A RANK-SUM TEST OF WHETHER TWO MULTIVARIATE SAMPLES WERE DRAWN FROM THE SAME POPULATION: PRELIMINARY REPORT

D. S. Robson

BU-159-M February, 1964

Abstract

A discrete or qualitative variable $g(x)$ defined on a finite set of N points $\{x_i\}$, $i=1,\cdots,N$ generates a partition into k disjoint subsets of $n_1,\cdots,n_k$ points, respectively, $n_1+\cdots+n_k=N$, with each subset representing an equivalence class, $x_i \approx x_j$ iff $g(x_i) = g(x_j)$. Under a statistical null hypothesis such a partition is random and each of the $N!/\prod_1^k n_i!$ possible partitions is assigned equal probability. Alternative hypotheses to be considered here are characterized by some form of segregation where the distance $d(x_i,x_j)$ between two points tends to be smaller for equivalent points than for non-equivalent points. A nonparametric measure of segregation is the proportion of times like neighbors are nearer than unlike neighbors, where each of the N points is considered to have N-1 neighbors. The mean ($= \frac{1}{2}$) and variance of this statistic are derived under the null hypothesis.

Intended extensions include the asymptotic theory of random partitions and the development of measures of attraction, repulsion, and buffering between equivalence classes. The hope is to provide answers to such questions as: if $g(x_i) < g(x_j) < g(x_k)$ then does the x_j class tend to fall in between the x_i and x_k classes?

A RANK-SUM TEST OF WHETHER TWO MULTIVARIATE SAMPLES WERE DRAWN FROM THE SAME POPULATION: PRELIMINARY REPORT

D. S. Robson*

BU-159-M

February, 1964

Introduction

The problem considered here was first posed as the question of whether birds of a feather nest together. Two species of birds having the common habit of nesting on cliff ledges were observed to share the same cliff. A map of the cliff was made and nest cites identified as to species were indicated. The nests of the two species were so interspersed that a studied inspection of the map gave no clear indication of the expected prejudice against neighbors of another feather. The statistical problem thus born was that of devising a test of the hypothesis that the two species were randomly mixed over the occupied cliff ledges.

In more general terms, we are given a configuration of $N=m+n$ points (in a metric space) which have been partitioned into two sets, A and B, of m and n points, respectively, and a procedure is required for testing the hypothesis of random partitioning against the alternative hypothesis that A and B are the result of some form of segregating operation. Such a nonparametric test procedure, being based only on the conditional distribution of partitions for the given configuration of $m+n$ points, should have application to the general problem of testing whether two multivariate samples were drawn from the same population.

---

*Biometrics Unit, Plant Breeding Department, Cornell University, Ithaca, N.Y.

Construction of the test statistic

A function of the partition which might be expected to manifest any tendency for segregation is the proportion of times that individuals have like neighbors nearer than unlike neighbors. Any particular member a_i of the set A has m-1 like neighbors and n unlike neighbors, thus generating n(m-1) comparisons of the relative nearness of a- and b- neighbors. For all m members of A the total number of comparisons generated is then mn(m-1); similarly, the set B generates mn(n-1) comparisons, so the test statistic is the proportion of the mn(m+n-2) = mn(N-2) comparisons in which like neighbor is nearer than unlike.

Computation of this statistic involves the measurement of the relative distance between the N(N-1)/2 pairs of points in the configuration. These measurements are most conveniently displayed in the form of a symmetric matrix $\{d_{ij}\}$ with entries

$$d_{ij} = \text{distance from point } i \text{ to point } j$$

or in the directly usable form of a (nonsymmetric) matrix $\{r_{ij}\}$ where

$$r_{ij} = \text{the rank of } d_{ij} \text{ in the set } \{d_{i1}, d_{i2}, \cdots, d_{i\,m+n}\} .$$

Since $d_{ii}=0$ then r_{ii} may be taken as zero, also, so that each row of $\{r_{ij}\}$ contains the diagonal element 0 and all of the integers from 1 to N-1. For later convenience the first m rows and columns of $\{r_{ij}\}$ may be chosen to represent the points in A while the last n=N-m rows and columns correspond to the points in B. Thus, $\{d_{ij}\}$ is the partitioned matrix

$$\{d_{ij}\} = \left\{\begin{array}{c|c} \{d(a_i,a_j)\} & \{d(a_i,b_j)\} \\ i=1,\cdots,m;\ j=1,\cdots,m & i=1,\cdots,m;\ j=m+1,\cdots,N \\ \hline \{d(b_i,a_j)\} & \{d(b_i,b_j\} \\ i=m+1,\cdots,N;\ j=1,\cdots,m & i=m+1,\cdots,N;\ j=m+1,\cdots,N \end{array}\right\}$$

and $\{r_{ij}\}$ is the corresponding partitioned matrix of rank-orders within the rows of $\{d_{ij}\}$.

The test statistic is then simply obtained from the partitioned matrix $\{r_{ij}\}$ as

$$S = \frac{S_a+S_b}{mn(N-2)} - \frac{1}{2} - \frac{2}{N-2}$$

where

S_a = sum of the entries in the submatrix $\{r(a_i,b_j)\}$

S_b = sum of the entries in the submatrix $\{r(b_i,a_j)\}$.

To confirm that the statistic S is the proportion of times like neighbor is nearer than unlike, first note if the entries in the i^{th} row of $\{r(a_i,b_j)\}$ are rearranged in order of increasing magnitude as

$$r(a_i,b_{j_{(1)}}) < \cdots < r(a_i,b_{j_{(k)}}) < \cdots < r(a_i,b_{j_{(n)}})$$

then $r(a_i,b_{j_{(k)}}) - k$ is the number of a-points whose distance from a_i is less than $d(a_i,b_{j_{(k)}})$. Hence,

$$\sum_{k=1}^{n}[r(a_i,b_{j_{(k)}})-k] = \sum_{j=m+1}^{m+n} r(a_i,b_j) - \frac{n(m+1)}{2}$$

is the total number of comparisons in which the distance from a_i to a is less than the distance from a_i to b. Summing this count over all m rows of the submatrix $\{r(a_i,b_j)\}$ gives

$$\sum_{i=1}^{m} \left[\sum_{j=m+1}^{m+n} r(a_i,b_j) - \frac{n(n+1)}{2} \right] = S_a - \frac{mn(n+1)}{2}$$

and combining this with the corresponding quantity derived from $\{r(b_i,a_j)\}$ gives $mn(N-2)S$ as the number of times like neighbor is nearer than unlike.

Conditional mean and variance of the test statistic

The mean value of S when averaged over the $\binom{N}{m}$ possible partitions of the given configuration of points is 1/2. For purposes of demonstrating this an alternative expression for S is somewhat more convenient; namely, replacing S_a by $S_a = mN(N-1)/2 - S_a^*$, where S_a^* is then the sum of entries in the square submatrix $\{r(a_i,a_j)\}$. Under randomization, each entry r_{ij} of $\{r_{ij}\}$ then appears in $\binom{N-2}{m-2}$ of the $\binom{N}{m}$ equally likely S_a^*'s; hence, the average value of S_a^* is

$$\mathcal{E}(S_a^*) = \frac{\binom{N-2}{m-2}}{\binom{N}{m}} \sum_{i=1}^{N} \sum_{j=1}^{N} r_{ij} = \frac{\binom{N-2}{m-2}}{\binom{N}{m}} \frac{N^2(N-1)}{2} = \frac{Nm(m-1)}{2}$$

since

$$\sum_{i=1}^{N} \sum_{j=1}^{N} r_{ij} = \sum_{i=1}^{N} r_{i\cdot} = \sum_{i=1}^{N} \left[\frac{N(N-1)}{2} \right] = \frac{N^2(N-1)}{2} \quad .$$

Therefore

$$\mathcal{E}(S_a) - \frac{mn(n+1)}{2} = \frac{mn(m-1)}{2}$$

and combining this with the corresponding result for S_b gives $\mathcal{E}(S) = 1/2$.

A similar argument leads to the variance of S_a^*; since

r_{ij} appears in $\binom{N-2}{m-2}$ of the possible S_a^*'s

r_{ij} and r_{ik} appear in $\binom{N-3}{m-3}$ of the possible S_a^*'s

r_{ij} and r_{hk} appear in $\binom{N-4}{m-4}$ of the possible S_a^*'s

then

$$\mathcal{E}(S_a^{*2}) = \frac{1}{\binom{N}{m}} \sum_{i=1}^{N} \sum_{j=1}^{N} r_{ij} \left\{ \binom{N-4}{m-4} r_{..} \right.$$

$$+ \left[\binom{N-4}{m-4} - \binom{N-3}{m-3} \right] (r_{i.} + r_{j.} + r_{.i} + r_{.j})$$

$$\left. + \left[\binom{N-2}{m-2} - 2 \binom{N-3}{m-3} + \binom{N-4}{m-4} \right] (r_{ij} + r_{ji}) \right\}$$

$$= \frac{\binom{N-4}{m-3}}{\binom{N}{m}} \sum_{j=1}^{N} r_{.j}^2 + \frac{\binom{N-4}{m-2}}{\binom{N}{m}} \sum_{i<j}^{N} (r_{ij} + r_{ji})^2$$

$$+ \frac{N^2 m^2 (m-1)}{4} \left[(m-1) - \frac{n}{N-2} \right]$$

Hence, the variance of S_a^* is

$$\mathrm{Var}(S_a^*) = \frac{nm(m-1)}{N-2}\left\{\frac{1}{N(N-1)(N-3)}\left[(m-2)\sum_{j=1}^{N} r_{\cdot j}^2 + (n-1)\sum_{i<j}^{N}(r_{ij}+r_{ji})^2\right] - \frac{N^2 m}{4}\right\} .$$

A corresponding formula holds for $\mathrm{Var}(S_b^*)$, and the covariance of S_a^* and S_b^* is, by the same argument,

$$\mathrm{Cov}(S_a^*, S_b^*) = \frac{mn(m-1)(n-1)}{N-2}\left\{\frac{N^2}{4} - \frac{2}{N(N-1)(N-3)}\left[\sum_{j=1}^{N} r_{\cdot j}^2 - \sum_{i<j}^{N}(r_{ij}+r_{ji})^2\right]\right\} .$$

The variance of S therefore takes the form

$$\mathrm{Var}(S) = \frac{1}{mn(N-2)^3}\left\{\frac{1}{N(N-1)(N-3)}\left[4(m-1)(n-1)\sum_{i<j}^{N}(r_{ij}+r_{ji})^2 - \{(N-2)-(m-n)^2\}\sum_{j=1}^{N} r_{\cdot j}^2\right] - \frac{N^2}{4}\left[(N-2)+(n-m)^2\right]\right\} .$$

When $n = m = N/2$ this reduces to

$$\mathrm{Var}(S) = \frac{4}{N^3(N-1)(N-2)(N-3)}\left\{\sum_{i<j}^{N}(r_{ij}+r_{ji})^2 - \frac{1}{N-2}\sum_{j=1}^{N} r_{\cdot j}^2\right\} - \frac{1}{(N-2)^2} .$$

An upper bound on the variance of S for the case $m = n = N/2$ is obtained from the inequalities

$$\sum_{j=1}^{N} r_{\cdot j}^2 \geq r_{\cdot\cdot}^2/N = N^3(N-1)^2/4$$

$$\sum_{i<j}^{N} (r_{ij} + r_{ji})^2 \leq 2 \sum_{i<j}^{N} (r_{ij}^2 + r_{ji}^2) = N^2(N-1)(2N-1)/3$$

giving

$$\mathrm{Var}(S) \leq \frac{2}{3N(N-3)}$$

for any configuration of N points randomly partitioned into two equal-sized subsets.

The effect of clustering of like points

Segregation of the two kinds of points might be expressed through the phenomena of clustering or colonizing into small groups of like kind which might then be randomly mixed in the configuration. Formally, a cluster of size k may be defined as a closed set k points arranged so that the k-1 nearest neighbors of each point in the set are the remaining points of the set. An indication of the effect of clustering of like points may be obtained by regarding the points of any given configuration of size N as tiny clusters of k points, all points in a cluster being of the same kind as the original point; that is, we shall suppose that each point a_i is now replaced by a tightly packed cluster A_i consisting of the points $a_{i_1}, a_{i_2}, \cdots, a_{i_k}$ and similarly b_j is

replaced by $B_j = \{b_{j_1}, b_{j_2}, \cdots, b_{j_k}\}$. If S_N denotes the value of the test statistic for the original configuration of N points and the original partition into a- and b-points then for the enlarged configuration of Nk points and the same partition applied to clusters, the test statistic takes the value

$$S_{Nk} = S_N + \frac{2(k-1)}{Nk-2}(1 - S_N) .$$

For if $h_N(a_i, b_j)$ is the number of a-points in the original configuration whose distance from a_i is less than $d(a_i, b_j)$, then for a_{i_ν} in A_i and b_{j_μ} in B_j the corresponding number is

$$h_{Nk}(a_{i_\nu}, b_{j_\mu}) = (k-1) + kh_N(a_i, b_j)$$

Summing this count over all points in A_i and B_j gives

$$h_{Nk}(A_i, B_j) = k^2(k-1) + k^3 h_N(a_i, b_j)$$

and summing further over all m of the A-clusters and all n of the B-clusters gives

$$h_{Nk} = mnk^2(k-1) + k^3 h_N$$

as the number of times (a,a)-neighbors are closer than (a,b)-neighbors. Similarly, if g_N denotes the number of times (b,b)-neighbors are closer than (b,a)-neighbors in the original configuration then

$$g_{Nk} + h_{Nk} = 2mnk^2(k-1) + k^3(g_N + h_N)$$

$$= mnk^2[2(k-1) + k(N-2)S_N]$$

and the indicated relation between S_{Nk} and S_N then follows from the definition

$$S_{Nk} = \frac{g_{Nk}+h_{Nk}}{mnk^2(Nk-2)}$$

As k is increased, S_{Nk} increases to the limiting value

$$S_{Nk} \to S_N + \frac{2}{N}(1 - S_N) \quad .$$

Thus, if an observed configuration consists of a random mixture of N tight and pure clusters of the two kinds then S_N will be approximately 1/2 and S_{Nk} will be approximately 1/2 + 1/N. If k is large, this deviation of 1/N will be statistically significant since the variance of S_{Nk}, bounded above by

$$\mathrm{Var}(S_{Nk}) \leq \frac{2}{3Nk(Nk-2)}$$

approaches zero as k increases. In the most conspicuous case of segregation the configuration consists simply of N=2 clusters, one of each kind, and $S_{Nk} = 1$.