USE OF DISCRIMINANT AND ALLIED FUNCTIONS IN MULTIVARIATE ANALYSIS

By C. RADHAKRISHNA RAO

Indian Statistical Institute

SUMMARY. The use of discriminant function is justified only when it is known that the individual to be classified belongs to a certain subset of alternative groups. In this paper statistical methods are developed for examining whether a given individual can be supposed to have arisen from one of the populations in the subset and then using the discriminant function.

The use of certain other linear functions in problems of classification of groups, prediction, etc. has also been discussed.

1. Introduction

That the likelihood ratio is the best criterion for discrimination between two alternative hypotheses is justified by the fact that it is a sufficient statistic when only the two alternatives are considered (Smith, 1947). But, in practice, when it is not certain that an observed specimen belongs to one or the other of two given groups, there is need to consider the possibility of its belonging to some unknown group. The likelihood ratio which is equivalent to the discriminant function may not be sufficient when other alternatives are considered. Elsewhere (Rao, 1960), I have considered an example which gave rise to some controversy because such a possibility was ignored and a discriminant function was used to decide between two given alternatives (Bronowski and Long, 1951; Yates and Healy, 1951).

It is, therefore, important first to examine whether the discriminant function constructed on the basis of two given alternatives is *adequate* for drawing inference on an observed specimen. Only when this is confirmed by the observations on the specimen can we actually use the discriminant function for coming to a decision. The paper is devoted to development of suitable tests for this purpose.

Concepts of size and shape and their measurement based on multiple variables are also considered. These functions have been found useful in the study of inter-relationships and in the problem of discrimination between groups.

2. Sufficiency of the discriminant function over a wider set of alternatives

Let us consider two multivariate normal distributions with parameters (μ_1, Λ) and (μ_2, Λ) where μ_1, μ_2 are vectors of mean values and Λ, the common dispersion matrix. The log likelihood ratio, which is the linear discriminant function based on these two alternatives, is proportional to

$$\delta' \Lambda^{-1} x, \qquad \dots \quad (2.1)$$

where $\delta = \mu_1 - \mu_2$, Λ^{-1}, the reciprocal of Λ and x, the vector of observations. Lemma 1, which may be considered as a generalization of Smith's result, shows that the discriminant function (2.1) is not only sufficient for the alternatives from which it is derived but also for a wider set.

Lemma 1 : *The discriminant function* $\delta' \Lambda^{-1} x$ *is sufficient for the set of alternatives* $(a_1\mu_1 + a_2\mu_2, \Lambda)$, *where* $a_1 + a_2 = 1$, i.e., *the probability distribution of* x *given* $\delta' \Lambda^{-1} x$ *is the same for all expected values of* x *lying on the line joining* μ_1 *and* μ_2.

To prove the lemma it is enough to show that the ratio

$$P(x|a_1\mu_1+a_2\mu_2) \div P(\delta'\,\Lambda^{-1}\,x|a_1\mu_1+a_2\mu_2) \quad \ldots \quad (2.2)$$

of normal densities is independent of (a_1, a_2) provided $a_1+a_2 = 1$. The logarithm of (2.2) is proportional to

$$(x-a_1\delta-\mu_2)'\,\Lambda^{-1}(x-a_1\delta-\mu_2)-\frac{[\delta'\,\Lambda^{-1}(x-a_1\delta-\mu_2)]^2}{\delta'\,\Lambda^{-1}\delta} \quad \ldots \quad (2.3)$$

where a_2 has been replaced by $(1-a_1)$. The coefficients of a_1^2 and a_1 in (2.3) are obviously zero establishing the required result.

Lemma 2 which gives an extension of the result of Lemma 1 to the case of k multivariate normal populations with parameters (μ_i, Λ), $i = 1, \ldots, k$, can be proved on the same lines. Let $L_1, \ldots, L_{k-1}$ be $(k-1)$ independent likelihood ratios or $(k-1)$ independent discriminant functions arising out of some $(k-1)$ pairs of populations.

Lemma 2 : *$L_1, \ldots, L_{k-1}$ are sufficient for the parameter set $a_1\mu_1+\cdots+a_k\mu_k$ where $a_1+\cdots+a_k = 1$.*

3. Test prior to the application of discriminant function

As observed earlier, in any practical problem it is of some importance to consider the possibility of an observed specimen belonging to a group other than any one of those specified. For this purpose, it appears more natural first to test the hypothesis that the discriminant function is sufficient for drawing an inference on the observed specimen. Only if there is no evidence against this hypothesis could we use the discriminant function for further study, otherwise the question of classification of the observed specimen as a member of one of the two given groups does not arise. Again, we do not use the discriminant function just to decide between the two given alternatives only but also allow yet another possibility of the observed specimen belonging to a third group with its mean values lying on the line joining the mean values of the two given groups.

We shall first consider the case where all the parameters are known and the result of Lemma 1 regarding the sufficiency of the discriminant function is strictly valid.

The probability density of the observation x for an arbitrary mean μ is

$$P(x|\mu) = \text{const.} \exp\{-\tfrac{1}{2}(x-\mu)'\Lambda^{-1}(x-\mu)\}. \quad \ldots \quad (3.1)$$

The criterion for testing the hypothesis that $\mu = a_1\mu_1+a_2\mu_2$ where μ_1 and μ_2 are specified, is provided by the likelihood ratio,

$$-2\log_e \frac{\sup\limits_{a_1+a_2=1} P(x|a_1\mu_1+a_2\mu_2)}{\sup\limits_{\mu} P(x|\mu)} \quad \ldots \quad (3.2)$$

$$= \inf_{a_1} (x-\mu_2-a_1\delta)'\,\Lambda^{-1}\,(x-\mu_2-a_1\delta)$$

$$= (x-\mu_2)'\,\Lambda^{-1}\,(x-\mu_2)-\frac{[\delta'\,\Lambda^{-1}\,(x-\mu_2)]^2}{\delta'\,\Lambda^{-1}\,\delta} \quad \ldots \quad (3.3)$$

$$= (x-\mu_1)'\,\Lambda^{-1}\,(x-\mu_1)-\frac{[\delta'\,\Lambda^{-1}\,(x-\mu_1)]^2}{\delta'\,\Lambda^{-1}\delta} \text{ (by symmetry)} \quad \ldots \quad (3.4)$$

The statistic (3.3) is the difference between

$$\chi^2 = (x-\mu_2)' \Lambda^{-1} (x-\mu_2) \quad \ldots \quad (3.5)$$

which is a χ^2 on p d.f. and provides a test of the hypothesis that the individual belongs to the second population when nothing is known about the alternatives, and

$$\chi_1^2 = \frac{[\delta' \Lambda^{-1} (x-\mu_2)]^2}{\delta' \Lambda^{-1} \delta} \quad \ldots \quad (3.6)$$

which is a χ^2 on 1 d.f. and is useful for examining whether the individual belongs to the second population given that the alternatives for mean values are points on the line joining μ_1 and μ_2. We may then write the required test criterion (3.3) as

$$\chi_2^2 = \chi^2 - \chi_1^2 \quad \ldots \quad (3.7)$$

which is a χ^2 on $(p-1)$ d.f. If χ_2^2 is significantly large, there is evidence of the observed specimen belonging to a third group. On the other hand, if the value of χ_2^2 is not large, there is some assurance that the use of the discriminant function will not be misleading. The closeness of the observed specimen to one or the other of two groups is then judged by the discriminant function or same as chi-squares

$$\frac{[\delta' \Lambda^{-1} (x-\mu_2)]^2}{\delta' \Lambda^{-1} \delta} \text{ and } \frac{[\delta' \Lambda^{-1} (x-\mu_1)]^2}{\delta' \Lambda^{-1} \delta} \quad \ldots \quad (3.8)$$

each with 1 d.f. As observed earlier, both these chi-squares may be large in a particular situation, indicating another possibility of the observed specimen belonging to a third group.

For example, in the classification of the Highdown Skull (Rao, 1952, pp. 294-296), considering the Bronze Age population we find $\chi^2 = 12.694$ and $\chi_1^2 = 3.993$ giving a difference, $\chi_2^2 = 8.701$ on 5 d.f. Considering the alternative of Maiden Castle population, $\chi^2 = 9.091$ and $\chi_1^2 = .390$ giving the same difference, $\chi_2^2 = 8.701$ on 5 d.f. which is not significantly high. We may then use χ_1^2 values on 1 d.f. for arriving at a decision. The χ_1^2 for Bronze Age is significant at the 5% level and that for Maiden Castle is small, which suggests a classification of the Highdown Skull as a member of the latter population.

Bronowski and Long (1952) amended the procedure, suggested by them in the earlier paper (1951), of using the discriminant function without a preliminary test of its sufficiency, to admit the possibility of an observed specimen belonging to a third group. Their new approach, however, does not make use of the discriminant function at all.

4. Tests when the number of alternatives is more than two

As before we consider the case where the alternative distributions are completely specified, i.e., the parameters involved are all known. In view of the result of Lemma 2, it may be examined first, whether $(k-1)$ independent likelihood ratios or discriminant functions arising out of k specified distributions are sufficient or not. Or in other words, whether the observed specimen can be considered to have arisen from a population whose mean is on the hyper plane determined by the means of the k specified populations. If, as before, x stands for the vector of observations, $\mu_1, ..., \mu_k$ for mean values and Λ, the common dispersion matrix, the test criterion is

$$\chi_2^2 = \min_{\Sigma a=1} (x-\Sigma a_i\mu_i)' \Lambda^{-1}(x-\Sigma a_i\mu_i) \quad ... \quad (4.1)$$

which is a χ^2 on $(p-k+1)$ d.f. A method of computing a statistic of the form (4.1) is discussed in an earlier paper (Rao, 1959). If the value of χ_2^2 in (4.1) is not significantly high we consider only $(k-1)$ discriminant functions instead of p variables. For each group we then have a χ^2 on $(k-1)$ d.f. to judge the affinities of the observed specimen with that particular group.

5. Tests when the common dispersion matrix is estimated

In case Λ is estimated by S, which has Wishart's distribution on n d.f. the statistic

$$\frac{n-p+k}{n(p-k+1)} \chi_2^2 \quad ... \quad (5.1)$$

which is constant times (4.1) with Λ^{-1} replaced by S^{-1}, has F distribution on $(p-k+1)$ and $(n-p+k)$ d.f. In the case $k = 2$, the F statistic is

$$\frac{n-p+2}{n(p-1)} \chi_2^2 = \frac{n-p+2}{n(p-1)} \left\{(x-\mu_2)'S^{-1}(x-\mu_2) - \frac{[\delta'S^{-1}(x-\mu_2)]^2}{\delta'S^{-1}\delta}\right\} \quad ... \quad (5.2)$$

with $p-1$ and $n-p+2$ d.f. These results follow from a general distribution problem solved in an earlier paper (Rao, 1959).

There is the further problem of testing the hypothesis that the specimen belongs to a particular group, using the estimated discriminant function or functions in the case of more than two alternatives. For $k = 2$, the test criterion for the hypothesis that the specimen belongs to (say) the second group is

$$\frac{n-p+1}{1} \left\{ \frac{[\delta'S^{-1}(x-\mu_2)]^2}{\delta'S^{-1}\delta} \div (n+\chi_2^2)\right\} \quad ... \quad (5.3)$$

which is F on 1 and $(n-p+1)$ d.f. A similar statistic with an F distribution on $(k-1)$ and $(n-k+3)$ d.f. can be constructed in the case of k alternatives.

No exact tests of the types (5.2) and (5.3) seem to be available when the mean values of the two alternative populations are themselves estimated. An approximate test of the type (5.2) is provided by the smallest root of the determinantal equation defined by Fisher (1939). The further problem of using an estimated discriminant function as in (5.3) is not satisfactorily solved.

6. Size and shape functions

Penrose (1947) defined certain linear functions of measurements as representing the size and shape of an organism. More general functions were constructed by the author (Rao, 1958) to examine differences in size and shape between groups of organisms. These concepts are further generalised in this section.

Let us consider a linear function

$$X = a_1 x_1 + \dots + a_k x_k + a_{k+1} x_{k+1} + \dots + a_{k+m} x_{k+m} \qquad \dots \quad (6.1)$$

of $(k+m)$ variables $x_1, \dots, x_{k+m}$, with unit variance, i.e.,

$$\Sigma\, \Sigma\, a_i a_j \lambda_{ij} = 1 \qquad \dots \quad (6.2)$$

where λ_{ij} is the covariance between x_i and x_j.

We wish to choose the coefficients $a_1, \dots, a_{k+m}$ in such a way that a given increase in X produces on the average maximum changes, in given directions and given ratios, in a subset of the variables (say) $x_1, \dots, x_s$. If the regression of x_i on X is taken to be linear then the average change in x_i for a unit change in X is exactly the regression coefficient of x_i on X. Let the ratios of specified changes in $x_1, \dots, x_s$ including signs be $\rho_1 : \rho_2 : \dots : \rho_s$. Using a multiplier α, we write the equations

$$\left.\begin{array}{ll} \Sigma\, a_i \lambda_{ij} = \alpha\, \rho_j, & j = 1, \dots, s \\ \Sigma\, a_i \lambda_{ij} = \alpha\, \gamma_j, & j = 1, \dots, k-s \end{array}\right\} \qquad \dots \quad (6.3)$$

where $\gamma_1, \dots, \gamma_{k-s}$ are arbitrary constants. The algebraic problem is that of maximising α with the restriction (6.2) on a_i, by suitably choosing γ_j. Let Λ^{-1} be partitioned

$$\Lambda^{-1} = (\Lambda_1 \vdots \Lambda_2)$$

such that Λ_1 is of order $(k \times s)$, Λ_2 of order $(k \times \overline{s-k})$ and ρ and γ are the column vectors of elements ρ_i and γ_j respectively. The solution for a, the column vector of a_i is

$$a = (\Lambda_1 \vdots \Lambda_2)\binom{\rho}{\gamma} = \alpha(\Lambda_1\rho + \Lambda_2\gamma). \qquad \dots \quad (6.4)$$

Using (6.2)

$$\alpha^2 = 1 \div (\Lambda_1\rho + \Lambda_2\gamma)'\Lambda(\Lambda_1\rho + \Lambda_2\gamma). \qquad \dots \quad (6.5)$$

α is a maximum when the denominator in (6.5) is a minimum, i.e., γ is chosen to satisfy

$$\Lambda_2'\Lambda\Lambda_2\gamma = \Lambda_2'\Lambda\Lambda_1\rho. \qquad \dots \quad (6.6)$$

Having chosen γ, we determine α from (6.5) and then a from (6.4) to obtain the desired linear function.

An alternative way of specifying the linear function (6.1) is as follows. The equations ensuring a given ratio of the regression coefficients of x_i on X, $i = 1, \dots, s$ are

$$\Sigma\, a_i \lambda_{ij} = \alpha\, \rho_j, \quad j = 1, \dots, s. \qquad \dots \quad (6.7)$$

The residual variance of x_i given x is

$$\lambda_{ii} - \alpha^2 \rho_i^2 / a'\Lambda a. \qquad \dots \quad (6.8)$$

We may choose a, α, subject to restrictions (6.7) to minimise the sum of the residual variance for the s variables $x_1, \dots, x_s$

$$\sum_1^s (\lambda_{ii} - \alpha^2 \rho_i^2 / a'\Delta a)$$

or maximise $\alpha^2 / a'\Delta a$. The equations leading to the determination of a are same as (6.4) and (6.6), thus establishing the equivalence of the two problems.

If ρ_i are all positive we may call (6.1) a size function with respect to the characters $x_1, \dots, x_s$, in the sense that an increase in the value of size results in an increase, on the average, in each of the characters $x_1, \dots, x_s$. By choosing some of the ρ_i to be positive and others negative we obtain a function with the property that an increase in its value increases the value of some characters and decreases the value of some others. Such a function may be called a shape function.

It may be noted that in the construction of a size or a shape function for $x_1, ..., x_s$ we have used other measurements $x_{s+1}, ..., x_k$ also, and further that the linear function (6.1) constructed by the method indicated is invariant for linear transformations of the extra variables $x_{s+1}, ..., x_k$.

The present approach admits the possibility of constructing a linear function of any set of anthropometric measurements to represent shape of the head and to discriminate between long and broad headed people, instead of using the cephalic index. Such a linear function was used in an earlier paper (Rao, 1950) to study differences in head shape of castes and tribes belonging to different States in India.

A closely related problem is that of determining a linear function X of k variables such that the sum of the residual variances of $x_1, ..., x_s$ given X is a minimum, without placing any restrictions on the regression coefficients. Since the sum of residual variances is

$$\sum_{i=1}^{s} \left\{ \lambda_{ii} - \frac{(\Sigma \lambda_{ij} a_j)^2}{a'\Lambda a} \right\}$$

the problem is same as that of maximising

$$\sum_{i=1}^{s} \frac{(\Sigma \lambda_{ij} a_j)^2}{a'\Lambda a} = \frac{a'Ba}{a'\Lambda a} \qquad \ldots \quad (6.9)$$

where $B = A'A$, A being the $(s \times k)$ matrix obtained by taking the first s rows of the matrix Λ. The vector a maximising the ratio (6.9) is then the latent vector corresponding to the maximum latent root of the determinantal equation

$$|B - \lambda \Lambda| = 0. \qquad \ldots \quad (6.10)$$

In fact, by considering the first, second, third, ... latent vectors we obtain a series of linear functions $X_1, X_2, X_3, ...$ such that $X_1, ..., X_i$ (for any i) is an optimum set of i linear functions which can best predict the subset of variables $x_1, ..., x_s$. The determination of such functions is likely to be useful in specifying the sizes of ready-made garments, shoes, etc.

References

BRONOWSKI, J. and LONG, W. M. (1951): Statistical methods in anthropology. *Nature*, 168, 794.

——— (1952): *Amer. J. Phys. Anthrop.* 10, n.s. 385.

FISHER, R. A. (1939): The sampling distribution of some statistics obtained from non-linear regression. *Ann. Eugen.*, 9, 238.

PENROSE, L. S. (1947): Some notes on discrimination. *Ann. Eugen.*, 13, 228–237.

RAO, C. R. (1952): *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, New York.

——— (1958): Appendix to Majumdar, D. N. and Rao, C. R. (1958): Bengal Anthropometric Survey, 1945: A Statistical Study. *Sankhyā*, 19, 201.

——— (1959): Some problems involving linear hypothesis in multivariate analysis. *Biometrika*, 46, 49.

——— (1960): Multivariate analysis: An indispensable aid in applied research. *Sankhyā*, 23, 317–338.

SMITH, C. A. B. (1947): Some examples of discrimination. *Ann. Eugen.*, 13, 272–282.

YATES, F and HEALY, M. J. R. (1951): Statistical methods in anthropology. *Nature*, 168, 1116.

Paper received : March, 1962.