TESTS WITH DISCRIMINANT FUNCTIONS IN MULTIVARIATE ANALYSIS

By C. RADHAKRISHNA RAO
Statistical Laboratory, Calcutta.

Introduction

In a paper (Rao : 1946), the author has considered a general method of arriving at suitable studentised statistics which are to be used in tests of linear hypotheses when the observed set of variables, whose expectations are linear functions of unknown parameters, are independent and have the same variance. The principle is to take a linear compound, subject to some restrictions of parametric functions and maximise the ratio of its estimate to the standard error of this estimate.

These methods can be extended to the case of observations from multivariate correlated populations and it is found that tests of significance can, in many cases, be carried out with the use of published tables of t and F alone. The general problem of distribution connected with the statistics arrived at by the above principle has been discussed below and solutions to a few problems which appear to be new have been given.

2. The problem of distribution

In the problems considered by Fisher (1938, 1940), Bartlet (1939), the tests of significance concerning discriminant functions were derived by drawing an analogy with the general regression problem involving pseudovariates. In cases where the introduction of pseudovariates is not possible we may use a standard distribution derived below. Let

$$\begin{matrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ x_{1s} & x_{2s} & \dots & x_{ks} \end{matrix} \quad \dots \quad (2.1)$$

be s sets of observations on k variates $x_1, \dots x_k$ characterised by the probability differential

$$\text{Const. } e^{-\frac{1}{2}[(x_1-m)^2+x^2_2+\dots+x^2_k]}dx_1dx_2\dots dx_k \quad \dots \quad (2.2)$$

We may represent the s observations on the i-th variate $x_{i1}, \dots\dots x_{is}$ by a point P_i in an Euclidean space of s dimensions or by the vector $\overrightarrow{OP_i}$ where O is the origin. The whole sample of observations (2.1) may then be represented by vectors $\overrightarrow{OP_1}, \dots\dots \overrightarrow{OP_k}$ in this space. Let $\overrightarrow{OA}$ represent a vector of unit length along a line which makes equal angles with the coordinate axes. The vector $\overrightarrow{OP_i}$ consists of two components one along $\overrightarrow{OA}$ and the other orthogonal to $\overrightarrow{OA}$ so that products of vectors $\overrightarrow{OP_i}$ and $\overrightarrow{OP_j}$ contain contributions due to components in these two directions which may be represented by y_{ij} and b_{ij} respectively. In this case

$$\left.\begin{aligned} y_{ij} &= s\bar{x}_i\bar{x}_j \\ b_{ij} &= \Sigma(x_{ir}-\bar{x}_i)(x_{jr}-\bar{x}_j) \end{aligned}\right\} \quad \dots \quad (2.3)$$

The configuration of the observations (2.1) is diagramatically represented below

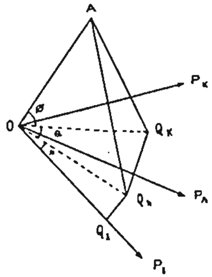

Fig 1.

where Q_i represents the foot of the perpendicular from A to the subspace constituted by the vectors $\vec{OP}_1, \ldots\ldots \vec{OP}_i$. The angles $A\hat{O}Q_k$, $Q_k\hat{O}Q_r$ and $Q_r\hat{O}Q_1$ are represented by ϕ, θ and ψ respectively. The magnitude of AQ_i is given by

$$AQ_i^2 = \frac{|\,b_{pq}\,|}{|\,b_{pq}+s\bar{x}_p\bar{x}_q\,|} \qquad (p,q=1,2,\ldots i) \qquad \ldots (2.4)$$
$$= \frac{1}{1+V_i}$$

where $V_i = \sum_1^i \sum_1^i b_i^{pq} s\bar{x}_p\bar{x}_q$ and b_i^{pq} are the elements of matrix (b_i^{pq}) reciprocal to (b_{pq}), $p, q = 1, 2, \ldots\ldots i$.

We are interested in the distributions of statistics V_r and $(1+V_k)/(1+V_r)$ given by

$$\left.\begin{aligned} V_r &= \frac{\cos^2\phi\cos^2\theta}{1-\cos^2\phi\cos^2\theta} \\ \text{and}\quad \frac{1+V_k}{1+V_r} &= 1+\cot^2\phi\sin^2\theta \end{aligned}\right\} \qquad \ldots (2.5)$$

We may find the distributions of V_r and $U=[(1+V_k)/(1+V_r)]-1$. The joint distribution of ϕ, θ, ψ and t the length of $\vec{OP}_1$ is derivable by taking the product of the volumes contributed by allowing infinitesimal increments to ϕ, θ, ψ and t as

$$\text{Const.}\ e^{-\frac{1}{2}t^2+\sqrt{s}\,m\,t\cos\phi\cos\theta\cos\psi}\ t^{s-1}\,dt$$
$$(\sin\phi)^{s-k-1}d\phi\,(\cos\phi\sin\theta)^{k-r-1}\cos\phi\,d\theta \qquad \ldots (2.6)$$
$$(\cos\phi\cos\theta\sin\psi)^{r-2}\cos\phi\cos\theta\,d\psi$$

Integrating over ψ the above reduces to

$$\text{Const.}\ e^{-\frac{1}{2}t^2}\ t^{s-1-(r-2)/2}\ I_{\frac{r-2}{2}}(2\sqrt{s}\,mt\cos\phi\cos\theta)$$
$$(\sin\phi)^{s-k-1}(\cos\phi)^{k-1}(\cos\theta)^{r-1}(\sin\theta)^{k-r-1}\,dt\,d\phi\,d\theta \qquad \ldots (2.7)$$

Intergrating over t which varies from 0 to ∞ we get the distribution of ϕ and θ as

$$\text{Const.}\ (\sin\phi)^{s-k-1}(\cos\phi)^{k-1}(\sin\theta)^{k-r-1}(\cos\theta)^{r-1}$$
$$ {}_1F_1(s/2,\ r/2,\ 2sm^2\cos^2\phi\cos^2\theta)\ d\phi\ d\theta \qquad \text{.. (2.8)}$$

Changing over to the variables V_r and U we get their joint distribution as

$$\text{Const.}\ \frac{V_r^{\frac{r}{2}-1}}{(1+V_r)^{s/2}}\ {}_1F_1\left(s/2,\ r/2,\ 2sm^2\ \frac{V_r}{1+V_r}\right) dV_r \qquad \text{.. (2.9)}$$

$$\times\ \frac{U^{\frac{k-r}{2}-1}}{(1+U)^{\frac{s-r}{2}}}\ dU \qquad \text{.. (2.10)}$$

which shows that V_r and U are independently distributed. Since the distribution of V_r and U are directly derivable from the joint distribution of means, variances and covariances given by

$$\text{Const.}\ e^{-\frac{1}{2}\sum_i(s\bar{x}_i^2+b_{ii})+s\bar{x}_1 m}\ |\,b_{ij}\,|^{\frac{s-1}{2}-\frac{k+1}{2}}\ \pi\, dx_i\ \pi\, db_{ij} \qquad \text{.. (2.11)}$$

we arrive at the following lemma.

Lemma. If the variables $z_1, \ldots\ldots z_k$ and $c_{ij}\ (i,j=1,2,\ldots k)$ are distributed as

$$\text{Const.}\ e^{-\frac{1}{2}\sum(z_i^2+c_{ii})+f z_1}\ |\,c_{ij}\,|^{\frac{q}{2}-\frac{k+2}{2}}\ \pi\, dz_i\ \pi\pi\, dc_{ij} \qquad \text{.. (2.12)}$$

then

(*i*) the statistic $V_r=\sum_1^r\sum_1^r c^{pq} z_p z_q$ is distributed as

$$\text{Const.}\ \frac{V_r^{\frac{r}{2}-1}}{(1+V_r)^{\frac{q+1}{2}}}\ {}_1F_1\left((q+1)/2,\ r/2,\ 2f^2\ \frac{V_r}{1+V_r}\right) dV_r \qquad \text{.. (2.13)}$$

so that when $f=0$ the statistic $V_r(q+1-r)/r$ can be used as the variance ratio with r and $(q+1-r)$ degrees of freedom and

(*ii*) the statistic $U=[(1+V_k)/(1+V_r)]-1$ is distributed as

$$\text{Const.}\ \frac{U^{\frac{k-r}{2}-1}}{(1+U)^{\frac{q+1-r}{2}}}\ dU \qquad \text{.. (2.14)}$$

so that $U(q+1-k)/(k-r)$ can be used as a variance ratio with $(k-r)$ and $(q+1-k)$ degrees of freedom.

3. Generalisation of Student's t

Students' test connected with pairs of observations admits generalisation in two directions.

The first is to test whether the means of p correlated variates are the same on the basis of a sample of size n.

If $x_{1i}, \ldots\ldots x_{pi}$ are the observations on the variates corresponding to the i-th sample we replace the observations by a linear compound $x_i=l_1x_{1i}+\ldots\ldots+l_px_{pi}$ subject to the restriction $l_1+l_2+\ldots+l_p=0$. The problem, formally, reduces to testing whether the mean value of the variate x is zero. The appropriate statistic for this is $c=\sqrt{n}\,\bar{x}/s$ where

$$\left.\begin{aligned} \bar{x} &= l_1\bar{x}_1+\ldots\ldots+l_p\bar{x}_p \\ \bar{x}_i &= (x_{i1}+\ldots\ldots+x_{in})/n \\ (n-1)s^2 &= \Sigma\,\Sigma\, l_i\, l_j\, b_{ij} \\ b_{ij} &= \Sigma\,(x_{ir}-\bar{x}_i)(x_{jr}-\bar{x}_j) \end{aligned}\right\} \qquad \text{.. (3.1)}$$

The compounding coefficients $l_1, \ldots\ldots, l_p$ are chosen so as to maximise this statistic or a constant times this statistic. Denoting the maximum value of $n\bar{x}^2/(n-1)s^2$ by V_{p-1} we get V_{p-1} as the root of the determinantal equation.

$$\begin{vmatrix} n\bar{x}_1\bar{x}_1 - V_{p-1}b_{11} & \ldots & n\bar{x}_1\bar{x}_p - V_{p-1}b_{1p} & 1 \\ \ldots & \ldots & \ldots & \cdot \\ n\bar{x}_p\bar{x}_1 - V_{p-1}b_{p1} & \ldots & n\bar{x}_p\bar{x}_p - V_{p-1}b_{pp} & 1 \\ 1 & \ldots & 1 & 0 \end{vmatrix} = 0 \qquad \ldots (3.2)$$

To find the value of V_{p-1} we may follow an alternative procedure which leads to the problem of distribution as well. By arbitrary choice of constants we can construct $(p-1)$ linear combinations

$$y_i = m_{i1} x_1 + \ldots + m_{ip} x_p \qquad \ldots (3.3)$$

such that $\Sigma m_{ij}=0$ for all $i=1, 2, \ldots\ldots (p-1)$. Choosing a linear compound $l_1x_1+\ldots+l_px_p$ of $x_1, x_2 \ldots x_p$ such that $\Sigma l=0$ is same as choosing an arbitrary linear compound of $y_1, \ldots y_{p-1}$ as defined in (3.3). If we choose the linear compound $\lambda_1 y_1+\ldots+\lambda_{p-1} y_{p-1}$ where λ's are free and construct the statistic

$$v = \frac{\lambda_1 \bar{y}_1 + \ldots + \lambda_{p-1} \bar{y}_{p-1} \sqrt{n}}{\sqrt{\Sigma \Sigma \lambda_i \lambda_j c_{ij}}} \qquad \ldots (3.4)$$

where $c_{ij}=\Sigma(y_{ir}-\bar{y}_i)(y_{jr}-\bar{y}_j)$ we get the maximum value V'_{p-1} of v^2 as the root of the equation

$$| n\bar{y}_i\bar{y}_j - V'_{p-1} c_{ij} | = 0, \; i, j, = 1, 2, \ldots (p-1) \qquad \ldots (3.5)$$

or

$$V'_{p-1} = \sum_i^{p-1} \sum_j^{p-1} n\, c^{ij} \bar{y}_i \bar{y}_j \qquad \ldots (3.6)$$

where c^{ij} are the elements of the matrix reciprocal to (c_{ij}), $i, j=1, 2 \ldots (p-1)$. The maximised values V_{p-1} and V'_{p-1} in the two cases must necessarily be the same so that $V_{p-1}= V'_{p-1}$. It immediately follows that the statistic V_{p-1} or V'_{p-1} is invariant for any system of parameters m_{ij} such that $\sum_j m_{ij}=0$ chosen to construct the $(p-1)$ variates $y_1, \ldots\ldots, y_{p-1}$. In general we may choose a set which introduces simplicity in the evaluation of the statistic V_{p-1}.

The distribution problem is also simplified for we need only find the distribution of V_{p-1} from the joint distribution of $\bar{y}$'s and c_{ij}. As the statistic is invariant under linear transformations of the variates we can assume without loss of generality that the variates are distributed independently with unit variances and that the mean values of all y's except y_1 are zeroes. The mean value of y_1 denoted by m is zero on the null hypothesis. The actual value of m in terms of $m_1, \ldots m_p$ the mean values of $x_1, \ldots x_p$ and their variances and covariances α_{ij} is given by the root of the equation.

$$\begin{vmatrix} m_i m_j - m^2 \alpha_{ij} & \vdots & 1 \\ \ldots\ldots\ldots\ldots & \vdots & \ldots \\ 1 & \vdots & 0 \end{vmatrix} = 0 \qquad \ldots (3.7)$$

Since the probability density $\bar{y}$'s and c_{ij} is given by

$$\text{Const. } e^{-\frac{n}{2}\Sigma(\bar{y}_i^2 + c_{ii}) + n\, m\, \bar{y}_1} \, |c_{ij}|^{\frac{n}{2} - \frac{p}{2}} \qquad .. \quad (3.8)$$

We get the distribution of V_{p-1} as (2·13) with $r=p-1$, $q=n-1$ and $f=\sqrt{n}\, m$, so that to test the null hypothesis $m=0$ we may use the statistic $V_{p-1}(n-p+1)/(p-1)$ as the variance ratio with $(p-1)$ and $(n-p+1)$ degrees of freedom. It is also instructive to ascertain the values of $l_1, \ldots l_p$ of the compounding coefficients of the original variables so that we have a knowledge of a contrast leading to maximum discrepancy. The coefficients $l_1 \ldots l_p$ are obtainable from the linear equations

$$l_1(n\,\bar{x}_1\,\bar{x}_i - V_{p-1}\, b_{1i}) + \ldots + l_p(n\bar{x}_p\bar{x}_i - V_{p-1} b_{pi}) + \mu = 0, \quad i = 1, 2, \ldots p \qquad .. \quad (3.9)$$

and

$$l_1 + l_2 + \ldots + l_p = 0$$

where μ is also a variable to be solved for simultaneously.

On the other hand we can also test for the significance of t specified contrasts

$$y_i = l_{1i}\, x_1 + \ldots + l_{pi}\, x_p, \quad \sum_j l_{ji} = 0, \quad i = 1, 2 \ldots t \qquad .. \quad (3.10)$$

We need only take a linear compound of y's and maximise a certain statistic. If we denote by

$$d^{ij} = \Sigma(y_{ir} - \bar{y}_i)(y_{jr} - \bar{y}_j) \qquad .. \quad (3.11)$$

then the appropriate statistic is

$$V_t = n \Sigma \Sigma\, d^{ij}\, \bar{y}_i\, \bar{y}_j \qquad .. \quad (3.12)$$

where d^{ij} are the elements of the matrix reciprocal to (d_{ij}), $i, j = 1, \ldots t$. The quantities d_{ij} are derivable as linear combinations of b_{ij}'s defined in (3.1). To test for the significance of V_t we use the statistic $V_t(n-t)/t$ as the variance ratio with t and $(n-t)$ degrees of freedom.

If the further question is asked as to whether the contrasts (3.10) explain away the the differences among the p variates we have to compare V_t and V_{p-1} as derived from specified contrasts and all the available contrasts respectively. The distribution derived in (2.14) gives that the statistic

$$\frac{n-p+1}{p-t-1} \left(\frac{1+V_t}{1+V_{p-1}} - 1 \right) \qquad (3.13)$$

can be used as a variance ratio with $(p-t-1)$ and $(n-p+1)$ degrees of freedom.

As an illustrative example we may consider the following problem connected with the testing of bias in using small sample plots in crop-cutting experiments.

The design, due to P. C. Mahalanobis, consisted in locating a random point in a field and constructing four concentric circles of radii 2 ft, 4 ft., 6 ft., and 8 ft. respectively. The inner circle is harvested first and the yield is recorded. The first annular, the second and third annular rings are harvested and the yields are separately recorded. From these by suitable addition we can get the yields as given by circular sample cuts of radii 2ft, 4 ft, 6 ft, and 8 ft at a chosen point. . The yield rates (in some unit of weight per unit area) as given by the four circles are represented by c_1, c_2, c_3 and c_4 respectively. These variables are correlated and are subject to different errors depending on the correlation within a field. The problem is to test whether the mean yield rates as given by sample cuts of various sizes are the same.

Two dimensional charts representing the scatter of any pair of yield rates such as c_1 and c_2 gave that the arrays of c_2 for given c_1 are not homoscedastic and variance increases with increase in c_1. The data are then split into groups by the yield rate determined by c_1, so that within a group the arrays are nearly homoscedastic and the test for bias is carried on in each group. Incidentally the nature of bias for different intensities of yield rates can be studied.

In this case we may consider the variables

$$y_1=c_2-c_1, \; y_2=c_3-c_2, \; y_3=c_4-c_3$$

The mean values and the d_{ij} matrix constructed from 38 observations in the range of yield rate 0 to 10, are given below

$$\bar{y}_1 = -\cdot 1 \;, \quad \bar{y}_2 = \cdot 43, \quad \bar{y}_3 = \cdot 24$$

$$d_{ij} = \begin{pmatrix} 90\cdot 49 & -23\cdot 56 & -14\cdot 06 \\ -23\cdot 56 & 80\cdot 26 & -45\cdot 98 \\ -14\cdot 06 & -45\cdot 98 & 93\cdot 94 \end{pmatrix}$$

$$V_3 = \cdot 23$$

so that $35V_3/3=2\cdot 68$ can be considered as a variance ratio with 3 and 35 degrees of freedom. The result is not significant showing that in this group of yield rates there is no evidence of bias.

The second generalisation of students' t is concerned with testing, on the basis of a sample of size n from a $2p$ variate population containing the variables $y_1, y_2 \ldots y_{2p}$ whether $E(y_i)=E(y_{i+p})$ for $i=1, 2 \ldots p$.

From the $2p$ variates we construct the p variates $z_i=y_i-y_{i+p}(i=1, 2 \ldots p)$ in which case the test reduces to that of testing whether p correlated variables have assigned mean values viz, zeroes in this case ; a problem which has been considered by Hotelling (1935). We use the statistic

$$\frac{n-p}{p} \; V_p = \frac{n-p}{p} \, n \, \Sigma \, \Sigma \, \bar{z}_i \, \bar{z}_p \, d^{ij} \qquad (3.14)$$

as the variance ratio with p and $(n-p)$ degrees of freedom.

The problem where the p variates of the first set are uncorrelated with the variates of the second set and the dispersion matrices of the two sets are identical can be answered with the use of Mahalanobis' Generalised distance.

In a special case where the estimates of variances and covariances come out as constant multipliers of a stochastic variable the problem can be answered with the use of Mahalanobis' Generalised distance but the distribution in this case is different from the one to be used above. This, statistic has been termed by the author as the Mahalanobis' distance of the second kind in (Rao : 1944). The class of hypotheses arising out of these problems can be appropriately treated as tests in least squares fully discussed by the author in (Rao : 1946).

4. A PROBLEM IN THE CLASSIFICATION OF THREE POPULATIONS

An important problem that arises in the classification of three multivariate populations π_1 π_2 and π_3 is to test whether the population π_2 is nearer to one of π_1 and π_3 when it is known that π_1 and π_3 are different. If the variates corresponding to the three populations π_1, π_2 and π_3 are represented by $x_1, \dots x_p$; $y_1 \dots y_p$; and $z_1, \dots z_p$ then we replace the p variates by a linear combination of these variates defined by $x = \Sigma lx$, $y = \Sigma ly$, $z = \Sigma lz$. The problem is thus formally reduced to the case of a univariate classification.

There are two contrasts arising out of the three mean values of the variables x,y and z which give the inequalities in the mean values of the three populations. One such contrast is supplied by E(x) = E(z) which gives the differences in the mean values of π_1 and π_3. Another contrast is 2E(y) − E(x) − E(z) which determines the nearness of π_2 to one of π_1 and π_3. We take this contrast and choose the compounding coefficients so as to maximise the ratio of its estimate to the standard error.

If n_1, n_2 and n_3 are sample sizes available for the populations π_1, π_2 and π_3 then assuming equality of variances and covariances we can build up the estimates of variances and covariances from the quantities

$$c_{ij} = \sum_{r=1}^{n_1} (x_{ir} - \bar{x}_i)\, x_{jr} + \sum_{r=1}^{n_2} (y_{ir} - \bar{y}_i)\, y_{jr} + \sum_{r=1}^{n_3} (z_{ir} - \bar{z}_i)\, z_{jr} \qquad \dots \quad (4.1)$$

The ratio to be maximised is

$$\frac{\sqrt{n}\; \Sigma\, l_i\, (2\bar{y}_i - \bar{x}_i - \bar{z}_i)}{\sqrt{\Sigma \Sigma\, l_i\, l_j\, c_{ij}}} \qquad \dots \quad (4.2)$$

where $1/n = 4/n_2 + 1/n_1 + 1/n_3$. The maximum value of the square of (4.2) comes out as

$$V_p = \sqrt{n}\, \Sigma \Sigma\, c^{ij}\, (2\,\bar{y}_i - \bar{x}_i - \bar{z}_i)(2\bar{y}_j - \bar{x}_j - \bar{z}_j) \qquad \dots \quad (4.3)$$

where c^{ij} are the elements of the matrix reciprocal to (c_{ij}). Since c_{ij} are determined with $(n_1 + n_2 + n_3 - 3)$ degrees of freedom it follows that the statistic

$$\frac{n_1 + n_2 + n_3 - 2 - p}{p}\; V_p \qquad \dots \quad (4.4)$$

can be used as the variance ratio with p and $(n_1 + n_2 + n_3 - p - 2)$ degrees of freedom.

When the test gives a significant result the nearness to π_1, or π_3 has to be determined by the evaluation of Mahalanobis' generalised distances between π_1, π_2 and π_3, π_2. If the former is greater than the latter then π_2 is nearer to π_3 and viceversa.

This test can be extended to answer another type of problem considered by Wald (1944) connected with the classification of an individual into one of two groups, π_1 and π_3 from which samples of sizes n_1 and n_3 are available. We need construct the statistic (4.3) with $n_2 = 1$ and test for the significance of V_p. If V_p is significant then the individual can be classified as belonging to π_1 or π_2. If not we may have to be doubtful about the classification of the individual. This appears to be a symmetrical test when the individual is known to belong to one of the two groups. The exact nature of this test, however, requires further investigation. The same statistic (4·3) cad be used when a group of n_3 individuals are to be classified with one of two groups π_1 and π_2 from which samples of sizes n_1 and n_2 are available.

References

1. Bartlett, M. S. (1939). The standard errors of discriminant function coefficients. *J. R. S. S. Suppl.* 6, 169-73.

2. Fisher, R. A. (1938). The statistical utilisation of multiple measurements. *Ann. of Eugen.*, 8, 376-80.

3. ——— (1940). The precision of discriminant functions. *Ann. of Eugen.*, 10, 422-29.

4. Rao, C. R. (1944) Generalised variance of populations *Proc Ind. Sc. Cong.*

5. ——— (1946). On linear combinations of observations and the general theory of least squares, *Sankhya*, 7.

6. Wald, A. (1945). On a statistical problem arising in the classification of an individual in to one of two groups. *Ann. of Math. Stat.* 15, 145-62.

Paper received : 9 September 1945