Inference on Discriminant Function Coefficients

C. RADHAKRISHNA RAO, *Indian Statistical Institute*

1. INTRODUCTION

Some years ago, the author developed some tests for examining hypotheses such as "the coefficients of some specified characters in the linear discriminant function are zero" (Rao, 1946; 1948; 1949) and "the coefficients of two given characters are in a specified ratio" (Rao, 1952). These tests were meant to be generalizations of an earlier test by Fisher (1940) for a proposed (assigned) discriminant function (i.e., when the proportions of all the coefficients are specified). During the last few years, the author has received a number of queries regarding the theory and application of these tests. The present paper is an attempt to answer these queries and to propose tests for other hypotheses of interest in the use of discriminant functions.

It must be mentioned that the individual coefficients in the linear discriminant function are not definite population parameters, as only the ratios of the coefficients are unique. For this reason estimates of the individual coefficients and their

standard errors are not meaningful. We can only draw inferences on the ratios of the coefficients.

All the tests considered in the paper are special cases of a test for examining the sufficiency of a given subset out of a larger set of variables, for purposes of discrimination between two populations. *The hypothesis of sufficiency* is explicitly defined as follows. Let $(X_1, ..., X_p)$ be a p dimensional random variable. Then the subset $(X_1, ..., X_q)$ is said to be sufficient for discrimination between two populations if the conditional distributions of $(X_1, ..., X_p)$ given $(X_1, ..., X_q)$ are the same for both the populations. We may also describe the hypothesis of sufficiency of $(X_1, ..., X_q)$ as the *absence of additional information* contained in $(X_{q+1}, ..., X_p)$ when the variables $(X_1, ..., X_q)$ are already available. We consider this general problem in Section 2.

The problem of inference on the coefficients of a genetic selection index as developed by Smith (1936) needs an entirely different approach. A test is described for examining the adequacy of a *straight selection function* or any proposed selection function. Other questions such as the adequacy of a subset of the phenotypic observations for assessing some well defined genetic worth of an individual need further study.

2. NOTATIONS AND PRELIMINARY RESULTS

Let $\boldsymbol{X}' = (X_1, ..., X_p)$ be a p dimensional random variable and consider the partition $\boldsymbol{X}' = (\boldsymbol{X}_1' : \boldsymbol{X}_2')$ where $\boldsymbol{X}_1$ consists of the first q components of $\boldsymbol{X}$ and $\boldsymbol{X}_2$, the rest of the components of $\boldsymbol{X}$. Assuming the first two moments of $\boldsymbol{X}$ to exist, we can write the corresponding partitions of $E(\boldsymbol{X})$, the expectation of $\boldsymbol{X}$, and $D(\boldsymbol{X})$, the dispersion matrix of $\boldsymbol{X}$, as

$$E(\boldsymbol{X}') = \boldsymbol{\mu}' = (E(\boldsymbol{X}_1') : E(\boldsymbol{X}_2')), \tag{1}$$

$$D(\boldsymbol{X}) = \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \tag{2}$$

Let H_1 and H_2 be two simple hypotheses specifying the means and dispersion matrices as

$$E(\boldsymbol{X}' \mid H_1) = \boldsymbol{\mu}_1' = (\boldsymbol{\mu}_{11}' \vdots \boldsymbol{\mu}_{12}'), \quad D(\boldsymbol{X}) = \boldsymbol{\Sigma}, \tag{3}$$

$$E(\boldsymbol{X}' \mid H_2) = \boldsymbol{\mu}_2' = (\boldsymbol{\mu}_{21}' \vdots \boldsymbol{\mu}_{22}'), \quad D(\boldsymbol{X}) = \boldsymbol{\Sigma}. \tag{4}$$

If $\boldsymbol{\delta}' = \boldsymbol{\mu}_1' - \boldsymbol{\mu}_2' = (\boldsymbol{\delta}_1' \vdots \boldsymbol{\delta}_2')$, then the linear discriminant function between H_1 and H_2 based on $\boldsymbol{X}$ is defined as $\boldsymbol{\delta}' \boldsymbol{\Sigma}^{-1}\boldsymbol{X}$, while that based on $\boldsymbol{X}_1$ alone is $\boldsymbol{\delta}_1' \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$. We shall represent the linear discriminant function $\boldsymbol{\delta}' \boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ by $\boldsymbol{\lambda}'\boldsymbol{X} = \lambda_1 X_1 + \ldots + \lambda_p X_p$ and consider some hypotheses on $\boldsymbol{\lambda}$.

Note that the Mahalanobis distance between H_1 and H_2 based on $\boldsymbol{X}$ is $\Delta_p^2 = \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ while that based on $\boldsymbol{X}_1$ is $\Delta_q^2 = \boldsymbol{\delta}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1$.

With the above notations, the following statements are equivalent.

(a) $\boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 = \boldsymbol{0}$, i.e., the random variable $\boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$ obtained by subtracting from $\boldsymbol{X}_2$ its regression on $\boldsymbol{X}_1$ has the same expected value for both the populations.

(b) The coefficients $\lambda_{q+1}, \ldots, \lambda_p$ of the components $\boldsymbol{X}_{q+1}, \ldots, \boldsymbol{X}_p$, in the linear discriminant function based on $\boldsymbol{X}$ are all zero.

(c) $\Delta_p^2 = \Delta_q^2$, i.e., there is no additional distance contributed by the variables $X_{q+1}, \ldots, X_p$.

(d) Every linear function of $\boldsymbol{X}$ uncorrelated with $\boldsymbol{X}_1$ has the same expected value for both the populations.

(e) If $\boldsymbol{X} \sim \mathcal{N}_p$ (i.e., distributed as p-variate normal), then the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{X}_1$ is the same for both the populations, which is the same as saying that $\boldsymbol{X}_1$ is sufficient for discrimination between the populations.

The compounding vector $\boldsymbol{\lambda}$ of $\boldsymbol{X}$ in the linear discrimination function $\boldsymbol{\delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ is

$$(\boldsymbol{\lambda}_1' \vdots \boldsymbol{\lambda}_2') = (\boldsymbol{\delta}_1' \vdots \boldsymbol{\delta}_2') \cdot \begin{pmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{pmatrix} \tag{5}$$

where $\boldsymbol{\lambda}_2' = (\lambda_{q+1}, \ldots, \lambda_p)$ is the vector of coefficients of the components of $\boldsymbol{X}_2$. From (5) $\boldsymbol{\lambda}_2' = \boldsymbol{\delta}_1'\boldsymbol{\Sigma}^{12} + \boldsymbol{\delta}_2'\boldsymbol{\Sigma}^{22}$ and by the statement (b), $\boldsymbol{\lambda}_2 = \mathbf{0}$. By virtue of the algebraic equivalence

$$\boldsymbol{\delta}_1'\boldsymbol{\Sigma}^{12} + \boldsymbol{\delta}_2'\boldsymbol{\Sigma}^{22} = \mathbf{0} \Longleftrightarrow \boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 = \mathbf{0}, \tag{6}$$

this proves the equivalence of statements (a) and (b).

It is easy to prove the identity

$$\begin{aligned}\Delta_p^2 &= \boldsymbol{\delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \\ &= \boldsymbol{\delta}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 + (\boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1)'(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1)\end{aligned} \tag{7}$$

If $\Delta_p^2 = \Delta_q^2 = \boldsymbol{\delta}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1$, the second term in (7), which is a positive definite quadratic form is zero. Hence $(\boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}^{-1}\boldsymbol{\delta}_1) = \mathbf{0}$ and vice-versa, i.e., (a) $\Longleftrightarrow$ (c).

Let $\boldsymbol{L}_1'\boldsymbol{X}_1 + \boldsymbol{L}_2'\boldsymbol{X}_2$ be a linear function of $\boldsymbol{X}$. Then the statement (d) implies $\boldsymbol{L}_1'\boldsymbol{\Sigma}_{11} + \boldsymbol{L}_2'\boldsymbol{\Sigma}_{21} = \mathbf{0}$, i.e., $\boldsymbol{L}_1 = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{L}_2$. Consider $\boldsymbol{L}_1'\boldsymbol{\delta}_1 + \boldsymbol{L}_2'\boldsymbol{\delta}_2 = \boldsymbol{L}_2'(-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2) = \mathbf{0}$ for all $\boldsymbol{L}_2 \Longleftrightarrow \boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 = \mathbf{0}$, i.e., (d) $\Longleftrightarrow$ (a).

To prove (c) $\Longleftrightarrow$ (a) we need only consider the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1$, which is $(p-q)$ variate normal with

$$E(\boldsymbol{X}_2 \mid \boldsymbol{X}_1, H_1) - E(\boldsymbol{X}_2 \mid \boldsymbol{X}_1, H_2) = \boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 \tag{8}$$

$$D(\boldsymbol{X}_2 \mid \boldsymbol{X}_1, H_1) = D(\boldsymbol{X}_2 \mid \boldsymbol{X}_1, H_2) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{9}$$

Sufficiency of $\boldsymbol{X}_1$ is equivalent to $\boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 = \mathbf{0}$, which proves the desired result.

3. TEST FOR SUFFICIENCY OF A SUBSET

We shall develop a test for the hypothesis that "$\boldsymbol{X}_1$ is sufficient" or "$\boldsymbol{X}_2$ has no additional information in the presence of $\boldsymbol{X}_1$" on the basis of samples of sizes n_1 and n_2 from the two populations. Let

$$\boldsymbol{d}' = (\boldsymbol{d}_1' \vdots \boldsymbol{d}_2') \tag{10}$$

represent the difference in sample means of X_1 and X_2 and

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \tag{11}$$

the pooled sum of products (S. P. matrix) within samples on (n_1+n_2-2) degrees of freedom (d.f.).

Let us observe that under the assumption of normality

$$E(X_2 \mid X_1, H_1) = \alpha_1+\Gamma X_1, \tag{12}$$

$$E(X_2 \mid X_1, H_2) = \alpha_2+\Gamma X_1,$$

where Γ represents the matrix of regression coefficients. If the dispersion matrix of X is the same under H_1 and H_2, then $D(X_2 \mid X_1, H_1) = D(X_2 \mid X_1, H_2)$. Hence the hypothesis under test is $\alpha_1 = \alpha_2$.

The formulation (12) may be recognized as the multivariate extension of the Gauss-Markoff set up, with $(p-k)$ variables (the components of X_2). So, no new problem arises in the consideration of test criteria. One obtains, to begin with, the dispersion matrices due to deviation from hypothesis and due to error. We also notice that the set up (12) involves two sets of parameters, with the null hypothesis concerning only one of the sets. In such case the computations are simple, involving what is known as covariance adjustment (see the discussion on page 119 in Rao (1952) in the unvariate case and on pages 468-69 in Rao (1965) in the multivariate case).

The S.P. matrix within populations jointly for X_1, X_2 is

$$W = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = S \tag{13}$$

on (n_1+n_2-2) d.f. and that between populations is

$$B = \frac{n_1 n_2}{n_1+n_2} \begin{pmatrix} d_1 d_1' & d_1 d_2' \\ d_2 d_1' & d_2 d_2' \end{pmatrix} \tag{14}$$

on 1 d.f., giving the total S.P. matrix

$$T = W+B = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & B_{22} \end{pmatrix}. \tag{15}$$

We now compute the residual S.P. matrices for X_2 making covariance adjustment for X_1, obtaining

$$S_{22}-S_{21}S_{11}^{-1}S_{12} \quad \text{with} \quad \text{d.f.} = n_1+n_2-2-q \tag{16}$$

for within, and

$$T_{22}-T_{21}T_{11}^{-1}T_{12} \text{ with d.f.} = n_1+n_2-1-q \tag{17}$$

for the total.

Now applying the Wilks Λ criterion which is the likelihood ratio test applied on the conditional distributions of X_2 given X_1 we have

$$\Lambda = \frac{|S_{22}-S_{21}S_{11}^{-1}S_{12}|}{|T_{22}-T_{21}T_{11}^{-1}T_{12}|} = \frac{|S|}{|S_{11}|}\,\frac{T}{|T_{11}|} \tag{18}$$

$$= \frac{|S|}{|T|}\cdot\frac{|S_{11}|}{|T_{11}|}$$

It is well known that in the special case when the matrix B, has 1 d.f., the statistic

$$\frac{n_1+n_2-p-1}{p-q}\left(\frac{1}{\Lambda}-1\right) \tag{19}$$

has an F distribution on $(p-q)$ and (n_1+n_2-p-1) d.f.

Defining

$$D_p^2 = (n_1+n_2-2)d'S^{-1}d, \quad D_q^2 = (n_1+n_2-2)d_1'S_{11}^{-1}d_1 \tag{20}$$

which are estimates of Δ_p^2 and Δ_q^2 and $c = n_1n_2/(n_1+n_2)(n_1+n_2-2)$ we can write the F statistic (19) as

$$\frac{n_1+n_2-p-1}{p-q}\cdot\frac{c(D_p^2-D_q^2)}{1+cD_q^2} \tag{21}$$

the form in which the test was originally expressed (Rao, 1949; 1952).

Optimum properties of the test (21) have been recently investigated by Giri (1964).

4. INFERENCE ON DISCRIMINANT FUNCTION COEFFICIENTS

The test criterion (21) was developed for testing the hypothesis that the coefficients of specified variables in the linear discriminant function are zero. We can apply the same test by a suitable transformation of the variables in drawing inferences of various types on the coefficients of a linear discriminant function.

4.1. Test for a proposed discriminant function (Fisher, 1940)

Let $\lambda'X$ be the assigned discriminant function. Then the null hypotheses under test can be written as

$$\lambda'X \propto \delta'\Sigma^{-1}X \rightarrow \delta \propto \Sigma\lambda, (\lambda \text{ given}). \tag{22}$$

We make the transformation (assuming $\lambda_1 \neq 0$ without loss of generality)

$$Y_1 = \lambda'X,\ Y_2 = X_2, \ldots, Y_p = X_q. \tag{23}$$

Then the null hypothesis says that in the discriminant based on $Y_1, \ldots, Y_p$, the coefficients of $Y_2, \ldots, Y_p$ are zero. Hence the test (21) applies with $q = 1$. We need the values of D_1^2 and D_p^2 based on the Y values. Since D^2 is invariant under a linear transformation

$$D_p^2(Y_1, \ldots, Y_p) = D_p^2(X_1, \ldots, X_p) = (n_1+n_2-2)d'S^{-1}d \tag{24}$$

$$D_1^2(Y_1) = (n_1+n_2-2)(\lambda'd)^2/\lambda'S\lambda \tag{25}$$

so that D_1^2 and D_p^2 are expressed in terms of the statistics d and S defined in (10) and (11) in terms of the original variables (X_1,

..., X_p). Substituting the values (24), (25) for D_p^2 and D_1^2 we obtain the statistic

$$\frac{n_1+n_2-p-1}{p-1}\ \frac{cD_p^2-cD_1^2}{1+cD_1^2} \tag{26}$$

which is an F statistic on $(p-1)$ and (n_1+n_2-p-1) d.f.

If we are using α probability level of significance, the test (26) can be written as

$$\frac{(\lambda'd)^2}{\lambda'S\lambda} \geqslant \frac{c(n_1+n_2-p-1)D_p^2-(p-1)F_\alpha}{c(p-1)F_\alpha+c(n_1+n_2-p-1)} \tag{27}$$

where F_α is the upper α probability value of F on $(p-1)$ and (n_1+n_2-p-1) d.f. The inequality (27) provides a cone with vertex at the origin, within which the direction vector λ of the coefficients of the true discriminant function will lie with probability $1-\alpha$.

4.2. Test for a given ratio of the coefficients of two variables

Let the ratio of the coefficients of X_1 and X_2 be ρ, i.e. $\lambda_1/\lambda_2 = \rho$. For a given ρ, the discriminant function can be written

$$\lambda_2(\rho X_1+X_2)+\lambda_3X_3+...+\lambda_pX_p \tag{28}$$

where $\lambda_2, ..., \lambda_p$ are unknown. Equating (28) to $\delta\Sigma^{-1}X$ the null hypothesis can be written as

$$\delta \propto \Sigma \begin{pmatrix} \rho \\ \cdots \\ b \end{pmatrix} \tag{29}$$

where $\rho' = (\rho, 1)$ and the vector b is unknown.

Let us make the transformation

$$Y_1 = \rho X_1+X_2, \quad Y_2 = X_3, ..., Y_{p-1} = X_p, Y_p = X_1. \tag{30}$$

Then the null hypothesis (29) says that in the discriminant function based on $Y_1, \ldots, Y_p$, the coefficient of Y_p is zero. Hence the test (21) applies with $q = p-1$. We need the values of $D_p^2(Y_1, \ldots, Y_p) = D_p^2(X_1, \ldots, X_p)$ and $D_{p-1}^2(Y_1, \ldots, Y_{p-1})$. To compute D_{p-1}^2, consider the partition of the original variable $\boldsymbol{X}' = (\boldsymbol{X}_1' \vdots \boldsymbol{X}_2')$ where $\boldsymbol{X}_1' = (X_1, X_2)$ and $\boldsymbol{X}_2' = (X_3, \ldots, X_p)$ with the corresponding partition

$$\boldsymbol{d}' = (\boldsymbol{d}_1' \vdots \boldsymbol{d}_2') \tag{31}$$

of the sample mean differences and the partition

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix} \tag{32}$$

of the within pooled dispersion matrix. Then

$$(n_1+n_2-2)^{-1}D_{p-1}^2(Y_1, \ldots, Y_{p-1}) \tag{33}$$

$$= \boldsymbol{d}_2'\boldsymbol{S}_{22}^{-1}\boldsymbol{d}_2 + \frac{[\boldsymbol{\rho}'(\boldsymbol{d}_1 - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{d}_2)]^2}{\boldsymbol{\rho}'(\boldsymbol{S}_{11} - \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21})\boldsymbol{\rho}}$$

and the variance ratio test on 1 and (n_1+n_2-p-1) d.f. for an assigned ρ is

$$\frac{n_1+n_2-p-1}{1} \quad \frac{c(D_p^2 - D_{p-1}^2)}{1+c\,D_{p-1}^2} \tag{34}$$

where $(n_1+n_2-2)^{-1}\ D_p^2 = \boldsymbol{d}'\boldsymbol{S}^{-1}\boldsymbol{d}$ and D_{p-1}^2 is as given in (33). To obtain a confidence interval for ρ, we consider the equation

$$\frac{n_1+n_2-p-1}{1} \quad \frac{cD_p^2 - cD_{p-1}^2}{1+cD_{p-1}^2} = F_\alpha \tag{35}$$

which is quadratic in ρ. The confidence bounds for ρ are obtained by determining the roots of the equation (35).

4.3. Test for assigned ratios of the coefficients of several variables

Let $\rho_1, \ldots, \rho_k$ be the assigned ratios of the coefficients of the first k variables $X_1, \ldots, X_k$, where some of the ρ_i may be zero. In such a case the discriminant function is of the form

$$\lambda_k(\rho_1 X_1 + \ldots + \rho_k X_k) + \lambda_{k+1} X_{k+1} + \ldots + \lambda_p X_p \tag{36}$$

where $\lambda_k, \ldots, \lambda_p$ are unknown. Equating (36) to $\boldsymbol{\delta}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$, the null hypothesis can be written as

$$\boldsymbol{\delta} \propto \boldsymbol{\Sigma} \begin{pmatrix} \boldsymbol{\rho} \\ \ldots \\ \boldsymbol{b} \end{pmatrix} \tag{37}$$

where $\boldsymbol{\rho}' = (\rho_1, \ldots, \rho_k)$ and $\boldsymbol{b}$ is unknown.

Let $\rho_k \neq 0$ without loss of generality. Then consider the transformation

$$\begin{aligned} Y_1 &= \rho_1 X_1 + \ldots + \rho_k X_k, \\ Y_2 &= X_{k+1}, \ \ldots, \ Y_{p-k+1} = X_p, \\ Y_{p-k+2} &= X_1, \ \ldots, \ Y_p = X_{k-1}. \end{aligned} \tag{38}$$

The hypothesis (37) says that in the discriminant function based on $Y_1, \ldots, Y_p$, the coefficients of the last $(k-1)$ variables are all zero. Hence the test (21) is applicable with the value of $q = p-k+1$. As in the earlier case, we shall express the values of D_p^2 and D_q^2 in terms of the original variables. Consider the partition $(\boldsymbol{X}_1', \boldsymbol{X}_2')$ of $\boldsymbol{X}'$ where $\boldsymbol{X}_1$ consists of the components $X_1, \ldots, X_k$ and $\boldsymbol{X}_2$ of the rest. Corresponding to such a partition of the random variable, we have the following partition of the sample mean difference and the within S.P. matrix.

$$\boldsymbol{d}' = (\boldsymbol{d}_1' \vdots \boldsymbol{d}_2'), \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}. \tag{39}$$

Then

$$(n_1+n_2-2)^{-1} D_p^2 = \boldsymbol{d}'\boldsymbol{S}^{-1}\boldsymbol{d} \tag{40}$$

$$(41)\quad (n_1+n_2-2)^{-1} D_q^2 = \boldsymbol{d}_2' \mathbf{S}_{22}^{-1}\boldsymbol{d}_2 + \frac{[\boldsymbol{\rho}'(\boldsymbol{d}-\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\boldsymbol{d}_2)]^2}{\boldsymbol{\rho}'(\mathbf{S}_{11}-\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})\boldsymbol{\rho}}$$

substituting the expressions (40), (41) in (21), we find that the test criterion is an explicit function of $\boldsymbol{\rho}$. Hence we can test for any assigned value of $\boldsymbol{\rho}$ or determine the confidence zone of $\boldsymbol{\rho}$.

4.4. Test whether $\boldsymbol{\lambda}$ belongs to a given linear manifold

Let $\boldsymbol{A}$ be a $(p \times k)$ matrix providing a basis of the given manifold. Then the hypothesis $\lambda \epsilon M(A) \Longrightarrow \boldsymbol{\delta} = \boldsymbol{\Sigma A \theta}$ where $\boldsymbol{\theta}$ is a $k \times 1$ vector. The values of $\boldsymbol{\delta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$ are unknown but the hypothesis only specifies a relationship among them through the known matrix $\boldsymbol{A}$. Consider the transformation

$$(42)\qquad \boldsymbol{Y}_1 = \boldsymbol{A}'\boldsymbol{X}, \qquad \boldsymbol{Y}_2 = \boldsymbol{B}'\boldsymbol{X}$$

where $\boldsymbol{B}$ is chosen such that $\boldsymbol{B}'\boldsymbol{\Sigma A} = \mathbf{0}$, i.e., the linear functions in $\boldsymbol{B}'\boldsymbol{X}$ are uncorrelated with those in $\boldsymbol{A}'\boldsymbol{X}$. If $\boldsymbol{\delta} = \boldsymbol{\Sigma A \theta}$, then

$$(43)\qquad E(\boldsymbol{B}'\boldsymbol{X} \mid H_1) - E(\boldsymbol{B}'\boldsymbol{X} \mid H_2) = \boldsymbol{B}'\boldsymbol{\delta} = \boldsymbol{B}'\boldsymbol{\Sigma A \theta} = \mathbf{0}.$$

Thus, according to statement (e) in Section 2, $\boldsymbol{Y}_1 = \boldsymbol{A}'\boldsymbol{X}$ is sufficient for discrimination between the hypotheses H_1 and H_2 or the coefficients of the components of $\boldsymbol{Y}_2$ in the discriminant function expressed in terms of $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$ are all zero. Hence the test (21) applies with $q = k$. The matrix $\boldsymbol{B}$ in (42) is arbitrary subject to the condition $\boldsymbol{B}'\boldsymbol{\Sigma}\,\boldsymbol{A} = \mathbf{0}$, but we do not need to know $\boldsymbol{B}$ in order to evaluate the test criterion (21). We observe that

$$(44)\qquad D_p^2(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = D_p^2(\boldsymbol{X}) = (n_1+n_2-2)\boldsymbol{d}'\mathbf{S}^{-1}\boldsymbol{d}$$

$$(45)\qquad D_k^2(\boldsymbol{Y}_1) = D_k^2(\boldsymbol{A}'\boldsymbol{X}) = (n_1+n_2-2)\boldsymbol{d}'\boldsymbol{A}(\boldsymbol{A}'\mathbf{S}\boldsymbol{A})^{-1}\boldsymbol{A}\boldsymbol{d}$$

which depend only on $\boldsymbol{A}$, $\boldsymbol{d}$ and $\mathbf{S}$, where $\boldsymbol{d}$ and $\mathbf{S}$ are as defined in (10) and (11).

5. DISCRIMINANT FUNCTION FOR GENETIC SELECTION

Consider an observable variable X which has the decomposition

$$X = \gamma + \epsilon \tag{46}$$

in terms of two unobservable variables γ and ϵ which are uncorrelated. The variable γ denotes the conceptual genotypic measurements and ϵ denotes the environmental effects so that X may be considered as representing phenotypic measurements. Under the set up (46), Smith (1936) considered the problem of predicting a linear function $a'\gamma$ representing the genetic worth of an individual on the basis of phenotypic measurements X. Let $D(X) = \Sigma$, $D(\gamma) = \Gamma$ and $D(\epsilon) = E$. Since γ and ϵ are uncorrelated,

$$\Sigma = \Gamma + E \tag{47}$$

$$\text{Cov}(X, \gamma) = \Gamma, \quad \text{Cov}(X, a'\gamma) = \Gamma a$$

giving the regression of $a'\gamma$ on X as $\lambda' X$ where

$$\Sigma \lambda = \Gamma a \quad \text{or} \quad \lambda = \Sigma^{-1}\Gamma a. \tag{48}$$

If Σ and Γ are known, then the best predictor of $a'\gamma$ or the best selection index for $a'\gamma$ is the regression function of $a'\gamma$ on X.

As an alternative to the regression function $a'\Gamma\,\Sigma^{-1}X$ we may consider the *straight selection function* $a'X$, which is simpler to compute and which does not involve Γ and Σ. It is, therefore, of interest to find the conditions under which $a'\Gamma\Sigma^{-1}X$ and $a'X$ are equivalent. Now, $a'\Gamma\Sigma^{-1}X \propto a'X$ implies that there exists a constant μ such that

$$\Sigma^{-1}\Gamma a = \mu\, a \text{ or } \Gamma a = \mu\, \Sigma a, \tag{49}$$

or a is an eigen vector of the determinantal equation

$$|\Gamma - \mu\, \Sigma| = 0, \tag{50}$$

The condition (49) is the same as,

$$\boldsymbol{E}\boldsymbol{a} = \frac{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}\boldsymbol{\Sigma}\boldsymbol{a}, \tag{51}$$

writing $$\boldsymbol{\Gamma} = \boldsymbol{\Sigma} - \boldsymbol{E}.$$

Let a_1, the first component of $\boldsymbol{a}$, be non-zero. Then we consider the transformation from $\boldsymbol{X}$ to $\boldsymbol{Y}$,

$$Y_1 = \boldsymbol{a}'\boldsymbol{X},\ Y_2 = X_2, \ldots, Y_p = X_p \tag{52}$$

with the corresponding decomposition $\boldsymbol{Y} = \boldsymbol{\gamma}^* + \boldsymbol{\epsilon}^*$ and the dispersion matrices $\boldsymbol{\Sigma}^*$ and $\boldsymbol{E}^*$. In terms of $\boldsymbol{\Sigma}^*$ and $\boldsymbol{E}^*$, the condition (49) is equivalent to

$$\frac{\Sigma^*_{j1}}{\Sigma^*_{11}} = \frac{E^*_{j1}}{E^*_{11}},\quad j = 2, \ldots, p. \tag{53}$$

i.e., the regression coefficient of ϵ^*_j on ϵ^*_1 is the same as that of Y_j on Y_1. We shall consider a test of the hypothesis (53) in Section 5.1 on the basis of independent estimators of $\boldsymbol{\Sigma}^*$ and $\boldsymbol{E}^*$ having Wishart distributions.

A more general hypothesis of interest is the assignment of the ratios of the coefficients of a subset of the phenotypic measurements X_i in the selection index (regression function). In such a case the relationship between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ can be written as

$$\boldsymbol{\Gamma}\boldsymbol{a} = \mu\,\boldsymbol{\Sigma}\begin{pmatrix} \boldsymbol{\rho} \\ \cdots \\ \boldsymbol{b} \end{pmatrix} \tag{54}$$

where $\boldsymbol{\rho}$ is the vector of assigned ratios of the coefficients of the (say) first q variables and $\boldsymbol{b}$ is unknown. There does not seem to be a simple test of the hypothesis (54). However, one can determine the likelihood ratio test based on independent estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{E}$ and thus provide an asymptotic test with the usual chi-square approximation.

5.1. A test for the hypothesis (53)

Let T and W be independent random matrices such that

$$T \sim \omega_p(\Sigma^*, k) \tag{55}$$

$$W \sim \omega_p(E^*, m) \tag{56}$$

where $\omega(A, b)$ represents Wishart distribution of a $p \times p$ random matrix with the hypothetical matrix A and degrees of freedom b. Given T_{11}, the conditional distributions

$$\left(\frac{T_{12}}{T_{11}}, \ldots, \frac{T_{1p}}{T_{11}}\right) \sim \mathcal{N}_p\left[\nu_1, \frac{1}{T_{11}}(\Sigma^*_{ij.1})\right] \tag{57}$$

$$(T_{ij.1}) = \left(T_{ij} - \frac{T_{i1}\,T_{j1}}{T_{11}}\right) \sim \omega_p[(\Sigma^*_{ij.1}), k-1)] \tag{58}$$

are independent. Similarly

$$\left(\frac{E_{12}}{E_{11}}, \ldots, \frac{E_{1p}}{E_{11}}\right) \sim \mathcal{N}_p\left[\nu_2, \frac{1}{E_{11}}(E^*_{ij.1})\right] \tag{59}$$

$$(E_{ij.1}) = \left(E_{ij} - \frac{E_{i1}E_{j1}}{E_{11}}\right) \sim \omega_p[(E^*_{ij.1}), m-1] \tag{60}$$

are independent. The hypothesis (53) under test is the same as the hypothesis

$$\nu_1 = \nu_2 \tag{61}$$

i.e., the means of the two normal distributions (57) and (59) are equal. Since estimates of the dispersion matrices are available, it is possible to provide an appropriate test of the hypothesis (61).

If $(\Sigma^*_{ij.1})$, $(E^*_{ij.1})$, the true residual dispersion matrices, are equal, then we have the standard test based on the likelihood ratio

$$\Lambda = \frac{|(T_{ij.1}) + (W_{ij.1})|\,|(T_{11} + W_{11})}{|(T_{ij}) + (W_{ij})|} \tag{62}$$

where

$$\frac{m+k-p}{p-1} \quad \frac{1-\Lambda}{\Lambda} \tag{63}$$

has an F distribution on $(p-1)$ and $(m+k-p)$ d.f.

If the residual dispersion matrices $(\Sigma^*_{ij\cdot 1})$ and $(E^*_{ij\cdot 1})$ are not equal, then we have the following approximate test. The differences

$$d_1 = \frac{T_{12}}{T_{11}} - \frac{W_{12}}{T_{11}}, \ldots, d_{p-1} = \frac{T_{1p}}{T_{11}} - \frac{W_{1p}}{W_{11}} \tag{64}$$

have the estimated dispersion matrix

$$(C_{ij}) = \frac{1}{(k-1)T_{11}}(T_{ij\cdot 1}) + \frac{1}{(m-1)W_{11}}(W_{ij\cdot 1}). \tag{65}$$

If (C^{ij}) is the reciprocal of (C_{ij}), then the statistic for testing the significance of the differences $d_1, \ldots, d_{p-1}$ is

$$\Sigma\Sigma\, C^{ij} d_i d_j \tag{66}$$

which can be used approximately as a chi square on $(p-1)$ d.f.

Note: The choice of a test of the hypothesis (53) or (61) depends on whether the residual dispersion matrices $(\Sigma^*_{ij\cdot 1})$ and $(E^*_{ij\cdot 1})$ are equal or not. Since we have the estimates $(T_{ij\cdot 1})$ and $(W_{ij\cdot 1})$ the hypothesis of equality $(\Sigma^*_{ij\cdot 1}) = (E^*_{ij\cdot 1})$ can be subjected to a suitable test.

In problems of genetic selection we have certain families (or individual lines) out of which some have to be selected on the basis of performance of individuals within family. We obtain observations on a certain number of characteristics from each of n individuals in a family. These observations (after transformation to the Y_i variables as in (51)) provide an analysis of dispersion (S.P. matrices) as between and within families, which correspond to the T and W matrices used in the test. For a numerical application of the tests (63) and (66) and further discussion, the reader is referred to a paper by the author (Rao, 1953).

References

Fisher, R. A. (1940). "The Precision of Discriminant Functions," *Ann. Eugenics*, **10**, 422-429.

Giri, N. (1964). "On the Likelihood Ratio Test of a Normal Multivariate Testing Problem," *Ann. Math. Statist.*, **35**, 181-189.

Rao, C. R. (1946). "Tests with Discriminant Functions in Multivariate Analysis," *Sankhyā*, **7**, 407-414.

Rao, C. R. (1948). "Tests of Significance in Multivariate Analysis," *Biometrika*, **35**, 58-79.

Rao, C. R. (1949). "On Some Problems Arising Out of Discrimination with Multiple Characters," *Sankhyā*, **9**, 343-364.

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

Rao, C. R. (1953). "Discrimination Functions for Genetic Differentiation and Selection," *Sankhyā*, **12**, 229-246.

Rao, C. R. (1965). *Linear Statistical Inference with Applications*, John Wiley and Sons, New York.

Smith, Fairfield H. (1936). "A Discriminant Function for Plant Selection," *Ann. Eugenics*, **7**, 240.

(Received Jan. 18, 1966.)