A SURVEY OF DISCRIMINATION TECHNIQUES

by 4589

RONALD L. IMAN

B. S., Kansas State University, 1962

M. A., Kansas State Teachers College, 1965

A MASTER'S REPORT

Submitted in Partial Fulfillment of the

Requirements for the Degree

MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND COMPUTER SCIENCE

KANSAS STATE UNIVERSITY

MANHATTAN, KANSAS

1970

Approved by:

Major Professor

ACKNOWLEDGEMENT

The writer wishes to express his sincere appreciation to his major professor, Dr. Ray A. Waller, for his helpful suggestions in directing the writing of this report.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION AND BACKGROUND

This report is concerned with the problems of differentiating between two or more populations on the basis of measurements of one or more variables obtained from a sample of each of the populations. The procedure for accomplishing this differentiation is defined as follows:

> Discrimination. Two or more populations are known to exist and a sample is taken from each. The problem then is to set up a rule, based on the sample values, such that some future observation(s) can be allocated to the correct population(s), when it is not known from which population it emanates. It is desirable to do this identification with the minimum amount of misclassification.

The history of discriminant analysis essentially has its beginning in 1919 when Pearson tackled the problem of determining the "Coefficient of Racial Likeness." So extensive has the work been since that time that by 1950, in a survey of discriminatory analysis and related topics, Hodges gives an up-to-date bibliography containing over 250 listings. Hence, this report will not attempt to update Hodges' bibliography, but rather acquaint the reader with various techniques of solving the problem of differentiation between two or more populations.

CHAPTER II

DISTRIBUTION FREE TECHNIQUES

Very little work has actually been done on distribution free techniques, although in retrospect, in its most basic form it has to be the oldest discrimination technique. Kendall and Stuart (1968) give an entirely distribution free technique for the classical Iris data used by Fisher (1936). Fisher's data is given in Table I and it contains the measurements in centimeters of four variables: (1) sepal length, (2) sepal width, (3) petal length and (4) petal width. These measurements were made on fifty flowers from each of three varieties of Iris: (1) setosa, (2) versicolor and (3) virginica. Examination of Table I shows that the petal width of setosa has a mean of .246, range of .2 - .6 and variance of .0109, while versicolor petal width has a mean of 1.326, range of 1.0 - 1.8 and variance of .0383. Hence, for these two populations, petal width is an absolute discriminant, i.e. correct population identification could be made on the basis of petal width alone with a mistake rarely being made.

The problem is not quite as simple when considering versicolor and virginica since no absolute discriminant exists for these populations. Petal width has an overlap in the range 1.4 - 1.8 with 62 cases falling outside the common range. Petal length has 63 cases falling outside of the common range. Thus, petal length is the "best first discriminator," i.e. "best first filter." The frequency distribution for petal length and petal width is given in Table II. Examination of Table II shows the common range for petal length to be 4.5 - 5.1.

This allows for the following set of rules of discrimination on the first pass:

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7	2·8	6·7	2·0
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3	2·7	4·9	1·8
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7	3·3	5·7	2·1
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2	6·0	1·8
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8	4·8	1·8
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0	4·9	1·8
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8	5·6	2·1
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

TABLE I

Multiple Measurements on Three Species of Iris

Variate Values	Petal Length Vers.	Virg.	Variate Values	Petal Width Vers.	Virg.
4.3	25				
4.4	4		1.0	7	
4.5	7	1	1.1	3	
4.6	3	-	1.2	5	
4.7	5	-	1.3	13	
4.8	2	2	1.4	7	1
4.9	2	3	1.5	10	2
5.0	1	3	1.6	3	1
5.1	1	7	1.7	1	1
5.2		2	1.8	1	11
5.3		2	1.9		5
5.4		2	2.0		6
5.5		3	2.1		6
5.6		6	2.2		3
5.7		3	2.3		8
5.8		3	2.4		3
5.9		13	2.5		3
	50	50		50	50

TABLE II

Frequency Distributions of Petal Length and Petal Width
for Iris Versicolor and Iris Virginica

Variate Values	Petal Width	
	Vers.	Virg.
1.2	1	
1.3	2	
1.4	4	
1.5	9	2
1.6	3	-
1.7	1	1
1.8	1	5
1.9		3
2.0		3
2.1		-
2.2		-
2.3		1
2.4		1
	21	16

TABLE III

Frequency Distribution of 37 Cases Not Distinguished by Petal Length

petal length $\leq$ 4.4 allot to versicolor,

petal length $\geq$ 5.2 allot to virginica,

4.5 $\leq$ petal length $\leq$ 5.1 refer to next variable.

This set of rules allows for correct identification of 63 of the 100 flowers. The frequency distribution for the petal width of the remaining 37 flowers is given in Table III.

The common range for petal width is now 1.5 - 1.8. The second set of rules (second filter) is:

petal width $\leq$ 1.4 allot to versicolor,

petal width $\geq$ 1.9 allot to virginica,

1.5 $\leq$ petal width $\leq$ 1.8 refer to next variable.

This second set of rules identifies 15 of the remaining 37 cases, while leaving 22 cases undecided. The frequency distributions for sepal length and sepal width are given in Table IV.

Table IV indicates that sepal width would be the next discriminator (filter) with the following set of rules:

sepal width $\geq$ 3.1 allot to versicolor,

sepal width < 3.1 refer to next variable.

This variable identifies 6 more flowers leaving 16 unidentified. Table V gives the frequency distribution for sepal length on the remaining 16 cases.

Examination of Table V allows for the fourth and final set of rules:

sepal length $\geq$ 6.4 allot to versicolor,

sepal length $\leq$ 5.3 allot to virginica,

5.4 $\leq$ sepal length $\leq$ 6.3 undecided.

This last set of rules identifies only 3 of the remaining 16 flowers, or the four sets of rules together have identified 87 of the 100 flowers.

Variate Values	Sepal Length Vers.	Virg.	Variate Values	Sepal Width Vers.	Virg.
4.9		1	2.2	1	1
-	-	-	2.3	-	-
5.4	1	-	2.4	-	-
5.5	-	-	2.5	1	1
5.6	1	-	2.6	-	-
5.7	-	-	2.7	1	1
5.8	-	-	2.8	1	2
5.9	1	1	2.9	1	-
6.0	3	2	3.0	3	3
6.1	-	1	3.1	2	
6.2	1	1	3.2	2	
6.3	2	2	3.3	1	
6.4	1	-	3.4	1	
6.5	1	-			
6.6	-	-			
6.7	2	-			
6.8	-	-			
6.9	1	-			
	14	8		14	8

TABLE IV

Frequency Distributions of 22 Cases Not Distinguished by Petal Length and Petal Width

Variate Values	Sepal Length	
	Vers.	Virg.
4.9		1
-	-	-
5.4	1	-
5.5	-	-
5.6	1	-
5.7	-	-
5.8	-	-
5.9	-	1
6.0	2	2
6.1	-	1
6.2	1	1
6.3	1	2
6.4	-	-
6.5	1	-
6.6	-	-
6.7	1	-
	8	8

TABLE V

Distribution of 16 Cases Not Distinguished by
Petal Length, Petal Width, Sepal Width

This discrimination technique while quite elementary, is none the less entirely distribution free, since it relies only on the rank order of the variate values and involves nothing more complicated than counting. It would seem that other procedures should work better than the one presented here; however, it would possibly be at the expense of sacrificing the distribution-free nature of the procedure.

CHAPTER III

FISHER'S LINEAR DISCRIMINANT FUNCTION

The first clear statement of the problem of discrimination and the first proposed solution to that problem, were given by R. A. Fisher in the middle 1930's. Fisher's work was first published by Barnard (1935) and by Martin (1936) before Fisher's own paper in the Annuals of Eugenics in 1936.

The general situation given by Fisher is as follows. We are given two populations π_1 and π_2 and a sample of observations from each population. The problem then is to set up a rule, based on the sample values, that will correctly allocate a future observation to the correct population when it is not known from which population it emanates. If only one measurement is obtained from the sample values, i.e., the univariate case, then the discrimination rule could be something as simple as a function of the sample means and variances. A separation point could then be determined such that observations falling below this point would be identified as π_1 and above this point would be π_2.

Fisher has reduced the multivariate case to the above univariate case by representing the set of k measurements, i.e. $\{x_1, x_2, \ldots, x_k\}$, as a single value Y, which is a linear combination of the sample measurements, i.e.

$$Y = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_k x_k.$$

Of course, it is desirable to choose the coefficients $\lambda_1, \lambda_2, \ldots, \lambda_k$ such that the optimal identification is obtained.

Fisher shows how to obtain these coefficients such that the ratio,

$$\frac{\text{difference of sample means}}{\text{standard error within samples}}$$

i.e.

$$\frac{\left| \bar{Y}_{\pi_1} - \bar{Y}_{\pi_2} \right|}{\sqrt{\sum_{i=1}^{k_1} (Y_{i\pi_1} - \bar{Y}_{\pi_1})^2 + \sum_{i=1}^{k_2} (Y_{i\pi_2} - \bar{Y}_{\pi_2})^2}}$$

is maximized. To further explain, it is desirable to obtain the maximum amount of separation of the sample means with the minimum amount of overlap.

In addition to Fisher's work showing how to obtain the coefficients, they are also derived by Goulden (1952), Kendall and Stuart (1968), York (1963) as well as others, so the derivation will not be repeated in this report.

Fisher's example uses the data given in Table I of Chapter II. It is unfortunate that Fisher chose the species setosa and versicolor for his application, since it was established in Chapter II that petal width is an absolute discriminant for these two species, hence no refined statistical technique is needed to distinguish between these two populations. Fisher's equation for these two populations is,

$$Y = x_1 + 5.9037\, x_2 - 7.1299\, x_3 - 10.1036\, x_4 \qquad (1)$$

where x_1 = sepal length,

x_2 = sepal width,

x_3 = petal length

and x_4 = petal width.

The mean value of Y for versicolor, obtained by substituting the variable means in equation (1), is -21.482 and for setosa is +12.335.

The midpoint of these means is -4.574, (i.e. Fisher's separation point.) Thus, for any value of Y below -4.574, allot to versicolor and above -4.574 allot to setosa. Values of exactly -4.574 would be equally likely to be in either class. Using this point as a separation criteria, the following results are obtained:

		Predicted Population	
		Setosa	Versicolor
Actual Population	Setosa	50	0
	Versicolor	0	50

When applying the linear discriminant function technique to setosa and virginica, the results are:

		Predicted Population	
		Setosa	Virginica
Actual Population	Setosa	50	0
	Virginica	0	50

The remaining pair, versicolor and virginica, are free of absolute discriminants. Therefore, a good test of the discriminating ability of the linear discriminant function is provided. The results obtained are:

		Predicted Population	
		Versicolor	Virginica
Actual Population	Versicolor	48	2
	Virginica	4	46

Fisher makes no attempt to justify on probabilistic grounds his definition of optimum separation, nor his restriction to linear combinations

of the measurements. However, it would appear that when the populations are normal (or multivariate normal) with the same variance (or same dispersion Matrix), then the linear discriminant function has certain desirable properties. Otherwise it may not be best. Chapter IV will investigate the multivariate normal case in the case where the dispersion matrices are equal and where they are unequal.

CHAPTER IV

CLASSIFICATION USING MULTIVARIATE ANALYSIS

This chapter will compare Monte Carlo results using data from a multivariate normal distribution where the dispersion (variance-covariance) matrices are equal and hence are pooled against the same procedure when the dispersion matrices are not pooled. Also, pooling versus non-pooling is compared when dispersion matrices are not equal.

POOLED DISPERSION MATRICES

Anderson (1958) presents a procedure for pooling dispersion matrices. It is given here for only two populations A and B, but is readily extendable to more than two populations.

The dispersion (variance-covariance) matrix for the first population A is given by:

$$\Sigma_1 = \frac{1}{N_1 - 1} \begin{bmatrix} \sum(X - \bar{X}_{A1})^2 & \sum(X - \bar{X}_{A1})(X - \bar{X}_{A2}) \\ \sum(X - \bar{X}_{A2})(X - \bar{X}_{A1}) & \sum(X - \bar{X}_{A2})^2 \end{bmatrix}$$

or the equivalent form

$$\Sigma_1 = \frac{1}{N_1 - 1} \begin{bmatrix} \sum x_{A1}^2 & \sum x_{A1}\, x_{A2} \\ \sum x_{A2}\, x_{A1} & \sum x_{A2}^2 \end{bmatrix}$$

and the dispersion matrix for B is

$$\Sigma_2 = \frac{1}{N_2 - 1} \begin{bmatrix} \sum x_{B1}^2 & \sum x_{B1}\, x_{B2} \\ \sum x_{B2}\, x_{B1} & \sum x_{B2}^2 \end{bmatrix}$$

where $x_{ij} = X_{ij} - \bar{X}_{ij}$, $i = A,B$; $j = 1,2$.

These matrices are assumed to be the same for all populations and, therefore, they are pooled to obtain the best estimate of the common dispersion matrix. The pooled matrix $\sum$ is obtained as follows:

$$\sum = \frac{1}{N_1 + N_2 - 2} \begin{bmatrix} \sum x_{A1}^2 + \sum x_{B1}^2 & \sum x_{A1}x_{A2} + \sum x_{B1}x_{B2} \\ \sum x_{A2}x_{A1} + \sum x_{B2}x_{B1} & \sum x_{A2}^2 + \sum x_{B2}^2 \end{bmatrix}$$

It is this point which makes the pooled dispersion technique differ from the individual dispersion technique in which the dispersion matrices are not assumed to be the same and hence are not pooled.

The i^{th} density is given by:

$$p_i(\chi) = \frac{1}{(2\pi)^{m/2}|\sum|^{1/2}} \exp\{(-1/2)[\chi - \mu^{(i)}]' \sum^{-1} [\chi - \mu^{(i)}]\}$$

where, $\mu^{(i)'} = [\mu_1^{(i)} \quad \mu_2^{(i)}]$ and $\chi' = [x_1 \quad x_2]$

i.e., $\mu^{(i)'}$ is a vector of the means of the variables of the i^{th} population and χ is a vector of observations.

The ratio of the densities of populations A and B is

$$\frac{p_1(\chi)}{p_2(\chi)} = \frac{\exp\{(-1/2)[\chi - \mu^{(1)}]' \sum^{-1} [\chi - \mu^{(1)}]\}}{\exp\{(-1/2)[\chi - \mu^{(2)}]' \sum^{-1} [\chi - \mu^{(2)}]\}}$$

$$= \exp\{(-1/2)[(\chi - \mu^{(1)})' \sum^{-1} (\chi - \mu^{(1)}) - (\chi - \mu^{(2)})' \sum^{-1} (\chi - \mu^{(2)})]$$

The region of classification into A, R_1, is the set of χ's for which the above ratio $\geq k$ (for suitable chosen k), i.e.,

$$\frac{P_1(\chi)}{P_2(\chi)} = \exp\{(-1/2)[(\chi - \mu^{(1)})'\sum^{-1}(\chi - \mu^{(1)}) - (\chi - \mu^{(2)})'\sum^{-1}(\chi - \mu^{(2)})]\} \geq k$$

Since the logarithmic function is monotonic increasing, the inequality can be rewritten as

$$(-1/2)[(\chi - \mu^{(1)})'\sum^{-1}(\chi - \mu^{(1)}) - (\chi - \mu^{(2)})'\sum^{-1}(\chi - \mu^{(2)})] \geq \log k.$$

The left side can be rewritten as

$$\chi \sum^{-1}(\mu^{(1)} - \mu^{(2)}) - (1/2)(\mu^{(1)} + \mu^{(2)})'\sum^{-1}(\mu^{(1)} - \mu^{(2)}). \qquad (1)$$

The first term of (1) is the discriminant function and is a linear function of the components of the observation vector. That is, upon expanding the first term,

$$X_1[\, x_1^2(\mu_1^{(1)} - \mu_1^{(2)}) + \sum x_2 x_1(\mu_2^{(1)} - \mu_2^{(2)})]$$

$$+ X_2[\sum x_1 x_2(\mu_1^{(1)} - \mu_1^{(2)}) + \sum x_2^2(\mu_2^{(1)} - \mu_2^{(2)})]$$

the linear combination of the variables X_1 and X_2 becomes apparent. The second term of equation (1) is a constant.

The best regions of classification are then given by:

$$R_1: \chi'\sum^{-1}(\mu^{(1)} - \mu^{(2)}) - (1/2)(\mu^{(1)} + \mu^{(2)})'\sum^{-1}(\mu^{(1)} - \mu^{(2)}) \geq \log k$$

$$R_2: \chi'\sum^{-1}(\mu^{(1)} - \mu^{(2)}) - (1/2)(\mu^{(1)} + \mu^{(2)})'\sum^{-1}(\mu^{(1)} - \mu^{(2)}) < \log k$$

INDIVIDUAL DISPERSION MATRIX

A different approach can be utilized in the multivariate procedure by not assuming equal dispersion matrices and instead utilize the individual dispersion matrix from each population.

Assuming that the data used in the derivation of the discriminant functions comes from a multivariate normal distribution, the density function for the j^{th} population is given by:

$$f_j(\underline{x}) = (2\Pi)^{M/2}(\det S_j)^{-1/2}\exp[(-1/2)(\underline{x} - \bar{\underline{x}}_j)' S_j^{-1}(\underline{x} - \bar{\underline{x}}_j)]. \qquad (2)$$

Where: M is the number of variables,

$\underline{x}$ is a vector of variables (dimension 1 x M),

$\bar{\underline{x}}_j$ is a vector of the means of the variables in the j^{th} class (dimension 1 x M),

S_j is the variance-covariance (dispersion) matrix for the j^{th} class (dimension M x M), where $S_j(N_j - 1)$ is given by:

$$\begin{bmatrix} \sum(x_{1j} - \bar{x}_j)^2 & \sum(x_{1j} - \bar{x}_j)(x_{2j} - \bar{x}_j) & \cdots & \sum(x_{1j} - \bar{x}_j)(x_{mj} - \bar{x}_j) \\ \sum(x_{2j} - x_j)(x_{1j} - x_j) & \sum(x_{2j} - \bar{x}_j)^2 & \cdots & \sum(x_{2j} - \bar{x}_j)(x_{mj} - \bar{x}_j) \\ \cdots & \cdots & \cdots & \cdots \\ \sum(x_{mj} - \bar{x}_j)(x_{1j} - \bar{x}_j) & \sum(x_{mj} - \bar{x}_j)(x_{2j} - \bar{x}_j) & \cdots & \sum(x_{mj} - \bar{x}_j)^2 \end{bmatrix}$$

NOTE: S_j is a symmetric matrix.

$\det S_j$ is the determinant of the matrix S_j.

Again, it is at this point that the two procedures differ. The individual matrices can be used as just indicated or a pooled estimate can be used if the dispersion matrices are assumed to be equal. Once the density functions are obtained for all k populations, the conditional probability that a sample, say $\underline{x}$, belongs to the j^{th} class is given by:

$$P_j(\underline{x}) = \frac{f_j(\underline{x})}{\sum_{r=1}^{k} f_r(\underline{x})}, \qquad (3)$$

EXAMPLE COMPARING THE TWO TECHNIQUES

To compare the individual and pooled dispersion matrix techniques, normal sample data was obtained as follows: random samples of size 50 were selected from 16 known normal populations. To insure that the samples were indeed normal, a chi-square goodness of fit test was run on each of the 16 samples. Results of the chi-square tests indicated that all samples were accepted as normal even with a probability of rejection of normality on each sample as large as 0.09 (i.e., min $\hat{\alpha} = 0.09$).

The 16 normal samples were then grouped to represent four populations, each identified by four variables. The samples were grouped as follows:

	Population			
Variable	1	2	3	4
1	N(0,1)	N(1,1)	N(2,1)	N(-1,1)
2	N(2,4)	N(1,4)	N(0,4)	N(1/2,4)
3	N(0,4)	N(2,1)	N(3,4)	N(1,1)
4	$N(3000,10^6)$	$N(4000,10^6)$	$N(5000,10^6)$	$N(3600,10^6)$

The populations from which the samples were taken were selected in such a way as to provide a great deal of overlap of the corresponding variables from population to population. Figure 1 emphasizes the amount of overlap involved in using these variables and makes it clear that no absolute discriminants were used; i.e., variables from population-to-population overlap.

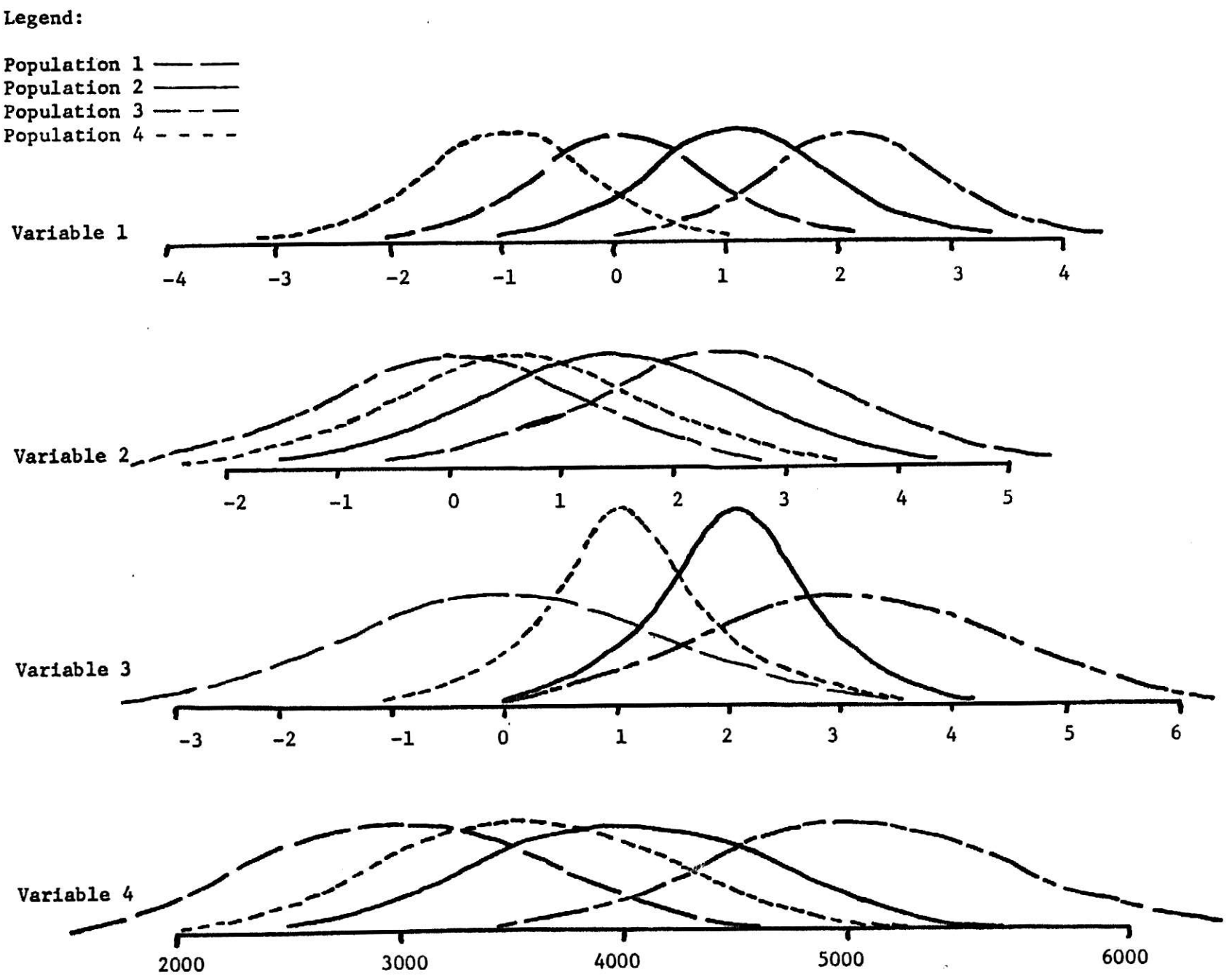


Figure 1. Normal Variables Used.

The purpose of the following example is to illustrate how the two techniques reacted under the following conditions:

(1) having the same number of populations as variables,

(2) having more populations than variables,

(3) having fewer populations than variables,

(4) having equal dispersion matrices (thus satisfying the basic criterion for pooling) and

(5) having unequal dispersion matrices.

Figure 1 illustrates that condition (4) is not satisfied if variable number 3 is used.

Table I summarizes the results of running 25 combinations of populations and variables. The populations used appear in the first column, the variables used in column two, sample size in column three, percentage of correct decisions using the individual dispersion matrices technique in column four and percentage of correct decisions using the pooled dispersion technique in column five. An asterisk on a percentage indicates that a statistically significant difference exists between the two procedures at the 10 per cent level. The interpretation of this difference is simply that we can be at least 90 per cent confident that the procedure with the higher percentage is superior. If no asterisk is present, then the procedures are not significantly different. In the 25 cases considered here, the individual dispersion matrix technique is superior in 12 cases, and in the remaining 13 cases no significant difference exists.

To further explain the summary given in Table I, the complete results for line one in Table I would be written with the following format.

INDIVIDUAL DISPERSION MATRIX

Predicted Population**

Actual Population	1	2	3	4
1	38	5	2	5
2	4	39	5	2
3	1	6	42	1
4	7	5	1	37

Total Correct: 156
Per cent Correct: 78
Variables used: 1,2,3,4

POOLED DISPERSION MATRIX

Predicted Population

Actual Population	1	2	3	4
1	36	7	0	7
2	3	33	11	3
3	1	10	38	1
4	7	5	0	38

Total Correct: 145
Per Cent Correct: 725
Variables Used: 1,2,3,4

**i.e. Population with largest probability.

The average percentage correctly identified for each variation of each procedure is given below. The summary of percentages of the two procedures makes it clear that the best discrimination occurs when the number of variables exceeds the number of populations, and discriminating power deteriorates as the number of variables becomes less than the number of populations.

INDIVIDUAL DISPERSION

Populations	Variables	% Correct
4	4	79.0
4	3	70.0
4	2	60.4
3	4	83.7
3	3	78.8
3	2	69.9

POOLED DISPERSION

Populations	Variables	% Correct
4	4	74.5
4	3	64.6
4	2	55.4
3	4	78.8
3	3	71.0
3	2	64.4

POPULATIONS	VARIABLES	SAMPLE SIZE	INDIVIDUAL DISPERSION PER CENT CORRECT	POOLED DISPERSION PER CENT CORRECT
1,2,3,4	1,2,3,4	200	78.0	72.5
1,2,3,4	1,2,3	200	76.0*	69.0
1,2,3,4	1,2,4	200	66.0	67.5
1,2,3,4	1,3,4	200	75.0	69.5
1,2,3,4	2,3,4	200	63.0*	52.5
1,2,3,4	1,2	200	62.0	58.5
1,2,3,4	1,3	200	73.0	67.5
1,2,3,4	1,4	200	62.5	60.5
1,2,3,4	2,3	200	58.5*	50.5
1,2,3,4	2,4	200	46.5	46.5
1,2,3,4	3,4	200	60.0*	49.0
1,2,3	1,2,3,4	150	84.0*	77.3
1,2,4	1,2,3,4	150	80.0	78.7
1,3,4	1,2,3,4	150	86.0	82.0
2,3,4	1,2,3,4	150	84.7*	77.3
1,2,3	1,2,3	150	82.7*	73.3
1,2,3	1,2,4	150	71.3	72.7
1,2,3	1,3,4	150	84.0*	75.3
1,2,3	2,3,4	150	77.3*	62.7
1,2,3	1,2	150	66.0	63.3
1,2,3	1,3	150	82.7*	72.7
1,2,3	1,4	150	66.7	66.7
1,2,3	2,3	150	68.0*	58.0
1,2,3	2,4	150	64.0	62.0
1,2,3	3,4	150	72.0*	64.0

TABLE I

Monte Carlo Results Using Data From a Multivariate Normal Distribution

*Indicates significant difference at 10% level.

CONCLUSIONS FROM EXAMPLE

Great care should be exercised in drawing conclusions from this example. It should be kept in mind that the assumption for homogeneity of variance-covariance matrices, as required for using the pooled dispersion technique, is not satisfied when variable 3 is used (this is clear from Figure 1). Therefore, when variable 3 is used, the individual dispersion should be expected to do a better job of identification than pooled dispersion. Of the 25 tests that were run on each of the procedures, 17 involved variable 3. The individual dispersion matrix technique had a higher percentage of correct identification in all 17 cases. In the remaining 8 cases, where pooling would have been justified, the individual percentage was higher than pooling in 4 cases, the same as pooling in 2 cases, and less than pooling in 2 cases. Of these last 8 cases, no significant difference was observed.

COMPARISON OF MULTIVARIATE ANALYSIS AND LINEAR DISCRIMINANT FUNCTION

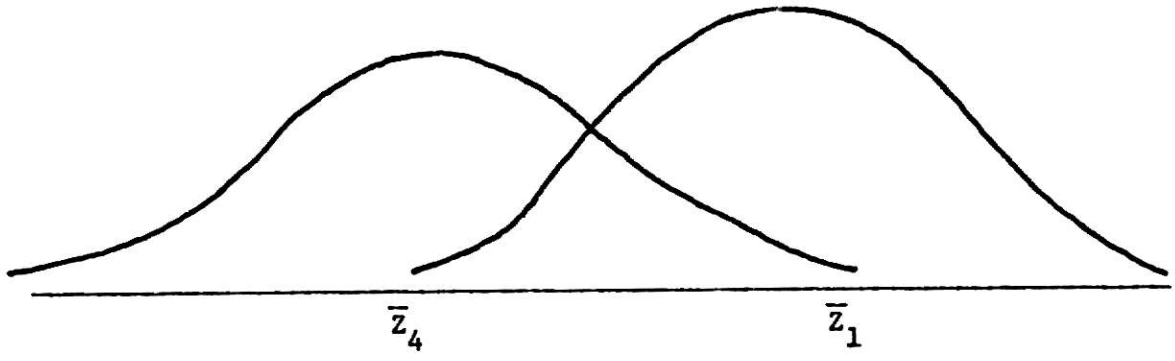
The MA (Multivariate Analysis) model can be compared to the LDF (Linear Discriminant Function) model if separation is limited to two populations. To make the comparison, normal data from the previous example was used. Populations 1 and 4 were selected for the comparison.

The LDF model provides for separation into two populations by considering a linear combination of the available variables; i.e., a linear function $Z = \lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_k X_k$ is obtained. This linear function is derived in such a manner so as to be the linear function that does the best possible job of separation of the two populations. If the original variables $X_1, X_2, \ldots, X_k$ are each normally distributed, then the linear combination

of them (i.e., the linear function Z) will be normally distributed also. Therefore, each of the two populations comprised of k variables is transformed to two normal populations. Each of these new normal populations has a density function with different parameters of the following form:

$$f(Z) = \frac{1}{\sqrt{\Pi 2}\ \sigma} \exp\{-1/2(\frac{Z - \mu}{\sigma})^2\}.$$

The graphs of these two functions will be similar to the following:



It is then desired to obtain the point of intersection of these two graphs, since it is at this point where the probabilities of classification are equal, and either side of this point a higher probability will be assigned to one population than the other. Thus, a basis for comparing LDF model is now available. For the populations considered here, 1 and 4, the equation to be solved for Z is:

$$\frac{1}{\sqrt{2\Pi}(1.326)} \exp\left\{-\frac{1}{2}\left(\frac{Z - 1.758}{1.326}\right)^2\right\} = \frac{1}{\sqrt{2\Pi}(1.472)} \exp\left\{-\frac{1}{2}\left(\frac{Z - 2.168}{1.472}\right)^2\right\}$$

The solution to this equation yields a Z value of -2.190. Therefore, all sample points yielding a Z value greater than -2.190 will be classified as belonging to population 1 and all others to population 4. For the 100 sample

points considered here (50 from population 1 and 50 from population 4) the following results are obtained:

LDF

		Predicted Population	
		1	4
Actual Population	1	39	11
	4	10	40

Total Correct: 79%

MA (individual dispersion matrices)

		Predicted Population	
		1	4
Actual Population	1	43	7
	4	8	42

Total Correct: 85%

This difference, while not statistically significant at the 10% level ($\hat{\alpha} = .13$), would indicate that on the original separation MA is at least as good as LDF.

The comparison of these two models is perhaps more easily understood if a geometric interpretation is considered. The LDF model, being a linear function, says that the best separation of two populations can be made by use of a straight line for a boundary. The MA model, being quadratic in form, says that better separation can be made through use of a curved line (or a curved surface in case of more than two populations). Since a straight line is a special case of a curved line, there are cases in which separation with a straight line boundary will be as good as a curve, but a straight line in general will not allow for fluctuations in the populations.

When the multivariate technique is applied to the Fisher Iris data, the following results are obtained:

POOLED DISPERSION MATRICES

		Predicted Population		
		Setosa	Versicolor	Virginica
Actual Population	Setosa	50	0	0
	Versicolor	0	48	2
	Virginica	0	1	49

INDIVIDUAL DISPERSION MATRICES

		Predicted Population		
		Setosa	Versicolor	Virginica
Actual Population	Setosa	50	0	0
	Versicolor	0	48	2
	Virginica	0	1	49

The pooled and individual dispersion techniques yield identical results on the Fisher data, even to the extent of missing the same test points. However, the probability, of belonging to each population, assigned to the test points differs for the two techniques. This difference, while not critical on the Fisher data, could be when the two or more populations are difficult to distinguish.

CHAPTER V

CLASSIFICATION BY USING A MODIFIED LEAST SQUARES APPROACH

The technique used in this chapter is a modification of the classical least squares procedure. In the classical least squares techniques, sample measurements are made on one or more independent variables $x_1, x_2,..,x_k$ and a certain outcome is observed (i.e. the dependent variable Y). Based upon these sample values an equation of the form,

$$Y = c_1 + c_2x_1 + c_3x_2 + \ldots + c_{k+1}\,x_k,$$

is derived in a manner that gives the best possible fit to the data.

In attempting to apply the least squares technique to the Fisher data, the independent variables are easily identified, i.e. petal length, petal width, sepal length and sepal width. However, there is no observable outcome or dependent variable, but rather a particular class of flowers is observed. So the modified technique used here will arbitrarily assign discrete values as classification variables for each class. To explain further, when independent variables are obtained from setosa, the dependent variable will be assigned a discrete value, say 1. Likewise, say a 2 for versicolor and a 3 for virginica. The ratio of the regression sum of squares to the total sum of squares, represented as R^2 appears with each set of results, thus providing some measure of the degree of fit for the model. Using the values of 1,2, and 3, the following results are obtained when 1.5 and 2.5 are used as separation points for the three populations:

Predicted Population

Actual Population	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	2	48

$R^2 = .9304$

Changing the arbitrary classification variables respectively to 1, 5, and 6 gives the following results when 3.5 and 5.5 are used as separation points.

Predicted Population

Actual Population	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	3
Virginica	0	2	47

$R^2 = .9622$

When classification variables are powers of e such as e^0, e^1, and e^2 with separation points of $(e^0 + e^1)/2$ and $(e^1 + e^2)/2$, the results are:

Predicted Population

Actual Population	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	45	5
Virginica	0	3	47

$R^2 = .8026$

Using this procedure to separate the populations pair wise, perfect separation is obtained for setosa (1) and versicolor (2), as well as for setosa (1) and virginica (3). When the pair virginica (3) and versicolor (2) are compared the following results are obtained:

		Predicted Population	
		Versicolor	Virginica
Actual Population	Versicolor	48	2
	Virginica	1	49

$R^2 = .7839$

It is now apparent that different choices of the arbitrary classification variables give different results. It remains to establish a procedure that will yield the optimum classification variables for a given set of data. That is, for a given set of data, what set of classification variable values will give the optimum separation.

CHAPTER VI

SUMMARY

Four methods of discrimination were presented in the previous chapters. They were, in order of presentation:

(1) Kendall and Stuart's Filtering Technique

(2) Fisher's Linear Discriminant Function

(3) Multivariate Analysis

  (a) Pooled Dispersion Matrices

  (b) Individual Dispersion Matrices

(4) Modified Least Squares Approach

  (a) All populations

  (b) Pair wise

Procedures (1) and (4) as illustrated were distribution free. It is conceivable, however, that circumstances could require confidence intervals on the arbitrary discrete values used by (4) in which case it would be necessary to specify an error structure which would in turn remove the distribution free property of (4). All procedures were demonstrated using the Fisher data given in Table I of Chapter II, with the results summarized as follows:

Procedure	Number Right	Number Unidentifiable or Wrong
(1)	137	13
(2)	144	6
(3a)	147	3
(3b)	147	3
(4a)	146	4
(4b)	147	3

Casual examination of the results would indicate procedure (1) to be of little value. However, in the presence of an absolute discriminant, no refined techniques are needed, and procedure (1) is the preferred technique.

Procedure (2) appears to be optimal only if assumptions of normality and homogeneity of variance are made in the univariate case. The assumption of homogeneity of variance is necessitated by the fact that Fisher's separation point is simply the midpoint between the population means. This assumption can be relaxed if the procedure outlined on Page 24 is followed for determining a separation point. Since Fisher treated the multivariate case as an univariate case by considering a linear combination of the variables involved, it is necessary for the linear combination to be normally distributed with the homogeneity of variance requirement remaining the same.

If the data has a multivariate normal distribution, then either procedure (3a) or (3b) may be used. However, the Monte Carlo study in Chapter IV would seem to indicate that results vary as to whether the dispersion matrices are pooled or used individually. In the 25 cases considered in Chapter IV, with sample sizes of either 150 or 200, the assumption of homogeneity of dispersion matrices was not met in 17 of the cases. Of these 17 cases, the technique using individual dispersion matrices recorded a higher percentage of correct classifications in all 17 cases with 12 of these differences being statistically significant at the 10% level. In the remaining 8 cases where pooling would have been justified, no significant difference was detected, but the results show (3b) having a higher percentage in 4 cases, a lower percentage in 2 cases, and the same percentage in 2 cases. These results would seem to indicate that the individual dispersion matrices technique may be at least as good as the pooled technique, even when pooling is justified.

Procedures (4a) and (4b) as used here have the advantage of being distribution free, while at the same time having the disadvantage of not yet having a procedure for the selection of the values of the classification variables for three or more populations, i.e. (4a). However, the dichotomous case represented by (4b) appears to be highly effective and no problem of selection of classification values exists since any pair of values gives the same results. Some evidence of the effectiveness of (4b) is observed when its results are compared with (2). In this case (4b) missed only 3 test points while (2) missed 6 test points.

None of the procedures considered here gave perfect results, but all were consistent on the test points they could not identify. To explain further, procedure (1) could not identify 13 test points. Procedure (2) missed 6 points all of which were among the 13 points missed by (1). Procedure (4a) missed 4 points all of which were among the 6 points missed by (2). Procedures (3a), (3b) and (4b) all missed the same 3 points, all of which were among the 4 points missed by (4a). Examination of the points missed indicates all to be outliers, hence it would appear doubtful that any discrimination technique would be able to identify these points without misclassification of some other points.

BIBLIOGRAPHY

Anderson, T. W. An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons, 1958, 126-153.

Barnard, M. M. "The Secular Variations of Skull Characters in Four Series of Egyptian Skulls," Annuals of Eugenics. Vol. 6, (1935), 353-371.

Fisher, R. A. "The Use of Multiple Measurements in Taxonomic Problems," Annuals of Eugenics, Vol. 3, Part II, (1936), 179-188.

Goulden, Cyril H. Methods of Statistical Analysis. New York: John Wiley and Sons, 1952, 378-393.

Hodges, Joseph L., Jr. "Discriminatory Analysis," Project Number 21-49-004, Report No. 1, School of Aviation Medicine, U.S.A.F., Randolph AFB, Texas, (1950).

Kendall, Maurice G. and Alan Stuart. The Advanced Theory of Statistics, Vol. 3, New York: Hafner Publishing Company, 1968, 331-335.

Martin, E. S. "A Study of an Egyptian Series of Mandibles, with Special Reference to Mathematical Methods of Sexing," Biometrics, Vol. 28, (1936), 149-178.

York, Leroy J. "Discriminant Functions," Unpublished Master's Report, Kansas State University, Department of Statistics, (1963), 13-18.

A SURVEY OF DISCRIMINATION TECHNIQUES

by

RONALD L. IMAN

B. S., Kansas State University, 1962

M. A., Kansas State Teachers College, 1965

AN ABSTRACT OF A MASTER'S REPORT

Submitted in Partial Fulfillment of the

Requirements for the Degree

MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND COMPUTER SCIENCE

KANSAS STATE UNIVERSITY

MANHATTAN, KANSAS

1970

ABSTRACT

This report presents four procedures for discriminating between two or more populations. Discrimination is defined as follows.

Discrimination. Two or more populations are known to exist and a sample is taken from each. The problem then is to set up a rule, based on the sample values, such that some future observation(s) can be allocated to the correct population(s), when it is not known from which population it emanates. It is desirable to do this identification with the minimum amount of misclassification.

The procedures presented in this report for accomplishing discrimination were in order of appearance:

(1) Kendall and Stuart's Filtering Technique

(2) Fisher's Linear Discriminant Function

(3) Multivariate Analysis

  (a) Pooled Dispersion Matrices

  (b) Individual Dispersion Matrices

(4) Modified Least Squares Approach

  (a) All populations

  (b) Pair wise

Procedures (1) and (4) as illustrated in this report were distribution free. Procedure (4) is a new technique developed by the author, and results indicate that it may have certain optimal properties in the dichotomous case. However, questions remain unanswered regarding the choice of classification variables when more than two populations are to be identified.

Procedures (3a) and (3b) were the subject of a Monte Carlo study of discriminating among 3 or 4 populations using 2, 3, or 4 variables. Results of the study show that (3b) may be at least as good as (3a) even when dispersion matrices are homogeneous.

All procedures were demonstrated using the classical Iris data of Fisher.