# SOME TESTS USED IN DISCRIMINATORY ANALYSIS

A Thesis

Presented to

the Department of

Actuarial Mathematics and Statistics

University of Manitoba

by

Prem Nath Bhalla

August 1966

# ACKNOWLEDGEMENTS

This study was supported by a teaching fellowship from the Faculty of Graduate Studies and Research, University of Manitoba. I am extremely grateful and express my sincere appreciation to Dr. G. I. Paul, Department of Actuarial Mathematics and Statistics, University of Manitoba, for entrusting me with the problem; for his most valuable guidance and the constructive criticism throughout the period of this study and during the writing of this thesis. The kindly assistance of Dr. A. G. Robinson, Professor and Acting Head, Department of Entomology, for supplying the data on aphid species, is hereby gratefully acknowledged.

I am thankful to Prof. E. R. Vogt, Head, Department of Actuarial Mathematics and Statistics, for providing me the facilities and for his constant encouragement. My thanks are also due to Prof. H. J. Boom, Department of Actuarial Mathematics and Statistics and to Dr. G. O. Losey, Department of Mathematics and Astronomy, for their helpful suggestions.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

An experimenter is generally concerned with the problem of deciding whether the two populations are different with regard to the measurements on individuals in several characters. Some tests that were developed for this purpose are examined with the object of evaluating differences among them, and also with the view of determining their applicability in a taxonomic problem. The following procedures on discriminatory analysis were applied to the data from the department of Entomology on two species of aphids to test the null hypothesis that the mean vector of one multivariate normal population is equal to the mean vector of another multivariate normal population assuming that both the populations have a common dispersion matrix:

(1) Fisher's discriminant function

(2) Hotelling's $T^2$

(3) Mahalanobis' $D^2$

These statistics differ from each other by a constant multiplier but the test statistic in each case is the same and is distributed as a variance ratio.

In classifying an individual having the observed measurements to one of the two species, Fisher's classification procedure has been used. Measurements on characters which do not contribute significantly to the increase in distance between the two populations were eliminated by using Rao's test and the best discriminant function was formed with measurements on the characters; total body length, length of the cornicle and length of cauda.

The most striking observation of the study is that we can form the best discriminant function for classification even without applying Rao's test and using only Fisher's discriminant function. This is achieved by forming discriminant functions with different combinations of characters, and the discriminant function formed including those particular characters which has the minimum probability of misclassification is the best discriminant function.

As a final operation the assumption of the equality of the dispersion matrices in the two populations was tested and the assumption was confirmed.

# CHAPTER I.

## INTRODUCTION

Research workers are often concerned with the statistical analysis of data that consist of sets of measurements on a number of individuals or objects. The mathematical model on which the analysis is based is known as the multivariate normal distribution or a combination of multivariate normal distributions.

Multivariate analysis not only provides a new field of development for mathematical theory but is also an important tool in the hands of research workers for the analysis and interpretation of the statistical data.

A number of problems arising in multivariate populations are analogues of problems arising in univariate populations. In the univariate case we may test the hypothesis that the mean of a variable has a given value. In the multivariate case we may test the hypothesis that the mean vector is a given vector. In the univariate case concerning two samples we may test the null hypothesis that the two populations have the same mean with the same known or unknown variance. Analogous to this in the multivariate case is the test of the null hypothesis that the mean vector of one multivariate normal population is equal to the mean vector of another multivariate normal population with the same known or unknown dispersion matrix.

Generally a research worker is also faced with a problem of correctly assigning an individual to a class known to belong to one of two or more classes. This is done by observing various measurements on the individual followed by an appropriate statistical analysis of the measurements. Some measurements whose ranges for the two classes are completely disjoint are ideal for classification purposes. But it generally happens in practice that there is a certain range of values where individuals from both classes may appear. This prevents one from being correct in all cases for the classification of an individual. If we know the distribution of the characteristics in both classes we can devise a procedure in which the probability of misclassification is least.

In the past many years different authors have devised different tests in discriminatory analysis. As a result the research worker is faced with a problem of choosing the most suitable procedure for testing the null hypothesis that the mean vector of one multivariate normal population is equal to the mean vector of another multivariate normal population with the same known or unknown dispersion matrix.

The problem undertaken in this thesis is to find the best test, among Fisher's discriminant function, Hotelling's $T^2$ and Mahalanobis' $D^2$, in discriminatory analysis for the purpose of testing the equality of the mean vectors in two multivariate normal populations. The main assumptions made in completing the test are that the populations are multivariate normal and have a common dispersion matrix. Similarities among these methods or tests are investigated, and a discussion of

Fisher's classification statistic is made, eliminating non-discriminating characters, and the best linear compound for assigning an individual to one of the two groups is developed.

An example, using data from the Department of Entomology, is provided to illustrate the behaviour and properties of the tests.

## REVIEW OF THE LITERATURE

The idea to generalize the univariate analysis of variance technique to the case of the multivariate is not a recent one. This subject may be regarded as beginning with Pearson's $C^2$, the coefficient of racial likeness. This was a test of divergence between the two samples rather than a measure of the actual magnitude of divergence. The first work to be published on Pearson's coefficient of racial likeness was by Tildesley (1921) on Burmese skulls. Wishart (1928) took up the case of the simultaneous sampling distribution of the variances and covariances in the samples from multivariate normal populations. Roy (1939) devised some tests for the significance of the difference in the dispersion matrices of two multivariate normal populations. Bartlett (1938) also developed a test for the equality of dispersion matrices.

Mahalanobis started his work on the same subject as he felt that Pearson's coefficient of racial likeness provided a test of significance but did not measure the magnitude of the difference. Mahalanobis (1927) found the $D^2$-statistic as an alternative measure of the distance which he applied to discuss racial mixtures in Bengal.

Mahalanobis (1930) gave theoretical formulae for tests and measures of group divergence.

Bose (1936) obtained the exact distribution of the $D^2$-statistic in the non-null case when both populations have the same known dispersion matrix. Bose and Roy (1938) studentised the distribution of the $D^2$-statistic. Fisher (1936) gave a generalization of the $D^2$-statistic for more than two populations. Further development of the theory is due to Rao and Roy (1946-48, 1948c, 1949, 1950). The distribution of the ratio of $D^2$ calculated on (p+r) characters over $D^2$ calculated on p characters, was found by Rao. The test based on this is useful in deciding whether the inclusion of r more characters to p basic characters results in an increase in the measure of divergence between the two populations.

Hotelling (1931) found the distribution of the quantity $T^2$ for a multivariate normal population which is the natural extension of Student's distribution to a sample from a univariate normal population. Theorems on the efficiency of the $T^2$-test have been established by Hsu (1938) and Wald (1943), who have shown that under various conditions no better test is possible.

Fisher (1936) introduced discriminant function analysis and resolved the problem of tests of significance in multivariate populations. He showed that a set of multiple measurements may be used to provide a discriminant function, linear in the observations, having the property that better than any other linear function it will discriminate between any two chosen classes, for example between two species. The problem is reduced to the case of a single variate by using a linear compound of the several variables, where the compounding coefficients are chosen to maximize the value of the statistic suitable for a single variate. Discriminant function analysis can be used for the classification of an individual in to one of two populations provided we know the parameters characterizing the two populations.

Wald (1944) suggested the use of classification statistics in situations where such knowledge of parameters about the population is absent. Anderson (1951) and Rao (1954) also developed their classification statistic and John (1960) found the exact distribution of these classification statistics.

## DEFINITIONS OF TERMS AND PROCEDURES

The following terms and procedures are used throughout the study with their definitions and symbolic representation.

### Population

A population consists of all possible values of a variable. It is generally required that the parameters of the population may be obtained so that the population may be fully described.

### Parameter

A parameter is usually an unknown quantity in the population which may vary over a set of values and which occurs in expressions defining frequency functions.

Sample

The selection of the part of the population to represent the whole is known as sampling and the part selected is known as the sample. A sample is obtained for the purpose of studying the characteristics of the population and the inferences about the population are made on the basis of the sample.

Estimate of a Parameter

An estimate is a particular value yielded by an estimator, which in turn is a rule or method of estimating a parameter of a parent population. The latter is usually expressed as a function of sample values.

Statistic

Any function of the sample values, used as an estimate of a parameter of the population, is called a statistic.

Arithmatic Mean

The arithmetic mean is the arithmetic average and is the result obtained when the sum of the values of the individuals in the data is divided by the number of individuals in the data. The arithmetic mean is calculated by the use of the formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $\Sigma$ is the summation notation.

Range

The range of a sample is the difference between the largest and the smallest of observations and gives some idea of variability present.

$$W = x\ \max - x\ \min.$$

Variance

The population variance is the average value of the squared deviations of the individual variates from the population mean. Symbolically it is represented by $\sigma^2$ and written as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$$

The sample variance is an unbiased estimate of the population variance and is represented symbolically by $\hat{\sigma}^2$ where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}$$

The variance is a measure of the dispersion of the variates about the arithmetic mean.

Covariance or Mean Product

The population covariance is the average of the sum of the products of deviations of the two variables $x_1$ and $x_2$ from their respective means and is written as

$$\sigma_{12} = \frac{\sum_{i=1}^{N} (x_{1i}-\mu_1)\ (x_{2i}-\mu_2)}{N}$$

The estimate of the population value from the sample is denoted by $\hat{\sigma}_{12}$ and symbolically written as

$$\hat{\sigma}^2_{12} = \frac{\sum_{i=1}^{n} (x_{1i}-\bar{x}_1)\ (x_{2i}-\bar{x}_2)}{n-1}$$

<u>Standard Deviation</u>

This is defined as the square root of the variance and is denoted by $\sigma$ for population and $\hat{\sigma}$ for a sample.

<u>Standard Error</u>

It is possible to obtain a large number of samples from a population and it is expected that the means calculated from these samples are subject to random variation. A measure of this variation is known as the standard error of the sample arithmetic mean. The variability among means of random samples of a given size is related to the variability of the parent population by a definite mathematical relation. If $\hat{\sigma}_{\bar{x}}$ stands for standard error of the means of samples of n individuals then

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

<u>Expected Value</u>

The expected value of the function g(x) of a continuous random variable x whose frequency function is f(x) is given by

$$E\left[g(x)\right] = \int_{-\infty}^{\infty} g(x)\ f(x)\ dx.$$

Unbiased Estimate

The statistic

$$t\ (x_1,\ x_2\ \ldots\ x_n)$$

is called an unbiased estimator of the parameter $\theta$ if

$$E\ (t) = \theta$$

This definition states that the random variable t possesses a distribution whose mean is the parameter $\theta$.

When estimating the mean $\mu$ of the population we set $t=\bar{x}$ and find

$$\begin{aligned} E(\bar{x}) &= E\,\frac{1}{n}\,(x_1+x_2\ \ldots + x_n) = \frac{1}{n}\left[E\ (x_1)+E(x_2)\ \ldots + E(x_n)\right] \\ &= \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right] = \mu \end{aligned}$$

Matrix

An mxn matrix is a rectangular array of real numbers and is written as

$$A = a_{ij} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$$j = 1,\ 2\ \ldots\ n; \quad i = 1,\ 2\ \ldots\ m.$$

If $m = n$ then the matrix is called a square matrix.

## Sum of Two Matrices

If A and B are both mxn matrices then their sum is defined as the mxn matrix

$$A + B = \left[ a_{ij} \right] + \left[ b_{ij} \right] = \left[ a_{ij} + b_{ij} \right]$$

## Product of a Matrix by a Real Number

The product of a matrix A by a real number x is

$$xA = Ax = \left[ xa_{ij} \right]$$

## Multiplication of Two Matrices

If the number of columns of matrix A is the same as the number of rows of matrix B

$$A = \left[ a_{ij} \right] \quad i = 1 \ldots r, \quad j = 1 \ldots m$$

$$B = \left[ b_{jk} \right] \quad j = 1 \ldots m, \quad k = 1 \ldots n$$

$$A \times B = \left[ a_{ij} \; b_{jk} \right] = \sum_{j=1}^{m} a_{ij} b_{jk}$$

AB is a matrix with r rows and n columns. If A and B are square matrices then $AB \neq BA$ in general.

## Transpose of a Matrix A

The transpose of the mxn matrix $A = \left[ a_{ij} \right]$ is defined to be the matrix $A' = \left[ b_{ij} \right]$

where $b_{ij} = a_{ji}$ for $i=1, \ldots n, \quad j=1 \ldots m$.

## Symmetric Matrix

Matrix A is called symmetric if

$$A = A' \text{ i.e. } a_{ij} = a_{ji}$$

Vector

A vector

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{bmatrix}$$

with m components can be treated as a matrix with m rows and one column. The above matrix operations hold for vectors also.

Transpose of a Vector

Transpose of $\underline{x}$ is written as $\underline{x}'$ and symbolically denoted as

$$\underline{x}' = \begin{bmatrix} x_1 \ x_2 & \cdots & x_m \end{bmatrix}$$

Identity Matrix

The pxp matrix

$$I_p = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \delta_{ij} \end{bmatrix}$$

where

$$\delta_{ij} = \begin{Bmatrix} 1 & \text{if} & i = j \\ 0 & \text{if} & i \neq j \end{Bmatrix}$$

is called the identity matrix of order p

Determinant of a Square Matrix

If A is a pxp matrix then determinant

$$A = \left| A \right| = \sum_{\sigma} (\mathrm{Sgn}\sigma) a_{1},\sigma(1) a_{2},\sigma(2) \ldots a_{p},\sigma(p)$$

where $\sigma$ runs through all permutations of $\{1, 2, \ldots p\}$ and

$$\mathrm{Sgn}\sigma = \begin{cases} 1 & \text{if } \sigma \text{ is an even permutation} \\ -1 & \text{if } \sigma \text{ is an odd permutation} \end{cases}$$

Minor of an Element

The minor of an element $a_{ij}$ in a matrix

$$A = \left[ a_{kl} \right]$$

is the determinant of a sub-matrix of A obtained by deleting the $i^{th}$ row and $j^{th}$ column.

Cofactor of $a_{ij}$

The cofactor of $a_{ij}$ is the number where

$$A_{ij} = (-1)^{i+j} \text{ x minor of } A_{ij}$$

Non-Singular Matrix

A square matrix A is non-singular if A has an inverse. We then have the theorem: A square matrix A is non-singular if and only if

$$\left| A \right| \neq 0$$

Inverse of a Matrix A

If $\left| A \right| \neq 0$, then there exists a unique matrix B such that

$$AB = BA = 1$$

and matrix B is called the inverse or reciprocal of matrix A and is represented by $A^{-1}$.

Quadratic Form

Let $A = \left[ a_{ij} \right]$ be a symmetric pxp matrix then the function $Q(x)$ defined on all p-dimensional vectors $\underline{x}$ by

$$Q(x) = \underline{x}' A \underline{x} = \sum_{i,j=1}^{p} a_{ij} x_i x_j$$

is called a quadratic form with matrix A

where $\underline{x}' = \left[ x_1, x_2 \dots x_p \right]$

Positive Definite Quadratic Form

The matrix A and the quadratic form associated with it are called positive definite when $\underline{x}' A \underline{x} > 0$ for $x \neq 0$.

Positive Semi-definite Quadratic Form

If $\underline{x}' A \underline{x} \geqslant 0$ for all x then A and the quadratic form are known as positive semi-definite.

Probability Density Function

If there is a probability $f(x)dx$ that a certain value x will take values between x and x+dx, then $f(x)$ is called a p.d.f. with properties

$$f(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P\left[ X \in a \right] = F(a) = \int_{-\infty}^{a} f(x)dx$$

Univariate Normal Density Function

A random variable x of the continuous type $(-\infty < x < \infty)$ is said to be normally distributed with mean $\mu$ and variance $\sigma^2$, or briefly $N(\mu, \sigma^2)$, if it has a p.d.f. of the type

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[ - \frac{(x-\mu)^2}{2\sigma^2} \right]$$

<u>Multivariate Normal Density Function</u>

A set of p random variables $(x_1, x_2 \ldots x_p)$, $(-\infty < x_i < \infty)$ is said to follow a p-variate normal distribution, or briefly $N(\mu, \Sigma)$, if the joint probability density function is of the type

$$f\,(x_1, x_2 \;\ldots\; x_p) = \frac{1}{|\Sigma|^{1/2}(2)^{p/2}} \exp\left[ -\frac{1}{2}\,(\underline{x} - \underline{\mu})' \;\Sigma^{-1}(\underline{x} - \underline{\mu}) \right]$$

where random variables $x_1, x_2 \ldots x_p$ constitute a random vector,

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{bmatrix}$$

and p means of the distribution constitute the population mean vector

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_p \end{bmatrix}$$

$\Sigma$ is the population dispersion matrix which is positive definite and symmetric and written as

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

It is also well known that

$$E(\underline{x}) = \underline{\mu}$$

and

$$E(\underline{x}-\bar{\underline{x}})\ (\underline{x}-\bar{\underline{x}})' = \Sigma\ .$$

# CHAPTER II

One of the first problems in discriminatory analysis is the construction of a suitable function that can be used in discriminating among two or more populations.

## FISHER'S DISCRIMINANT FUNCTION

It was shown by Fisher (1936) that a set of multiple measurements may be used to provide a discriminant function, linear in the observations, having the property that, better than any other linear function, it will discriminate between the two classes when two or more populations are measured on several characters $x_1$, $x_2$ ... $x_p$.

We shall consider the question what linear function of p variables

$$X = l_1x_1+l_2x_2 \dots + l_px_p$$

will maximize the ratio of the difference between the class means to the standard deviation within the classes. We may represent the differences between the means in two samples as

$$d_p = \bar{x}_p^{(1)} - \bar{x}_p^{(2)} \qquad p=1,2, \dots, p$$

The sums of squares and products are represented by

$$S_{ij} = \left\{ \sum_{\alpha=1}^{n_1} (x_{i\alpha}^{(1)} - \bar{x}_i^{(1)})(x_{j\alpha}^{(1)} - \bar{x}_j^{(1)}) + \sum_{\alpha=1}^{n_2} (x_{i\alpha}^{(2)} - \bar{x}_i^{(2)})(x_{j\alpha}^{(2)} - \bar{x}^{(2)}) \right\}$$

where $n_1$ and $n_2$ are the sample sizes from the two populations, and i and j take independent values 1,2, ..., p. Then for a linear function X of measurements, as defined above, the difference between the means of X in the two classes is

$$D' = l_1 d_1 + l_2 d_2 + \dots + l_p d_p$$

while the variance within the classes is proportional to

$$S' = \sum_{i=1}^{p} \sum_{j=1}^{p} l_i \, l_j \, S_{ij}$$

The particular linear function which best discriminates between the classes will be one for which the ratio $\frac{D'^2}{S'}$ is greatest, for variation of the coefficients $l_1 \, l_2 \dots l_p$ independently of each other.

This gives for each l

$$\frac{D'}{S'^2} \left\{ 2S' \frac{\partial D'}{\partial l} - D' \frac{\partial S'}{\partial l} \right\} = 0$$

$$\frac{1}{2} \frac{\partial S'}{\partial l} = \frac{S'}{D'} \frac{\partial D'}{\partial l}$$

$\frac{S'}{D'}$ is a factor constant for p unknown coefficients.

Consequently the required coefficients are

$$S_{11} l_1 + S_{12} l_2 + \dots + S_{1p} l_p = d_1$$

$$S_{21} l_1 + S_{22} l_2 + \dots + S_{2p} l_p = d_2$$

$$\cdot \qquad \cdot \qquad \dots \quad \cdot \qquad\qquad \cdot$$

$$S_{p1} l_1 + S_{p2} l_2 + \dots + S_{pp} l_p = d_p$$

that is

$$\begin{bmatrix} S_{11} & S_{12} & \cdot & \cdot & \cdot & S_{1p} \\ S_{21} & S_{22} & \cdot & \cdot & \cdot & S_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ S_{p1} & S_{p2} & \cdot & \cdot & \cdot & S_{pp} \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ \cdot \\ l_p \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ d_p \end{bmatrix}$$

and

$$\begin{bmatrix} l_1 \\ l_2 \\ \cdot \\ l_p \end{bmatrix} = \begin{bmatrix} S^{11} & S^{12} & \cdot & \cdot & \cdot & S^{1p} \\ S^{21} & S^{22} & \cdot & \cdot & \cdot & S^{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ S^{p1} & S^{p2} & \cdot & \cdot & \cdot & S^{pp} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ d_p \end{bmatrix}$$

The significance of this can be tested from the results given in the analysis of variance table. The total variation is sub-divided into a portion that arises from the effectiveness of the linear function and a portion that is ascribable to uncontrolled sources. The sum of squares ascribable to the linear function is

$$(\bar{X}_A - \bar{X}_B)^2 \frac{n_1 \, n_2}{N_{1+N2}} \qquad \text{and}$$

$$D' = l_1 d_1 + l_2 d_2 + \cdot \cdot \cdot + l_p d_p$$

is the sum of squares within the species or that portion ascribable to uncontrolled sources.

Analysis of Variance Table.

| Serial number | Source | Degrees of freedom | Sum of squares |
|---|---|---|---|
| 1 | Between classes | p | $\frac{n_1 n_2}{n_1+n_2}(l_1 d_1+\ldots l_p d_p)^2$ |
| 2 | Within classes | $n_1+n_2-p-1$ | $\Sigma\, l_i l_j S_{ij} = D'$ |

An F-test is constructed by taking the ratio of the mean squares from these two sources (linear function in the numerator) and if this ratio is significant then one concludes that the function is effective as a discriminant; otherwise it is ineffective.

HOTELLING'S $T^2$-STATISTIC FOR TWO-SAMPLE PROBLEM

We consider two sets of p-variates

$(X_1^{(1)}, X_2^{(1)} \ldots X_p^{(1)})$ and $(X_1^{(2)}, X_2^{(2)} \ldots X_p^{(2)})$ with the following assumptions:

1. It is assumed that each of the sets $(X_1^{(1)}, X_2^{(1)} \ldots X_p^{(1)})$ and $(X_1^{(2)}, X_2^{(2)} \ldots X_p^{(2)})$ has a p-variate normal distribution with mean vectors $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$ respectively.

2. The two sets are independent of each other.

3. The dispersion matrices of the variates $(X_1^{(1)}, X_2^{(1)} \ldots X_p^{(1)})$ and $(X_1^{(2)}, X_2^{(2)} \ldots X_p^{(2)})$ are of the same order i.e. $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$.

The normal population with mean values $\mu_1^{(1)}, \mu_2^{(1)} \ldots \mu_p^{(1)}$ and dispersion matrix $\Sigma^{(1)}$ is denoted by population $P_1$ and the normal population with mean values $\mu_1^{(2)}, \mu_2^{(2)} \ldots \mu_p^{(2)}$, and dispersion matrix $\Sigma^{(2)}$ by $P_2$.

A sample of size $n_1$ is drawn from the population $P_1$ and a sample of size $n_2$ is drawn from the population $P_2$. We denote $x_{i\alpha}^{(1)}$ as the $\alpha^{th}$ observation of $x_i^{(1)}$ $(i=1,2 \ldots p; \alpha=1,2 \ldots n_1)$, and $x_{i\alpha}^{(2)}$ as the $\alpha^{th}$ observation of $x_i^{(2)}$ $(i=1, \ldots p; \alpha=1,2 \ldots n_2)$.

We wish to test the null hypothesis that the mean vector of one multivariate normal population is equal to the mean vector of the other multivariate normal population where the dispersion matrices are assumed to be equal but unknown.

Let $x_1^{(1)}, x_2^{(1)} \ldots x_{n_1}^{(1)}$ be a sample from $N(\underline{\mu}^{(1)}, \Sigma^{(1)})$

and $x_1^{(2)}, x_2^{(2)} \ldots x_{n_2}^{(2)}$ be a sample from $N(\underline{\mu}^{(2)}, \Sigma^{(2)})$.

($\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$ from assumption 3).

$\bar{x}^{(1)}$ is distributed as $N(\underline{\mu}^{(1)}, \frac{1}{n_1} \Sigma)$ and

$\bar{x}^{(2)}$ is distributed as $N(\underline{\mu}^{(2)}, \frac{1}{n_2} \Sigma)$.

$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\ (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$ is distributed according to $(\underline{0}, \Sigma)$

The estimate of the variance covariance is

$$\hat{\sigma}_{ij} = \frac{1}{n_1+n_2-2} \left\{ \sum_{\alpha=1}^{n_1} (x_{i\alpha}^{(1)} - \bar{x}_i^{(1)})(x_{j\alpha}^{(1)} - \bar{x}_j^{(1)}) + \sum_{\alpha=1}^{n_2} (x_{i\alpha}^{(2)} - \bar{x}_i^{(2)})(x_{j\alpha}^{(2)} - \bar{x}_j^{(2)}) \right\}$$

the R.H.S. expression being the sum of the corrected sum of products for two samples divided by $n_1 + n_2 - 2$.

The estimate of the common dispersion matrix $\Sigma$ is

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdots & \hat{\sigma}_{pp} \end{bmatrix}$$

The inverse of matrix $\hat{\Sigma}$ is represented as

$$\hat{\Sigma}^{-1} = \begin{bmatrix} \hat{\sigma}^{11} & \hat{\sigma}^{12} & . & . & . & \hat{\sigma}^{1p} \\ \hat{\sigma}^{21} & \hat{\sigma}^{22} & . & . & . & \hat{\sigma}^{2p} \\ . & . & . & . & . & . \\ \hat{\sigma}^{p1} & \hat{\sigma}^{p2} & . & . & . & \hat{\sigma}^{pp} \end{bmatrix}$$

Then $$T^2 = \frac{n_1 n_2}{n_1+n_2} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})' \hat{\Sigma}^{-1} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$$

is distributed as $T^2$ with $n_1 + n_2 - 2$ degree of freedom.

The critical region is

$$T^2 \geqslant \frac{(n_1+n_2-2)}{n_1+n_2-p-1} \; p \; F_{p,n_1+n_2-p-1}$$

for a given significance level.

We may write

$$T^2 \; \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p} \geqslant F_{p,\, n_1+n_2-p-1}$$

MAHALANOBIS' $D^2$-STATISTICS

We again consider two sets of p-variates $(X_1^{(1)}, X_2^{(1)} \dots X_p^{(1)})$ and $(X_1^{(2)}, X_2^{(2)} \dots X_p^{(2)})$, with the same assumptions of Hotelling's $T^2$-statistic.

Let $x_1^{(1)}, x_2^{(1)} \dots x_{n_1}^{(1)}$ be a sample from $N(\underline{\mu}^{(1)}, \Sigma^{(1)})$ and $x_1^{(2)}, x_2^{(2)} \dots x_{n_2}^{(2)}$ be a sample from $N(\underline{\mu}^{(2)}, \Sigma^{(2)})$ and on each individual p-measurements are taken.

The estimate of variance covariance is

$$\hat{\sigma}_{ij} = \frac{1}{n_1+n_2-2}\left\{\sum_{\alpha=1}^{n_1} (x_{i\alpha}^{(1)}-\bar{x}_i^{(1)})(x_{j\alpha}^{(1)}-\bar{x}_j^{(1)}) + \sum_{\alpha=1}^{n_2} (x_{i\alpha}^{(2)}-\bar{x}_i^{(2)})(x_{j\alpha}^{(2)}-\bar{x}_j^{(2)})\right\}$$

The estimate of the common dispersion matrix $\Sigma$

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdot & \cdot & \cdot & \hat{\sigma}_{1p} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdot & \cdot & \cdot & \hat{\sigma}_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hat{\sigma}_{p1} & \hat{\sigma}_{p2} & \cdot & \cdot & \cdot & \hat{\sigma}_{pp} \end{bmatrix}$$

The inverse of matrix $\hat{\Sigma}$ is represented as

$$\hat{\Sigma}^{-1} = \begin{bmatrix} \hat{\sigma}^{11} & \hat{\sigma}^{12} & . & . & . & \hat{\sigma}^{1p} \\ \hat{\sigma}^{21} & \hat{\sigma}^{22} & . & . & . & \hat{\sigma}^{2p} \\ . & . & . & . & . & . \\ \hat{\sigma}^{p1} & \hat{\sigma}^{p2} & . & . & . & \hat{\sigma}^{pp} \end{bmatrix}$$

Mahalanobis' distance between the two populations as estimated from the samples on the basis of p-characters is

$$D^2 p = (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})' \hat{\Sigma}^{-1} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$$

To test the hypothesis specifying no difference in mean values of the p characters for the two populations with identical dispersion matrices, the statistic

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

can be used as a variance ratio with p and $(n_1 + n_2 - p - 1)$ degrees of freedom.

## APPLICATION OF THE PROCEDURES

The data used in the analysis were supplied by Dr. A. G. Robinson, Department of Entomology, University of Manitoba, and consisted of a series of measurements taken on two species of aphids.

The two species of aphids are morphologically similar but occur on different host plants. They are both greenish in colour and are described as different species. The species (A) *Dactynotus erigeronensis* occurs mainly on the host plant *Erigeron canadensis*. The other species (B) *Dactynotus richardsi* was found on the host plant *Grindelia squarrosa*. The measurements on the following characters were taken by the same person and recorded in millimetres.

$x_1$: Total body length

$x_2$: Length of the cornicle

$x_3$: Reticulate area of the cornicle

$x_4$: Length of cauda

$x_5$: Length of the hind tibia

$x_6$: Length of the antennal segment V

Table 2-1 shows six measurements on each of the twenty five aphids of species A and Table 2-2 shows six measurements on each of the fifteen aphids of the species B. Table 2-3 gives a summary of the comparison of the estimates of the statistical parameters of six characters for the two species of aphids.

TABLE 2-1 Measurement~~s~~ on Species A

| No. | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| 1 | 2.20 | 0.80 | 0.32 | 0.36 | 1.32 | 0.30 |
| 2 | 2.04 | 0.80 | 0.28 | 0.38 | 1.44 | 0.36 |
| 3 | 2.42 | 0.90 | 0.28 | 0.38 | 1.50 | 0.38 |
| 4 | 2.52 | 0.96 | 0.32 | 0.40 | 1.54 | 0.36 |
| 5 | 2.18 | 0.82 | 0.28 | 0.36 | 1.36 | 0.30 |
| 6 | 2.32 | 0.80 | 0.28 | 0.36 | 1.36 | 0.30 |
| 7 | 2.40 | 0.84 | 0.26 | 0.42 | 1.44 | 0.34 |
| 8 | 2.56 | 0.88 | 0.26 | 0.36 | 1.46 | 0.32 |
| 9 | 2.56 | 0.94 | 0.24 | 0.42 | 1.56 | 0.32 |
| 10 | 2.40 | 0.84 | 0.32 | 0.40 | 1.42 | 0.32 |
| 11 | 2.56 | 0.94 | 0.28 | 0.40 | 1.56 | 0.36 |
| 12 | 2.42 | 0.78 | 0.20 | 0.38 | 1.36 | 0.34 |
| 13 | 2.30 | 0.92 | 0.26 | 0.44 | 1.56 | 0.40 |
| 14 | 2.54 | 0.88 | 0.26 | 0.42 | 1.20 | 0.36 |
| 15 | 2.24 | 0.82 | 0.26 | 0.34 | 1.40 | 0.30 |
| 16 | 2.18 | 0.92 | 0.30 | 0.40 | 1.60 | 0.38 |
| 17 | 2.32 | 0.92 | 0.24 | 0.40 | 1.56 | 0.36 |
| 18 | 2.32 | 0.92 | 0.30 | 0.38 | 1.56 | 0.36 |
| 19 | 2.32 | 0.84 | 0.28 | 0.40 | 1.42 | 0.36 |
| 20 | 2.40 | 0.88 | 0.24 | 0.36 | 1.52 | 0.40 |
| 21 | 2.28 | 0.78 | 0.24 | 0.36 | 1.32 | 0.30 |
| 22 | 2.32 | 0.80 | 0.24 | 0.40 | 1.46 | 0.32 |
| 23 | 2.16 | 0.84 | 0.26 | 0.36 | 1.52 | 0.36 |
| 24 | 2.30 | 0.88 | 0.24 | 0.44 | 1.48 | 0.36 |
| 25 | 2.32 | 0.84 | 0.22 | 0.36 | 1.36 | 0.32 |

TABLE 2-2 Measurements on Species B

| No. | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| 1 | 1.78 | 0.70 | 0.21 | 0.38 | 1.30 | 0.32 |
| 2 | 1.93 | 0.76 | 0.24 | 0.40 | 1.34 | 0.32 |
| 3 | 2.02 | 0.72 | 0.28 | 0.42 | 1.32 | 0.31 |
| 4 | 2.02 | 0.65 | 0.23 | 0.40 | 1.23 | 0.31 |
| 5 | 2.02 | 0.76 | 0.24 | 0.42 | 1.36 | 0.35 |
| 6 | 2.00 | 0.77 | 0.27 | 0.38 | 1.40 | 0.35 |
| 7 | 1.74 | 0.58 | 0.21 | 0.35 | 1.14 | 0.27 |
| 8 | 1.58 | 0.52 | 0.18 | 0.28 | 1.00 | 0.25 |
| 9 | 1.82 | 0.60 | 0.20 | 0.35 | 1.20 | 0.28 |
| 10 | 1.93 | 0.64 | 0.21 | 0.35 | 1.22 | 0.30 |
| 11 | 2.02 | 0.68 | 0.24 | 0.37 | 1.27 | 0.31 |
| 12 | 1.84 | 0.66 | 0.20 | 0.36 | 1.18 | 0.33 |
| 13 | 2.02 | 0.57 | 0.18 | 0.35 | 1.14 | 0.28 |
| 14 | 1.82 | 0.58 | 0.19 | 0.32 | 1.07 | 0.26 |
| 15 | 2.02 | 0.66 | 0.24 | 0.37 | 1.23 | 0.30 |

TABLE 2-3 Comparison of Estimates of Statistical Parameters for Six Characters of Aphids Species A and B

| Character | Species A | | Species B | | Variance Ratio | Diff. of Means | S.E. of the diff. | Diff. of Means / S.E. of diff. |
|---|---|---|---|---|---|---|---|---|
| | Range | Mean ± S.E. | Range | Mean ± S.E. | | | | |
| 1 | 2.04–2.56 | 2.3432 ±0.027486 | 1.58–2.02 | 1.9040 ±0.034972 | 1.0297 | 0.4392 | 0.04464596 | 9.8374** |
| 2 | 0.78–0.96 | 0.8616 ±0.011008 | 0.52–0.77 | 0.6567 ±0.019776 | 1.9359 | 0.2049 | 0.02084528 | 9.8296** |
| 3 | 0.20–0.32 | 0.2664 ±0.006191 | 0.18–0.28 | 0.2213 ±0.007868 | 2.0124 | 0.0451 | 0.01005245 | 4.4865** |
| 4 | 0.36–0.44 | 0.3872 ±0.005523 | 0.28–0.42 | 0.3667 ±0.009537 | 1.7892 | 0.0205 | 0.01024609 | 2.0007 |
| 5 | 1.20–1.60 | 1.4512 ±0.019904 | 1.00–1.40 | 1.2267 ±0.028494 | 1.2296 | 0.2245 | 0.03385001 | 6.6322** |
| 6 | 0.30–0.40 | 0.3432 ±0.006285 | 0.25–0.35 | 0.3027 ±0.007653 | 1.1240 | 0.0405 | 0.010816 | 4.0067** |

* Significant at 0.05 level of probability.
** Significant at 0.01 level of probability.

The ranges in species A and species B for the characters $x_1$ and $x_2$ are non-overlapping and the ranges for characters $x_3$, $x_4$, $x_5$ and $x_6$ are overlapping. The mean values for all the characters in species A are larger than those in species B. The difference between the means of the two species for characters $x_1$, $x_2$, $x_3$, $x_5$, and $x_6$ is significantly different at the 1% level, and, if we consider a one-tailed t-test we find that the difference between the means of the two species for character $x_4$ is significant at the 5% level. The variances for all the characters in species A and in species B are of the same order. The above information is useful in building up a discriminant function.

## FISHER'S DISCRIMINANT FUNCTION

We begin with a linear function composed of six variables

$$X = l_1x_1+l_2x_2+l_3x_3+l_4x_4+l_5x_5+l_6x_6.$$

TABLE 2-4 Observed Means for Two Species and Their Difference.(mm.)

| Character | Species A | Species B | Difference $d=\bar{x}^{(1)}-\bar{x}^{(2)}$ |
|---|---|---|---|
| $x_1$ | 2.3432 | 1.9040 | 0.4392 |
| $x_2$ | 0.8616 | 0.6567 | 0.2049 |
| $x_3$ | 0.2664 | 0.2213 | 0.0451 |
| $x_4$ | 0.3872 | 0.3667 | 0.0205 |
| $x_5$ | 1.4512 | 1.2267 | 0.2245 |
| $x_6$ | 0.3432 | 0.3027 | 0.0405 |

The matrix of pooled sums of squares and products of six measurements, within species is

$$S = \begin{bmatrix} 0.7101 & 0.1775 & 0.0243 & 0.0803 & 0.1595 & 0.0421 \\ 0.1775 & 0.1548 & 0.0362 & 0.0517 & 0.2066 & 0.0544 \\ 0.0243 & 0.0362 & 0.0360 & 0.0119 & 0.0489 & 0.0084 \\ 0.0803 & 0.0517 & 0.0119 & 0.0374 & 0.0701 & 0.0221 \\ 0.1595 & 0.2066 & 0.0489 & 0.0701 & 0.4082 & 0.0842 \\ 0.0421 & 0.0544 & 0.0084 & 0.0221 & 0.0842 & 0.0364 \end{bmatrix}$$

The numerical value of the matrix inverse to the matrix S is

$$S^{-1} = \begin{bmatrix} 2.4979 & -5.0063 & 2.0476 & -3.3179 & 1.1797 & 3.4058 \\ -5.0063 & 40.0137 & -14.7177 & -6.8111 & -11.0466 & -20.9260 \\ 2.0476 & -14.7177 & 39.6721 & -4.3724 & -0.1156 & 13.3944 \\ -3.3179 & -6.8111 & -4.3724 & 56.5083 & -0.8781 & -17.2516 \\ 1.1797 & -11.0466 & -0.1156 & -0.8781 & 8.6162 & -4.2263 \\ 3.4058 & -20.9260 & 13.3944 & -17.2516 & -4.2263 & 71.9667 \end{bmatrix}$$

The values of the coefficients in the discriminant function are obtained as follows:

$$\underline{l} = S^{-1}\,\underline{d} \qquad \text{where} \qquad \underline{d} = \underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)}$$

$$\begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \\ l_6 \end{bmatrix} = \begin{bmatrix} 2.4979 & -5.0063 & 2.0476 & -3.3179 & 1.1797 & 3.4058 \\ -5.0063 & 40.0137 & -14.7177 & -6.8111 & -11.0466 & -20.9260 \\ 2.0476 & -14.7177 & 39.6721 & -4.3724 & -0.1156 & 13.3944 \\ -3.3179 & -6.8111 & -4.3724 & 56.5083 & -0.8781 & -17.2516 \\ 1.1797 & -11.0466 & -0.1156 & -0.8781 & 8.6162 & -4.2263 \\ 3.4058 & -20.9260 & 13.3944 & -17.2516 & -4.2263 & 71.9667 \end{bmatrix} \begin{bmatrix} 0.4392 \\ 0.2049 \\ 0.0451 \\ 0.0205 \\ 0.2245 \\ 0.0405 \end{bmatrix} = \begin{bmatrix} 0.49839417 \\ 1.86917965 \\ 0.09974770 \\ -2.78741441 \\ -0.00536696 \\ -0.57563340 \end{bmatrix}$$

Substituting these values of $l_1$, $l_2$ ... $l_6$ in the discriminant function we get

$$X_1 = x_1 (0.49839417) + x_2(1.86917965) + x_3(0.09974770)$$
$$- x_4(2.78741441) - x_5(0.00536696) - x_6(0.57563340)$$

Sum of squares within the species is

$$D = l_1d_1 + l_2d_2 + \dots + l_6d_6 = 0.52472822$$

Sum of squares between the species is

$$\frac{n_1n_2}{n_1+n_2}(\bar{X}_A-\bar{X}_B)^2 = \frac{n_1n_2}{n_1+n_2}(l_1d_1 + l_2d_2 \dots + l_6d_6)^2$$

$$= \frac{25x15}{25+15}(0.52472822)^2 = 9.375\ (0.52472822)^2 = 2.58130973.$$

TABLE 2-5 Analysis of Variance Table

| Source | d.f. | S.S. | M.S. | F. |
|---|---|---|---|---|
| Between species | 6 | 2.58130973 | 0.43021828 | 27.0563** |
| Within species | 33 | 0.52472822 | 0.01590085 | |

The observed value of F with 6 and 33 degrees of freedom is highly significant, therefore, we conclude that the two species are from populations with different mean vectors.

MAHALANOBIS' $D^2$

The estimate of the common dispersion matrix $\Sigma$ is

$$\hat{\Sigma} = \frac{1}{38} S = \frac{1}{38}\left[\, S_{ij} \,\right]$$

where S is the matrix of pooled sum of squares and products of six measurements used in the case of Fisher's discriminant function.

$$\hat{\Sigma}^{-1} = 38\ S^{-1} = 38\left[\, S^{ij} \,\right]$$

The distance between the two populations as estimated from the samples on the basis of six measurements is

$$D_6^2 = (\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)})' \hat{\Sigma}^{-1} (\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)})$$

$$\text{where } \bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)} = \begin{bmatrix} 0.4392 \\ 0.2049 \\ 0.0451 \\ 0.0205 \\ 0.2245 \\ 0.0405 \end{bmatrix} = \underline{d}$$

$$D_6^2 = \begin{bmatrix} 0.4392 & 0.2049 & 0.0451 & 0.0205 & 0.2245 & 0.0405 \end{bmatrix} \hat{\Sigma}^{-1} \begin{bmatrix} 0.4392 \\ 0.2049 \\ 0.0451 \\ 0.0205 \\ 0.2245 \\ 0.0405 \end{bmatrix}$$

$$= 19.93967236.$$

To test the hypothesis specifying no difference in mean values of the six characters for two populations, with the same dispersion matrix, the statistic used is

$$\frac{n_1 n_2}{n_1+n_2} \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p} D_6^2$$

$$= 9.375 \times \frac{33}{38} \times \frac{1}{6} \; (19.93967236)$$

$$= 27.0563$$

This is the variance ratio with 6 and 33 degrees of freedom, found previously, and is significant at the 1% level of significance, which shows that the two species are drawn from populations with different mean vectors.

## HOTELLING'S $T^2$-STATISTICS

As in the case of Mahalanobis' $D^2$ the estimate of the dispersion matrix required in Hotelling's method is

$$\hat{\Sigma} = \frac{1}{38} S = \frac{1}{38} \left[ S_{ij} \right]$$

where S is the matrix of pooled sum of squares and products of six measurements used in the case of Fisher's discriminant function.

Inverse of $\Sigma$ is

$$\hat{\Sigma}^{-1} = 38\, S^{-1} = 38 \left[ S^{ij} \right]$$

$$\text{Hotelling's } T^2 = \frac{n_1 n_2}{n_1+n_2} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})' \hat{\Sigma}^{-1} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$$

$$= (9.375)(19.93967236)$$

$$T^2 = 186.93442828$$

is distributed as $T^2$ with (25+15-2) degrees of freedom.

The critical region is

$$T^2 \geqslant \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,\ n_1+n_2-p-1}$$

$$T^2 \geqslant \frac{(38)6}{33} F_{6,\ 33}$$

$$\text{or} \quad T^2 \times \frac{33}{38} \times \frac{1}{6} \geqslant F_{6,\ 33}$$

$$= 27.0563 > F_{6,\ 33}$$

This observed value of F is identical to the values of F found in the case of the discriminant function and in the case of the test on Mahalanobis' $D^2$.

# CHAPTER III

## CLASSIFICATION PROCEDURE

The practical utility of the discriminant function X is in classifying an individual having the observed measurements into one of the two groups. Consider the problem of classifying a given collection of individuals as belonging to one or the other of the two populations $P_1$ and $P_2$ based on measurements taken on each individual with respect to p-characters. We assume further that the usual assumptions necessary for the analysis are satisfied.

According to the method suggested by Fisher (1936) an individual with measurements $x_1$, $x_2$ ... $x_p$ is assigned to $P_1$ or $P_2$ according to whether

$$x' \, \Sigma^{-1} \, (\underline{\mu}^{(1)} - \underline{\mu}^{(2)}) \gtrless \tfrac{1}{2}(\underline{\mu}^{(1)} + \underline{\mu}^{(2)})' \, \Sigma^{-1} \, (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})$$

It is to be noted that Fisher's method requires the knowledge of $\underline{\mu}^{(1)}$, $\underline{\mu}^{(2)}$ and $\Sigma$. These parametric values are replaced in the analysis by their estimates which are calculated from the samples. The final step in the classification procedure is to substitute the mean values of the p characters in

$$X = l_1 x_1 + l_2 x_2 + \dots + l_p x_p$$

$$X_{\underline{\mu}^{(1)}} = \underline{\mu}^{(1)'} \Sigma^{-1} (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})$$

$$X_{\underline{\mu}^{(2)}} = \underline{\mu}^{(2)'} \Sigma^{-1} (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})$$

$$\frac{X_{\underline{\mu}^{(1)}} + X_{\underline{\mu}^{(2)}}}{2} = \frac{1}{2} (\underline{\mu}^{(1)} + \underline{\mu}^{(2)})' \Sigma^{-1} (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})$$

This value is the partition value for the groups, and enables us to assign an individual to group one if the observed value of X exceeds $\dfrac{X_{\underline{\mu}^{(1)}} + X_{\underline{\mu}^{(2)}}}{2}$ ; otherwise the individual is assigned to group two.

The value of the discriminant function obtained from the data on aphids by substituting the calculated values of $l_1, l_2 \ldots l_6$ is

$$X = x_1(0.49839417) + x_2(1.86917965) + x_3(0.09974770)$$
$$-x_4(2.78741441) - x_5(0.00536696) - x_6(0.57563340)$$

$$X_{\underline{\mu}^{(1)}} = \underline{\mu}^{(1)'} \Sigma^{-1} (\underline{\mu}^{(1)} - \underline{\mu}^{(2)})$$

$$= 2.3432\ (0.49839417) + 0.8616\ (1.86917965) + 0.2664\ (0.09974770)$$
$$- 0.3872\ (2.78741441) - 1.4512\ (0.00536696) - 0.3932\ (0.57563340)$$
$$= 1.52026242$$

$$X_{\underline{\mu}^{(2)}} = 1.9040\ (0.49839417) + 0.6567\ (1.86917965) + 0.2213\ (0.09974770)$$
$$- 0.3667\ (2.78741441) - 1.2267\ (0.00536696) - 0.3027\ (0.57563340)$$
$$= 0.99553420$$

The partition value between the two groups

$$\frac{X_{\underline{\mu}}(1) + X_{\underline{\mu}}(2)}{2} = \frac{2.51579662}{2} = 1.25789831$$

and the difference between the mean values that were substituted into the discriminant function is

$$X_{\underline{\mu}}(1) - X_{\underline{\mu}}(2) = 0.52472822$$

The mean sum of squares within species is

$$\frac{0.52472822}{33} = 0.01590085$$

with standard error equal to 0.12609857.

The probability of misclassification is calculated as

$$\text{Prob. } \left(Z \geqslant \frac{X_{\underline{\mu}}(1) - X_{\underline{\mu}}(2)}{2} \div \text{S.E.}\right).$$

This probability is obtained from normal distribution tables; i.e.

$$\text{Prob. } \left(Z \geqslant \frac{0.26236411}{0.12609857} = 2.080627\right)$$

and the probability of misclassification is 1.873%.

Suppose we wish to assign an aphid with the following measurements (in mm.) on six characters to one of the species: The values

$$x_1 = 2.02 \quad x_2 = 0.65 \quad x_3 = 0.23$$
$$x_4 = 0.40 \quad x_5 = 1.23 \quad x_6 = 0.31$$

are substituted into the discriminant function as follows for that individual

$$X = x_1(0.49839417) + x_2(1.86917965) + x_3(0.09974770)$$
$$-x_4(2.78741441) - x_5(0.00536696) - x_6(0.57563340)$$

with the result that X = 0.94465149. This observed value does not exceed the partition value, therefore the aphid is taken as belonging to species B.

ADDITIONAL INFORMATION SUPPLIED BY SOME CHARACTERS

One of the problems facing a statistician in the study of discriminatory analysis tests is the nature and number of characters which may be of use in discriminating between the two classes. It is useful to test the significance of the addition of the distance obtained by including extra characters. The particular character or number of characters which do not increase the distance between the two classes in the populations should not be included in the function. These characters should be ignored as they weaken the test of significance. It was felt at one time that the more the number of characters under study the better the discriminant function. Rao (1948-1949), however, considered a particular model and showed that the distance function tends to a finite limit with the increase in the number of characters.

In classification problems where the measurements are taken for the purpose of assigning an individual to a particular group, the error of misclassification depends upon the distance between the two groups. The additional characters observed may increase the distance to a very little extent and it may not be worthwhile to have the measurements on extra characters. Suppose initially that we have p characters and q more characters are added and that samples of $n_1$ and $n_2$ individuals are available from two classes containing measurements on all the p+q characters. If $D^2_{p+q}$ is of the same order as that of $D^2_p$, then the ratio

$$R = \frac{1 + \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)} D^2_{p+q}}{1 + \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)} D^2_p}$$

is about unity. A high value of this ratio would indicate that $D^2_{p+q}$ is significantly greater than $D^2_p$ so that q characters supply some additional information. The actual test to use is

$$\frac{n_1+n_2-p-q-1}{q} (R-1)$$

as a variance ratio with q and $(n_1+n_2-p-q-1)$ degrees of freedom. The additional distance is expressed as $D^2_{p+q} - D^2_p$.

APPLICATION OF THE PROCEDURE

(i) Suppose we wish to know the contribution to the increase in distance by addition of the characters $x_5$ and $x_6$ to the first four characters $x_1$, $x_2$, $x_3$, and $x_4$.

Now

$$R = \frac{1 + \frac{25\text{x}15}{(25+15)(25+15-2)} D^2_6}{1 + \frac{25\text{x}15}{(25+15)(25+15-2)} D^2_4}$$

$$R-1 = \frac{1+\frac{9.375}{38} \times 19.93967236}{1+\frac{9.375}{38} \times 19.75836638} - 1 = 0.00761415$$

The test of significance for the additional two characters $x_5$ and $x_6$ is

$$\frac{(25+15-4-2-1)}{2}(0.00761415) = \frac{33}{2}(0.00761415)$$

$$= 0.12563347.$$

The tabular value of F for 2 and 33 degrees of freedom at the 5% probability level is 3.285. The calculated value is not significant and we can therefore safely eliminate measurements on characters $x_5$ and $x_6$.

The additional distance is

$$D_6^2 - D_4^2 = 19.93967236 - 19.75836638$$

$$= 0.18130698.$$

In the calculation of a revised discriminant function the new values of the coefficients omitting measurements on $x_5$ and $x_6$ are

$$l_1 = 0.53210334, \qquad l_2 = 1.64439462,$$

$$l_3 = 0.21004046, \qquad l_4 = -2.93424693$$

and the discriminant function is

$$X = x_1(0.53210334) + x_2(1.64439462) + x_3(0.21004046)$$

$$- x_4(2.93424693).$$

The partition value between the two groups is 1.32347081, and the probability of misclassification is obtained by referring to tables of the normal distribution; that is to say,

$$\text{Prob. } \left(Z \geqslant \frac{0.25997850}{0.12188482} = 2.13298506\right)$$

and the probability of misclassification is 1.647%.

(ii) In the calculation of a further revised function, omitting $x_4$ we consider the first three characters $x_1$, $x_2$, $x_3$ with p=3 and q=1, then

$$R-1 = 0.35972180$$

the observed $F_{1,35} = \frac{(25+15-3-1-1)}{1} = 35x0.35972180 = 12.590263$

and the tabular value of $F_{1,35}$ at 1% level = 7.42

with the result that the calculated value of $F_{1,35} = 12.59$ is highly significant. This means that the increase in distance by including $x_4$ in the discriminant function is highly significant.

The revised discriminant function using $x_1$, $x_2$ and $x_3$ only, is

$$X = x_1(0.40840402)+x_2(0.81957764)+x_3(0.15297426)$$

with partition value between the two groups equal to 1.52677192.

The probability of misclassification obtained as before is 3.710%.

Discriminant functions using other combinations of the p variables, the estimated distances between the populations, and the associated probabilities of misclassification are summarized in Table 3-1.

TABLE 3-1 Probability of misclassification.

| Sr.No | Char.under study | Dist.bet.grps | Prob.of miscl. in % |
|---|---|---|---|
| 1 | $x_1$, $x_2$ | 13.436236 | 3.527 |
| 2 | $x_1$, $x_2$, $x_3$ | 13.459662 | 3.710 |
| 3 | $x_1$, $x_2$, $x_4$ | 19.712971 | 1.321 |
| 4 | $x_1, x_2, x_3, x_4$ | 19.758366 | 1.647 |
| 5 | $x_1, x_2, x_3, x_4, x_5, x_6$ | 19.939672 | 1.873 |

From Table No. 3-1 it is observed that the discriminant function with characters $x_1$, $x_2$, $x_4$ gives the smallest probability of misclassification. Therefore these characters will form the best discriminant function:

$$X = x_1(0.52471888)+x_2(1.70005422)-x_4(2.92854904)$$

A TEST FOR THE EQUALITY OF DISPERSION MATRICES

Consider h p-variate normal populations represented by samples. We may think of the first population as represented by swarm of points in p-dimensional (cartesian) space centered at a point characterized by a vector $\underline{\mu}^{(1)}$ and dispersed about this point in an ellipsoidal shape characterized by the dispersion matrix $\Sigma^{(1)}$. The second population in the same p-dimensional space is also represented by swarm of points which overlaps the first swarm to a greater or lesser extent.

The second vector of means $\underline{\mu}^{(2)}$, will in all probability lie at some non-zero distance away from $\underline{\mu}^{(1)}$ and spread of swarm measured by $\Sigma^{(2)}$ may or may not differ from that of the first swarm.

Once we have assumed that the p-variates are distributed normally, the whole swarm of points of any population may be replaced by the vector $\underline{\mu}^{(h)}$ and the matrix $\Sigma^{(h)}$, $h=1,2$.

The first point to make is that if $\Sigma^{(1)} \neq \Sigma^{(2)}$ the corresponding ellipsoid will be oriented in different directions whether or not the mean vectors $\underline{\mu}^{(1)}$ and $\underline{\mu}^{(2)}$ coincide.

Generally one does not visualize the possible difference in dispersion about the means but one assumes that the correlation between any two characters is approximately the same in each group. Unfortunately there is no criteria available to test the hypothesis

$$\overset{(1)}{\underset{(R)}{\Sigma}} = \overset{(2)}{\underset{(R)}{\Sigma}} \; . \; . \; . = \overset{(h)}{\underset{(R)}{\Sigma}} \text{ ———————— I}$$

though the hypothesis

$$\Sigma^{(1)} = \Sigma^{(2)} \quad . \quad . \quad . \quad = \quad \Sigma^{(h)} \text{ ———————— } \mathrm{II}$$

can be tested relatively easily and accurately if all samples are p-variate normal.

If relation (II) is rejected because of difference in the dispersion matrices and relation (I) is accepted by inspection of the corresponding elements of $\Sigma_{(R)}$, the technique given below based on the truth of (II) is correctly applicable with standardization of the variates. The test of equality of dispersion matrices, $\Sigma^{(1)} = \Sigma^{(2)} \; . \; . \; . = \Sigma^{(1)}$ is independent of the mean vectors.

If $\hat{\Sigma}^{(1)}$ is the estimated dispersion matrix of the $1^{th}$ sample supposed to be based on $n_{(1)}$ observations on p-tuples, then the set $\hat{\Sigma}^{(1)}$, $1=1,2 \; . \; . \; . \; h$ is to be compared with the estimate of $\Sigma$ based on all

$$n = n_1 + n_2 \; . \; . \; . + n_h \text{ observations.}$$

The comparison of dispersion matrices is done by comparing their determinants. The required test criterion is

$$-2 \left[ 1 - \left\{ \sum_{1=1}^{h} \frac{1}{n_{1-1}} - \frac{1}{n-h} \right\} \frac{2p^2+2p-1}{6(p+1)(h-1)} \right] \ln \left\{ \frac{\prod_{1=1}^{h} \left| \hat{\Sigma}^{(1)} \right|^{(n_{1-1})/2}}{\left| \hat{\Sigma} \right|^{(n-h)/2}} \right\}$$

and is distributed approximately as chi-square with $(h-1)p(p+1)/2$ degrees of freedom.

(i) Consider the test of the equality of dispersion matrices of populations to which the samples of species A and species B belong. Let the characters that enter into the dispersion matrices be $x_1$, $x_2$ and $x_4$.

$$\left|\hat{\Sigma}^{(1)}\right| = \left(\frac{1}{24}\right)^3 \left|S^{(1)}\right| = \left(\frac{1}{24}\right)^3 (10^{-6} \times 339.3730)$$

$$\left|\hat{\Sigma}^{(2)}\right| = \left(\frac{1}{12}\right)^3 \left|S^{(2)}\right| = \left(\frac{1}{12}\right)^3 (10^{-6} \times 51.0698)$$

$$\left|\hat{\Sigma}\right| = \left(\frac{1}{38}\right)^3 \left|S\right| = \left(\frac{1}{38}\right)^3 (10^{-6} \times 1510.6193)$$

The required test criterion is

$$-2\left[1-\left\{\frac{1}{24}+\frac{1}{14}-\frac{1}{38}\right\}\frac{2x9+2x3-1}{6(3+1)(2-1)}\right] \text{In}\left\{\frac{\left|\hat{\Sigma}^{(1)}\right|^{12} \cdot \left|\hat{\Sigma}^{(2)}\right|^{7}}{\left|\hat{\Sigma}\right|^{19}}\right\}$$

$$-2\left[1-0.09401\right]\left\{12\log(10^{-6}\times 339.370)+7\log(10^{-6}\times 51.0498) -19\log(10^{-6}\times 1510.62) -36\log 24-21\log 14+57\log 38\right\}$$

$$-2\left[0.90599\right]\left\{-1.7902 \times 2.302585\right\}$$

$$= 7.469140$$

The table value of Chi-square with $\frac{(2-1)3\cdot(3+1)}{2} = 6$ degrees of freedom at the 5% level is 12.592. The observed value of Chi-square is not significant and we conclude that the sizes and orientation of the concentric density ellipsoids are the same for the two species.

(ii) Consider the test of equality of dispersion matrices of the two populations of species A and species B in the case where measurements that enter into the dispersion matrices are on the six characters $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$.

$$\left|\hat{\Sigma}^{(1)}\right| = \left(\frac{1}{24}\right)^6 \left|S^{(1)}\right| \quad ; \quad = \left(\frac{1}{24}\right)^6 \left(10^{-12} x\ 6866.49\right)$$

$$\left|\hat{\Sigma}^{(2)}\right| = \left(\frac{1}{14}\right)^6 \left|S^{(2)}\right| \quad = \left(\frac{1}{14}\right)^6 \left(10^{-12} x\ 4.70519\right)$$

$$\left|\hat{\Sigma}\right| = \left(\frac{1}{38}\right)^6 \left|S\right| \quad = \left(\frac{1}{38}\right)^6 \left(10^{-12} x\ 67557.23\right)$$

The required test criterion is

$$-2\left[1-\left\{\frac{1}{24}+\frac{1}{14}-\frac{1}{38}\right\}\frac{2x36+3x6-1}{6(6+1)(2-1)}\right] \text{In}\left\{\frac{\left|\Sigma^{(1)}\right|^{12} \cdot \left|\Sigma^{(2)}\right|^{7}}{\left|\Sigma\right|^{19}}\right\}$$

$$-2\left[\ 0.8162\right]\Big\{ 12\log(10^{-12}\ x\ 6866.99) + 7\log(10^{-12}\ x\ 4.70519)$$

$$-19\log(10^{-12}\ x\ 67557.23) - 72\log 12$$

$$-42\log 14 + 114\log 38\Big\}$$

$$= (-1.6324x-8.434518)\ x\ 2.302585 = 31.703141.$$

Table value of Chi-square for $\frac{(2-1)6(6+1)}{2} = 21$ degrees of freedom is 32.671 at the 5% level. The observed value is not significant and we conclude that the dispersion matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are of the same order. This establishes the validity of assumption 3 in this treatise that the populations have equal dispersion matrices.

# CHAPTER IV

## SUMMARY AND CONCLUSIONS

The three procedures, Fisher's discriminant function, Mahalanobis' $D^2$ and Hotelling's $T^2$, applied for discriminatory analysis differ from each other only by constant multipliers. But the test statistic used in each of the three cases is the same and is distributed as a variance ratio. The assumptions that the two populations are multivariate normal and have a common dispersion matrix are the same in both Mahalanobis' $D^2$ and Hotelling's $T^2$. Fisher originally did not assume the equality of dispersion matrices but he mentioned that if the dispersion matrices in the two populations are different then the probability of misclassification would be little more than the calculated value. If one is interested mainly in testing the hypothesis that the mean vector of one multivariate normal population is the same as that of the other multivariate normal population with identical dispersion matrices then any of the three tests are equally good.

The idea of including the maximum number of characters in forming a discriminant function is not a convincing one, as the increase in the number of characters does not always have minimum probability of misclassification; however, there is always an increase in the distance between the two populations. It is due to the fact that some of the additional characters included increase the mean sum of squares within the species as well as the mean sum of squares between the species but the latter sum of squares is not increased as much as the former.

If a research worker is interested both in testing the hypothesis of equality of mean vectors in two multivariate normal populations and in classifying individuals into one of the two multivariate normal populations, then Fisher's discriminant function is an ideal one. It was felt before that Fisher's discriminant function in combination with Mahalanobis' $D^2$ is useful for finding the best discriminant function with optimum number of characters to be included. The characters which do not contribute significantly to the increase in distance between two populations can be eliminated. This is done by applying Rao's test of significance for an increase in distance by including additional characters to the existing characters. Even by using this technique we may not achieve the objective of forming the best discriminant function with minimum probability of misclassification.

Perhaps it has not been observed before that we can form the best discriminant function without applying Mahalanobis' $D^2$. The probability of misclassification can be calculated by forming discriminant functions with different combinations of particular characters and the discriminant function having the smallest probability of misclassification would be selected as the best discriminant function.

In connection with the example from the Department of Entomology it was observed that the best discriminant function is formed by the characters:

$x_1$: total body length. $x_2$: length of the cornicle. $x_4$: length of cauda.

If one wishes to construct the discriminant function for the purpose of classifying members in these species then the variables on which the measurements should be taken are $x_1$, $x_2$ and $x_4$. These characters do not have the maximum distance between the two species but they do give rise to a discriminant function with the minimum probability of misclassification.

In conclusion, we feel that Fisher's discriminant function is an ideal one because it is not only useful for discriminating between two species but can also be utilized for forming the best discriminant function with minimum probability of misclassification, provided the assumption of equality of dispersion matrices in the two populations is satisfied.

# BIBLIOGRAPHY


Anderson, T. W. (1951): Classification by multivariate analysis. *Psychometrika*, **16**: 31.

________, (1958): *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, New York.

Bartlett, M. S. (1938): Further aspect of the theory of multiple regression. *Proc. Camb. Phil. Soc.*, **34**: 33.

________, (1939a): A note on a test of significance in multivariate analysis. *Proc. Camb. Phil. Soc.*, **35**: 180.

Bhattacharya, A. (1946): On measure of divergence of two multinomial populations. *Sankhyā*, **7**: 401.

Bose, R. C. (1936): On the exact distribution and moment coefficients of the $D^2$-statistic. *Sankhyā*, **2**: 143.

Bose, R. C. and Roy, S. N. (1938): The exact distribution of the studentised $D^2$-statistic. *Sankhyā*, **4**: 535.

Fisher, R. A. (1936): The use of multiple measurements in taxanomic problems. *Ann. Eugenics*, **7**: 179.

________, (1938): The statistical utilization of multiple measurements. *Ann. Eugenics*, **8**: 376

________, (1940): The precision of discriminant functions. *Ann. Eugenics*, **10**: 422.

________, (1956): *Statistical Methods and Scientific Inference*. Oliver and Boyd, London.

Hotteling, H. (1931): The generalization of Student's ratio. *Ann. Math. Stat.*, **2**: 360.

________, (1952): A generalized T-test and measure of multivariate dispersion. Second Berkeley Symposium. *Univ. California Press* (1952): 23.

Hsu, P. L. (1938): Notes on Hotelling's generalized T. *Ann. Math. Stat.*, 9: 231-243.

John, S. (1960): On classification problems. *Sankhyā*, **22**: 301.

________, (1960): On some classification statistics. Sankhyā, 22: 309.

Kendall, M. G. (1957): A Course in Multivariate Analysis. Hafner Publishing Co., New York, 185 pp.

Mahalanobis, P. C. (1927): Analysis of race mixtures in Bengal. Jour. Asiatic Soc. Bengal, 23: 3.

________, (1930): Tests and measures of divergence. Jour. Asiatic Soc. Bengal, 26: 541.

________, (1936): On the generalized distance in statistics. Proc. Nat. Inst. Sci. India, 12: 49.

Majumdar, D. N. and Rao, C. R. (1958): Bengal anthrometric survey, 1945. A statistical study. Sankhyā, 19: 201.

Pearson, K. (1926): On coefficient of racial likeness. Biometrika, 13: 105.

Rao, C. R. (1946): Tests with discriminant functions in multivariate analysis. Sankhyā, 7: 407.

________, (1948a): The utilization of multiple measurements in problems in biological classification. J. Roy. Stat. Soc. (B), 10: 159.

________, (1948b): Tests of significance in multivariate analysis. Biom., 35: 58.

________, (1948c): On distance between two populations. Sankhyā, 9: 246.

________, (1949): On some problems arising out of discrimination with multiple characters. Sankhyā, 9: 343.

________, (1950): Statistical inference applied to classificatory problems. Sankhyā, 10: 229.

________, (1952): Advanced Statistical Methods in Biometric Research. John Wiley and Sons, New York.

________, (1954a): A general theory of discrimination when the information about alternative population is based on samples. Ann. Math. Stat., 25: 651.

________, (1954b): On the use and interpretation of distance function in statistics. Bull. Int. Statist. Inst., 34: 90.

________, (1955b): Analysis of dispersion for multiply classified data with unequal number in cells. Sankhyā, 15: 253.

________, (1960): Multivariate Analysis: An indispensable statistical aid in applied research. Sankhyā, 22: 317-338.

________, (1962): Use of discriminant and allied functions in multivariate analysis. Sankhyā 24 (A): 149.

________, and Shaw, D. C. (1948): On a formula for the prediction of cranical capacity. Biometrics, 4: 247.

________, and Slater, P. (1949): Multivariate analysis applied to differences between neurotic groups. British J. Psy. (Stat. Sec.), 2: 17.

Roy, S. G. and Mukherjee, G. D. (1964): Use of Mahalanobis' $D^2$ statistics in phase discrimination in desert locust males. Sankhyā, 26: 237.

Roy, S. N. (1939): p-statistics on some generalization in analysis of variance appropriate to multivariate problems. Sankhyā, 4: 391.

________, (1950): Univariate and multivariate analysis as problem in testing of composite hypothesis. Sankhyā, 10: 29.

Seal, H. L. (1964): Multivariate Statistical Analysis for Biologists. Methuen and Co., Ltd., London.

Tildesley, M. L. (1921): On Burmese skulls. Biometrika, 13: 247.

Wald, A. (1943): Tests of statistical hypothesis concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc., 54: 426.

________, (1944): On statistical problem arising in the classification of an individual into one of the two groups. Ann. Math. Stat., 15: 145.

Wilks, S. S. (1932): Certain generalization in analysis of variance. Biom., 24: 471.

Wishart, J. (1928): The generalized product moment distribution in samples from a normal multivariate population. Biom., 20: 32.