Two-Sample Multivariate Test of Homogeneity

اختبار متعدد الأبعاد للتجانس

Ali S. Barakat

Dept. of Statistics, Faculty of Science, An-Najah National University, Nablus, Palestine.

E-mail: barakat@najah.edu

Received: (13/11/2000), Accepted: (4/3/2003)

Abstract

Given independent multivariate random samples $X_1, X_2, ..., X_{n_1}$ and $Y_1, Y_2, ..., Y_{n_2}$ from distributions F and G, a test is desired for H_o: F = G against general alternatives. Consider the $k \bullet (n_1+n_2)$ possible ways of choosing one observation from the combined samples and then one of its k nearest neighbors, and let S_k be the proportion of these choices in which the point and neighbor are in the same sample. SCHILLING proposed S_k as a test statistic, but did not indicate how to determine k. BARAKAT, QUADE, and SALAMA proposed a test statistic $W = N\sum kS_k$, which is equivalent to a sum of N Wilkoxon rank sums. The limiting distribution of the test has not been found yet.

We suggest as a test statistic $T_m = \Sigma\ \Sigma h(m,j)$ و

Where h (m,j) = I{j^{th} nearest neighbor of the median m is a y}.

The limiting distribution of T_m is normal. A simulation with multivariate normal data suggests that our test is generally more powerful than Schilling's test using k = 1, 2 or 3.

ملخص

لقد قدم Schilling اختبارا للتجانس باستخدام المسافة بين نقطة البداية والنقاط القريبة منها وكان عدد هذه النقاط محدودا. تم تقديم اختبار آخر من قبل بركات وقويد وسلامه آخذين بعين الاعتبار موقع النقطة القريبة واستخدام كل النقاط القريبة وليس عددا محدودا منها ولكن هذا الاختبار لم يعرف توزيعه حتى الآن. في هذا البحث نقترح اختبارا للتجانس وذلك بدءا بالنقطة التي تمثل الوسيط ومن ثم النقاط القريبة منها مرتبة حسب المسافة والأخذ بعين الاعتبار موقع النقطة ولقد تم إثبات أن توزيع هذا الاختبار هو التوزيع الطبيعي، ويتميز كذلك بأنه أقوى من اختبار Shchilling.

1. Introduction

Let X_1, X_2,......, X_{n_1} and Y_1, Y_2, ..., Y_{n_2} be two independent random samples in R^d, from distributions F and G, respectively. The problem under consideration is to test the hypothesis H_o: F = G, against the general alternative H_a: $F \neq G$.

Let Z_1, Z_2, ..., Z_N, where $N = n_1+n_2$, is the combined sample such that

$$Z_i = \begin{cases} X_i & \text{if} \quad i = 1,2,\ldots,n_1 \\ Y_{i-n_1} & \text{if} \quad i = n_1+1, n_1+2,\ldots..N \end{cases}$$

Let $\| . \|$ be the Euclidean norm, and define "the" k-th nearest neighbor to Z_i as that point Z_j satisfying $\|Z_{j'} - Z_i\| \prec \|Z_j - Z_i\|$ for exactly (k-1) values of j' $(1 \leq j' \leq N,\ j' \neq i, j)$; we assume that there will be no ties.

Interest in statistical procedures based on such nearest neighbors has grown as high-speed computers have made the application of these techniques practicable, since the idea of making inferences about an object based on nearby objects appears to be a fundamental mechanism of human perception.

Schilling's approach [1] is as follow. Let:

h(i,j) = I{k-th nearest neighbor, Z_j, of Z_i and Z_i are from different samples} for k = 1, 2, ..., N-1 where I {E} is the indicator function of the event E and $N=n_1+n_2$. Count the number of k-nearest neighbor to Z_i which are in the same sample, viz.

$$T_{ik} = \sum_{j=1}^{k} [1 - h(i,j)]$$

Summing these counts over all observations yields what may be called "Schilling total", of order k:

$$T_k = \sum_i \sum_j [1 - h(i,j)] = \sum T_{ik}$$

His test statistic is

$$S_k = T_k \ / \ NK$$

which is the proportion of all k-nearest neighbor comparisons in which a point and its neighbor are members of the same sample. Schilling shows that the asymptotic distribution of S_k under H_o is normal.

Schilling's work suggests that the choice of order is not of great importance; nevertheless, it is arbitrary, and he gives no guidance for chosing it. BARAKAT, QUADE, and SALAMA [2] proposed the sum of the Schilling totals as a test statistic, which is equivalent to a certain weighted average of the Schilling proportions:

$$W = \sum T_k = N\sum kS_k$$

Simulations with multivariate normal data show that W is generally more powerful than S_k using k=1,2, or 3. The asymptotic distribution of W has not been known.

We propose the following test statistic:

$$T_m = \sum_{k=1}^{N} \sum_{j=1}^{k} h(m,j) = \sum_{k} T_{mk}$$

where m is the median of the combined sample and

h(m,j) = I{j-th nearest neighbor of the median is a y}.

In this test order is of great importance and all nearest neighbors to the median are used, i.e. this test uses all nearest neighbors to the median and it takes into account the position of each nearest neighbor to the median.

Under the alternative hypothesis, we expect T_m to have too small or too large values because of a lack of complete mixing of the two samples. Hence too small or too large values of T_m are significant.

2. Illustrative Example for Computing the Test Statistic T_m

Let $X_1 = (3,1,9)$, $X_2 = (2,5,8)$, and $X_3 = (4,6,1)$ be the first sample and $Y_1 = (5,9,4)$, $Y_2 = (1,10,6)$, $Y_3 = (2,3,5)$ and $Y_4 = (4,8,3)$ be the second sample. The combined sample is

$Z_1, Z_2, \ldots\ldots, Z_7$

where:

$$Z_i = \begin{cases} X_i & for \quad i = 1,2,3 \\ Y_{i-3} & for \quad i = 4,5,6,7 \end{cases}$$

Find the median for the combined sample m which is equal to (3,6,5). Then calculate $\|Z_i - m\|$, i = 1,2,....,7. The combined order arrangement of $\|Z_i - m\|$ from smallest to largest will give us the k-th nearest neighbor to m.

K	:	1	2	3	4	5	6	7
k-th nearest neighbor	:	Z_7	Z_6	Z_2	Z_4	Z_3	Z_5	Z_1
X or Y	:	Y	Y	X	Y	X	Y	X
h(m,k)	:	1	1	0	1	0	1	0
T_{mk}	:	1	2	2	3	3	4	4

Therefore,

$$T_m = \sum_{k=1}^{7} T_{mk} = 1+2+........+4 = 19$$

3. The Null-Hypothesis Distribution of T_m

If the two samples are maximally seperated then we obtain

$$\max(T_m) = \frac{n_2(1+n_1+N)}{2}$$

and

$$\min(T_m) = \frac{n_2(n_2+1)}{2}$$

To obtain the expectation and the variance of T_m under H_o we need the following result:

Result 1: i. $E[h(m,j)] = \frac{n_2}{N}$

ii. $Var[h(m,j)] = \frac{n_1 n_2}{N}$

iii. $Cov[h(m,j), h(m,j')] = \frac{-n_1 n_2}{N^2(N-1)}$

Proof:

Since $\Pr[h(m,j) = S] = \left(\frac{n_2}{N}\right)^{s}\left(\frac{n_1}{N}\right)^{1-s}$, $s = 0,1$

is the Bernoulli distribution, then [3]

i. $E[h(m,j)] = \frac{n_2}{N}$

ii. $Var[h(m,j)] = \frac{n_2}{N} \bullet \frac{n_1}{N} = \frac{n_1 n_2}{N^2}$

iii. $E[h(m,j), h(m,j'), j \neq j'] = \Pr[h(m,j) = 1 \cap h(m,j') = 1]$

$$= \frac{\binom{n_2}{2}}{\binom{N}{2}}$$

$$= \frac{n_2(n_2 - 1)}{N(N-1)}$$

so,

$$Cov[h(m,j), h(m,j')] = \frac{-n_1 n_2}{N^2(N-1)}$$

Result 2: i. $E(T_m) = \frac{n_2(N+1)}{2}$

ii. $Var(T_m) = \frac{n_1 n_2 (N+1)}{12}$

Proof:

i. $E(T_m) = E\left[\sum_{k=1}^{N} T_{mk}\right] = E\left[\sum_{k=1}^{N}\sum_{j=1}^{k} h(m,j)\right] = \sum_{k=1}^{N} k\frac{n_2}{N} = \frac{n_2(N+1)}{2}$

ii. $$\mathrm{Var}[T_m] = \mathrm{Var}\left[\sum_{k=1}^{N}\sum_{j=1}^{k} h(m,j)\right]$$

$$= \mathrm{Var}\left[\sum_{j=1}^{N}(N-j+1)h(m,j)\right]$$

$$= \sum_{j}(N-j+1)^2 \mathrm{Var}[h(m,j)] + \sum\sum_{j\neq j'} (N-j+1)(N-j+1)\mathrm{Cov}[h(m,j),h(m,j')]$$

$$= \frac{n_1 n_2}{N^2}\sum(N-j+1)^2 - \frac{n_1 n_2}{N^2(N-1)}\sum\sum_{j\neq j'} (N-j+1)(N-j'+1)$$

$$= \frac{n_1 n_2}{N^2(N-1)}\left[N\sum_{j}(N-j+1)^2 - (\sum_{j}(N-j+1))^2\right]$$

$$= \frac{n_1 n_2}{N^2(N-1)}\left[\frac{N^2(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4}\right]$$

$$= \frac{n_1 n_2 (N+1)}{12}$$

Rresult 3: Under H_o, $T_m = \sum_{k=1}^{N} T_{mk}$ has the Wilcoxon-Mann-Whitney distribution.

Proof:

$$T_m = \sum_{k=1}^{N} T_{mk} = \sum_{k=1}^{N}\sum_{j=1}^{k} h(m,j)$$

$$= \sum_{j=1}^{N}(N = j+1)h(m,j)$$

$$= n_2(N+1) - W^*$$

where $W^* = \sum_{j=1}^{N} jh(m,j)$ is the Wilcoxon rank sum statistic. So, T_m has a Wilcoxon-Mann-Whitney distribution [4].

Result 4: Under H_o, T_m has an asymptotic distribution which is normal.

Proof: From result 3, since a linear relationship exists between T_m and W^*, the Wilcoxon rank sum statistic, the properties of the tests are the same, including normality, consistency and the minimum ARE of .864 relative to the t-test [4].

4. Monte Carlo Estimation of the Power of T_m

In this section we consider the power of the test based on T_m, Schilling's test for k=1,2, and 3 and Barakat's test W, against a location shift. Our procedure is similar to that used by BARAKAT, QUADE and SALAMA [2].

The power depends on the following factors:

1. The type I error: We set $\alpha = 0.05$
2. The sample sizes: We chose $n_1 = n_2 = 10$ and 50.
3. The number of dimensions: We chose d = 2,5, and 10
4. The common distribution of the two populations under H_o: We considered only the multivariate normal distributions, then compared the power of the nearest neighbor tests with that of Hotelling's T^2 test.
5. The magnitude and direction of the shift: For $\rho = .36$ we considered two directions, a "same direction shift" (SDS) and an "opposite direction shift" (ODS). The magnitude of the shift was calculated so as to give power .70 or .90 using Hotelling's T^2-test. We also used the SDS for $\rho = 0.0$ (but in this case there is essentially no difference). For each combination of sample size and dimension we generated 1000 sets of N($= n_1+n_2$) d-dimensional multivariate normal observations using the IMSL programs RNMUN.

To compute the power we added the appropriate shift values to the last n members of each set of the generated sets of N d-dimensional multivariate numbers for each combination of sample size, dimension, correlation and shift, producing two samples differing by a shift, and calculated the five test statistics. The estimated power of any test statistic is then the proportion of the 1000 pairs of samples for which it exceeded its critical value.

Our results are shown in Table 1. As we expected, the power of the new test T_m is, in every case, at least as large as that of BARAKAT, QUADE, and SALAMA (1996) which is more powerful than that of Schilling's test S_k for k=1,2, and 3.

Table (1): Estimated power of the (S_k, k=1,2,3), W, and T_m tests for multinormal Data. Two samples of size n, in d dimensions, with common correlation ρ

n	ρ	d	Hotelling's T^2 power =.70					Hotelling's T^2 power =.90				
			S_1	S_2	S_3	W	T_m	S_1	S_2	S_3	W	T_m
10		2	.248	.426	.495	.723	.839	.444	.691	.732	.901	.930
	.00	5	.353	.560	.610	.845	.780	.573	.796	.837	.963	.926
		10	.699	.784	.838	.962	.944	.902	.960	.967	.999	.988
		2	.269	.455	.520	.762	.819	.466	.700	.758	.922	.932
	.36	5	.593	.685	.795	.924	.986	.814	.885	.937	.989	.991
	SDS	10	.784	.929	.966	.995	1.000	.945	.994	.997	1.000	1.000
		2	.252	.410	.444	.650	.678	.438	.640	.701	.871	.868
	.36	5	.461	.497	.595	.674	.615	.669	.739	.814	.916	.900
	ODS	10	.487	.693	.747	.852	.875	.760	.908	.936	.984	.942
50		2	.144	.231	.276	.746	.730	.205	.364	.458	.898	.862
	.00	5	.188	.227	.280	.726	.641	.290	.375	.458	.917	.741
		10	.205	.264	.298	.769	.722	.319	.430	.474	.932	.806
		2	.175	.227	.290	.769	.766	.258	.380	.462	.923	.992
	.36	5	.187	.278	.350	.862	.783	.313	.463	.579	.972	.826
	SDS	10	.319	.406	.487	.950	.865	.468	.424	.717	.997	.890
		2	.157	.216	.268	.633	.607	.246	.350	.443	.852	.787
	.36	5	.184	.229	.288	.597	.756	.257	.371	.445	.842	.848
	ODS	10	.211	.227	.265	.540	.558	.312	.374	.418	.782	.708

In some cases, especially with smaller sample sizes and higher dimensionality, W and T_m have greater powers than Hotelling's T^2. The power for ρ = 0.36 with SDS is somewhat greater, and that for ρ = 0.36 with ODS is somewhat less, than that for ρ = 0.0.

Acknowledgement

The author wishes to thank Miss Samia Balalem for help in manuscript preparation

References

1] SCHILLING, M.F., “Multivariate two-sample tests based on nearest neighbors”, *Journal of the American Statistical Association,* **81**,.(1986), 799-809.

2] BARAKAT, A.S., QUADE, D., and SALAMA, I.A. “Multivariate homogeneity testing using an extended concept of nearest neighbors”, *Biometrical Journal* **38**, (1996), 605-612.

3] SEN, P.K. and SALAMA, I.A., “The Spearman Footrule and a Markov Chain Proporty”,. *Statistics and Probability Letters* **1**, (1983). 285-289.

4] Gibbons, J.D., “Nonparametric Statistical Inference” Marcel Dekker, New York, (1985).