ASSIGNING A POPULATION TO ONE OF THE CLUSTERS OF HOMOGENEOUS *P*-VARIATE POPULATIONS*

By

O.P. BAGAI

Panjab University, Chandigarh

(Received : May, 1977)

1. INTRODUCTION

Suppose we are given k p-variate normal populations, assumed to have the same covariance matrix. From these populations, samples respectively of sizes $N_1, N_2,..., N_k$ are drawn, and observations X_{irh}, $(i=1, 2,..., p;\ r=1, 2,..., k;$ and $h=1, 2,..., N_r)$ are made. Let $W=(w_{ij})$ and $B=(b_{ij})$ be the within and between mean product matrices with n_2 and n_1 degrees of freedom, where w_{ij} and b_{ij} are respectively defined as follows :

$$n_2\ w_{ij}=\sum_{r=1}^{k}\sum_{h=1}^{N_r}(X_{irh}-\bar{X}_{ir})\ (X_{jrh}-\bar{X}_{jr}) \qquad ...(1.1)$$

$$n_1\ b_{ij}=\sum_{r=1}^{k}N_r\ (\bar{X}_{ir}-\bar{X}_1)\ (\bar{X}_{jr}-\bar{X}_j) \qquad ...(1.2)$$

where $\bar{X}_{ir}$ is the mean of the ith trait from the rth population and $n_1=k-1$, $n_2=\sum_{r=1}^{k}(N_r-1)$.

Rao (1952) has considered assigning an individual to one of the k populations. Assuming that an individual with measurement $(X_1, X_2,..., Xp)$ belongs *a priori* to one of the populations, be computes (ignoring *a priori* probabilities) the linear discriminant scores for the rth population as follows :

$$\hat{L}_r=\sum_{j=1}^{p}\sum_{i=1}^{p}(w^{ij}\bar{X}_{ir})X_j-\tfrac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(w^{ij}\bar{X}_{ir}\bar{X}_{jr}) \qquad ...(1.3)$$

*This paper is in continuation of author's earlier paper of 1976. The reader is therefore referred to first read that paper so as to acquaint himself with the methodology, language and terminology followed in this paper.

2. PRELIMINARIES

Since, by definition, all the populations included in a cluster have identical mean vectors, we can consider a cluster as a single population whose mean vector is estimated to be the pooled mean vector of all those populations which are included in the cluster. Thus, if there are clusters, we shall imagine them as c distinct populations with their estimated mean vectors as the pooled mean vectors of those populations which are included in the respective clusters. Let the estimated mean vectors of the c (so-called) populations be given in the matrix form as follows :

$$\bar{z}^t\ (c\times p)=\begin{bmatrix} \bar{z}_{11}, \bar{z}_{21}, \ \dots\ , \bar{z}_{p1} \\ \bar{z}_{12}, \bar{z}_{22}, \ \dots\ , \bar{z}_{p2} \\ z_{1c}, z_{2c}, \ \dots\ , \bar{z}_{pc} \end{bmatrix} \qquad \dots(2.1)$$

To simplify the manipulation, we again resort to the technique of significant discriminant scores as explained and employed by the author [1]. That is, we use the matrise K^t (as obtained by us in the paper of 1976) and premultiply it to the matrix $\bar{z}^t\ (c\times p)$ and obtain the corresponding matrix $\bar{U}^t$ defined as :—

$$\bar{U}^t\ (c\times p')=(U_{ir})^t \quad \begin{matrix} i=1, 2, \dots, p' \\ r=1, 2, \dots, c \end{matrix} \qquad \dots(2.2)$$

where $p'\ (\leqslant p)$ is the number of significant discriminant scores.

Further, if $(\bar{X}_1, \dots, \bar{X}p)$ be the mean vector of the sample from the new population which we are assigning to one of the clusters, the corresponding significant discriminant scores are similarly obtained and are denoted as follows :—

$$(\bar{V}_1, \bar{V}_2, \dots, \bar{V}_p) \qquad \dots(2.3)$$

3. DISCUSSION OF THE TWO APPROACHES

(i) *Approach I. Use of $\hat{L}$—functions*

Since the c so-called populations are normally distributed, we use Rao's procedure as discussed above in section 1. We first compute $\hat{L}$—functions in the form already defined in (1.3) or in the form obtained below by the use of significant discriminant scores, namely.

$$\hat{L}_r=\sum_{i=1}^{p'} \bar{U}_{ir}\ \bar{V}_i-\tfrac{1}{2}\sum_{i=1}^{p'} \bar{U}_{ir}^2\ ,\ r=1, 2, \dots, c \qquad \dots(3.1)$$

Following Rao, we would assign, ignoring the *a priori* probabilities, the new population to the sth ($s \leqslant c$) 'so-called population' (or cluster) if

$$\hat{L}_s - \hat{L}_r \geqslant 0 \text{ for all } r=1, 2, \ldots, s-1, s+1, \ldots, c. \qquad \ldots(3.2)$$

(*ii*) *Approach II. Use of* T_k^2 *–Statistic*

In the above method, we have not been able to assign a probability to our decision. To do this, we propose the following :

Step 1

Let the size of of the sample drawn from the new normally distributed population be N. Include it in each cluster so that the number of populations involved in each cluster increases by one.

Step 2

Modify the statistic T_k^2 , given as in (1.4), with the help of significant discriminant scores and write the modified T_k^2 as follows :

$$T_k^2 = \sum_{i=1}^{P'} \sum_{r=1}^{k} N_r (\bar{Y}_{ir} - \bar{Y}_i)^2 \qquad \ldots(3.3)$$

which is the same as obtained earlier by the author ([1], (2.6), pp 40).

Step 3

Compute the statistics $T_{k_r}^2 + 1$ for $r=1, 2, \ldots, c$, where k_r is the number of populations in the rth cluster.

Step 4

Include the new population in the sth cluster, if computed $T_{k_s}^2 + 1 \leq$ tabular $T_{\alpha_{k_s}}^2 + 1$, where $T_{\alpha_{k_s}}^2 + 1$ is the tabular value as discussed in author's earlier paper [1], where the relevant tabular values are provided in Table I at page 41 of the author's paper [1].

Note : Since we allow overlappings, we shall include the new population in all those clusters for which the computed T^2 is non-significant.

4. ILLUSTRATION

Data for illustration is the same as taken up earlier by the author in his paper ([1], pp 46). It constitutes 14 species of trees with four measurements X_1, X_2, X_3 and X_4 on each tree. For simplicity sake, the fourteen species have been specified (as before) by numbers : 1, 2, ..., 14, instead of those by their respective names. The clusters

where the matrix (w^{ij}) is the inverse of the matrix (w_{ij}). He, then, suggests assigning the individual to the 5th population if $\hat{L}_s$ is greater than every other $\hat{L}_r$ for $r\,(\neq s)=1, 2,\ldots,s-1, s+1$.

Author [1] has considered forming of clusters of homogeneous populations from amongst the k given populations. As many as three alternative procedures were suggested-each consisting of two stages, where the first stage predicts a cluster and the second corrects the predicted cluster with a pre-assigned level of significance. To demonstrate theory, a numerical data was taken. It consisted of 14 species (*i.e.* populations) of trees and one of the three procedures was applied to obtain 7 clusters of homogeneous species on the basis of four measurements (on each observation) characterising the static bending property of a tree.

In what follows, we take up the problem of assigning a stray individual (or an other population) to one of the clusters where it is *a priori* known that it belongs to one of the clusters. Two alternative approaches have been suggested. The first approach deals with the method of L-functions as given above in (1.3) and the second with the use of Hotellings' T^2_k—statistic for k samples defined as follows:

$$T_k^2 = n_1\, tr\,(w^{-1}B)$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p} w^{ij}\left[\sum_{r=1}^{k} N_r\,(\bar{X}_{ir}-\bar{X}_i)\,(\bar{X}_{jr}-\bar{X}_j)\right] \quad ...(1.4)$$

where $\bar{X}_i$ is the pooled mean of the ith trait of all the k samples and the matrix W^{-1} is the inverse of the matrix W. The distribution of T_k^2, under the hypothesis of homogeneity of mean vectors, is known in the classical case to be the central chi-square with $p\,(k-1)$ degrees of freedom and in the studentized case to be an asymptotic expression as given by the author [1], expression (2.7), pp. 40).

An illustration is presented to demonstrate both the procedures. The data for illustration is the same as already taken up by the author [1], where 7 clusters from amongst 14 populations (*i.e.* species) of trees were concluded on the basis of their static bending property.

Finally, the closeness of the new specy to a particular cluster has also been verified through pictorial representation of the 14 points and the new point representing respectively the 14 populations and the new specy.

(of homogeneous populations) concluded by the author (1976, pp 55) were as follows :

(*a*) 2, 5, 6 and 8 ;
(*b*) 2, 7, 8 and 10 ;
(*c*) 2, 9 and 10 ;
(*d*) 2, 4, 10 and 11 ;
(*e*) 9, 12, 13 and 14 ;
(*f*) 1, by itself ,
(*g*) 3, by itself.

Consider, now, another shipment of trees of "Black cottonwood" having been arrived later on. The problem is to assign this new lot to one of the above 7 clusters on the basis of the same four measurements. Let the sample be :

Sample size	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$
61	962	4.70	2287	4102

Let the corresponding significant discriminant scores, as explained in section 2 and as given in (2. 3), be :

Sample size	$\bar{V}_1$	$\bar{V}_2$	$\bar{V}_3$	...(4.1)
61	0.4794140	1.1417523	0.4540478	

Demonstration of Approach I

Imagining each cluster to be a single population whose mean vector is estimated as the pooled mean vector of the populations (species) involved in the corresponding cluster, we write below the mean vectors of each of the seven clusters with the use of (2.1) and (2.2) :

Cluster	*Size*	$\bar{U}_1$	$\bar{U}_2$	$\bar{U}_3$	
(*a*)	931	0.66968541	1.27604842	0.62721413	
(*b*)	984	0.76536770	1.19428189	0.71449228	
(*c*)	368	1.08072219	1.17054579	0.58144678	...(4.2)
(*d*)	587	1.01019297	0.92978089	0.42782876	
(*e*)	1266	1.29464479	1.30553049	0.70800378	
(*f*)	264	0.94597083	1.72039748	0.34593229	
(*g*)	158	1.74328671	1.21889759	0.50048469	

Using (4.1), (4.2) and (3.1), we obtain the following $\hat{L}$-functions :

$\hat{L}_{(a)}=0.82768514$ $\hat{L}_{(d)}=0.69443164$

$\hat{L}_{(b)}=0.79361765$ $\hat{L}_{(b)}=0.58770051$

$\hat{L}_{(c)}=0.72048219$ $\hat{L}_{(c)}=0.49183864$

and $\hat{L}_{(g)}=0.06075676$

Since $\hat{L}_{(a)}$ is greater than all the remaining $\hat{L}$-functions, we would assign the new specy "Black Cottonwood" to the cluster designated by (*a*) *i.e.*, to (2, 5, 6, 8).

Demonstration of Approach II

Combining the new specy of "Black Cottonwood" with each of the sets of populations already in clusters, we compute T_k^2-values by the formula (3.3) and write them as follows :

T_5^2 (for 2, 5, 6, 8 and new specy) $=24.70$

T_5^2 (for 2, 7, 8, 10 and new specy) $=29.91$

T_4^2 (for 2, 9, 10 and new specy) $=30.94$

T_5^2 (for 2, 4, 10, 11 and new specy) $=34.39$

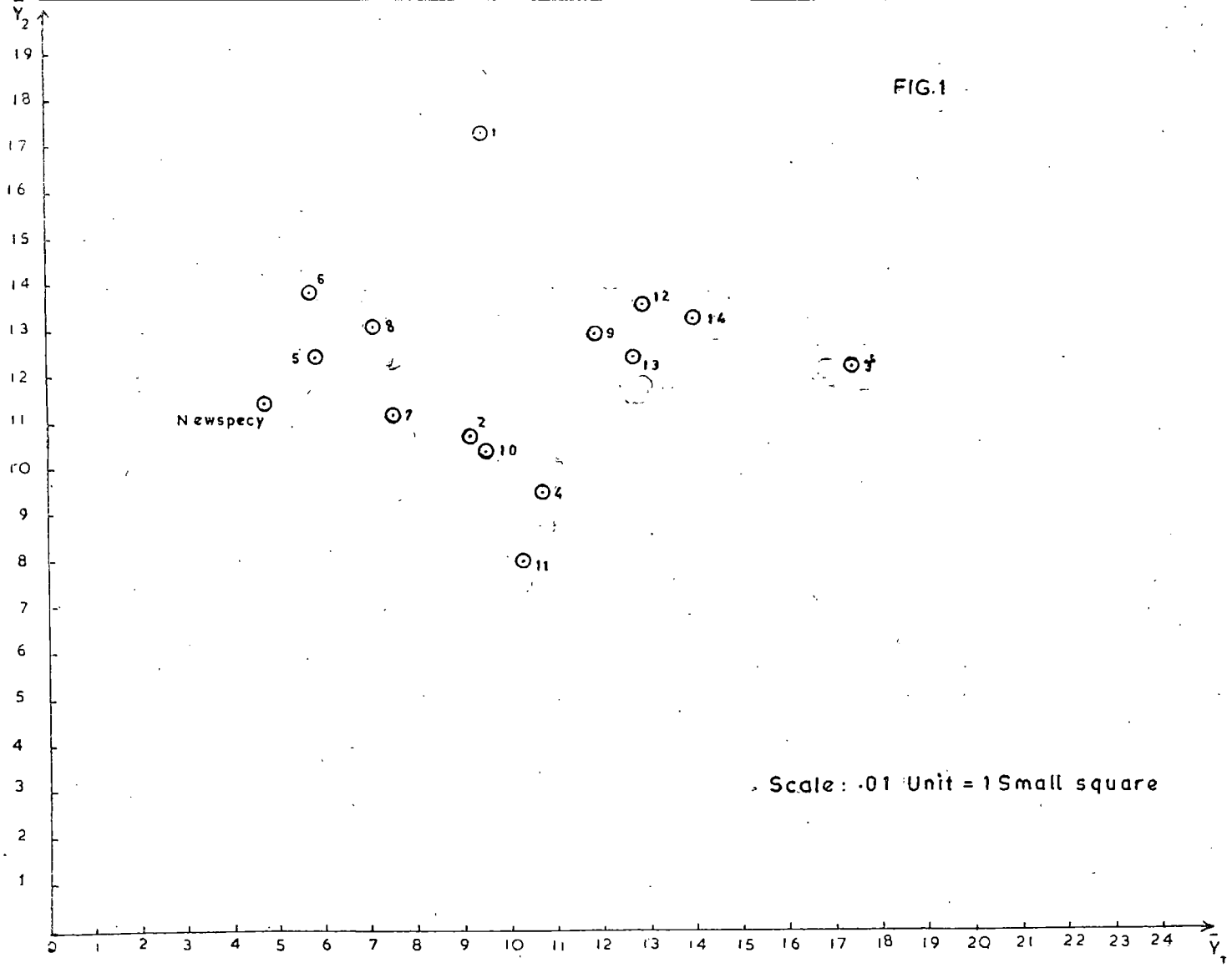
T_2^2 (for 1 and new specy) $=37.44$

T_5^2 (for 9, 12, 13, 14 and new specy) $=43.99$

T_2^2 (for 3 and new specy) $=77.66$

Now referring to the tabular values T_k^2 for $k=2, 3, 4$ and 5 at page 41 of the author's earlier paper (1976), we find that the T_5^2 (2, 5, 6, 8 and new specy) is the only one non-significant. Hence, the new specy would naturally be assigned to the cluster of species : 2, 5, 6 and 8.

FIG.1



Pictorial Verification

The closeness of the new specy to various clusters can also be verified from Fig. 1. Depending upon only the two most significant discriminant scores, we plot the coordinates of 14 species, *i.e.*, $(\bar{Y}_1, \bar{Y}_2)$, and of the new specy, *i. e.* $(\bar{V}_1, \bar{V}_2)$ as in (4.1), on the plane graph paper. We again find the new specy to be lying closer to the cluster of the species, 2, 5, 6 and 8.

ACKNOWLEDGMENT

The author is grateful to the refree for making a valuable suggestion in improving the paper.

REFERENCES

[1] Bagai, O.P. (1976) : Multivariate analogues of multiple comparisons methods. *Journal of Indian Society of Agricultural Statistics* ; 28, No. 2 ; 37-56.

[2] Rao, C.R. (1952) : *Advances Statistical Methods in Biometric Research*, N.Y. Wiley.