MULTIVARIATE ANALOGUES OF MULTIPLE COMPARISONS METHODS

By

O.P. Bagai

Punjab University, Chandigarh,

(Received : September, 1275)

1. Introduction

In Anthropology and Biological Sciences, situations arise when certain multivariate populations are found to be heterogeneous, and there is a need to find out which subsets of the populations are most alike and which are least alike (or, in the words of Rao [9], we want to find out the clusters of like populations).

Rao and Tocher made a subjective approach to this problem which, in fact, was not based on probabilistic considerations. Working on the principle of minimum average distance, they suggested a technique based on the criterion that any two groups belonging to the same cluster should at least on the average show a smaller Mahalanobis distance than those belonging to different clusters.

A graphical approach to the same problem was given by Rao [9] on the basis of significant discriminant scores. He suggested plotting the scores in a space whose dimensionality is equal to the number of significant eigen values. In situations where there are more than two significant eigen values, Rao [9] suggested having pairwise plane representations of the points. Relying mostly on the plane representation of the most significant scores, he proposed to form clusters of population which lie pictorially close to one another.

In what follows, we have proposed a procedure for forming clusters of like populations where we seek a departure from Rao's and Tocher's subjective approach. We, instead, suggest two stages. *Stage I* is a sort of prediction where we make use of Rao's graphical approach. *Stage II* corrects a predicted cluster where a new

definition of a cluster is first given and then as many as three alternative statistics are proposed. Further, in each, unlike Rao and Tocher, we are able to give probability to our decision. The test procedure for all the three alternatives is discussed in detail where only one of them is demonstrated by an illustrative example.

Again, with regard to the level of significance, we follow Duncan [3] and propose the $r-$mean vectors ($r=2, 3,...k$) significance level, α_r, for a preassigned α to be

$$\alpha_r=1-(1-\alpha)^{r-1} \quad (r=2, 3,..., k) \qquad ...(1.1)$$

where $(r-1)$ is the number of independent comparisons which can be specified among the r mean vectors. Since the statistic used in demonstrating the illustrative example involves a central chi-square in both the studentized and classical cases, we have computed the corresponding tabular chi-square values against

$$\alpha_r \ (r=2(1)\ 20) \text{ for preassigned } \alpha=.05 \text{ and } .01(1)$$

2. Definition of a cluster and statement of problem :

Definition of a cluster

A cluster of populations is a group of populations having the same mean vector.

Statement of Problem

Suppose we are given k p-variate normally distributed populations assumed to have the same dispersion matrix Σ,

Let X_{irh} ($i=2,.....,p$; $r=1, 2.....,k$; $h=1, 2,... N_r$ and $p<k$) be the observation of the ith trait on the h th individual from the r th sample of size N drawn from the r th population. Let $\bar{X}^t$ be the $k\times p$ matrix of k sample mean vectors. Further, let $B=(b_{ij})$ and $W=(w_{ij})$ be the between and within independent mean product (M.P) matrices with n_1 and n_2 degrees of freedom (D.F), respectively, computed on the basis of k p-variate samples, where,

$$n_1 b_{ij}=\sum_{r=1}^{k} N_r (\bar{X}_{ir}-\bar{X}_i)(\bar{X}_{jr}-\bar{X}_j) \qquad ...(2.1)$$

$$n_2 w_{ij}=\sum_{r=1}^{k} \sum_{h=1}^{N_r} (X_{irh}-\bar{X}_{ir})(X_{jrh}-\bar{X}_{jr}) \qquad .. (2.2)$$

and

$$n_1=k-1,\ n_2=\sum_{r=1}^{k}(N_r-1)$$

Suppose further that the hypothesis of homogeneity of mean vectors of the populations has been rejected by the use of Wilks' [11] $\triangle$-statistic (Rao [9] *p.* 260) and Barlett's [2] approximation to its probability.

After concluding the over-all heterogeneity of mean vectors of populations, our job now is to find which subsets of populations form clusters. To do this, we first introduce the following statistics : -

The first statistic D_2^2 is the Mahalanobis' distance between two populalions and is computed as follows :—

$$(\underline{\bar{X}}_1-\underline{\bar{X}})^t\ w^{-1}\ (\underline{\bar{X}}_1-\underline{\bar{X}}_2)$$

where w^{-1} is the inverse of w.

For testing the hypothesis of equality of the mean vectors involved in a predicted cluster, we propose in the first alternative of stage II an analogue of Duncan's stage 2 (Federer [4], pp. 19—40 of the multiple F test. He computed the variance of the means involved in a predicted cluster of like means and tested it against his least significance sums of squares with level of significance based on d.f. in multivariate situations, as the analogue of the "Variance of the k means involved in a cluster" we propose the statistic T^2 (Hotelling, [7] defined as follows :—

$$T_k^2=n_1\ tr.\ (w^{-1}\ B) \qquad \ldots(2.4)$$

where again, w^{-1} is the inverse of w.

The distribution of T_k^2 under the null hypothesis, is known in the classical case to be the central chi-square with $p(k_1-1)$ d.f. and in the studentized case to be an asymptotic expression involving chi-squares which we write below in (2.7).

Further, since we have made frequent use of both the studentized D_2^2 and T_r^2 ($r=2;\ 3.....k$), we propose to modify the expressions in (2.3) and (2.4) to an easily workable form with the use

of significant discriminant scores (Rao, [9]. Taking, therefore, $Y^t(k \times p)$ as the matrix of significant discriminant scores whose first column gives the discriminant score corresponding to the largest, the second to the second largest, and so forth, we reduce the studentized Statistics D_2^2 and T_k^2,respectively, in (2.3) and (2.4) to

$$D_2^2=\sum_{i=1}^{p'}(\bar{Y}_{i1}-\bar{Y}_{i2})^2 \qquad ...(2.5)$$

and

$$T_k^2=\sum_{i=1}^{p}\sum_{r=1}^{k} N_r(\bar{Y}_{ir}-\bar{Y}_i)^2 \qquad ...(2.6)$$

where $$\bar{Y}_i=\left(\sum_{r=1}^{k} N_r \bar{Y}_{ir}\right) \Big/ \left(\sum_{r=1}^{k} N_r\right)$$

and where the first p' scores are the most significant ones.

Since each of the statistics T_r^2 ($r=2, 3.....,k$) involves chi-squares for both the studentized and classical cases and further since the level of significance (defined in in (1.1) changes with the change in the value of r, we need, therefore, to find some more tabular chi-square values which so far have not been computed. To do this we first find α_r for r : 2(1) 20 from the formula (1.1) for both $\alpha=.01$ and .05 and then find the corresponding normal variates by the linear interpolation formula with the help of Table I of Hartley and Pearson, [6]. Finally adopting Aitken's iterative method for interpolation, the new chi-square values values are computed against the normal variates. These tabular values at various significance levels α_r for $r=2(1)$ 20 and d.f.$=1(1)$ 30 (10) 100 for preassigned $\alpha=.01$ and .05 have been found by the author [1].

Finally, to find the tabular values of studentized T_r^2 for any r, we use the formula (1 to [8]), for $r=2, 3,.....,k$,

$$T_r^2=\chi^2+\frac{1}{2n_2}\left(-\frac{p+n_1+1}{n_1 p}\chi^4\right)+... \qquad ...(2.7)$$

where χ^2 is central chi-square with $p(r-1)$ d.f.

Note : It may be pointed out that we propose to use p instead of p' for defining the degrees of freedom, since the affect of all, p

correlated variates has been taken care of by the p' discriminant scores.

Since the illustration presented for demonstration concerns the studentized T_r^2, its tabular values for $r=2(1)\ 5$, $n_1=1(1)4$, $p=4$, and $n_2=29$ at 5% and 1% significance levels are tabulated approximately as given below in Table I : —

TABLE I

$r=(n+1)$	d.f. $=p(r-1)$ $=n_1\ p$	$\chi^2(.05)_r$	$\chi^2(.01)_r$	$(T^2.05)_r$	$(T^2.01)_r$
2.	4	9,4877	13,2767	12,7371	18,2030
3.	8	13,4428	18,1825	16,7783	24,0936
4.	12	17,1889	22,7748	21,7064	29,9103
5.	16	20,8200	27,1912	25,6131	25,6187

3. The proposed stages for Forming Clusters

We propose two stages for the purpose. Stage I comprises three steps wherein we predict the possible clusters. Stage II then corrects the prediction on some probabilistic basis. Three alternative methods have been proposed for stage II which are as follows :

(*i*) The Dudcan-Hotelling test.

(*ii*) The 'Extreme Distance from the Mean', *E*-test.

(*iii*) The 'Largest Distance', R-test.

Stage I : **Prediction :**

Step 1 : Compute (2^k) Mahalanobis distances by the formula (2.5) between all the pairs of k populations and set up the table of distance, where the distances of each population from the remaining ones are arranged in order of increasing magnitude, Such a table (like that in Table 5) will enable us to visualize which of the populations are closer to each other and which are farther away.

Step 2 : Represent graphically the significant discriminent scores of each population. For $p>2$, they be represented pairwise on plane graph paper. Relying largely on the plane representations of the most significant discriminant scores, visualize which of the populations lie closer together and which of them lie farther apart.

Step 3 : Step 3 deals with the prediction of the clusters obtained on the basis of the first two steps. Keeping in view the table of distances and the graphic plane representations, estimates roughly the 'would be' clusters – closeness being the only criterion for the populations to form a predicted cluster.

The following two points are worth noting .—

(*i*) That a wide range be allowed in selecting subsets of clusters (Since giving a narrow range might result in the loss of a population lying actually in a cluster),

(*ii*) That over-lappings be allowed (Since sometimes one is uncertain as to whether to include one (or more) population (*s*) in one or the other cluster (*s*)).

Stege II **Alternative I**

Correction by the Duncan—Hotelling Test

No generality is lost if we explain the procedure for only one predicted cluster having k_1 populations in the following steps :—

(*i*) Compute the Statistic $T^2_{k_1}$ by the formula (2.6).

(*ii*) Compare the computed $T^2_{k_1}$ with the tabular $T^2_{\alpha_{k_1}}$ where

α_{k_1} is already defined in (1.1)

(*iii*) If $T^2_{k_1}$ is less than or equal to $T^2_{\alpha_{k_1}}$, all the k_1 populations are concluded to form a cluster. Otherwise, split the k_1 populations into k_1 sets of (k_1-1) populations each.

(*iv*) Compare the computed $T^2_{(k_1-1)}$ values for each of the k_1 sets with the tabular $T^2_{\alpha_{k_1-2}}$. Of these some may be significant and some may not be. Those nonsignificant will yield clusters with the corresponding number of populations involved in them. Those for which $T^2_{(k_1-1)}$—values are significant are further split into (k_1-1) sets of (k_1-2) populations each and their corresponding $T^2_{k_1-2}$ — values are compared with the tabular $T^2_{\alpha_{k_1-2}}$. In this way, the process is continued till we arrive at the clusters of the type defined.

The working criterion analogous to Duncan's can be presented as follows :—

"A group of k_1 populations will form a cluster if $T^2_{k_1}$ computed for the mean vectors of the k_1 populations is non significant and also the T^2 of each and every set of populations of which the k_1 populations form a subset is significant according to α_r—level T_r^2—test, for some pre-assigned α, where r is the number of populations involved in the set".

Alternative II **Correction by the E Test**

For the second alternative procedure, the proposed test statistic is the "Extreme Mahalanobis distance from the Mean"—which we name as E-test. The level of significance is again based on degrees of freedom (d.f.) and is as defined in (1.1). The exact distribution of the E-statistic is not known. Siotani [10] has found the approximate distribution of this statistic for the k p-variate normal populations and has computed the tabular values at 5% and 1% significance levels for some particular values of p. Following Siotani's tables, the required tabular values at revised significance levels can be computed, and with these approximate tabular values in hand, we discuss the procedure for the E-test as follows :—

Suppose again, without loss of generality, that the predicted cluster contains k_1 populations, To correct it, we propose the following steps :—

(*i*) Compute the statistic $E_i(i=1, 2, \ldots, k_1)$, the Mahalanobis' distance between the mean vectors of the ith population and the grand mean vector of the k_1 populations.

(*ii*) Without losing generality, let E_{k_1} be the largest of all the computed $E_i(i=1, 2, \ldots\ldots, k_1)$.

(*iii*) Compare the E_{k_1} with the tabular $E_{\alpha_{k_1}}$, where α_{k_1} is as defined already in (1.1) and α is some pre-assigned significance level.

(*iv*) If E_{k_1} is less than or equal to $E_{\alpha_{k_1}}$, all the k_1 populations involved are concluded to form a cluster. Otherwise, split the k_1 populations into k_1 sets of (k_1-1) populations each.

(v) Compare the extreme distance of each set of (k_1-1) populations from their respective grand mean vectors with the tabular $E_{\alpha_{(k_1-1)}}$. Out of them some may be significant and some may not be. Those non-significant will yield clusters with the corresponding populations involved in them. Those for which the extreme E's are significant, are further split into sets of (k_1-2) each and their corresponding extreme E's are then compared against the tabular $E_{\alpha_{(k_1-2)}}$. In this way the process is contained till we arrive at the clusters of the type defined.

Thus a working criterion analogous to Duncan's can be stated as follows :—

"A group of k_1 populations will form a cluster if the extreme distance E_{k_1} (assumed to be the largest amongst all the k_1 distances between the mean vectors of the individual populations and their grand mean vector) is non-significant and if, furthermore, such extreme E's of each and every new set of populations of which the k_1 populations form a subset, is significant according to α_r—level E-test for some pre-assigned α, where r is the number of populations in the set".

Alternative III. **Correction by the R-test**

Lastly, the third alternative procedure, the proposed test statistic is the 'Largest Mahalanobis distance' which we call, for brevity, as R-test. The exact or approximate distribution of the R-statistic is not known. We have been able to find the distribution (Bagai, [1] of the R-statistic in the classical bivariate case in the form of definite integral for any number of populations, but still its tabular values have not baen computed. Again the level of significance is proposed to be based on degrees of freedom and is as given in (1.1).

In discussing the procedure for this alternative, we again follow Duncan and extend his procedure for the range test (Federer, [4] to the Multivariate case. Without loss of generality, suppose that a predicted cluster consists of k_1 populations. The procedure to

correct this prediction is described in detail in the following steps :—

(*i*) Compute $\binom{k_1}{2}$ Mahalanobis distances R_{rs} ($r \neq s = 1, 2, \ldots, k_1$) between the r th and s th populations.

(*ii*) Again, no generality will be lost, if we suppose that the distance R_{1k_1} between the first and the k_1 th populations is the largest amongst $\binom{k_1}{2}$ distances.

(*iii*) Compare the computed R_{1k_1} with the tabular $R_{\alpha_{k_1}}$ where α_{k_1} is as defined in (1.1) and α is a pre-assigned level of significance. If R_{1k_1} is less than or equal to $R_{\alpha_{k_1}}$, all the k_1 populations involved are considered to form a cluster. Otherwise, split the set of k_1 populations into k_1 sets of (k_1-1) populations each.

(*iv*) Compare the largest distance of each set of (k_1-1) populations with the tabular $R_{\alpha_{k_1-1}}$. Out of them some may be significant and some may not be. Those non-significant will yield clusters with the populations involved in them. Those for which the largest distance is significant are further split into sets of (k_1-2) and their respective largest distances are again compared against their corresponding tabular values $R_{\alpha_{k_1-2}}$. In this way, the process is continued till we arrive at the clusters of the type defined,

Thus the working criterion analogous to Duncan's can be summed up as follows :—

"A group of k_1 populations will form a cluster if the distance R_{1k_1} (assumed to be the largest amongst all $\binom{k_1}{2}$ distances) between the first and the k_1 th populations is non-significant and also the largest distance, amongst all possible distances between pairs of each and every new set of populations of which the k_1 populations form a subset, is significant according to α_r—level R-test, for some pre-assigned α, where r is the number of populations in the set".

4. Illustration Through Alternative I

To demonstrate the theory we present below an illustration where the samples have been drawn on the basis of nested sampling.

Description of the Data :

Data were taken from the Forest Productions Leboratory Division, Forestry Branch, Department of Northern Affairs and National Resources, Vancouver, B.C., Canada. Shipments of logs of various species of trees from various localities of Canada were received. The interest lay in clustering the species on the basis of their static bending properties. For this purpose, the following six measurements were taken at several locations on each tree :—

X_1 : Modulus of elasticity ;

X_1 : Work to the maximum limit ;

X_3 : Fibre Strength at proportional limit ;

X_4 : Modulus of rupture ;

X_5 : Specific gravity at oven dry ;

and X_6 : Work to the proportional limit.

Note : While finding the values of the determinants of the sum of product (S.P,) matrices to be used for tests of significance, it was found that they came out to be zeros, which enabled us to suspect that the variables were functionally dependent. The fact was actually varified when the physical interpretation was sought. The last two variables were found to be functionally dependent on the first four X_1, X_2, X_3, and X_4, We thus discarded X_5 and X_6 and continued our work on the variables X_1, X_2, X_3, and X_4,

The species taken for the purpose are listed as follows :—

(1) Yellow cedar,

(2) Lodge pols pine,

(3) Western larch,

(4) Western Yellow pine,

(5) Western white pine,

(6) Western white spruce,

(7) Sitka spruce,

(8) Amabilis fir,

(9) Western Aamloch,

(10) Engelman spruce,

(11) Western red cadar,

(12) Coast mature Doughlas fir,

(13) Interior mature Douglas fir,

(14) Coast second growth Douglas fir.

In what follows we will call each species by its corresponding number instead of specifying each time its name.

Description of the Model of Nested Sampling

We have the mixed model of nested sampling with fixed species and random localities and locations on trees. Further, the number of localities and locations is not uniform in all cases.

Let X_{ihjtl} be the observation of the i th character on the l th location of the t th tree belonging to the j th locality of the h th species. In place of observation X_{ihjtl} we are provided with the means $\bar{X}_{ihjt\cdot}$ alongwith the corresponding number of locations. The model for such data is :

$$\underline{\bar{X}}_{hjt\cdot} = \underline{\mu} + \underline{\xi}_h + \underline{\eta}_{j(h)} + \underline{\delta}_{t(hj)} + \underline{\bar{e}}_{hjt\cdot}$$

where

(*i*) $\underline{\bar{X}}_{hjt\cdot} \equiv (\bar{X}_{ihjt}, \ldots\ldots, \bar{X}_{qhjt})$ is a four dimensional mean vector of locations on the t th tree from the j th locality of the h th species.

(*ii*) $\underline{\mu}$ is the four dimensional mean vector of the populations; and $\underline{\bar{X}}\ldots$ is the corresponding sample statistic.

(*iii*) $\underline{\xi}_h$ is again tee four dimensional h th species fixed effect, but, for the sake of illustration, we will take it as random, distribued normally with mean vector zero and convariance matrix Σ_ξ .

(*iv*) $\underline{\eta}_{j(h)}$ is the four dimensional j th locality within h th species random effect, normally distributed with mean vector zero and covariance matrix Σ_η.

(*v*) $\underline{\delta}_{t(hj)}$ is the four dimensional t th tree within hth species from the j th locality random effect, normally distaibuted with mean vector zero and convariance matrix Σ_δ.

(*vi*) $\underline{e}_{hjt}$ is the four dimensional mean error vector of $\underline{e}_{hjtl}$ where each $\underline{e}_{hjtl}$ is random and normally distributed with mean vector zero and covariance matrix Σ_e.

(*vii*) Finally, $\underline{\xi}_h$, $\underline{\eta}_{j(h)}$ and $\underline{\delta}_{t(h,j)}$ are independent and $E(\underline{\xi}_h)=E(\underline{\eta}_{j(h)}=E(\underline{\delta}_{t(h,j)})=O$.

Our model is just the analogue of the univariate model on nested sampling with unequal call frequencies discussed by Ganguli (1941), We follow his method for finding the co-efficients of the expected M.P. matrices and end up with the Table 2 of analysis of variance.

TABLE 2

Source of variation	*d.f.*	*S.P. matrices*	*E(M.P. Matriccs)*
Species	13	A	$\Sigma_e+13.381\ \Sigma_\delta+81.27\ \Sigma_\eta+246\ \Sigma_\xi$
Locallties within species	29	B	$\Sigma_e+13.791\ \Sigma_\delta+81.26\ \Sigma_\eta$
Trees within localities	218	C	$\Sigma_e+13.372\ \Sigma_\delta$
Locations*	4248	D	Σ_e

*We do not have this row in our emample since we have only mean observations on each tree·

Here, $A=\left(\sum_h\left[n_h\ldots(\bar{X}_{i_1h}\ \ldots-\bar{X}_{i_1}\ \ldots)\quad(\bar{X}_{i_2h}\ \ldots-\bar{X}_{i_2}\ \ldots)\right]\right)$

and $(A/13)=\begin{bmatrix} 10675527 & 38557 & 30971647 & 53851101 \\ 38557 & 305 & 156717 & 273320 \\ 30971647 & 156717 & 121780733 & 201012595 \\ 53871101 & 373320 & 201012595 & 343055522 \end{bmatrix}$

$B=\left(\sum_h\sum_j\left[n_{hj}\ldots(\bar{X}_{i_1hj}\ldots-\bar{X}_{i_1h}\ldots)\quad(\bar{X}_{i_2hj}\ldots-\bar{X}_{i_2h}\ldots)\right]\right)$

and $(B/29)=\begin{bmatrix} 988308 & 1397 & 1936541 & 3167949 \\ 1397 & 21 & 6231 & 12721 \\ 1936541 & 6431 & 7821366 & 9469922 \\ 316749 & 12721 & 9469922 & 15396656 \end{bmatrix}$

and $C = \left(\sum_h \sum_j \sum_t \left[n_{hjt\cdot} \cdot (\bar{X}_{i_1\, hjt\cdot} \cdot - \bar{X}_{i_1\, hj\cdot} \cdot) \; (\bar{X}_{i_2\, hjt\cdot} - \bar{X}_{i_2\, hj\cdots}) \right]\right)$

and $(C/217) = \begin{bmatrix} 299438 & 558 & 593421 & 994326 \\ 558 & 77 & 1496 & 313 \\ 592421 & 1496 & 21011669 & 2575188 \\ 994325 & 313 & 2575188 & 4281234 \end{bmatrix}$

Note : Referring back to Table 2 showing the analysis of variance, we notice that the corresponding co-efficients in the formula for expected values are approximately equal. Thus will treat it as a problem of nested sampling with equal numbers in the sub-classes and will proceed with the usual procedure of tests of significance.

To test the locality effect, Wilk's criterion was applied to the independent S.P. matrices B and C, with 29 and 217 d.f. respectively and the locality effect was found to be significant be Bartlett's approximate test. Similarly the species effects were found to be significant upon taking the independent S.P. matrices A and B respectively with 13 and 29 d.f. from this, we thus conclude that the species are heterogeneous.

Start of the Problems

After concluding that the fourteen species are heterogeneous we proceed to the main problem of forming clusters as follows :—

We treat A and B respectly as the between and within matrices with 13 and 29 d.f. and present below in Table 3 the means $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, and $\bar{X}_4$ of the characters of the species alongwith the corresponding sizes :—

TABLE 3

Species No.	*Size*	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$
1	264	1311	8.04	3664	6527	0.95	1.27	0.35
2	78	1285	5.35	2989	5657	0.92	1.07	0.64
3	158	1648	7.88	5002	8609	1.74	1.22	0.50
4	212	1137	5.45	3334	5718	1.07	0.95	0.37
5	324	1183	5.13	2877	4818	0.59	1.25	0.57
6	93	1113	5.76	2644	4831	0.57	1.39	0.47
7	380	1368	4.84	3078	5408	0.76	1.12	0.78
8	436	1341	5.57	2999	5560	0.71	1.31	0.70
9	200	1477	6.68	4150	6952	1.19	1.29	0.56
10	90	1251	5.36	3079	5662	0.95	1.04	0 85
11	207	1046	4.87	3102	5302	1.03	0.80	0.34
12	458	1650	6.97	4491	7548	1.29	1.35	0.69
13	348	1647	6.59	4099	7351	1·27	1.24	0.79
14	260	1583	7.41	4285	7697	0.40	1.32	0,61

We solve for L (4×4) and Φ (4×4), the equations :—

$$L\left[\left(\frac{A}{13}\right)\left(\frac{B}{29}\right)^{-1}\right]=\Phi L \qquad \ldots(5.2)$$

by a suitable method, and get :—

$$L(4\times4)\equiv\begin{bmatrix} -0.001064336 & -0.158567182 & -0.00006704 & 0.000590678 \\ 0.001162664 & 0.369050519 & 0.000134634 & -0.000497864 \\ 0.001336923 & -0.069180685 & -0.000181461 & -0.000028873 \\ 0.000325045 & 0.023168191 & 0.000669918 & -0.000460445 \end{bmatrix}$$

and $\Phi(4\times4)=\begin{bmatrix} 25.94 & 0 & 0 & 0 \\ 0 & 11.84 & 0 & 0 \\ 0 & 0 & 5.65 & 0 \\ 0 & 0 & 0 & 1.65 \end{bmatrix}$

Applying Bartlett's modified first approximation, we test the significance of the eigen values of Φ *i.e.*, of 25.94, 11.84, 5.65 and 1.65, and find 1.65 to be non-significant at the 5% level. Discarding thus the last row of matrix $L(4\times4)$ which corresponds to 1.65, we get the matrix $K(3\times4)$. Now, if $\bar{X}^t$ (4×4) be the matrix of mean vectors of species given in the last four columns of Table 3, we get by the following formula :—

$$\bar{Y}^t\,(14\times3)=\bar{X}^t\,K^t$$

the matrix $\bar{Y}^t$ (14×3) of significant discriminant scores which is again presented in Table 3.

Note : The column under $\bar{Y}_1$ corresponds to the largest significant discriminant score, the column under $\bar{Y}_2$ to the second largest and that under $\bar{Y}_3$ to the third largest significant score.

Finally, we compute the distances between the $\binom{14}{2}$ pairs of species of trees by the formula (2 5) and present them in Table 4— called "Table of Distances", arranging the distances of each population from the remaining ones in order of increasing magnitude.

Since there are three significant discriminant scores, we should plot pairwise points *i.e.*, $(\bar{Y}_1, \bar{Y}_2)$, $(\bar{Y}_1, \bar{Y}_3)$ and $(\bar{Y}_2, \bar{Y}_3)$ on the plane graph papers and then keeping them in front we should look to the closeness of the points. To economise space, we, instead, plot, as in Fig. I, fourteen paired points of only most significant ones *i.e.*, $(\bar{Y}_1, \bar{Y}_2)$.

Forming of Culsters

Keeping before us Table 4 and Fig. I and then following the criteria discussed in step 3 of Stage I in section 3, we predict the following clusters :—

(2, 5, 6, 7, 8) (2, 5, 7, 8, 10) (2, 4, 10, 11)

(2, 4, 9, 10), (9, 12, 13 14) and 1 and 3 by themselves.

Stage II

We now correct the above predicted clusters. For each of which we have a tabular set up given below, and from them we obtain the corrected clusters which are listed at the end of table 8.

TABLE 5

Populations Involved	*Computed* T_r^2	*d. f.*	*Tabular* T^2 5%	1%	*Conclusion*	*Cluster*
2,5,6,7,8	34,47	16	25,6131	35,6187	Significant	—
2,5,6,8	19.89	12	21.7064	29.9100	Non-significant	2,5,6,8
2,5,6,7	41.33	,,	,,	,,	Significant	—
2,6,7,8	40.56	,,	,,	,,	-do-	—
2,5,7,8	34.21	,,	,,	,,	-do-	—
5,6.7,8	27.91	,,	,,	,,	-do-	—
2,5,7	20 50	8	16.7783	24.0936	-do-	—
2,6,7	22.83	,,	,,	,,	-do-	—
2,7,8	13.61	,,	,,	,,	Non-significant	2,7,8*
5,6 7	23.37	,,	,,	,,	Significant	—
5,7,8	20.44	,,	,,	,,	-do-	—
6,7,8	12.21	,,	,;	,,	-do-	—
6,7	17.38	4	12,1371	18.2030	-do-	—
5,7	15.35	,,	,,	,,	-do-	—

* The cluster (2, 7, 8) would be found included in a bigger cluster in the Table 7.

TABLE 6

2,4,10,11	15.76	12	21-7064	29.9100	Non-significant	2,4,10.11
9.12,13,14	16.62	12	21-7066	29.9100	-do-	9,12,13,14

TAALE 7

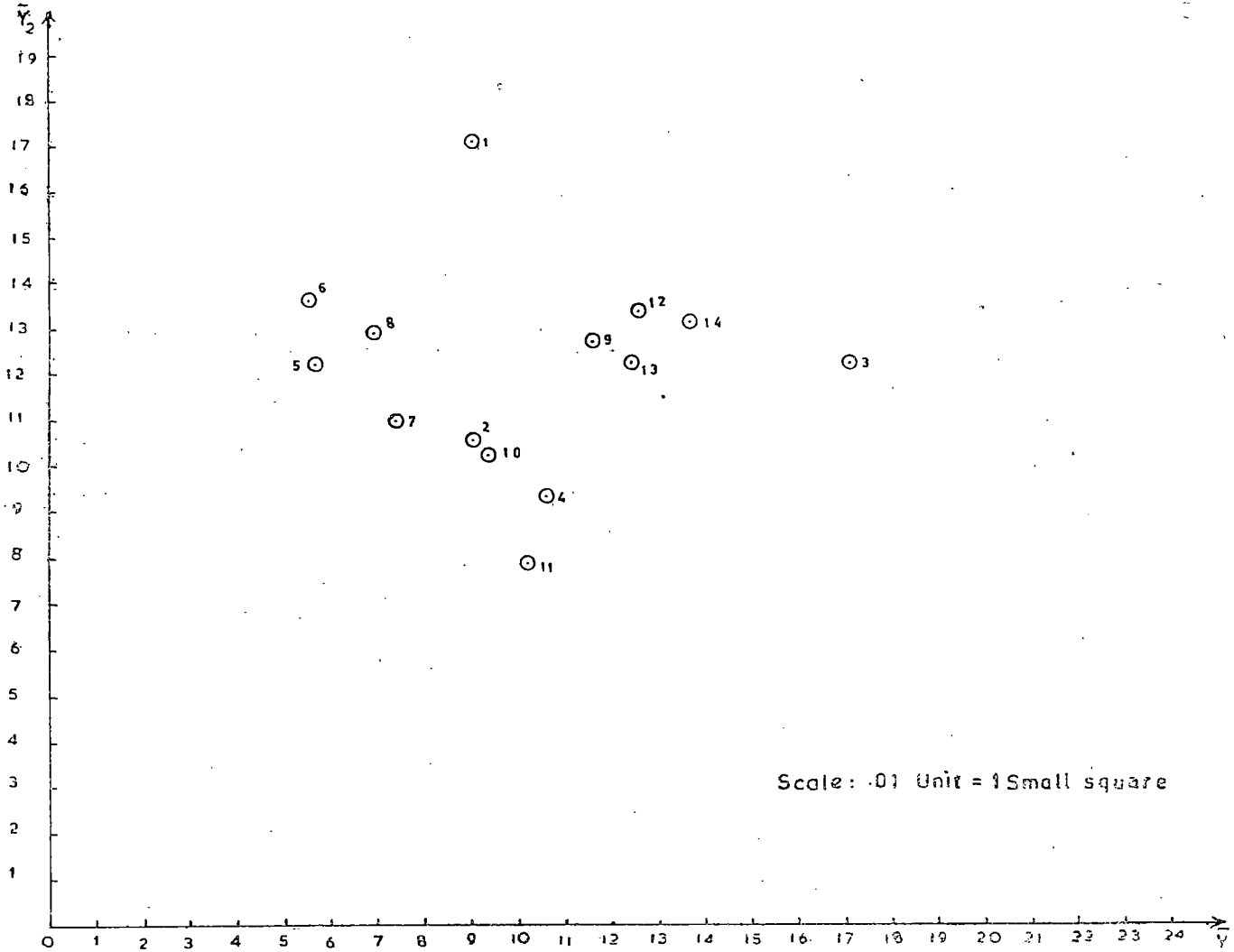
Populations involued	*Computed* T_r^2	*d. f.*	*Tabular* T^2 5%	1%	*Conclusion*	*Cluster*
1,5,7,8,10	34.98	16	25.6131	35.6187	Significant	—
2,5,7,10	27.62	12	21.7064	29.9100	-do-	—
2,5,8,10	25.30	,,	-do-	-do-	-do-	—
5,7,8,10	30.15	,,	-do-	-do-	-do-	—
2,5,7,8	26.31	,,	-do-	-do-	-do-	—
2,7,8,10	21.37	,,	-do-	-do-	Non-Significant	2,7,8,10
2,5,7	20.50	8	16.7783	24.0936	Significant	—
2,5,8	17.79	,,	-do-	-do-	Non-Significant	2,5,8*
2,5,10	18.90	,,	-do-	-do-	Significant	—
2,7,8	22.44	,,	-do-	-do-	-do-	—
5,7.10	23.62	,,	-do-	-do-	-do-	—
5.8,10	19.09	,,	-do-	-do-	-do-	—
5.7	15.35	4	12.1371	18.2030	-do-	—
5,10	12.98	,,	-do-	-do-	-do-	—

*We could exclude this because it has already been included in the bigger cluster (2,5,6,8).

TABLE 8

2,4,9,10	24.37	12	21.7064	29.9100	Significant	—
2,4,9	21.88	8	16.7783	24.0936	-do-	—
2,4,10	8.20	,,	-do-	-do-	Non-significant	2,4,10*
2,9,10	11.45	,,	-do-	-do-	-do-	2,9;10
4,9,10	20.42	,,	-do-	-do-	Significant	—
4,9	16.62	4	-do-	-do-	-do-	—

*We could exclude this because it has already been included in the bigger cluster (2,4,10,11).


$\tilde{Y}_2$
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 23 23 24
$\tilde{Y}$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Scale: ·01 Unit = 1 Small square

1	2	3	4	5	6	7	8	9	10	11	12	13	14
6/ .2669	10/ .0058	14/ .1375	11/ .0250	6/ .0296	5/ .0296	8/ .0436	5/ .0380	12/ .0306	2/ .0058	4/ .0250	13/ .0119	12/ .0119	12/ .0202
9/ .2936	7/ .472	12/ .2565	10/ .0657	8/ .0380	8/ .0800	2/ .0472	7/ .0436	14/ .0460	2/ .0657	10/ .1193	14/ .0202	14/ .0573	9/ .0460
8/ .3516	8/ .1048	9/ .3098	2/ .1106	7/ .0877	7/ .1983	10/ .0836	6/ .0800	13/ .0613	7/ .0836	2/ .1739	9/ .0306	9/ .0613	13/ .0573
12/ .3752	4/ .1106	13/ .3100	9/ .1616	2/ .1453	2/ .2427	5/ .0877	2/ .1048	10/ .1195	11/ .1193	9/ .3067	2/ .2185	2/ .1745	3/ .1375
5/ .4041	9/ .1292	4/ .5384	7/ .2929	10/ .1747	1/ .2669	6/ .1983	10/ .1464	2/ .1292	9/ .1195	13/ .3475	10/ .2223	10/ .1860	10/ .2698
14/ .4374	5/ .1453	10/ .6655	14/ .2951	4/ .3614	10/ .2723	9/ .2675	9/ .2572	4/ .1616	8/ .1464	7/ .3654	3/ .2565	7/ .2780	2/ .2946
2/ .5078	11/ .1739	11/ .7018	13/ .2984	9/ .3722	9/ .4043	13/ .2780	13/ .3245	8/ .2572	5/ .1747	5/ .4483	4/ .3060	4/ .2984	4/ .2951
10/ .5122	13/ .1745	2/ .7232	12/ .3060	1/ .4041	4/ .4457	4/ .2929	12/ .3409	7/ .2675	13/ .1860	14/ .4676	8/ .3409	3/ .3100	1/ .4374
13 .5284	12/ .2185	1/ .9112	5/ .3614	11/ .4483	12/ .5677	12/ .3470	1/ .3516	1/ .2936	12/ .2223	12/ .4840	7/ .3470	8/ .3245	11/ .4674
7/ .5802	6/ .2427	7/ 1.0615	8/ .3712	13/ .5151	11/ .5681	11/ .3654	4/ .3712	11/ .3067	14/ .2698	8/ .4941	1/ .3752	11/ .3475	7/ .4847
4/ .6041	14/ .2946	8/ 1.1198	6/ .4457	12/ .5237	13/ .6080	14/ .4847	14/ .4890	3/ .3098	6/ .2723	6/ .5681	11/ .4840	5/ .5151	8/ .4890
11/ .8504	1/ .5078	5/ 1.3460	3/ .5384	14/ .6718	14/ .7108	1/ .5802	11/ .4941	5/ .3722	1/ .5122	3/ .7018	5/ .5237	1/ .5284	5/ .6718
3/ .9111	3/ .7232	6/ 1.4003	1/ .6041	3/ 1.3460	3/ 1.4003	3/ 1.0615	3/ 1.1198	6/ .4043	3/ .6655	1/ .8504	6/ .5677	6/ .6020	6/ .7108

Thus, from tables 5 to 8, we arrive at the following clusters :

(2,5,6,8), (2,7,8,10), (2,9,10)
(2,4,10,11) (9,12,13,14), (1, by itself } ... ... A
and (3, by itself)

Further, it remains to prove that each and every set of populations of which these clusters form a subset is significant. To do this, we refer back to the Table 4 of distances and also to the Fig. 1 and form the following bigger clusters by incorporating in the corrected clusters the populations lying closest to them :—

(2,5,6,8,10), (2,4,7,8,10), (2,4,7,10,11),
(2,4,9,10,11), (2,9,12,13,14), (3,9,12,13,14),
(2,9,10,13), (1,6), (3,14)

We test the significance of these bigger clusters and find them all to be significant which confirms the conclusion made as in (A) above.

SUMMARY

Various approaches to multiple comparisons methods have by now been made in Univariate Analysis of Variance. The anlogues situations in MANOVA have been considered and three alternative methods have been given. One of the methods has been demonstrated through an illustrative example by taking 10 species of trees and finding amongst them homogeneous groups (or clusters) on the basis of their static bending property.

REFERENCES

[1] Bagai, O.P. (1960) : *Ph. D. Thesis.* University of B.C. Vancouver, Canada.

[2] Bartlett, M. S. (1947) : Multivariate Analysis *Supp. J Roy. Stat. Soc.* 9,176—197.

[3] Duncan, D. B. (1955) : Multiple range and multiple F tests *Biometries* 11, 1—42.

[4] Federer, W.T. (1955) : *Experimental Design* : New York, The Macmillan Co.

[5] Ganguli, M. (1941) : A note on nested sampling *Sankhya* 5, 449—52.

[6] Hartley, H. O. and Pearson, E.S. (1954) : *Biometrika Tables for Statisticians*, Vol. I, Cambridge University Press.

[7] Hotelling. H. (1950) : A generalized T^2 test and measure of multivariate dispersion. *Proc. of Second Berkeley Symposium* on *Math. Stat. and Prob.* 23—41. Berkley, University of California.

[8] Ito, K. (1956) : Asymptatic formula for the distribution of Hotelling generalized T^2— statistics *Ann. Math. Stat.* 27, 1091--1105.

[9] Rao, C.R. (1952) : *Advanced Statistical Methods in Biometric Research.* New York, Wiley.

[10] Siotani, M. (1958) : The extreme value of the generalized distance of the individual points in the multivariate normal samples. *Ann. Instt. Statis. math*, 10, 183—208.

[11] Wilks, S.S. (1932) : Certain generalizations in the analysis of Variance. *Biometrika* 26, 471—494.