DISCRIMINANT FUNCTIONS

by

LEROY JOSEPH YORK

B. S., Kansas State University, 1961

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

Approved by:

Dr. H. C. Fryer, Head

TABLE OF CONTENTS

INTRODUCTION

The problem of discrimination or classification arises when an investigator is given an item, I, which is known to have come from one of k specified categories or populations, and is asked to classify this item into the population from which it came. The classification problem becomes statistical when we further specify that the available evidence about I consists of observed values of a set of random variables. The basis for classification is dependent upon these observed random variables as well as the information available about the k populations. In most practical situations it may be assumed that there is a finite number of populations from which the individual may have come, and that each population is characterized by a probability distribution. Thus the item I, is a random observation from one of k specified probability distributions. In some problems the probability distributions for all of the k populations are completely known. In other problems the probability distributions may be known or assumed to be of a specified type, but only sample estimates of their parameters are available. In all cases let the observed random variables from the item I be a set of measurements of, say, p characteristics or quantities, taken from I.

Suppose an item I is known to have come from one of k populations $\pi_1, \ldots, \pi_k$. Denote the vector of p measurements taken from the item as $V' = (x_1, \ldots, x_p)$. Let R denote the total sample space which consists of all possible vector measurements; thus R is a p-dimensional space. The purpose of the discriminant function is to divide the R space into k mutually exclusive subspaces $R_1, \ldots, R_k$; such that if an observation falls in the region R_j, it would be classified as coming from population

π_j , $j = 1, \ldots, k$. The criterion for determining the regions $R_1, \ldots, R_k$, will be that of minimizing either the probability of misclassification or the cost of misclassification.

The history of discriminatory analysis may be regarded as beginning with the work of Karl Pearson in 1920. Pearson was faced with the problem of measuring the distance between two multivariate populations. Pearson proposed a statistic which he called the "coefficient of racial likeness" and denoted it as C^2 . His first work on C^2 assumed that the p measured characteristics were independent. Pearson later made an adjustment in the "coefficient of racial likeness" to account for the relationship between the p variates.

In 1925 P. C. Mahalanobis became interested in the subject and proposed an alternative measure which he called D^2 . In 1931 Hotelling generalized "Student's" t into a statistic which he called T^2 . Hotelling's T^2 and Mahalanobis' D^2 are in fact equivalent; however, it was some time before this equivalence was realized. In 1936 Fisher published his first paper on discriminant functions. The main difference between his approach and that of Mahalanobis was that the latter was measuring distance between groups whereas Fisher was merely concerned with dividing the sample space into two regions and classifying a sample value to one population or another on the basis of which region it fell into. Then in 1939 Welch linked up the theory of discriminatory functions and that of statistical tests. It is at this point where one is interested in directions between group and regions of classification and not just the distance between groups that this report begins.

CLASSIFICATION INTO ONE OF TWO POPULATIONS WITH KNOWN PROBABILITY DISTRIBUTIONS

A Priori Probabilities Are Known

Consider a situation in which the item I is known to belong to one of two populations, π_1 or π_2. Using any given classification procedure, the investigator could make two kinds of errors in classification. If the item belonged to π_1, it could be classified as coming from π_2, denoted by $P(2/1)$; and if the item belonged to π_2, it could be classified as coming from π_1, denoted as $P(1/2)$. Let the a priori probability of selecting an observation from population π_i be represented by p_i. Also let the density function of population π_i be represented by $f_i(v; \theta_i)$, where

$$f_i(v; \theta_i) = f_i(x_1, \ldots, x_p ; \theta_{i1}, \ldots, \theta_{ip}) .$$

If R, the total sample space, is subdivided into two mutually exclusive subspaces, R_1 and R_2, such that all observations in R_1 are classified as belonging to π_1, and all observations in R_2 are classified as belonging to π_2, then the probability of correctly classifying an observation is

$$p_1 \int_{R_1} f_1(v; \theta_1)\, dv + p_2 \int_{R_2} f_2(v; \theta_2)\, dv ,$$

where $dv = dx_1, \ldots, dx_p$. The probability of misclassifying an observation is then

$$M = p_1 \int_{R_2} f_1(v; \theta_1)\, dv + p_2 \int_{R_1} f_2(v; \theta_2)\, dv . \qquad (1)$$

Using the above procedure for classification, the problem becomes that of choosing R_1 and R_2 so as to minimize M.

Using the method of Bayes, the a posteriori probabilities that I

belongs to π_1 or π_2 may be computed. That is, the conditional probability that an observation came from a certain population given the observed values of the items' p measurements may be computed. For instance, given the observed values for the p variates of an item, the conditional probabilities that the item belongs to population π_1 or to population π_2, are

$$\frac{p_1 f_1(v;\ \theta_1)}{p_1 f_1(v;\ \theta_1) + p_2 f_2(v;\ \theta_2)} \quad \text{and} \quad \frac{p_2 f_2(v;\ \theta_2)}{p_1 f_1(v;\ \theta_1) + p_2 f_2(v;\ \theta_2)},$$

respectively.

For a given observation I, the probability of misclassification is minimized by assigning I to that population which has the largest conditional probability. That is, if

$$\frac{p_1 f_1(v;\ \theta_1)}{p_1 f_1(v;\ \theta_1) + p_2(v;\ \theta_2)} > \frac{p_2 f_2(v;\ \theta_2)}{p_1 f_1(v;\ \theta_1) + p_2 f_2(v;\ \theta_2)},$$

then I is classified as coming from population π_1. If the direction of the inequality is changed and the statement holds, I is classified as coming from population π_2. When neither inequality holds, the populations are equally probable and it makes no difference which one is chosen. To make the regions R_1 and R_2 mutually exclusive one can arbitrarily assign I to R_1 when

$$p_1 f_1(v;\ \theta_1) = p_2 f_2(v;\ \theta_2).$$

Thus the subspaces R_1 and R_2 are

$$R_1 : \quad \frac{f_1(v;\ \theta_1)}{f_2(v;\ \theta_2)} \geq \frac{p_2}{p_1}, \tag{2}$$

and

$$R_2 : \quad \frac{f_1(v;\ \theta_1)}{f_2(v;\ \theta_2)} < \frac{p_2}{p_1}, \tag{3}$$

respectively, where the symbol : stands for "defined by."

It should be noted that since the probability of misclassification is minimized at each point, i.e., for all I, it is minimized for the entire space R.

Now the question arises: Is this the "best" procedure? The best procedure is that one which minimizes the probability of misclassification. For any procedure, the probability of misclassification is given by equation (1), where the intersection of R_1 and R_2 is the null set and the union of R_1 and R_2 is the entire R space.

Equation (1) can be expressed as

$$M = \int_{R_2} [p_1 f_1(v; \theta_1) - p_2 f_2(v; \theta_2)] \, dv + \int_{R} p_2 f_2(v; \theta_2) \, dv . \qquad (4)$$

The second term on the right hand side of equation (4) is a constant, namely p_2. Therefore M will be a minimum when R_2 includes all the points V' such that $p_1 f_1(v; \theta_1) - p_2 f_2(v; \theta_2) < 0$, and excludes all the points V' for which $p_1 f_1(v; \theta_1) - p_2 f_2(v; \theta_2) > 0$. Thus the regions R_2 and R_1 are those defined by equations (2) and (3).

If it is assumed that

$$\Pr\left[\frac{f_1(v; \theta_1)}{f_2(v; \theta_2)} = \frac{p_2}{p_1} \middle/ \pi_i \right] = 0 \qquad i = 1, 2$$

then the Bayes' procedure is unique.

If the expected cost of misclassification which is given by

$$C(2/1) \, p_1 \int_{R_2} f_1(v; \theta_1) \, dv + C(1/2) \, p_2 \int_{R_1} f_2(v; \theta_2) \, dv ,$$

is to be minimized, where $C(2/1)$ is the cost of classifying an observation into π_2 when it actually comes from π_1; and $C(1/2)$ is the cost of

classifying an observation into π_1 when it actually comes from π_2, then our regions become:

$$R_1 : \quad \frac{f_1(v;\ \theta_1)}{f_2(v;\ \theta_2)} \geq \frac{p_2 C(1/2)}{p_1 C(2/1)} ,$$

and

$$R_2 : \quad \frac{f_1(v;\ \theta_1)}{f_2(v;\ \theta_2)} < \frac{p_2 C(1/2)}{p_1 C(2/1)} .$$

Classification Into One of Two Known Multivariate Normal Populations

Now let us apply the general procedure outlined above to the case in which the two populations are multivariate normal populations with equal covariance matrices. Let $N(\mu^{(1)}, \Sigma)$ and $N(\mu^{(2)}, \Sigma)$ represent the two populations, where $\mu^{(k)\prime} = (\mu_1^{(k)}, \ldots, \mu_p^{(k)})$ is the vector of means for the kth population, and Σ is the common covariance matrix with elements σ_{ij} $i, j = 1, 2, \ldots, p$. Then the density function of the kth population $(k = 1, 2)$ is:

$$f_k(v;\ \theta_k) = \frac{\left|\Sigma^{-1}\right|^{1/2}}{(2\pi)^{p/2}} \; e^{-\left[\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma^{ij}(x_i - \mu_{ki})(x_j - \mu_{kj})\right]} , \qquad (5)$$

where the σ^{ij} are the elements of the inverse matrix Σ^{-1}, and $\left|\Sigma^{-1}\right|$ is the determinant of the inverse matrix Σ^{-1}. The ratio of the two density functions is:

$$\frac{f_1(v;\ \theta_1)}{f_2(v;\ \theta_2)} = e^{-\frac{1}{2}\left[\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(x_i-\mu_{1i})(x_j-\mu_{1j}) - \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(x_i-\mu_{2i})(x_j-\mu_{2j})\right]}$$

or equivalently,

$$\frac{f_1(v;\,\theta_1)}{f_2(v;\,\theta_2)} = e^{-\frac{1}{2}\left[\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(x_i x_j - x_i\mu_{1j} - x_j\mu_{1i} + \mu_{1i}\mu_{1j} - x_i x_j + x_i\mu_{2j} + x_j\mu_{2i} - \mu_{2i}\mu_{2j})\right]},$$

$$= e^{-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_{1i}\mu_{1j} - \mu_{2i}\mu_{2j}) + \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_{1j} - \mu_{2j})\,x_i} \quad (6)$$

The region R_1 was defined as the set of V' for which (2) holds. Since the logarithmic function is monotonic increasing, the inequality, and hence the region R_1, can be written in terms of the logarithm of (2). Taking the logarithms of these inequalities we have:

$$R_1 : \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_{1j} - \mu_{2j})\,x_i - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_{1i}\mu_{1j} - \mu_{2i}\mu_{2j}) \geq \ln k \,, \quad (7)$$

and

$$R_2 : \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_{1j} - \mu_{2j})\,x_i - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_{1i}\mu_{1j} - \mu_{2i}\mu_{2j}) < \ln k \,, \quad (8)$$

for k suitably chosen. If the probability of misclassification is to be minimized and the a priori probabilities are known, then $k = \frac{p_2}{p_1}$. If the cost of misclassification, or the ratio of the cost of misclassification, is known and is to be minimized rather than just the probability of misclassification, then k becomes $p_2C(1/2) \,/\, p_1C(2/1)$. For the particular case of the two populations being equally likely to occur, i.e., $p_1 = p_2$, and the cost of misclassification being the same for each population, R_1 becomes

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\mu_{1j} - \mu_{2j})\,x_i \geq \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\mu_{1i}\mu_{1j} - \mu_{2i}\mu_{2j}) \,. \quad (9)$$

It should be noted that if the covariance matrices are not equal then the region R_1 is defined by the quadratic expression:

$$-\frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{p}\alpha^{ij}(x_i - \mu_{1i})(x_j - \mu_{1j}) + \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{p}\beta^{ij}(x_i - \mu_{2i})(x_j - \mu_{2j}) \geq \ln k \,, \quad (10)$$

where (α^{ij}) and (β^{ij}) are the inverses of the covariance matrices of the two populations.

No A Priori Probabilities Are Known

Using matrix notation one can express the region R_1 as:

$$R_1 : \quad V' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) \geq \ln k$$

where,

$$(\mu^{(1)} - \mu^{(2)})' = (\mu_{11} - \mu_{21}, \mu_{12} - \mu_{22}, \cdots, \mu_{1p} - \mu_{2p}) ,$$

is the vector of differences between the population means for the p characteristics, and all other terms are as defined earlier.

Given the a priori probabilities, the method of Bayes was used to determine k . When a priori probabilities do not exist or are unknown, a procedure must be sought for determining k . Anderson (1958) proves a series of theorems which enable him to state that the Bayes procedure R* , for which $P(1/2) = P(2/1)$, is a minimax procedure. A procedure is called minimax if the maximum expected loss is a minimum. Hence k is determined so that the expected losses due to misclassification are equal.

Consider the distribution of the ratio of the natural logarithm of the density functions and denote this ratio by U .

$$\frac{f(v;\ \theta_1)}{f(v;\ \theta_2)} = U = V' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2} (\mu^{(1)} + \mu^{(2)})' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) .$$

If V' is distributed according to $N(\mu^{(1)}, \Sigma)$ then

$$E(U) = \mu^{(1)'} \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$$

or equivalently,

$$E(U) = \frac{1}{2}(\mu^{(1)} - \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) .$$

Denote the variance of U by $\sigma^2(U)$.

Then

$$\sigma^2(U) = \sigma^2(L) , \quad \text{where } L = V' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)})$$

$$\sigma^2(L) = \sigma^2([x_1, \ldots, x_p] \begin{bmatrix} c_1 \\ \cdot \\ \cdot \\ \cdot \\ c_p \end{bmatrix})$$

where $$C = \left[\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})\right] = \begin{bmatrix} c_1 \\ \cdot \\ \cdot \\ \cdot \\ c_p \end{bmatrix}$$

i.e.,

$$c_i = \sigma^{i1}(\mu_{11} - \mu_{12}) + \sigma^{i2}(\mu_{12} - \mu_{22}) + \ldots + \sigma^{ip}(\mu_{1p} - \mu_{2p}) .$$

$$(i = 1, \ldots, p)$$

Note that L can be expressed as $\sum_{i=1}^{p} c_i x_i$.

Wilks (1962) states that

"If $(x_1, \ldots, x_p)$ has the p-variate distribution

$$N (\mu_i ; || \sigma_{ij} ||) , \qquad i, j, = 1, \ldots, p$$

then

$$L = c_1 x_1 + \ldots + c_p x_p$$

has the distribution

$$N (\sum_{i=1}^{p} c_i u_i ; \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j \sigma_{ij}) "$$

hence

$$\sigma^2(L) = \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j \sigma_{ij} = \sigma^2(U) .$$

Therefore, using matrix notation one has

$$\sigma^2(U) = C' \Sigma C$$

$$\sigma^2(U) = \left[\Sigma^{-1} (\mu^{(1)} - \mu^{(2)})\right]' \left[\Sigma\right] \left[\Sigma^{-1} (\mu^{(1)} - \mu^{(2)})\right]$$

$$\sigma^2(U) = (\mu^{(1)} - \mu^{(2)})' \left[\Sigma^{-1}\right] (\mu^{(1)} - \mu^{(2)}) .$$

Hence, if V' comes from π_1, U is distributed as $N(\frac{a}{2}\ ;\ a)$, where $a = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1} (\mu^{(1)} - \mu^{(2)})$. Similarly if V comes from π_2, U is distributed as $N(-\frac{a}{2}\ ;\ a)$. Thus the probability of misclassification, given that the observation comes from π_1 and π_2, is

$$P(2/1) = \frac{1}{\sqrt{2\pi a}} \int_{-\infty}^{d} e^{-\frac{1}{2a}(U - \frac{1}{2}a)^2} \, dU ,$$

and

$$P(1/2) = \frac{1}{\sqrt{2\pi a}} \int_{d}^{\infty} e^{-\frac{1}{2a}(U + \frac{1}{2}a)^2} \, dU ,$$

respectively, where $d = \ln k$.

Setting $P(2/1)$ equal to $P(1/2)$ and transforming the U variate to the standard normal scale we have

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{d - a/2}{\sqrt{a}}} e^{-\frac{1}{2}z^2} \, dz = \frac{1}{\sqrt{2\pi}} \int_{\frac{d + a/2}{\sqrt{a}}}^{\infty} e^{-\frac{1}{2}z^2} \, dz .$$

In order for this equation to hold, d must equal zero.

Hence, when no a priori probabilities are known the problem is treated in the same manner as if a priori probabilities are known and are equal. Thus, for this case the regions R_1 and R_2 are defined by equation (7) and (8)

respectively, where $k = 1$.

When $C(1/2) \neq C(2/1)$ and the cost of misclassification is to be a minimum then d can be determined by use of the normal tables and trial and error such that, $C(2/1)\ P(2/1) = C(1/2)\ P(1/2)$.

Unknown Parameters

So far we have assumed that the distributions of the two populations are completely known. In most practical situations the two populations are not completely known. However, they are known or assumed to be of a specified type and their parameters must be estimated from two samples, one from each population. The first step in this problem is that of testing the hypothesis that the two samples actually do come from different populations. For classification would be meaningless unless the two populations are distinguishable. To test this hypothesis we will employ the T^2 statistic, which is a direct generalization of the Student t , derived by Hotelling (1931). The T^2 statistic may be used to test the hypothesis that a multivariate sample came from a specified normal population or that two independent multivariate samples have been drawn from the same normal population. In the two-sample problem, the normal populations must have the same but unknown covariance matrices.

Let x_{kji} denote the value of the jth variate measured on the ith individual from the kth sample, where $k = 1, 2,$; $i = 1, 2, \ldots, n_k$; and $j = 1, 2, \ldots, p$. Let $\bar{x}_{1j}$ and $\bar{x}_{2j}$ be the arithmetic means of the values of the jth variates in the first and second samples, respectively, where

$$\bar{x}_{kj} = \sum_{k=1}^{n_k} \frac{x_{kji}}{n_k} .$$

Next define

$$d_j = \bar{x}_{1j} - \bar{x}_{2j} , \qquad n = n_1 + n_2 - 2 ,$$

and

$$ns_{jj'} = \sum_{i=1}^{n_1} (x_{1ji} - \bar{x}_{1j})(x_{1j'i} - \bar{x}_{1j'}) + \sum_{i=1}^{n_2} (x_{2ji} - \bar{x}_{2j})(x_{2j'i} - \bar{x}_{2j'}) .$$

Now form the estimate of our covariance matrix

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

Hotelling's T^2 statistic is,

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{p} \sum_{i=1}^{p} s^{ij} d_i d_j = \frac{n_1 n_2}{n_1 + n_2} D' S^{-1} D$$

where s^{ij} is the element in the ij position of the inverse matrix of S and $D' = (\bar{x}_{11} - \bar{x}_{21}, \ldots, \bar{x}_{1p} - \bar{x}_{2p})$. Hotelling proved that the quantity $\frac{n + 1 - p}{n \cdot p} T^2$, is distributed as the F-distribution with p and $n + 1 - p$ degrees of freedom. That is

$$\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{(n_1 + n_2) p (n_1 + n_2 - 2)} \sum_{j=1}^{p} \sum_{i=1}^{p} s^{ij} d_i d_j ,$$

has the F-distribution with p and $n_1 + n_2 - p - 1$ degrees of freedom.

Now the critical region may be selected from the tables of the F-distribution at whatever level of significance is desired. Once we have accepted the hypothesis that the two samples are from different populations, let us turn to the problem of classifying V' into the population from which it came.

That is, we have a sample $V_1^{(1)}, \ldots, V_{n_1}^{(1)}$ from $N(\mu^{(1)} \Sigma)$ and a sample $V_1^{(2)}, \ldots, V_{n_2}^{(2)}$ from $N(\mu^{(2)}, \Sigma)$ and, on the basis of this information and the measurements taken on V', we want to classify V' as coming from π_1 or π_2. Our maximum likelihood estimate of $\mu^{(1)'}$ is $\bar{X}^{(1)'} = (\bar{x}_{11}, \ldots, \bar{x}_{1p})$, of $\mu^{(2)'}$ is $\bar{X}^{(2)'} = (\bar{x}_{21}, \ldots, \bar{x}_{2p})$, and of σ_{ij} is s_{ij} where $\bar{x}_{kj}$ and s_{ij} are defined on the preceding page.

Substituting these maximum likelihood estimates for the parameters we use the same criterion for the classification of V' as we did in the situations in which the parameters were known.

Hence for the case where $p_1 = p_2$,

$$R_1: \quad V'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) \geq \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}), \quad (11)$$

$$R_2: \quad V'S^{-1}(\bar{X}^{(1)} - X^{(2)}) < \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}). \quad (12)$$

R. A. FISHER'S APPROACH TO THE CLASSIFICATION PROBLEM

In 1936 R. A. Fisher considered the problem of discrimination in a totally different manner and obtained similar results. Fisher's approach was as follows:

There are two populations π_1, and π_2. From each population there is a sample, n_1 items from π_1, and n_2 items from π_2. Also, there is an item, I, which could have come from either π_1 or π_2. The decision problem is then to assign I to one of the two populations, where the available information consists of measurements of p quantities which are made on

I and the $n_1 + n_2$ sample items.

If there is only one characteristic, then the problem of classification is very simple; all individuals having values of the characteristic exceeding a suitably determined value, could be assigned to one group, and all others to the other group.

Fisher dealt with the multivariate problem, i.e., $p > 1$, by reducing it to a univariate problem. To do this he replaced the p measurements for each individual, by a single measurement, say Y. Fisher considered only linear combinations of the p variates. Therefore one has

$$Y_{ki} = z_1 x_{k1i} + z_2 x_{k2i} + \ldots + z_p x_{kpi}$$

as the linear combination of the p measurements representing the ith individual from the kth sample. If one denotes the measurement of the jth trait on item I by x_j, then

$$Y_I = z_1 x_1 + z_2 x_2 + \ldots + z_p x_p ,$$

is the linear combination of the p measurements representing the individual I.

The proper choice of the z_i's may then be measured by the relative ease of classifying I through the use of the values of Y_I and the Y_{ki}'s. Fisher introduces the numerical measure of the ability to distinguish between the two populations as being the ratio of:

$$\frac{\text{the difference between sample means}}{\text{the standard deviation within samples}} .$$

He then was able to suggest a reasonable criterion for determining appropriate values of the z_i's. This was the linear function of the measurements that maximized the ratio of the difference between sample means to the standard deviation within sample means.

Mathematically, Fisher maximized the ratio

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2} \Big/ (n_1 + n_2 - 2)} \tag{13}$$

where,

$$\bar{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}$$

is the mean value of the new variable for the kth sample. Note that the constant factor of $n_1 + n_2 - 2$ may be omitted since constant factors do not affect the maximization problem. Also, the square of the ratio may be considered for ease of computation.

One can show that $\bar{Y}_1 - \bar{Y}_2 = \sum_{j=1}^{p} z_j d_j$, where $d_j = \bar{x}_{1j} - \bar{x}_{2j}$, is the difference between sample means for the jth trait. This difference, $\bar{Y}_1 - \bar{Y}_2$, may be denoted as B.

In a similar manner the sum of squares, due to the variability within samples which is

$$\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2 \tag{14}$$

can be shown to be

$$\sum_{j=1}^{p} \sum_{m=1}^{p} z_j z_m w_{jm} = T , \tag{15}$$

where w_{jm} is the pooled sum of products of deviations from the sample means of traits j and m; that is,

$$w_{jm} = \sum_{k=1}^{2} \sum_{i=1}^{n_k} (x_{kji} - \bar{x}_{kj})(x_{kmi} - \bar{x}_{km}) . \tag{16}$$

Now the problem is to determine the values of the z_j's for which the ratio

$$\frac{B^2}{T} = \left(\sum_{j=1}^{p} z_j d_j\right)^2 \Big/ \sum_{j=1}^{p}\sum_{m=1}^{p} z_j z_m w_{jm} \qquad (17)$$

is maximized. In order to find the maximization solution differentiate (17) with respect to z_r, and set the derivative equal to zero, $r = 1, 2, \ldots, p$. This gives the following:

$$\sum_{j=1}^{p} z_j w_{jr} = d_r \sum_{j=1}^{p}\sum_{m=1}^{p} z_j z_m w_{jm} \Big/ \sum_{j=1}^{p} z_j d_j , \quad r = 1, 2, \ldots, p .$$

Since $\sum_{j=1}^{p}\sum_{m=1}^{p} z_j z_m w_{jm} \Big/ \sum_{j=1}^{p} z_j d_j$ is a constant for any set of equations and one is interested only in a proportional solution it can be ignored, leaving

$$\sum_{j=1}^{p} z_j w_{jr} = d_r , \qquad r = 1, 2, \ldots, p ,$$

which is a set of p simultaneous linear equations:

$$\begin{aligned}
z_1 w_{11} + z_2 w_{12} + \ldots + z_p w_{1p} &= d_1 \\
z_1 w_{12} + z_2 w_{22} + \ldots + z_p w_{2p} &= d_2 \\
\cdot \quad \cdot \quad \ldots \quad \cdot \quad \cdot \\
\cdot \quad \cdot \quad \ldots \quad \cdot \quad \cdot \\
\cdot \quad \cdot \quad \ldots \quad \cdot \quad \cdot \\
z_1 w_{p1} + z_2 w_{p2} + \ldots + z_p w_{pp} &= d_p
\end{aligned}$$

Representing this system of equations in matrix notation, one has

$$(W)(Z) = (D) \qquad (18)$$

where

(W) is a $p \times p$ matrix; while

(Z) and (D) are $p \times 1$ column vectors.

Thus one has a set of simultaneous equations which when solved will give

the z multipliers which will maximize the ratio of B^2 / T. This is the ratio of the square of the difference between sample means to the variance within samples for the variable Y. Once the Z vector is computed one can easily compute the quantities $\bar{Y}_1$, $\bar{Y}_2$, and Y_I. The problem now becomes a univariate one and I is placed in π_1 or in π_2 depending upon which $\bar{Y}_k$ that Y_I is closest to. That is, if Y_I is closer to $\bar{Y}_1$ than to $\bar{Y}_2$, classify I as coming from population π_1; otherwise, classify I as coming from population π_2. For simplification of computations, note that,

$$\bar{Y}_k = z_1 \bar{x}_{k1} + z_2 \bar{x}_{k2} + \ldots + z_p \bar{x}_{kp} .$$

PROPERTIES OF THE DISCRIMINANT FUNCTION

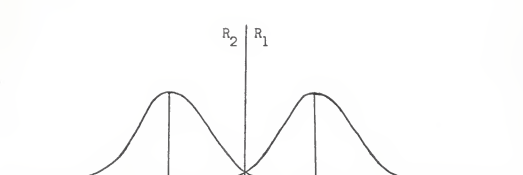
Comparison of Two Methods

It was interesting to note that if one took the case studied earlier where the density functions of the two populations were multivariate normals, with equal covariance matrices, and equal a priori probabilities; then it could be shown that the discriminating functions obtained by using the two different methods were, in fact, the same functions. Thus Welch's work put a theoretical basis under Fisher's discriminant function; at least in this special case.

Consider the case where $\bar{Y}_1$ was found to be the larger of the two sample means. Then using Fisher's method R_1 would be the set of Y_I for which

$$R_1 : \quad Y_I \geq \bar{Y}_1 - \frac{\bar{Y}_1 - \bar{Y}_2}{2} \quad \text{or} \quad \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2) \tag{19}$$

i.e.,



Note that $Y_I = \sum_{j=1}^{p} x_j z_j = V'(Z)$ and

$$\bar{Y}_k = \sum_{j=1}^{p} \bar{x}_{kj} z_j = (\bar{X}^{(k)})'(Z) ,$$

where $(Z)' = (z_1, \ldots, z_p)$ and $(\bar{X}^{(k)})' = (\bar{x}_{11}, \ldots, \bar{x}_{1p})$.

Hence the region R_1 of (19) can be written as

$$V'(Z) \geq (\bar{X}^{(1)})'(Z) - \frac{1}{2}\left[(\bar{X}^{(1)})'(Z) - (X^{(2)})'(Z)\right] . \qquad (20)$$

The Z vector is the solution of the matrix equation (18) and can be expressed as $(Z) = (W)^{-1}(D)$, hence, R_1 contains those items for which

$$V'(Z) \geq \left[(\bar{X}^{(1)})' - \frac{1}{2}(\bar{X}^{(1)})' + \frac{1}{2}(\bar{X}^{(2)})'\right](Z)$$

$$V'(W)^{-1}(D) \geq \left[\frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})'\right](W)^{-1}(D)$$

or

$$V'(W)^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) \geq \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})'(W)^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) . \qquad (21)$$

Multiplying both sides of equation (21) by the constant $\frac{1}{n_1 + n_2 - 2}$

equation (21) becomes (11), for $\frac{1}{n_1 + n_2 - 2}(W) = (S)$.

The Significance of a Discriminant Function

One may ask whether a particular discriminator is "significant." Questions of "significance" in discriminant functions have usually been discussed in terms of whether or not the parent populations are identical and hence whether or not a discriminant function is illusory. They are not so much a test of the functions as they are a test of the homogeneity of the populations, by use of the functions. If heterogeneity is found the function is significant in the sense that it discriminates between real differences in an optimal way. For making this test of significance Fisher suggested the use of an analysis of variance. The total sum of squares of deviations of all observations from their grand mean can be expressed as

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k}(Y_{ki} - \bar{\bar{Y}}_{..})^2 = \sum_{k=1}^{2}\sum_{i=1}^{n_k}(Y_{ki} - \bar{Y}_{k.})^2 + \sum_{k=1}^{2} n_k(\bar{Y}_{k.} - \bar{\bar{Y}}_{..})^2 \qquad (22)$$

The first component on the right side of equation (22) expresses the "within sample" variation and the second component expresses the "between sample" variation. Furthermore, the component representing the "between sample" variation can be written as:

$$\frac{(\sum_{i=1}^{n_1} Y_{1i})^2}{n_1} + \frac{(\sum_{i=1}^{n_2} Y_{2i})^2}{n_2} - \frac{(\sum_{i=1}^{n_1} Y_{1i} + \sum_{i=1}^{n_2} Y_{2i})^2}{n_1 + n_2}$$

or

$$(n_2 \sum_{i=1}^{n_1} Y_{1i} - n_1 \sum_{i=1}^{n_2} Y_{2i})^2 \;/\; n_1 n_2 (n_1 + n_2)$$

which for our discriminant problem is:

$$(\bar{Y}_1 - \bar{Y}_2)^2 \frac{n_1 n_2}{n_1 + n_2} = B^2 \frac{n_1 n_2}{n_1 + n_2} \; . \qquad (23)$$

Hence equation (23) represents the sum of squares due to the variability "between sample" means. The sum of squares due to variability "within samples" is given by equation (14).

If the system of equations (18), whose solution is the (Z) column vector, is multiplied on the left by the transpose of the (Z) column vector, i.e.,

$$(Z)' (W) (Z) = (Z)' (D) \tag{24}$$

the left hand side of the equation (24) is T the sum of squares due to variability "within samples," and the right hand side of the equation is B, the sum of squares due to the variability "between sample" means. Hence $T = B$. Fisher then concluded that if the measurements were normally distributed, or nearly normally distributed, then the linear compound of measurements, i.e., the Y_i, would be normally distributed. Therefore, if the variances of the two transformed groups are equal, the analysis of variance table would be:

Source of Variation	Degrees of Freedom	Sum of Squares
Between samples	p	$\frac{n_1 n_2}{n_1 + n_2} B^2$
Within samples	$n_1 + n_2 - p - 1$	$T = B$
Total	$n_1 + n_2 - 1$	$B(1 + \frac{n_1 n_2}{n_1 + n_2} B)$

(25)

This analysis of variance gives a means for testing the hypothesis that the two samples are actually from different populations. This situation would

be indicated by a significant F value with p and $n_1 + n_2 - p - 1$ degrees of freedom. A comparison of this analysis of variance and Hotelling's T^2 statistic, which was presented earlier in the report, will show that the two are identical testing procedures.

Computing Probabilities of Misclassification

The break-down of the sum of squares in the analysis of variance is of interest also in relation to the probabilities of misclassification. The within samples variation, divided by the within samples degrees of freedom, gives an estimate of the variance of Y . That is, the estimate of the variance of a single item Y is $B / (n_1 + n_2 - p - 1)$. Using the procedure (19) an element from group two is misclassified if its deviation from $\bar{Y}_2$, in the right direction, exceeds $1/2\ (\bar{Y}_1 - \bar{Y}_2)$. Also an element from group one is misclassified if its deviation from $\bar{Y}_1$, in the right direction, exceeds $1/2\ (\bar{Y}_1 - \bar{Y}_2)$. To find the probabilities of misclassification one simply needs to find the probability that Y will exceed the deviation which will cause misclassification. Fisher treats the ratio of $1/2\ (\bar{Y}_1 + \bar{Y}_2) - \bar{Y}_k$ to the standard error of Y , as being distributed as Student's t-distribution with $(n_1 + n_2 - p - 1)$ degrees of freedom $(k = 1 \text{ or } 2)$. Thus to find the probability of misclassifying an element which belongs to group two, the ratio of $1/2\ (\bar{Y}_1 + \bar{Y}_2) - \bar{Y}_2$ to the standard error of Y is computed. Then by comparing this ratio to the tabulated values of the appropriate t-distribution one determines the probability of misclassification. That is, $P(1/2)$, is equal to the probability of getting a t-variate, with appropriate degrees of freedom, which is greater than the ratio of $1/2\ (\bar{Y}_1 + \bar{Y}_2) - \bar{Y}_2$ to the

standard error of Y. The probability of misclassifying an element which belongs to group one, $P(2/1)$, is equal to the probability of getting a t variate, with appropriate degrees of freedom, which is less than the ratio of $1/2(\bar{Y}_1 + \bar{Y}_2) - \bar{Y}_1$ to the standard error of Y. It must be kept in mind that the deviation has to be in the right direction in order to cause misclassification. The total probability of misclassification is then $p_1 P(2/1) + p_2 P(1/2)$. In computing the probability of misclassification Fisher assumed that p_1 was equal to p_2. This of course may not be the true situation; and, if not, an adjustment must be made. A previous section showed that Welch's method and Fisher's method are actually equivalent. Using Welch's method, equations (7 and 8) gave the procedure for classification when $p_1 \neq p_2$ but gave no means for determining the probabilities of misclassification. Following Rao (1952), for the case where $p_1 \neq p_2$, one can express Welch's region R_1 in vector notation as

$$R_1 : V'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - 1/2(X^{(1)} + X^{(2)})'S^{-1}(X^{(1)} - X^{(2)}) \geq \ln p_2 - \ln p_1 ,$$

which can be written as:

$$R_1 : \qquad Y_I \geq 1/2(\bar{Y}_1 + \bar{Y}_2) + \ln p_2 - \ln p_1 .$$

Under these conditions $P(2/1)$ is determined by finding the probability that a t-variate with $(n_1 + n_2 - 2)$ degrees of freedom will be less than the ratio of $1/2(\bar{Y}_1 + \bar{Y}_2) + \ln p_2 - \ln p_1 - \bar{Y}_1$ to the standard error of Y. $P(1/2)$ is equal to the probability that a t-variate with $(n_1 + n_2 - 2)$ degrees of freedom will be equal to or exceed the ratio of $1/2(\bar{Y}_1 + \bar{Y}_2) + \ln p_2 - \ln p_1 - \bar{Y}_2$ to the standard error of Y. Again the deviation must occur in the right direction to cause misclassification. The standard error of Y is obtained from the analysis of variance of (25).

The "Doubtful" Region

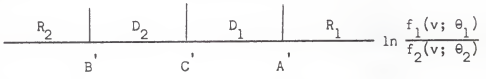
The classification problem is somewhat different than most problems of testing hypotheses. The general procedure for testing hypotheses is to arbitrarily set the probability of a type one error and then use some specified procedure to determine the region of rejection for the test. In the above approach to the classification problem the critical point was determined first, then the probability of misclassification using this critical point was determined. The critical point was chosen so that the probability of misclassification would be a minimum and was not arbitrarily set by the investigator. Only when $P(2/1)$ and $P(1/2)$ are small can the investigator assert with a high degree of confidence that any given individual is correctly classified. Rao (1952) gives a method by which $P(2/1)$ and $P(1/2)$ can be arbitrarily chosen by the investigator. To do this Rao divides the R space into three regions, R_1, R_2, and R_D. Individuals that fall in regions R_1 and R_2 are classified into population π_1 and π_2 respectively, and individuals falling into R_D remain in doubt, as to which population they belong.

These three regions are:

$$
\begin{aligned}
R_1 &: \quad \frac{f_1(v;\, \theta_1)}{f_2(v;\, \theta_2)} \geq A \\
R_D &: \quad B < \frac{f_1(v;\, \theta_1)}{f_2(v;\, \theta_2)} < A \qquad (26) \\
R_2 &: \quad \frac{f_1(v;\, \theta_1)}{f_2(v;\, \theta_2)} \leq B
\end{aligned}
$$

Then within certain limitations the quantities A and B can be chosen so that the probabilities of misclassification can be set at preassigned levels.

The diagram below shows the nature of the decisions that could be made after ascertaining the value of the ratio or its logarithm.



$$B' = \ln B \qquad A' = \ln A \qquad C' = \ln p_2 - \ln p_1$$

An individual in region R_1 can be assumed (at a given risk) to belong to population π_1 . The region R_2 has a similar meaning for population π_2 . Individuals falling into regions D_1 and D_2 may be provisionally assigned to π_1 or π_2 , respectively. The point B' is determined such that if an individual belongs to group 1 the probability of its Y value being equal to or less than B' is $P(2/1)$. Rao states that one can find this value of B' by setting the ratio of $(B' - \bar{Y}_1)$ to the standard error of Y equal to that ordinate of the appropriate t-distribution for which the probability of getting a smaller t value is equal to $P(2/1)$; and solving for B' . The point A' is determined such that if an element belongs to group 2 the probability of its value being equal to or greater than A' is $P(1/2)$. One can find the value of A' by setting the ratio of $(A' - \bar{Y}_2)$ to the standard error of Y equal to that ordinate of the appropriate t-distribution for which the probability of getting a larger t value is equal $P(1/2)$; and solving for A' . The standard error of Y can be obtained from the analysis of variance of (25). It should be noted that C' which equals zero when $p_1 = p_2$, is the critical value obtained when using just two regions R_1 and R_2 .

One might believe that the use of the doubtful region is the "ideal" situation for it gives a means for controlling the probabilities of misclassification. However in a practical application if $P(2/1)$ and $P(1/2)$ are

set too small then D_1 and D_2 will become too large, i.e., the doubtful region will contain too many individuals and one would be reserving judgment or not classifying too many of the individuals.

THE PROBLEM OF THREE OR MORE GROUPS

In the previous sections it was seen that if measurements on a certain number of variates are available for two groups, it is possible to construct a discriminant function which gives the maximum discrimination between them. This function is useful in assigning, with a certain degree of confidence, an individual to one or the other of the two groups to which it is known to belong. Let us now consider the problem of assigning an individual to one of k groups from which it is known to have come. Let $\pi_1, \ldots, \pi_k$ be k populations with density functions $f_1(v; \theta_1), \ldots, f_k(v; \theta_k)$ respectively. If the p-dimensional space is divided into k regions $R_1, \ldots, R_k$ such that if an observation falls into R_j it shall be classified as belonging to π_j, then the probability of misclassifying an observation from the ith population as coming from the jth population is

$$P(j/i) = \int_{R_j} f_i(v; \theta_i)\, dv .$$

If a priori probabilities $p_1, \ldots, p_k$ of an individual coming from $\pi_1, \ldots, \pi_k$ respectively exist, and the cost of misclassifying an observation coming from π_i as coming from π_j is $C(j/i)$, then the expected loss due to misclassification is:

$$\sum_{i=1}^{k} p_i \sum_{\substack{j=1 \\ j \neq i}}^{k} P(j/i)\, C(j/i) = \sum_{i=1}^{k} p_i \sum_{\substack{j=1 \\ j \neq i}}^{k} \int_{R_j} C(j/i)\, f_i(v; \theta_i)\, dv . \qquad (27)$$

The problem, then is that of choosing the regions $R_1, \ldots, R_k$ so that (27) is a minimum. Since a priori probabilities exist it is possible to define the conditional probability of an individual coming from a population, given the observed values of the individual's p variates. The conditional probability that an observation comes from π_i is

$$\Pr(\pi_i/V) = \frac{p_i f_i(v; \theta_i)}{\sum_{m=1}^{k} p_m f_m(v; \theta_m)} .$$

If one classifies an individual selected at random as belonging to π_j the expected loss is

$$\sum_{\substack{i=1 \\ i \neq j}}^{k} \frac{p_i f_i(v; \theta_i)}{\sum_{m=1}^{k} p_m f_m(v; \theta_m)} C(j/i) . \tag{28}$$

In order to minimize the expected loss one chooses that j for which equation (28), or equivalently for which

$$\sum_{\substack{i=1 \\ i \neq j}}^{k} p_i f_i(v; \theta_i) C(j/i) , \tag{29}$$

is a minimum.

These statements are summarized by a theorem due to T. W. Anderson (1958). "If p_i is the a priori probability of drawing an observation from population π_i with density $f_i(v; \theta_i)$, $(i = 1, \ldots, k)$, and if the cost of misclassifying an observation from π_i as from π_j is $C(j/i)$ then the regions of classification, $R_1, \ldots, R_k$, that minimize the expected cost are defined by assigning V to R_m if

$$\sum_{\substack{i=1 \\ i \neq m}}^{k} p_i f_i(v; \theta_i) C(m/i) < \sum_{\substack{i=1 \\ i \neq j}}^{k} p_i f_i(v; \theta_i) C(j/i) \tag{30}$$

$(j = 1, \ldots, k \quad j \neq m)$."

If the cost of misclassification is equal for all groups, that is $C(i/j) =$ constant for all i and j, (30) reduces to

$$R_m : \quad \frac{f_m(v; \theta_m)}{f_j(v; \theta_j)} > \frac{p_j}{p_m} \quad (j = 1, \ldots, k, j \neq m) . \tag{31}$$

In this case the observation V' is in R_m if m is the index for which $p_m f_m(v; \theta_m)$ is a maximum; that is, π_m is the most probable population.

The proof of Anderson's Theorem will be included at this point:

Note that the expected loss due to misclassification (27) can be written as

$$\sum_{j=1}^{k} \int_{R_j} \sum_{\substack{i=1 \\ i \neq j}}^{k} p_i \, C(j/i) \, f_i(v; \theta_i) \, dv . \tag{32}$$

By letting

$$h_j(v) = \sum_{\substack{i=1 \\ i \neq j}}^{k} p_i \, C(j/i) \, f_i(v; \theta_i)$$

one has

$$\sum_{j=1}^{k} \int_{R_j} h_j(v) \, dv = \int_{R} h(v) \, dv$$

as the expected loss due to misclassification, where $h(v) = h_j(v)$ for V' in R_j.

For the procedure described in the theorem, $h(v)$ is $h^*(v) = \min h_j(v)$, $j = 1, \ldots, k$. The difference between the expected loss for any procedure R and the procedure R^* is

$$\int_{R} [h(v) - h^*(v)] \, dv = \sum_{j=1}^{k} \int_{R_j} [h_j(v) - \min h_j(v)] \, dv . \tag{33}$$

Equation (33) is seen to be equal to or greater than zero. Therefore the expected loss incurred by using any other procedure must be equal to or greater than the expected loss incurred when using the Anderson theorem.

For further consideration of the k-group case, let us consider three populations—the generalization to k is immediate. Furthermore let us assume that the costs of misclassification are the same for all three groups. From (31) the regions for classification are:

$$\begin{aligned} R_1 &: \quad p_1 f_1(v;\theta_1) \geq p_2 f_2(v;\theta_2) \ , \ p_1 f_1(v;\theta_1) \geq p_3 f_3(v;\theta_3) \\ R_2 &: \quad p_2 f_2(v;\theta_2) \geq p_1 f_1(v;\theta_1) \ , \ p_2 f_2(v;\theta_2) \geq p_3 f_3(v;\theta_3) \\ R_3 &: \quad p_3 f_3(v;\theta_3) \geq p_1 f_1(v;\theta_1) \ , \ p_3 f_3(v;\theta_3) \geq p_2 f_2(v;\theta_2) \ . \end{aligned} \tag{34}$$

If the a priori probabilities are all equal and the regions are defined in terms of their logarithms they become

$$\begin{aligned} R_1 &: \quad u_{12} \geq 0 \ , \quad u_{13} \geq 0 \\ R_2 &: \quad u_{21} \geq 0 \ , \quad u_{23} \geq 0 \\ R_3 &: \quad u_{31} \geq 0 \ , \quad u_{32} \geq 0 \end{aligned} \tag{35}$$

where

$$u_{ij} = \ln \frac{f_i(v;\theta_i)}{f_j(v;\theta_j)} = \left[V' - \tfrac{1}{2}(\mu^{(i)} + \mu^{(j)})' \right] \Sigma^{-1}(\mu^{(i)} - \mu^{(j)}) \ .$$

These regions may also be used for classifying an observation when nothing is known about p_1 , p_2 , p_3 , the a priori probabilities. For computational purposes it is advantageous to express the regions R_1 , R_2 , R_3 in still another form. By referring to an earlier section of this report one sees that the statement $u_{ij} \geq 0$ is equivalent to the statement, $Y_{ij} \geq \frac{1}{2}(\bar{Y}_i + \bar{Y}_j)$ where $Y_{ij} = x_1 z_{ij1} + \ldots + x_p z_{ijp}$. That is Y_{ij} is a linear combination of the p measurements taken from an individual. Using matrix notation $Y_{ij} = V'(Z)_{ij}$ where the $(Z)_{ij}$ column vector is the solution to the system of equations $(\Sigma)(Z)_{ij} = (u^{(i)} - u^{(j)})$ and Σ is the common covariance

matrix, also $\bar{Y}_i = (\mu^{(i)})' (Z)_{ij}$ and $\bar{Y}_j = (\mu^{(j)})' (Z)_{ij}$. Using the above as well as the fact that $\mu_{ij} = -\mu_{ji}$ the regions can be expressed as

$$\begin{aligned} R_1 &: \quad Y_{12} \geq \tfrac{1}{2}(\bar{Y}_1 + \bar{Y}_2) \qquad Y_{13} \geq \tfrac{1}{2}(\bar{Y}_1 + \bar{Y}_3) \\ R_2 &: \quad Y_{12} < \tfrac{1}{2}(\bar{Y}_1 + \bar{Y}_2) \qquad Y_{23} \geq \tfrac{1}{2}(\bar{Y}_2 + \bar{Y}_3) \\ R_3 &: \quad Y_{13} < \tfrac{1}{2}(\bar{Y}_1 + \bar{Y}_3) \qquad Y_{23} < \tfrac{1}{2}(\bar{Y}_2 + \bar{Y}_3) \end{aligned} \tag{36}$$

It is most important to note that $\bar{Y}_i$ is the mean of the ith group using the discriminant function Y_{ij}, and $\bar{Y}_j$ is the mean of the jth group using the discriminant function Y_{ij}. That is, in region R_2 of (36) $\bar{Y}_2$ in the first statement is not equal to $\bar{Y}_2$ in the second statement. For $\bar{Y}_2$ in the first statement is equal to $(\mu^{(2)})' (Z)_{12}$ and $\bar{Y}_2$ in the second statement is equal to $(\mu^{(2)})' (Z)_{23}$. Thus when one speaks of $\bar{Y}_i$ and $\bar{Y}_j$ it must be with reference to a particular discriminant function.

A detailed investigation into the above procedure reveals that nothing more has been done than compute a simple (two-group) discriminant function for each possible combination of groups. That is, given an element at random we use the simple discriminant function Y_{12} to distinguish between π_1 and π_2 and the simple discriminant function Y_{13} to distinguish between π_1 and π_3. Similarly, by considering the other possible simple discriminant functions one can determine the regions R_2 and R_3.

In the k-group classification problem, as in the 2-group problem, the regions $R_1, \ldots, R_k$ were chosen so that the probability of misclassification would be a minimum. By using this method the investigator has no control over the error rate. So before one asserts that any given individual is correctly classified he would like to know what the error rate is. When

$k = 3$ the total probability of misclassification is the sum of $p_1 [P(2/1) + P(3/1)]$, $p_2 [P(1/2) + P(3/2)]$ and $p_3[P(1/3) + P(2/3)]$. To determine the probabilities of misclassification one needs to find the variances and covariances of the three discriminant functions. It can be shown that these values are readily obtainable from the mean values of the functions, that is,

$$\begin{aligned}
&\text{var } (Y_{12}) = \bar{Y}_1 - \bar{Y}_2 \\
&\text{var } (Y_{23}) = \bar{Y}_2 - \bar{Y}_3 \\
&\text{var } (Y_{13}) = \bar{Y}_1 - \bar{Y}_3 \\
&\text{cov } (Y_{12}, Y_{23}) = \bar{Y}_2 - \bar{Y}_3 \\
&\text{cov } (Y_{12}, Y_{13}) = \text{var } (Y_{12}) + \text{cov } (Y_{12}, Y_{23}) \\
&\text{cov } (Y_{23}, Y_{13}) = \text{var } (Y_{23}) + \text{cov } (Y_{12}, Y_{23})
\end{aligned} \tag{37}$$

where in the var (Y_{12}) , $\bar{Y}_1$ and $\bar{Y}_2$ are the means for groups one and two, respectively, using discriminant function Y_{12} . Similarly in the var (Y_{23}) , $\bar{Y}_2$ and $\bar{Y}_3$ are the means for groups two and three, respectively, using discriminant function Y_{23} . In the var (Y_{13}) , $\bar{Y}_1$ and $\bar{Y}_3$ are the means for groups one and three respectively, using discriminant function Y_{13} . In computing the covariance (Y_{12}, Y_{23}) , $\bar{Y}_2$ and $\bar{Y}_3$ are the means for groups two and three, respectively, using discriminant function Y_{12} . Using these variances and covariances one can obtain the correlations between the discriminant functions. Then using the variances and correlations one can determine the probability of misclassifying an observation given the population to which it belongs. The probability of correct classification for group one is:

$$\text{Pr}\left[Y_{12} \geq \tfrac{1}{2}(\bar{Y}_1 + \bar{Y}_2) \;\; ; \;\; Y_{13} \geq \tfrac{1}{2}(\bar{Y}_1 + \bar{Y}_3)\right].$$

If the p-variates are normally distributed and hence the Y_{ij} are normally distributed, one can use the bivariate normal distribution to calculate the above probability. This technique is illustrated by use of a numerical example in the appendix.

Once the probability of misclassification is obtained it may be so large that one could not have much confidence in a classification statement. This problem is theoretically resolved by the existence of a "doubtful region." A region such that if an observation lies in it judgment is withheld; that is, no classification is made. Rao (1952) using an extension of the Neyman-Pearson Fundamental Lemma has proved that there exist regions R_1 , R_2 , R_3 and a set of doubtful regions such that the probability of misclassification can be set at a predetermined level. The approach to this problem is similar to that used in the two-group problem where the probability of misclassification is set and then regions R_1 , R_2 , and R_D are determined. However, when there are more than two groups, the complexity of finding these regions for a particular problem makes their use prohibitive.

It has been assumed that the distributions of the three populations were completely known. When the parameters are unknown and must be estimated from samples one can substitute the maximum likelihood estimates for the unknown parameters. To determine the regions of classification one then treats these maximum likelihood estimates as if they were the parameters of the distribution.

UNEQUAL COVARIANCE MATRICES

Now let us consider the case in which the two multivariate normal populations with unequal mean vectors, also have unequal covariance matrices. Let

Σ_i represent the covariance matrix for the ith population $i = 1, 2$. Then the region R_1 of (2) is written as:

$$R_1 : \frac{\frac{\left|\Sigma_1^{-1}\right|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2} Q_1(x_1, \ldots, x_p)}}{\frac{\left|\Sigma_2^{-1}\right|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2} Q_2(x_1, \ldots, x_p)}} > \frac{p_2}{p_1} \qquad (38)$$

$$R_2 : \quad \text{otherwise}$$

where $Q_k(x_1, \ldots, x_p) = \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij} (x_i - \mu_{ki})(x_j - \mu_{kj})$.

Writing the regions R_1 and R_2 in terms of their logarithms one has

$$\begin{aligned} R_1 : \quad & Q_1(x_1, \ldots, x_p) \leq Q_2(x_1, \ldots, x_p) + \ln \frac{|\Sigma_1|}{|\Sigma_2|} + 2 \ln \frac{p_1}{p_2} \\ R_2 : \quad & Q_1(x_1, \ldots, x_p) > Q_2(x_1, \ldots, x_p) + \ln \frac{|\Sigma_1|}{|\Sigma_2|} + 2 \ln \frac{p_1}{p_2} \end{aligned} \qquad (39)$$

For the k-group problem following this procedure $R_1, \ldots, R_k$ of (30) become

$$R_j : \quad Q_j(x_1, \ldots, x_p) \leq Q_i(x_1, \ldots, x_p) - \log \frac{|\Sigma_j|}{|\Sigma_i|} + 2 \ln \frac{p_j}{p_i} \qquad (40)$$

$$(i, j = 1, 2, \ldots, k \quad j \neq i) .$$

Cooley and Lohnes (1962) were concerned with the general problem of discrimination, with emphasis on the problem of comparing the profile of an individual with that of a group. Their interest was that of being able to tell a prospective student for a given curriculum how favorably he compared with successful students in that field. For present purposes consider taking only two measurements from each individual, so that the group to which the individual is being compared can be considered to be a bivariate normal

population. One way to describe such a bivariate distribution is in terms of ellipses, each of which is the locus of points of a specified density. For the bivariate normal distribution, the size of the ellipse is determined by the value of the quadratic:

$$Q(x_1, x_2) = \sum_{i=1}^{2} \sum_{j=1}^{2} \sigma^{ij}(x_i - \mu_i)(x_j - \mu_j) . \tag{41}$$

Each individual is represented as a point in the sample space, and each point can be located on a particular ellipse by substituting the individuals observed values into equation (41).

If the individual is selected at random then $Q(x_1, x_2)$ is distributed as χ^2, with two degrees of freedom. Since the tabled probability of a given χ^2 is the likelihood of obtaining a larger value, it is also the proportion of sample points that would be expected to lie beyond the ellipse on which (x_1, x_2) lies. The ellipse used in this manner is called a centour, and it is a good index of the extent to which an individual resembles a particular group. The generalization of the centour method to the measurement of p variables on each individual is obvious. (41) becomes

$$Q_k(x_1, \ldots, x_p) = \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij}(x_i - \mu_{ki})(x_j - \mu_{kj}) \tag{42}$$

and is disbributed as χ^2 with p degrees of freedom.

Cooley and Lohnes suggest using the centour method for classifying individuals into one of k groups. The classification rule is to assign an individual to that group for which its centour is highest or in other words, its χ^2 is smallest. Thus for the two group case:

$$R_1 : \quad Q_1(x_1, \ldots, x_p) \leq Q_2(x_1, \ldots, x_p),$$

and (43)

$$R_2 : \quad Q_1(x_1, \ldots, x_p) > Q_2(x_1, \ldots, x_p) .$$

If there are k groups the regions are:

$$R_j : \quad Q_j(x_1, \ldots, x_p) \leq Q_i(x_1, \ldots, x_p)$$

$$(i, j = 1, \ldots, k \quad i \neq j)$$

These regions are seen to be special cases of regions (39) and (40).

When the covariance matrices are equal, an individual who lies in R_j in the sample space of this section will also lie in R_j in the discriminant space. Therefore under these conditions it is advantageous to use the discriminant space or discriminant function. However, when the covariance matrices are unequal the regions of the discriminant functions have not been given in a convenient form. Therefore when the covariance matrices for the k-groups are not equal one can use the sample space of this section, i.e., regions (39) and (40) for the classification of an individual.

APPENDIX

To illustrate the use of a discriminant function for classifying an individual into one of two groups, a numerical example has been taken from the book, Statistical Analysis in Biology by K. Mather. Mather was faced with the problem of classifying flies into one of two races. To do this he took a sample of eleven flies from each race and measured two traits on each fly. The observed values for both traits of the flies as well as the value of the discriminant function for each fly are given in table 1.

Table 1.

Race π_1			Race π_2		
Trait No. 1	Trait No. 2	Y_{1i}	Trait No. 1	Trait No. 2	Y_{2i}
6.36	5.25	2.546	6.00	4.88	2.394
5.92	5.12	*2.402	5.60	4.64	2.245
5.92	5.36	2.434	5.64	4.96	2.299
6.44	5.64	2.623	5.76	4.80	2.313
6.40	5.16	2.547	5.96	5.08	2.409
6.56	5.56	2.647	5.72	5.04	2.333
6.64	5.36	6.644	5.64	4.96	2.299
6.68	4.96	2.602	5.44	4.88	2.231
6.72	5.48	2.682	5.04	4.44	2.056
6.76	5.60	2.710	4.56	4.04	1.863
6.72	5.08	2.629	5.48	4.20	2.152

The discriminant function will be of the form $Y_I = z_1x_1 + z_2x_2$ where z_1 and z_2 are the solutions to the following equations:

$$2.628364z_1 + 1.277382z_2 = .934545$$

$$1.277382z_1 + 1.748655z_2 = .603636 .$$

Solving for z_1 and z_2 one has

$$Y_I = .291174x_1 + .132507x_2 .$$

Now

$\bar{Y}_1 = \bar{x}_{11}z_1 + \bar{x}_{12}z_2 = 2.5880$ and $\bar{Y}_2 = \bar{x}_{21}z_1 + \bar{x}_{22}z_2 = 2.2359$,

for $\bar{x}_{11} = 6.465454$, $\bar{x}_{12} = 5.323636$, $\bar{x}_{21} = 5.530909$ and $\bar{x}_{22} = 4.720000$.

The quantity $\frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)$ is equal to 2.41195 , therefore, the regions of classification are:

$$R_1 : \quad Y_I \geq 2.41195 \qquad R_2 : \quad Y_I < 2.41195 \tag{45}$$

Consider an individual whose traits are observed to be $x_1 = 6.12$ and $x_2 = 5.05$. For this individual $Y_I = 2.451145$, hence the individual is classified as belonging to race π_1 .

Next $d_1 = \bar{x}_{11} - \bar{x}_{21} = .934545$, $d_2 = \bar{x}_{12} - \bar{x}_{22} = .603636$, and

$B = z_1d_1 + z_2d_2 = .352101$.

Therefore the analysis of variance of Y may now be written as follows:

Source of Variation	d.f.	Ss	Ms	Variance Ratio
Between Races	2	.681863	.340932	18.397
Within Races	19	.352101	.018532	
Total	21	1.033964		

By consulting the tables of the F-distribution, it is seen that the probability of a variance ratio with 2 and 19 degrees of freedom exceeding 8.18 is .01 . Hence, the discriminant function is highly significant. Which indicates that if one were to apply the discriminant function to each member of the two groups, and then perform an analysis of variance to test the hypothesis that the two transformed groups have equal means he would find a significant difference between the group means.

Assuming that the two races are equally likely to occur, misclassification

of an individual will occur when its departure from the racial mean is greater than one half the difference between racial means, namely .17605 , provided that departure occurs in the right direction. A deviation of .17605 is 1.293 times the standard deviation of Y , as estimated by 19 degrees of freedom. Now $|t_{19}|$ exceeds 1.293 by chance about 20 per cent of the time. Since misclassification occurs in one direction only, the probability of misclassification using this discriminant function is .10 .

If the investigator wanted the probability of misclassification to be equal to .05 , then the regions R_1 , R_2 and R_D of (26) are determined. To find the region R_2 such that the probability of an individual from race one falling into R_2 is equal to 5 per cent one solves the equation $(B' - \bar{Y}_1) / .13613 = -2.093$ for B' . Similarly for R_1 one solves $(A' - \bar{Y}_2) / .13613 = +2.093$ for A' . Since $\bar{Y}_1 = 2.5880$ and $\bar{Y}_2 = 2.2359$ $B' = 2.3031$ and $A' = 2.5208$. Thus if the regions

R_1 : $Y_I \geq 2.5208$

R_D : $2.3031 < Y_I < 2.5208$

R_2 : $2.3031 \leq Y_I$

are used the probability of misclassifying an individual selected at random is .05 . Classifying all the individuals in the two samples by use of the discriminant function (45) it is seen that only one error would be made. The second individual in the sample from race π_1 , which is indicated with an asterick in Table 1, would be misclassified as coming from race π_2 .

To illustrate the 3-group classification problem let us examine an example from Rao (1952). Table 2 gives the mean values of each characteristic for the three groups.

Table 2.

Group	Mean Values of Characteristic Measured			
	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
A	164.51	86.43	25.49	51.24
B	160.53	81.47	23.84	48.62
C	158.17	81.16	21.44	36.72

The sample estimate of the covariance matrix is:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} 32.45 & 7.43 & 1.78 & 3.97 \\ ..... & 10.24 & 1.17 & 2.43 \\ ..... & ..... & 3.06 & 1.78 \\ ..... & ..... & .... & 12.25 \end{bmatrix} \end{array} = S$$

and the inverse of this covariance matrix is:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} .0371 & -.0245 & -.0088 & -.0059 \\ .0245 & .1212 & -.0248 & -.0125 \\ -.0088 & -.0248 & .3680 & -.0457 \\ -.0059 & -.0125 & -.0457 & .0927 \end{bmatrix} \end{array} = S^{-1}$$

$$Y_{ij} = V'(Z)_{ij} \quad \text{where} \quad (Z)_{ij} = (S^{-1})(\bar{X}^{(i)} - \bar{X}^{(j)}) ,$$

so that one wants to determine next the vectors of differences between group means. They are as follows:

$$(\bar{X}^{(1)} - \bar{X}^{(2)}) = \begin{bmatrix} 3.98 \\ 4.96 \\ 1.65 \\ 2.62 \end{bmatrix} \quad (\bar{X}^{(1)} - \bar{X}^{(3)}) = \begin{bmatrix} 6.34 \\ 5.27 \\ 4.05 \\ 4.52 \end{bmatrix} \quad (\bar{X}^{(2)} - \bar{X}^{(3)}) = \begin{bmatrix} 2.36 \\ 0.31 \\ 2.40 \\ 1.90 \end{bmatrix}$$

Now one can compute the $(Z)_{ij}$ column vectors which are:

$$(Z)_{12} = \begin{bmatrix} -.0039 \\ .4301 \\ .3293 \\ .0819 \end{bmatrix} \qquad (Z)_{13} = \begin{bmatrix} .0437 \\ .3265 \\ 1.0972 \\ .1305 \end{bmatrix} \qquad (Z)_{23} = \begin{bmatrix} .0476 \\ -.1036 \\ .7679 \\ .0486 \end{bmatrix}$$

The discriminant functions are:

$$Y_{12} = V'(Z)_{12} = -.0039\ x_1 + .4301\ x_2 + .3293\ x_3 + .0819\ x_4$$

$$Y_{13} = V'(Z)_{13} = .0437\ x_1 + .3265\ x_2 + 1.0972\ x_3 + .1305\ x_4$$

$$Y_{23} = V'(Z)_{23} = .0476\ x_i + .1036\ x_2 + .7679\ x_3 + .0486\ x_4\ .$$

To find the mean values of the discriminant function for the three groups one evaluates:

$$Y_{ij} = (\bar{X}^{(i)})(Z)_{ij} \quad \text{and} \quad Y_{ij} = (\bar{X}^{(j)})(Z)_{ij}\ ,$$

for all i and j.

Mean Values of the Discriminant Functions

Group	Discriminant Function Y_{12}	Y_{13}	Y_{23}
A	$\bar{Y}_1 = 49.1224$	$\bar{Y}_1 = 70.0630$	$\bar{Y}_1 = 20.9406$
B	$\bar{Y}_2 = 46.2467$	$\bar{Y}_2 = 66.1173$	$\bar{Y}_2 = 19.8706$
C	$\bar{Y}_3 = 45.1766$	$\bar{Y}_3 = 63.0317$	$\bar{Y}_3 = 17.8551$

Thus the regions of classification of (36) are:

R_1 : $Y_{12} \geq 47.6845$ $\quad$ $Y_{13} \geq 66.5473$

R_2 : $Y_{12} < 47.6845$ $\quad$ $Y_{23} \geq 18.8628$

R_3 : $Y_{13} < 66.5473$ $\quad$ $Y_{23} < 18.8628$

Consider an individual whose traits were observed to be

$x_1 = 162.00$ $\quad$ $x_2 = 84.00$ $\quad$ $x_3 = 24.00$ $\quad$ $x_4 = 49.00$

For this individual $Y_{12} = 47.4129$, $Y_{13} = 67.2327$, and $Y_{23} = 19.8198$. Since $47.4129 < 47.6845$ and $19.1898 \geq 18.8628$ the individual is assigned to group B .

To determine the probabilities of misclassification one needs the variances and covariances of Y_{12} , Y_{13} , and Y_{23} . Referring to (37) it follows that:

var $Y_{12} = 2.8757$ $\quad$ cov $(Y_{12}, Y_{23}) = 1.0701$

var $Y_{13} = 7.0313$ $\quad$ cov $(Y_{12}, Y_{13}) = 3.9458$

var $Y_{23} = 2.0155$ $\quad$ cov $(Y_{23}, Y_{13}) = 3.0856$

Thus the correlation matrix of Y_{12} , Y_{13} , and Y_{23} is:

	Y_{12}	Y_{13}	Y_{23}
Y_{12}	1.0000	.8810	.4459
Y_{13}	.....	1.0000	.8200
Y_{23}	.....	.....	1.0000

The probability of correctly classifying an individual from group A is:

$$P(Y_{12} \geq 47.6845 \ ; \ Y_{13} \geq 66.5473) ,$$

which gives the standard bivariate normal deviates:

$$h = \frac{47.6845 - 49.1224}{1.69} = -.85 \quad \text{and} \quad k = \frac{66.5473 - 70.0630}{2.65} = -1.33 .$$

Therefore, the probability of misclassifying an individual from group A is:

$$\Pr(h > .85) + \Pr(k > 1.33) - \Pr(h > .85,\ k > 1.33;\ r = .88) = .198$$
$$+ .092 - .085 = .205 .$$

The first two probabilities are obtained from the univariate normal tables while the third is taken from Pearson's tables for the bivariate normal distribution.

Similarly for group B the deviates are $h = .85$ and $k = -.71$, $r = .45$; so that the probability of misclassifying an individual from group B is:

$$\Pr(h > .85) + \Pr(k > .71) - \Pr(h > .85,\ k > .71;\ r = -.45) = .198$$
$$+ .239 - .013 = .424 .$$

For group C the deviates are $h = .71$ and $k = 1.33$; $r = .82$; so that the probability of misclassifying an individual from group C is:

$$\Pr(h > .71) + \Pr(k > 1.33) - \Pr(h > .71,\ k > 1.33;\ r = .82) = .239$$
$$+ .092 - .085 = .246 .$$

Assuming that $p_1 = p_2 = p_3$, the probability of misclassifying an individual taken at random from one of the three groups using these discriminant functions is: $1/3(.205) + 1/3(.424) + 1/3(.246) = .29$.

ACKNOWLEDGMENT

The writer wishes to express his appreciation to Dr. H. C. Fryer of the Department of Statistics, Kansas State University, for his helpful suggestions and advice during the preparation of this report.

REFERENCES

Anderson, T. W., Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York. (1958)

Brown, G. W., "Discriminant functions." Annals of Mathematical Statistics, 18: 514-528. (1947)

Cooley-Lohnes, Multivariate Procedures For the Behavorial Sciences. John Wiley and Sons, Inc., New York. (1962)

Fisher, R. A., Contributions to Mathematical Statistics. John Wiley and Sons, Inc., London. (1950)

Goulden, C. H., Methods of Statistical Analysis. John Wiley and Sons, Inc., London. (1950)

Hodges, J. L., Jr., "Discriminatory analysis." Project Number 21-49-004, Report No. 1, School of Aviation Medicine, U.S.A.F., Randolph AFB, Texas. (1955)

Kendall, M. G., A Course in Multivariate Analysis. Hafner Publishing Company, New York. (1957)

Mather, K., Statistical Analysis in Biology. Interscience Publishers, Inc., New York. (1947)

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

Smith, C. A. B., "Some examples of discrimination." Annals of Eugenics, 13: 272-282. (1947)

von Mises, R., "On the classification of observational data into distinct groups." Annals of Math. Stat., 16: 68-73. (1945)

Wald, A., "On a statistical problem arising in the classification of an individual into one of two groups." Annals of Math. Stat., 15: 145-162. (1944)

Welch, B. L., "Note on discriminant functions." Biometrika, 31: 218. (1939)

Wilks, S. W., Mathematical Statistics. John Wiley and Sons, Inc., New York. (1962)

DISCRIMINANT FUNCTIONS

by

LEROY JOSEPH YORK

B. S., Kansas State University, 1961

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

This report discusses the problem of classifying a single element into one of k populations from which it is known to have come. The basis for classification is whatever evidence is available about the element I and the k populations. The first case considered is that of classifying an element into one of two populations; the k population case follows.

The technique used is to divide the sample space into two regions, R_1 and R_2, such that if an observation belongs to R_1 it is classified as coming from population one; and if the observation belongs to R_2 it is classified as coming from population two.

If the probability distributions of the two populations are completely known and a priori probabilities of belonging to population one and two, respectively, exist, then the method of Bayes may be used to determine the regions of classification. When no a priori probabilities exist the minimax solution is obtained.

If the form of the probability distributions is known, but only maximum likelihood estimates of the parameters are available, one treats these estimates as the unknown parameters. These estimates are obtained from two samples, one from each population.

Fisher approached the classification problem by considering the linear function of the p measurements taken from the item that would maximize the ratio of the difference between sample means to the standard error within samples. Under certain conditions the two procedures result in the same regions of classification.

The question of significance of a discriminant function is also considered. This is discussed in terms of whether or not the parent populations are identical, and hence whether or not a discriminant function is illusory. Either Hotelling's T^2 statistic or the analysis of variance suggested by

Fisher may be used for this problem. The analysis of variance is also useful in determining the probability of misclassification. The probability of misclassification can be arbitrarily set by the investigator with the introduction of a "doubtful region." That is, a region for which judgment is withheld.

The problem of three or more groups is approached by the use of a set of discriminant functions. That is, a set of discriminant functions is obtained for the determination of classification between all possible pairs of groups.

When the covariance matrix of the p characteristics measured, is not the same for both populations; the discriminant functions are no longer linear functions. This problem of unequal covariance matrices is considered briefly.