean response over time. One of the difficulties in such experiments is that the subjects may drop out at various stages of the experiment before yielding all the observations planned. Under such possibilities, one has c_r vectors (each vector denoting a set of measurements on a specific subject) with all the r components when the subject participates in the complete experiment, c_{r-1} vectors with the last component missing when the subject fails to report for the last period, etc and c_2 vectors with only the first two components. It is assumed that a subject failing to report for measurement once does not return later (e.g. death of experimental subject). This type of data is called hierarchical because if the j th component is observed, then all the $(j-1)$ earlier components are also observed. It is reasonable to take $c_1 = 0$ because a subject reporting only for the first time does not give any information regarding the effect of the treatment and may be disregarded. Let

$$\sum_{m=2}^r c_m = N$$

where N is the total number of subjects involved in the experiment.

Let $\mathbf{x}_k^{(m)} = (X_{1k}, \dots, X_{mk})$ be an observation on a m-variate random vector with continuous distribution function $F_{mk}(x_1, \dots, x_m)$, $k = 1, 2, \dots, c_m$; $m = 2, 3, \dots, r$. Also, random vectors corresponding to different subjects are assumed to be independent. The test proposed by Bhapkar and Gore is for the hypothesis of permutation invariance, namely

$$(1.1) \quad H_0: F_{mk}(x_1, \dots, x_m) = F_{mk}(x_{j_1}, \dots, x_{j_m}) \\ k = 1, \dots, c_m; \quad m = 2, \dots, r$$

where (j_1, \dots, j_m) is any permutation of the first m positive integers. Define

$$(1.2) \quad U_i = \sum_{m=i}^r \sqrt{c_m} m(U_i^{(m)} - 1/m)$$

where $U_i^{(m)}$ is the proportion of m-dimensional vectors $\mathbf{x}_k^{(m)}$, $k = 1, \dots, c_m$ such that the ith component of the vector is largest among the m components if $c_m > 0$. Let $U_i^{(m)} = 0$ if $c_m = 0$ and let $\mathbf{U} = (U_1, \dots, U_r)N$. Then the test statistic S is given by

$$(1.3) \quad S = \mathbf{U}' \Sigma^* \mathbf{U}$$

where Σ^* is the generalised inverse of the covariance matrix of \mathbf{U} .

When the null hypothesis (1.1) is true, the expectation of $U_i^{(m)}$ is $1/m$. Thus S may be taken as a measure of the deviation from the null hypothesis. The test consists of rejecting the null hypothesis at a level α if S exceeds a predetermined constant S_α .

Bhaskar and Gore have shown that S has an asymptotic chi-square distribution with degrees of freedom equal to $\rho(3)$, the rank of 3. Hence, for large sample sizes, S_α may be approximated by the upper 100α percent point of the chi-square distribution with $\rho(3)$ degrees of freedom. In the following section, the exact significance points for small sample sizes will be obtained for some special values of r and c_i 's.

2. Numerical computation of S

When $r = 3$, $c_2 > 0$, the statistic (1.3) can be expressed as

$$(2.1) \quad S = (U_1^2 + U_2^2 + 2U_3^2)/5$$

and when $r = 4$, c_2 and c_3 are both greater than 0, the expression for the test statistic becomes

$$(2.2) \quad S = [7(U_1^2 + U_2^2) + 10U_3^2 + 19U_4^2 + 2U_3U_4]/63$$

From (1.2) and the definition of $U_i^{(m)}$, it follows that

$$\sum_{i=1}^r U_i = 0$$

so that (2.1) and (2.2) can be rewritten as

$$(2.3) \quad S = (2U_2^2 + 3U_3^2 + 2U_2U_3)/5$$

$$\text{and } S = [14U_2^2 + 17U_3^2 + 26U_4^2 + 14U_2(U_3 + U_4) + 16U_3U_4]/63$$

respectively.

Towards obtaining the exact significance points S_α for selected values of r and c_i 's we first compute the probability

function of S . Considering the case when $r = 3$, $c_2 = 2$, $c_3 = 1$ it follows from (2.3) that the values of S depends on U_2 and U_3 . But, from (1.2) it is clear that one only has to enumerate all possible values of $U_2^{(2)}$, $U_2^{(3)}$ and $U_3^{(3)}$ and the possible combinations; then the corresponding value of S can be calculated. Notice that

$$\sum_{i=1}^m U_i^{(m)} = 1$$

and since $c_3 = 1$, it follows that the only possible combinations of $U_2^{(3)}$ and $U_3^{(3)}$ are $(U_2^{(3)} = 0, U_3^{(3)} = 0)$, $(U_2^{(3)} = 1, U_3^{(3)} = 0)$ and $(U_2^{(3)} = 0, U_3^{(3)} = 1)$. [The first combination refers to the case when $U_1^{(3)} = 1$ or the first component of the only vector with three components is the largest; the second combination refers to the situation where the second component is the largest; the third possible combination is for the situation where the third component is the largest.] Each of these can be combined with the possible values of $U_2^{(2)}$ which are $(U_2^{(2)} = 0)$, $(U_2^{(2)} = 1/2)$ and $(U_2^{(2)} = 1)$ since $c_2 = 2$. [This is because in the 2-component vectors, either the second component is smaller than the first for both vectors giving $U_2^{(2)} = 0$, or the second component is larger than the first in only one of the two vectors giving $U_2^{(2)} = 1/2$.] Thus there are nine possible combinations.

To obtain $P(S = s)$ note that $U_2^{(2)}$ is independent of $U_i^{(3)}$ for all $i = 1, 2, 3$. Also, under H_0 ,

$$(2.4) \quad P(U_i^{(m)} = y/c_m) = \binom{c_m}{y} (1/m)^y (1 - 1/m)^{c_m - y}$$

where $y = 0, 1, \dots, c_m$. Observe that under H_0 , the probability

that the i th component of a m -dimensional vector is largest is $1/m$. Regarding this as the probability of a "success", equation (2.4) gives the probability of y successes out of c_m independent trials.

It can be easily checked that when $(U_2^{(2)} = 1/2, U_2^{(3)} = 0, U_3^{(3)} = 1)$ then $S = 2$. Hence by independence,

$$(2.5) \quad P(S = 2) = P(U_2^{(2)} = 1/2) P(U_2^{(3)} = 0, U_3^{(3)} = 1).$$

From (2.4), $P(U_2^{(2)} = 1/2) = 1/2$ and $P(U_1^{(3)} = 0) = 2/3$. Since the events $(U_2^{(3)} = 0, U_3^{(3)} = 1)$ and $(U_2^{(3)} = 1, U_3^{(3)} = 0)$ are equally likely, it follows that

$$P(U_2^{(3)} = 0, U_3^{(3)} = 1) = 1/3 = P(U_2^{(3)} = 1, U_3^{(3)} = 0).$$

Thus from (2.5),

$$P(S = 2) = 1/6 = 0.167.$$

The other probabilities are calculated in a similar fashion. When values of S are repeated, their probabilities are pooled.

3. Table of critical values of S

Given here are the values of S_α for $r = 3$ and $r = 4$ and for samples which satisfy the following conditions:

- (i) $c_i \# 4$ for $i = 2, \dots, r$
- (ii) $c_2 + \dots + c_r \# 8$
- (iii) $c_i \# c_j \# 4$ for $i \exists j, i, j = 2, \dots, r$.

For easier and clearer tabulation the following notation is

used. The pair (c_2, c_3) will refer to the situation where $r = 3$ and the triplet (c_2, c_3, c_4) will refer to the case where $r = 4$. For example, $(3,2)$ will indicate that $r = 3, c_2 = 3, c_3 = 2$ while $(4,3,1)$ indicates that $r = 4, c_2 = 4, c_3 = 3, c_4 = 1$.

It can be seen from Table 1 that even for these selected cases, the approximation with a chi-square distribution is impressive. For example for the case $(3,3,2)$ the upper 5 percent point is 7.215 which compares favourably with the chi-square value of 7.815 with 3 degrees of freedom, the asymptotic distribution of the test statistic S when $r = 4$.

References

Bhapkar, V.P. and Gore, A.P. (1973) "A distribution-free test for symmetry in hierarchical data" **Journal of Multivariate Analysis**, Vol. 3, p483-489.

TABLE 1. Upper-tail probabilities of S: $P = P(S > s)$

(1,1)		(3,3)		(2,1,1)	
s	P	s	P	s	P
3.000	.333	3.133	.186	5.000	.168
		4.200	.130	5.102	.126
	(2,1)	4.800	.102	5.444	.084
		5.533	.074	6.867	.042
2.800	.333	6.133	.046		
3.897	.167	7.200	.018		(3,1,1)
		9.000	.009		
	(3,1)				
			(4,3)		
				s	P
				5.194	.126
				5.536	.063
				5.667	.042
3.200	.166			7.583	
4.678	.083	2.986	.188		
.021					
	(4,1)	4.200	.132		
		4.786	.104		(4,1,1)
		5.571	.076		
		6.000	.062		
		6.400	.048		
		6.678	.029	4.746	.155
3.000	.251	7.600	.010	5.000	.145
3.600	.084			5.222	.114
5.400	.042	9.957	.005	5.889	.072
				5.937	.062
	(2,2)		(4,4)	6.000	.052
				8.222	.010
					(2,2,1)
2.800	.279	3.600	.152		
4.000	.168	3.750	.137		
4.800	.112	4.350	.131		
6.000	.056	4.500	.082		
		5.400	.057		
				4.714	.140
	(3,2)			5.600	.048
.126					4.867
		6.150	.039	5.159	.098
		6.900	.033	5.725	.084
		7.350	.027	6.206	.056
		8.000	.021	6.804	.042
		8.400	.016	7.248	.028
		9.600	.004	8.760	.014
		12.000	.002		
			(1,1,1)		
	(4,2)				
		s	P		
		5.222	.166		
		6.000	.083		
4.000	.182				
4.400	.140				
4.897	.084				
5.600	.084				
7.794	.014				

TABLE 1: (Continued)

(3,2,1)		(3,3,1) contd.		(2,2,2) contd.	
s	P	s	P	s	P
4.788	.154	7.810	.020	6.325	.098
4.840	.133	8.299	.018	6.357	.091
5.046	.119	8.462	.011	6.429	.077
5.047	.105	9.054	.004	6.540	.070
5.381	.084	11.251	.002	6.873	.067
5.522	.077			7.000	.060
6.728	.063	(4,3,1)		7.214	
.053					
6.857	.056			7.302	.046
6.878	.035	s	P	7.429	.039
7.471	.014			7.468	.032
9.565	.007	4.646	.156	7.714	.025
		4.659	.135	8.159	.022
(4,2,1)		4.765	.121	8.444	.019
s	P	4.862	.118	9.206	.012
		4.952	.117	10.000	.009
4.444	.140	5.143	.114	10.444	.006
4.714	.137	5.219	.107	12.000	.003
		5.224	.106		
4.854	.127	5.253	.101	(3,2,2)	
4.937	.120	5.468	.087		
5.189	.106	5.571	.080	s	P
5.507	.099	5.629	.077		
	5.594	.092	5.568	.074	
5.397	.155				
5.603	.078	5.690	.060	5.479	.145
5.725	.075	5.794	.046	5.503	.143
6.109	.054	6.095	.041	5.561	.133
6.804	.047	6.305	.038	5.578	.123
7.026	.037	6.460	.035	5.616	.113
7.203	.023	6.520	.034	5.914	.103
7.693	.020	7.045	.029	6.095	.100
7.779	.017	7.609	.022	6.159	.097
10.278	.003	8.016	.019	6.207	.087
		8.341	.016	6.300	.084
(3,3,1)		8.388	.015	6.462	.081
s	P	8.610	.012	6.503	.079
		9.276	.007	6.697	.069
4.841	.155	9.311	.006	6.703	.066
4.942	.134	12.021	.001	6.880	.056
4.982	.113			6.958	.053
5.066	.092	(2,2,2)		7.095	.048
5.143	.085			7.248	.045
5.646	.071	s	P	7.622	.040
5.667	.064			7.788	.037
5.884	.057	5.214	.157	7.840	.032
5.939	.050	5.429	.143	7.986	.029
5.979	.043	5.492	.136	8.023	.026
6.238	.036	5.778	.129	8.047	.023
7.045	.034	5.873	.119	8.381	.018
		6.135	.112	9.728	.016

7.349

.027

6.159

.105

9.789

.014

TABLE 1: (Continued)

(3,2,2) contd		(3,3,2)	
s	P	s	P
10.074	.009	4.971	.158
10.667	.004	4.984	.151
12.921	.002	5.481	.148
		5.630	.145
(4,2,2)		5.710	.140
		5.712	.137
s	P	5.724	.134
		5.777	.133
5.214	.159	5.822	.126
5.286	.149	6.000	.125
5.429	.142	6.050	.123
5.632	.137	6.074	.120
5.651	.136	6.146	.110
5.700	.129	6.296	.103
5.908	.126	6.381	.098
6.038	.119	6.504	.096
6.108	.118	6.661	.094
6.257	.111	6.667	.084
6.276	.104	6.810	.081
6.303	.097	6.873	.077
6.305	.095	6.881	.074
6.317	.093	7.028	.071
6.357	.091	7.106	.070
6.428	.081	7.175	.060
6.429	.074	7.185	.055
6.432	.069	7.215	.050
6.467	.068	7.295	.048
6.651	.065	7.323	.043
6.708	.058	7.466	.041
6.775	.055	7.526	.040
7.046	.053	7.550	.035
7.051	.051	7.620	.033
7.317	.044	7.638	.028
7.429	.042	8.133	.027
7.524	.037	8.143	.026
7.714	.035	8.646	.023
7.937	.032	8.657	.021
8.181	.029	8.667	.019
8.189	.027	8.697	.017
8.444	.025	8.884	.015
8.594	.020	9.238	.013
8.603	.017	9.533	.012
8.667	.016		
8.671	.014	10.015	.011
10.000	.012	10.228	.009
10.203	.009	10.810	.007
10.222	.008	11.481	.006
10.867	.005	11.959	.004
10.889	.002	12.551	.002
13.733	.001	14.859	.001