Discrimination models on two probability distributions
by Margaret Palmer Gessaman

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY in Mathematics
Montana State University

Abstract:
The discrimination problem for two probability distributions can be briefly described as follows: an individual is observed at random from a collection known to consist of elements from two distinct populations, and this individual is to be classified as to its parent population.

A discrimination model is constructed which requires control of the probability of misclassification under each distribution. In order to do this, a point must be included in the decision space which allows for reserving judgment, i.e. making no classification. The power for discrimination is defined as the overall probability that a decision is made.

If the level is chosen by the investigator, then the discrimination problem is the search for a discrimination procedure, if it exists, which is most powerful for discrimination at the chosen level.

When both distributions are completely known, a solution under quite general hypotheses is found.

A nonparametric discrimination model is given and a method for constructing procedures approximately of the required size is outlined. The theory of coverages is used in this construction. A consistency for sequences of discrimination procedures is defined. Under restricted models, sequences of discrimination procedures based on the theory of coverages are found which are consistent with an optimal procedure. When consistency does not seem possible, procedures which have some appeal are suggested.

DISCRIMINATION MODELS ON TWO PROBABILITY DISTRIBUTIONS

by

MARGARET PALMER GESSAMAN

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematics

Approved:

Head, Major Department

Chairman, Examining Committee

Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

December, 1966

ACKNOWLEDGMENT

The research in this thesis was supported by a research grant from the National Aeronautics and Space Administration.

The author is greatly indebted to Dr. C. P. Quesenberry for his guidance in all phases of this research and the long hours he spent in directing this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

The discrimination problem for two probability distributions can be briefly described as follows: an individual is observed at random from a collection known to consist of elements from two distinct populations, and this individual is to be classified as to its parent population.

A discrimination model is constructed which requires control of the probability of misclassification under each distribution. In order to do this, a point must be included in the decision space which allows for reserving judgment, i.e. making no classification. The power for discrimination is defined as the overall probability that a decision is made. If the level is chosen by the investigator, then the discrimination problem is the search for a discrimination procedure, if it exists, which is most powerful for discrimination at the chosen level.

When both distributions are completely known, a solution under quite general hypotheses is found.

A nonparametric discrimination model is given and a method for constructing procedures approximately of the required size is outlined. The theory of coverages is used in this construction. A consistency for sequences of discrimination procedures is defined. Under restricted models, sequences of discrimination procedures based on the theory of coverages are found which are consistent with an optimal procedure. When consistency does not seem possible, procedures which have some appeal are suggested.

CHAPTER I

INTRODUCTION

The discrimination problem for two distributions can be briefly described as follows: an individual is observed at random from a collection known to consist of elements from two distinct populations $\mathcal{P}_1$ and $\mathcal{P}_2$; and this individual is to be classified as to its parent population.

If the populations are distinct and there is an unambiguous means of defining the elements of the two populations, there might not appear to be an actual problem. However, here are some examples from three classes of problems, as suggested by M. G. Kendall (1965), which lend themselves to discrimination analysis:

(a) Lost information. Archaelogical and skeletal remains may be all an anthropologist has by which to classify members of two different races, whereas the problem would have been easy had they been encountered as living men.

(b) Diagnosis. The presence of certain suspected diseases could certainly be determined by surgery or postmortem. But it is preferable to diagnose the presence of the disease from external symptoms rather than to be driven to such extreme methods of diagnosis; and, in general, methods are not available to determine this presence in an unambiguous manner.

(c) Prediction. It may be possible to differentiate between two conditions when they occur, but it would be preferable to predict their occurrence in time to take necessary steps. Prediction of economic situations would be an example. Another

example from this class would be the removal of defective items, such as light bulbs, which could be definitely differentiated from nondefective items only by destroying the individual.

In order to make the classification of the individual, measurements are made on the individual and compared with the corresponding measurements taken on the elements of $\mathcal{P}_1$ and $\mathcal{P}_2$. A random variable X is defined on the k-dimensional Euclidean space, R^k, by the measurements made on the elements of $\mathcal{P}_1$ and a random variable Y is defined by the corresponding measurements made on the elements of $\mathcal{P}_2$. It is assumed the random variables X and Y are distributed over R^k according to distinct absolutely continuous probability distributions P_1 and P_2, respectively.

The k-dimensional measurements are taken on the observed individual, which then becomes an observation z of either the random variable X or the random variable Y. On the basis of the observation z a decision is to be made concerning the parent probability distribution of this observation.

A discrimination procedure Δ is a function on the space of the observation z, namely R^k, into the decision space. Use of a discrimination procedure Δ can lead to two types of error, i.e. two types of misclassification of z. The first type of error occurs if P_1 is the true distribution of z and it is classified as a P_2 random variable, and the second type of error occurs if P_2 is the true distribution of z and it is classified as a P_1 random variable.

In the classical discrimination model it is required that z be

classified either as a P_1 random variable or as a P_2 random variable. Thus the decision space contains exactly two elements. The solution of the problem under this model is a discrimination procedure, if it exists, which simultaneously minimizes the probabilities of the two types of error or which is a minimax solution if different "costs" are assigned to the two types of misclassification.

The discrimination problem can be reduced to subproblems; and one basis for reduction, due to Fix and Hodges (1951), is how much is known about the probability distributions; for example, consider the following subproblems:

(i) Both distributions are known completely.

(ii) P_1 and P_2 are known except for the values of one or more parameters.

(iii) Nothing is assumed about P_1 and P_2 except the existence of probability density functions, which follows from the absolute continuity of the distributions.

For the classical discrimination model, subproblem (i) has been completely solved by Welch (1939). The best known example of a solution to subproblem (i) is the linear discriminant function based on the work of R. A. Fisher (1936). There, it is assumed that P_1 and P_2 are k-variate normal distributions having identical dispersion matrices and distinct mean vectors.

Most attempts to solve subproblem (ii) have been made by estimating the unknown parameters and then proceeding as if the distributions were completely known. To the best knowledge of the author, the first

attempt to solve subproblem (iii) was made by Fix and Hodges (1951) and involved density function estimation at the value z.

However, suppose it is desired to control the level of the probability of misclassification under each distribution and a solution to the discrimination problem is sought among those discrimination procedures which give this control. In order to do this, it becomes necessary to allow a third decision, the decision to reserve judgment, i.e. not to classify. From the class of all discrimination procedures with the probability of misclassification under each distribution controlled at fixed levels an attempt will be made to find a procedure which maximizes the overall probability that z is classified.

Little work has been done on discrimination models which allow the investigator to reserve judgment. M. G. Kendall (1965) suggests the possibilities of (1) convex hulls based on samples from the two distributions if the distributions are to some extent unknown and (2) the use of order statistics to consider one component at a time for k-variate random variables. Similar notions could be applied to subproblem (i).

In this paper a discrimination model is introduced in which fixed levels for the probability of misclassification under each distribution are used to define classes of discrimination procedures. In this model a third decision, that of reserving judgment, is included in the decision space. For subproblem (i) a solution of the discrimination problem under this model is given. A nonparametric method for constructing discrimination procedures which give approximately the required level of control

on the probabilities of misclassification for large samples is given. A notion of consistency of sequences of discrimination procedures is introduced. For some restricted models of subproblem (iii) a sequence of discrimination procedures is found which is consistent with an optimal procedure. For subproblem (iii) the problem is not determinate and so some choices of discrimination procedures for this subproblem which have some desirable properties are suggested.

CHAPTER II

A DISCRIMINATION MODEL WITH FIXED LEVELS OF MISCLASSIFICATION FOR TWO COMPLETELY KNOWN PROBABILITY DISTRIBUTIONS

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two distinct populations. A random variable X is defined on the k-dimensional Euclidean space, R^k, by measurements taken on elements of $\mathcal{P}_1$, and a random variable Y is defined on R^k by the corresponding measurements taken on elements of $\mathcal{P}_2$. The random variables X and Y are distributed over R^k according to distinct absolutely continuous probability distributions P_1 and P_2, respectively. The assumption of absolute continuity guarantees the existence of probability density functions f_1 and f_2, respectively, with respect to Lebesgue measure λ on R^k. It is assumed that the probability distributions and their probability density functions are completely known.

An individual from a population consisting of elements from both $\mathcal{P}_1$ and $\mathcal{P}_2$ is observed. The k-dimensional measurements are made on the individual, which then becomes an observation z on either the random variable X or the random variable Y. On the basis of the observation z a decision is to be made whether to classify z as a P_1 random variable, to classify z as a P_2 random variable, or to reserve judgment, i.e. to make no classification. A discrimination procedure Δ will be a function from the space of the observations, namely R^k, into the decision space $D = (d_1, d_2, d_0)$ where

d_1 means classify z as a P_1 random variable,

d_2 means classify z as a P_2 random variable,

d_0 means reserve judgment.

The space of the observation z is thus partitioned by a discrimination

procedure Δ into the following sets:

$$(1) \qquad \begin{aligned} S_1 &= \{z; \Delta(z) = d_1\}, \\ S_2 &= \{z; \Delta(z) = d_2\}, \\ S_0 &= \{z; \Delta(z) = d_0\}. \end{aligned}$$

The discrimination procedure is completely determined by the sets S_1 and S_2, and conversely; thus a discrimination procedure will be identified by the notation $\Delta(S_1,S_2)$, where the sets defining the procedure will be indicated by the letter S with the subscripts corresponding to those in (1) and raised symbols to identify specific procedures.

Let π_1 and π_2 be the a priori probabilities that an observation comes from P_1 and P_2, respectively. It is assumed that π_1 and π_2 are nonzero, since otherwise the problem is trivial. The probability of an error of the first type is the probability that an observation comes from P_1 and is classified as a P_2 random variable; thus the probability of an error of the first type is $\pi_1 P_1(S_2)$. Similarly, the probability of an error of the second type is $\pi_2 P_2(S_1)$. The overall probability of error is $\pi_1 P_1(S_2) + \pi_2 P_2(S_1)$. If the values of $P_1(S_2)$ and $P_2(S_1)$ are controlled at fixed levels, note that a bound is also established on the overall probability of error. The values $P_1(S_2)$ and $P_2(S_1)$ for a discrimination procedure $\Delta(S_1,S_2)$ will be used to define its size.

Definition 2.1: A discrimination procedure $\Delta(S_1,S_2)$ is said to be of size-(α,β) for $\alpha \in (0,1)$ and $\beta \in (0,1)$ if $P_1(S_2) \leq \alpha$ and $P_2(S_1) \leq \beta$.

(In this paper, parentheses will be used to indicate an open interval and brackets to indicate a closed interval, while [) and (] will denote

closed-open and open-closed intervals, respectively.)

In other words, a discrimination procedure $\Delta(S_1,S_2)$ is said to be of size-(α,β) if, under each distribution, the probability of misclassification which results from using that procedure is no more than a given value.

The following definition of the exact size of discrimination procedures will be useful in the discussions which appear later in this chapter:

Definition 2.2: A discrimination procedure $\Delta(S_1,S_2)$ is said to be of exact size-(α,β) or to be of size equal to (α,β) if $P_1(S_2)=\alpha$ and $P_2(S_1)=\beta$.

The overall probability of correct classification for a discrimination procedure $\Delta(S_1,S_2)$ is $\pi_1P_1(S_1) + \pi_2P_2(S_2)$ and the overall probability of no classification is $\pi_1P_1(S_0) + \pi_2P_2(S_0)$. Suppose that for $\alpha\in(0,1)$ and $\beta\in(0,1)$, arbitrary but fixed constants, only those discrimination procedures of size-(α,β) are considered; then for each discrimination procedure in this class the overall probability of misclassification is at most $\gamma=\pi_1\alpha+\pi_2\beta$. The discriminiation problem under this model can be stated as follows: from among the class of size-(α,β) discrimination procedures find a procedure, if it exists, which maximizes the overall probability that the observation z is classified. This overall probability is

$$\{\pi_1P_1(S_1) + \pi_2P_2(S_2)\} + \{\pi_1P_1(S_2) + \pi_2P_2(S_1)\}.$$

Since $\sum_{i=0}^{2} P_1(S_i) = \sum_{i=0}^{2} P_2(S_i) = 1$, maximizing this probability is equivalent to minimizing the overall probability of no classification. It is this criterion which will be used to define the power for discrimination of a discrimination procedure. If the procedure which minimizes the overall probability for no classification also maximizes the probability of correct classification, such a procedure would have strong appeal. In Theorem 2.1 it will be shown that in some cases just such a discrimination procedure exists.

Definition 2.3: The power for discrimination of a discrimination procedure $\Delta(S_1,S_2)$ is defined to be the overall probability that under that procedure an observation z is classified, i.e.

$$1 - \pi_1 P_1(S_0) - \pi_2 P_2(S_0).$$

As a criterion to decide whether a given discrimination procedure is "better" for discrimination than another procedure or is "best" among all size-(α, β) procedures the concept of power for discrimination will be used as follows:

Definition 2.4: A discrimination procedure $\Delta(S_1,S_2)$ is said to be more powerful for discrimination at level-(α, β) than a discrination procedure $\Delta^*(S_1^*,S_2^*)$ if both are of size-(α, β) and

$$1 - \pi_1 P_1(S_0) - \pi_2 P_2(S_0) > 1 - \pi_1 P_1(S_0^*) - \pi_2 P_2(S_0^*).$$

Definition 2.5: A discrimination procedure $\Delta(S_1,S_2)$ is said to be most powerful for discrimination at level-(α, β) if for $\Delta^*(S_1^*,S_2^*)$

any size-(α, β) discrimination procedure, $\Delta^*(S_1^*, S_2^*)$ is not more powerful for discrimination at level-(α, β) than $\Delta(S_1, S_2)$.

The existence of a discrimination procedure which is most powerful at level-(α, β) will be considered next. If the ratio $f_1(x)/f_2(x)$ is a continuous random variable defined a.e. λ, then Theorem 2.1 will give the answer to this question, namely there always exists a discrimination procedure which is most powerful for discrimination at level-(α, β).

Let $\alpha \in (0,1)$ and $\beta \in (0,1)$ be arbitrary but fixed values. When P_1 and P_2 satisfy the hypotheses mentioned above, it is always possible to construct a size-(α, β) discrimination procedure by a method which will be outlined here. As will be shown in Theorem 2.1, there always exists at least one measurable subset of R^k whose probability under P_1 is exactly α; let A_1 be such a set. Similarly, there always exists at least one measurable subset of R^k whose probability under P_2 is exactly β; let A_2 be such a set. A discrimination procedure $\Delta(S_1, S_2)$ is defined as follows:

$$(2) \qquad \Delta(z) = \begin{cases} d_1 \text{ if } z \in \bar{A}_1 A_2 = S_1, \\ d_2 \text{ if } z \in A_1 \bar{A}_2 = S_2, \\ d_0 \text{ if } z \in A_1 A_2 \cup \bar{A}_1 \bar{A}_2 = S_0. \end{cases}$$

Since $\{A_1 A_2, \bar{A}_1 A_2, A_1 \bar{A}_2, \bar{A}_1 \bar{A}_2\}$ is a partition of R^k, this function is well-defined; and since

$$P_1(S_2) = P_1(A_1 \bar{A}_2) \leq P_1(A_1) = \alpha \text{ and}$$

$$P_2(S_1) = P_2(\bar{A}_1 A_2) \leq P_2(A_2) = \beta,$$

$\Delta(S_1,S_2)$ is a size-(α,β) discrimination procedure. No requirement is made of the sets A_1 and A_2 except that they be measurable sets of the proper size. However, it is apparent that some choices for these sets will give procedures which are more powerful for discrimination at level-(α,β) than others.

A lemma whose assertions will be used for key steps in the proof of Theorem 2.1 is set out here. Its proof follows directly from Definitions 2.3, 2.4, and 2.5.

Lemma 2.1: Let $\Delta(S_1,S_2)$ be a size-(α,β) discrimination procedure for $\alpha\in(0,1)$ and $\beta\in(0,1)$.

(a) $\Delta(S_1,S_2)$ has power one for discrimination if and only if $P_1(S_0) = P_2(S_0) = 0$.

(b) A sufficient condition that $\Delta(S_1,S_2)$ be most powerful for discrimination at level-(α,β) is that for $\Delta^*(S_1^*,S_2^*)$ any size-(α,β) discrimination procedure, $P_1(S_0) \leq P_1(S_0^*)$ and $P_2(S_0) \leq P_2(S_0^*)$.

(c) If there exists a size-(α,β) discrimination procedure $\Delta^*(S_1^*,S_2^*)$ such that $P_1(S_0) \geq P_1(S_0^*)$ and $P_2(S_0) \geq P_2(S_0^*)$ where at least one inequality is strict, then $\Delta(S_1,S_2)$ is not most powerful for discrimination at level-(α,β).

Theorem 2.1: Let P_1 and P_2 be distinct probability distributions defined on the k-dimensional Euclidean space, R^k, with probability density functions f_1 and f_2, respectively, with respect to Lebesgue measure λ on R^k such that the random variable $W = f_1(X)/f_2(X)$ is a continuous random variable defined a.e. λ.

(i) For arbitrary but fixed values $\alpha \in (0,1)$ and $\beta \in (0,1)$ there exist essentially unique positive constants c_1 and c_2 and sets A_1 and A_2 which define a discrimination procedure $\Delta(S_1,S_2)$ as in (2) such that

(3) $P_1(A_1)=\alpha$ and $P_2(A_2)=\beta$,

where

$$A_1 = \{x;\ f_1(x) < c_1 f_2(x)\} \text{ and}$$

(4)
$$A_2 = \{x;\ f_1(x) > c_2 f_2(x)\} \ .$$

(ii) Let $\Delta(S_1,S_2)$ be a discrimination procedure constructed as in (2) from sets A_1 and A_2 which satisfy (3) and (4).

(a) If $c_1 \leq c_2$, then $\Delta(S_1,S_2)$ is most powerful for discrimination at level-(α,β).

(b) If $c_1 > c_2$, then $\Delta(S_1,S_2)$ is not necessarily most powerful for discrimination at level-(α,β), but there exists a discrimination procedure with power one for discrimination which is of size-(α,β) and is constructed as in (2) from sets A_1 and A_2 which satisfy (4) for $c_1 = c_2 = c^*$.

(iii) Let c_1 and c_2 be the positive constants defined in (i).

(a) If $c_1 \leq c_2$, let $\Delta^*(S_1^*,S_2^*)$ be a procedure which is most powerful for discrimination at level-(α,β). Then $\Delta^*(S_1^*,S_2^*)$ is of exact size-(α,β) and $S_1^* = S_1$ a.e. λ, $S_2^* = S_2$ a.e. λ .

(b) If $c_1 > c_2$, then there exists a class of discrimination procedures with power one for discrimination and size-(α,β).

At least one of these procedures is constructed as in (2) from sets A_1 and A_2 which satisfy (4) for $c_1 = c_2 = c^*$.

In an attempt to make the ideas involved in the theorem clearer, a brief explanation of them will be given before the formal proof of the theorem. A reasonable choice for A_1 and A_2 should be concerned with the relative probabilities of X and Y for sets of values of z. In order to obtain a discrimination procedure which is most powerful for discrimination, the set A_1 should be composed of those values of z which are most likely to be observations on P_2 random variables rather than P_1 random variables; similarly, the set A_2 should be composed of those values of z which are most likely to be observations on P_1 random variables rather than P_2 random variables. The ratio $f_1(x)/f_2(x)$ of the probability density functions will be used to measure the relative probabilities at the values of z. Let A_1 and A_2 be the sets defined in Theorem 2.1 (i). If the sets A_1 and A_2 do not intersect, then the discrimination procedure $\Delta(S_1,S_2)$ constructed as in (2) from the sets A_1 and A_2 is most powerful for discrimination at level-(α, β) and any procedure most powerful for discrimination at level-(α, β) must agree with $\Delta(S_1,S_2)$ a.e. λ. If the sets A_1 and A_2 do intersect, this is not necessarily true, but a discrimination procedure can be found which is constructed as in (2) from sets of the same form as A_1 and A_2 and has power one for discrimination. These statements will be established in the proof of Theorem 2.1 (ii) and (iii).

Proof of the theorem; (i) The existence of essentially unique positive

constants c_1 and c_2 such that $P_1\{x;\ f_1(x) < c_1 f_2(x)\} = \alpha$ and $P_2\{x;\ f_1(x) > c_2 f_2(x)\} = \beta$ is treated in the existence proof for hypothesis testing functions of exact size α and β, respectively. A detailed treatment can be found in the proof of the Neyman-Pearson Fundamental Lemma given in Lehmann (1959). The sets A_1 and A_2 are measurable and define a discrimination procedure $\Delta(S_1,S_2)$ as in (2). Notice that since W is a continuous random variable, $\lambda\{x;\ W = c_1\} = \lambda\{x;\ W = c_2\} = 0$.

(ii) (a) Let $\Delta(S_1,S_2)$ be a discrimination procedure constructed as in (2) from sets A_1 and A_2 which satisfy (3) and (4) such that $c_1 \leq c_2$. Since $c_1 \leq c_2$, the following relations which will be used in the proof of this part are noted:

[1] A_1A_2 is empty and therefore $S_0 = \bar{A}_1\bar{A}_2$;

[2] $P_1(S_2) = P_1(A_1\bar{A}_2) = P_1(A_1) = \alpha$;

[3] $P_2(S_1) = P_2(\bar{A}_1A_2) = P_2(A_2) = \beta$.

Let $\Delta^*(S_1^*,S_2^*)$ be any discrimination procedure of size-(α,β). In order to show that $\Delta(S_1,S_2)$ is most powerful for discrimination at level-(α,β) it is sufficient, as was shown in Lemma 2.1 (b), to show that $P_1(S_0) \leq P_1(S_0^*)$ and $P_2(S_0) \leq P_2(S_0^*)$. The former will be established and the latter follows from symmetry.

Since the discrimination procedures partition R^k,

$$P_1(S_1) + P_1(S_2) + P_1(S_0) = P_1(S_1^*) + P_1(S_2^*) + P_1(S_0^*).$$

But $P_1(S_2) = \alpha$ and $P_1(S_2^*) \leq \alpha$, so that

$$(5) \quad P_1(S_1) + P_1(S_0) \leq P_1(S_1^*) + P_1(S_0^*).$$

From (5) it is seen that in order to show $P_1(S_0) \leq P_1(S_0^*)$ it would suffice to show $P_1(S_1) \geq P_1(S_1^*)$; that is, it would suffice to establish that the probability of correct classification when using $\Delta(S_1, S_2)$ is at least as great as when using any other size-(α, β) discrimination procedure.

Let $\psi(x)$ be the characteristic function of $S_1 = A_2$ and $\varphi(x)$ the characteristic function of S_1^*. Define the sets

$$S^+ = \{x;\ \psi(x) - \varphi(x) > 0\},$$
$$S^- = \{x;\ \psi(x) - \varphi(x) < 0\}.$$

If $x \in S^+$, then $\psi(x) = 1$ and $f_1(x) - c_2 f_2(x) > 0$. If $x \in S^-$, then $\psi(x) = 0$ and $f_1(x) - c_2(x) \leq 0$. In each case the product $(\psi - \varphi)(f_1 - c_2 f_2)$ is nonnegative. Consider the integral

$$(6) \int_{R^k} (\psi - \varphi)(f_1 - c_2 f_2)\, d\lambda = \int_{S^+ \cup S^-} (\psi - \varphi)(f_1 - c_2 f_2)\, d\lambda \geq 0.$$

But this integral can also be written as follows:

$$(7) \int_{R^k} (\psi - \varphi)(f_1 - c_2 f_2)\, d\lambda = P_1(S_1) - P_1(S_1^*) - c_2\{P_2(S_1) - P_2(S_1^*)\}.$$

Combining (6) and (7),

$$P_1(S_1) - P_1(S_1^*) \geq c_2\{P_2(S_1) - P_2(S_1^*)\} \geq 0,$$

since $P_2(S_1) = \beta$ and $P_2(S_1^*) \leq \beta$. Therefore, $P_1(S_1) \geq P_1(S_1^*)$, which

was sufficient to show $P_1(S_0) \leq P_1(S_0^*)$.

(b) Let $\Delta(S_1,S_2)$ be a discrimination procedure constructed as in (2) from sets A_1 and A_2 which satisfy (3) and (4) such that $c_1 > c_2$. An example which shows this procedure is not necessarily most powerful for discrimination at level-(α, β) can be found in Example 2.1 (ii). However, if $c^* \in [c_2,c_1]$, define sets $A_1^* = \{x;\ f_1(x) < c^* f_2(x)\}$ and $A_2^* = \{x;\ f_1(x) > c^* f_2(x)\}$ and construct a discrimination procedure $\Delta^*(S_1^*,S_2^*)$ as in (2) from these sets. Then

[4] $P_1(S_2^*) = P_1(A_1^*\overline{A_2^*}) = P_1(A_1^*) \leq P_1\{x;\ f_1(x) < c_1 f_2(x)\} = \alpha$,

[5] $P_2(S_1^*) = P_2(\overline{A_1^*}A_2^*) = P_2(A_2^*) \leq P_2\{x;\ f_1(x) > c_2 f_2(x)\} = \beta$,

[6] $P_1(S_0^*) = P_2(S_0^*) = 0$.

From Lemma 2.1 (a) the procedure $\Delta^*(S_1^*,S_2^*)$ has power one for discrimination at level-(α, β) and it is of size-(α, β) by definition.

(iii) (a) Let $\Delta^*(S_1^*,S_2^*)$ be most powerful for discrimination at level-(α, β). Let $c_1 \leq c_2$ be the positive constants defined in (i) and $\Delta(S_1,S_2)$ a discrimination procedure constructed as in (2) from sets A_1 and A_2 which satisfy (3) and (4). Under $\Delta(S_1,S_2)$ statements [1], [2], [3] hold as before.

Let $\psi(x)$ be the characteristic function of $S_1 = A_2$ and $\varphi(x)$ the characteristic function of S_1^*. Define the sets

$$S^+ = \{x;\ \psi(x) - \varphi(x) > 0\},\ S^- = \{x;\ \psi(x) - \varphi(x) < 0\},$$

$$S' = \{x;\ f_1(x) - c_2 f_2(x) \neq 0\},$$

$$S = (S^+ \cup S^-) \cap S'.$$

The probability of the set $S^+ \cup S^-$ is the probability of the set on which S_1 and S_1^* do not agree. On the set S the value of the product $(\psi - \varphi)(f_1 - c_2 f_2)$ is strictly positive. Then the integral

$$\int_{R^k} (\psi - \varphi)(f_1 - c_2 f_2)\, d\lambda = \int_S (\psi - \varphi)(f_1 - c_2 f_2)\, d\lambda$$

$$= P_1(S_1) - P_1(S_1^*) - c_2\{P_2(S_1) - P_2(S_1^*)\}$$

is strictly positive unless $\lambda(S) = 0$. Suppose $\lambda(S) > 0$; then $P_1(S_1) > P_1(S_1^*)$ which from (5) implies that $P_1(S_0) < P_1(S_0^*)$. From the proof of (ii) it is known that $P_2(S_0) \leq P_2(S_0^*)$ and therefore from Lemma 2.1 (c) $\Delta^*(S_1^*, S_2^*)$ is not most powerful for discrimination at level-(α, β), a contradiction. Thus $\lambda(S) = \lambda(S^+ \cup S^-) = 0$ and S_1 and S_1^* agree a.e. λ.

Also, since

$$\alpha = P_1(A_1) = P_1(S_2) = P_1(S_2^*) \leq \alpha \text{ and}$$

$$\beta = P_2(A_2) = P_2(S_1) = P_2(S_1^*) \leq \beta,$$

the procedure $\Delta^*(S_1^*, S_2^*)$ is of exact size-(α, β).

(b) Let $c_1 > c_2$ be the positive constants defined in (i). In (ii) (b) a discrimination procedure with power one for discrimination and size-(α, β) was constructed as in (2) from sets A_1 and A_2 which satisfy (4)

for $c_1 = c_2 = c^*$. Thus all procedures which are most powerful for discrimination at this level must have power one for discrimination. Note also that from Lemma 2.1 (a) the probability of no classification under each distribution for each procedure in this class is zero.

This completes the proof of the theorem.

For those probability distributions which satisfy the hypotheses of the theorem, the discrimination problem is completely solved. The positive constants c_1 and c_2 exist and can be found by numerical methods since the distributions are completely known. Let the discrimination procedure defined in (i) be denoted by $\Delta(c_1,c_2)$. If $c_1 \leq c_2$, $\Delta(c_1,c_2)$ is most powerful for discrimination at level-(α, β) and every procedure which is most powerful for discrimination at this level agrees with $\Delta(c_1,c_2)$ a.e. λ. It was shown that in this case $\Delta(c_1,c_2)$ also maximizes the overall probability of correct classification. If $c_1 > c_2$, a discrimination procedure can be constructed as in (2) from sets A_1 and A_2 which satisfy (4) in such a way that this procedure has power one for discrimination and is of size-(α, β). A few examples ~~will now be given to illustrate these concepts~~ in particular models.

Example 2.1: For θ a real-valued parameter, let $(P_\theta ; \theta \in \Omega)$ be a family of absolutely continuous probability distributions which has strictly monotone (increasing) density ratio in a continuous real-valued function $T(x)$, i.e. for $\theta > \theta'$, $f_\theta(x)/f_{\theta'}(x)$ is a strictly monotone (increasing) function of the random variable $T(x)$. Let f_θ and $f_{\theta'}$, for $\theta, \theta' \in \Omega$ where $\theta > \theta'$, be denoted by f_1 and f_2, respectively.

In the following discussion it will be assumed that the ratio $f_1(x)/f_2(x)$ satisfies the hypotheses of Theorem 2.1. Let $\alpha \in (0,1)$ and $\beta \in (0,1)$ be arbitrary but fixed constants, and let c_1 and c_2 be the constants defined in Theorem 2.1 (i).

If $c_1 \leq c_2$, the discrimination procedure $\Delta(c_1,c_2)$ defined by the sets $S_1 = \{x;\ f_1(x) > c_2 f_2(x)\}$ and $S_2 = \{x;\ f_1(x) < c_1 f_2(x)\}$ is most powerful for discrimination at level-(α, β). Since $f_1(x)/f_2(x)$ is a strictly increasing function of $T(x)$, this random variable can also be used to construct the discrimination procedure. It can be seen that $f_1(x)/f_2(x) < c_1$ if and only if $T(x) < c_1^*$ for c_1^* a unique real number. Therefore $P_1\{x;\ T(x) < c_1^*\} = \alpha$. Similarly there exists a unique constant c_2^* such that $f_1(x)/f_2(x) > c_2$ if and only if $T(x) > c_2^*$ and $P_2\{x;\ T(x) > c_2^*\} = \beta$. If P_1^* is the probability distribution induced on the space of $T(x)$ by P_1, then by definition of the induced probability distribution, $P_1^*\{t;\ t < c_1^*\} = P_1\{x;\ T(x) < c_1^*\} = \alpha$. Similarly, if P_2^* is the probability distribution induced on the space of $T(x)$ by P_2, then $P_2^*\{t;\ t > c_2^*\} = P_2\{x;\ T(x) > c_2^*\} = \beta$. Then the sets $S_2^* = \{t;\ t < c_1^*\}$ and $S_1^* = \{t;\ t > c_2^*\}$ where $P_1^*(S_2^*) = \alpha$ and $P_2^*(S_1^*) = \beta$ define a discrimination procedure which is equivalent to $\Delta(c_1,c_2)$. Note that c_1^* is the α^{th} percentile of the probability

distribution P_1^* and c_2^* is the $(1-\beta)^{th}$ percentile of the probability distribution P_2^*. Let them be denoted by t_α^* and $t_{1-\beta}^*$ respectively.

If $c_1 > c_2$ or equivalently $t_\alpha^* > t_{1-\beta}^*$, it is possible to construct a discrimination procedure with power one for discrimination at level-(α, β) and size-(α, β). In order to do this, select $t^* \varepsilon [t_{1-\beta}^*, t_\alpha^*]$ and define a discrimination procedure by the sets $S_1 = \{t;\ t > t^*\}$ and $S_2 = \{t;\ t < t^*\}$. The actual choice of the value t^* from the interval presents an interesting problem in its own right.

The class of univariate normal distributions with common variance σ^2 forms a family with real parameter μ of absolutely continuous, univariate probability distributions which has strictly monotone (increasing) density ratio in x. The following specific examples from such a family, will demonstrate the concepts of Theorem 2.1 and this example.

(i) Let P_1 and P_2 be two univariate normal probability distributions with common variance 1 and means 2 and 0 respectively. In this case,

$$\frac{f_1(x)}{f_2(x)} = \exp\left\{-\tfrac{1}{2}\left[(x-2)^2 - x^2\right]\right\} = \exp\{2x - 2\}.$$

This ratio is strictly increasing in x. From the above discussion it is known that the sets A_1 and A_2 as defined in Theorem 2.1 (i) can be defined directly in terms of x, i.e. there exist constants c_1 and c_2 such that $P_1\{x;\ x < c_1\} = \alpha$ and $P_2\{x;\ x > c_2\} = \beta$. If $\alpha = \beta = .1$, then $c_1 = 0.718$ and $c_2 = 1.282$. The discrimination

procedure constructed from A_1 and A_2 then is defined by the sets

$$S_1 = \{x;\ x > 1.282\},$$

$$S_2 = \{x;\ x < 0.718\},$$

$$S_0 = \{x;\ 0.718 \leq x \leq 1.282\}.$$

In other words, the discrimination procedure says

(1) if the observation z is greater than 1.282, classify z as a P_1 random variable;

(2) if the observation z is less than 0.718, classify z as a P_2 random variable;

(3) otherwise, do not classify z.

See Figure 2.1.

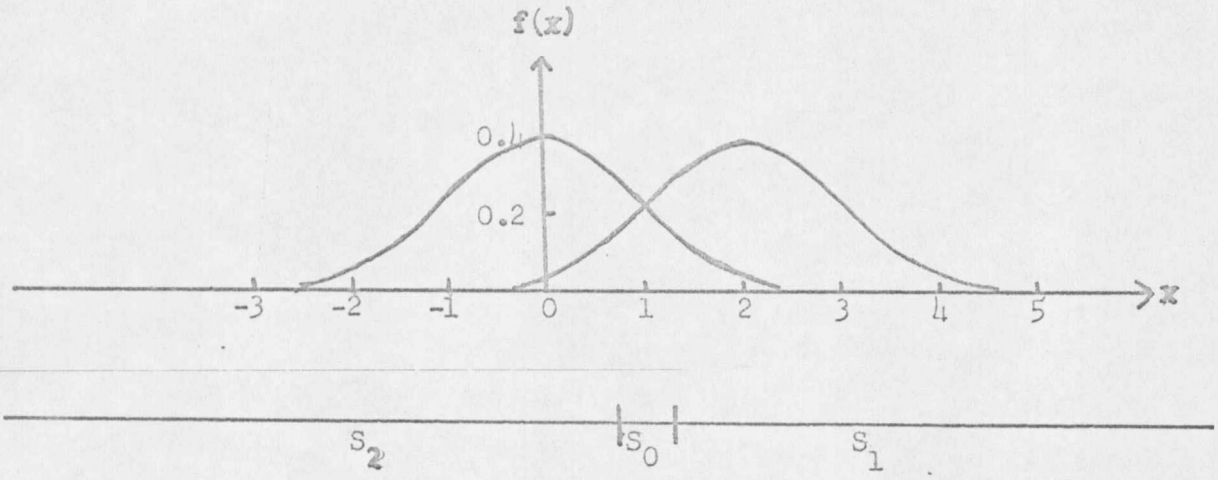

Figure 2.1: The size-(.1,.1) discrimination procedure $\Delta(c_1,c_2)$ for P_1:N(2,1) and P_2:N(0,1).

In this case, $c_1 < c_2$ and therefore this procedure is most powerful for discrimination at level-(.1,.1). The power for discrimination of this procedure is $(.86)\pi_1 + (.86)\pi_2 = .86$, and is independent of π_1 and π_2. Also any procedure which is most powerful at level-(.1,.1) will have exact size-(.1,.1) and agree with $\Delta(c_1,c_2)$ a.e. λ.

(ii) Let P_1 and P_2 be two univariate normal distributions with common variance 1 and means 3 and 0, respectively. As in (i) the selection of the sets A_1 and A_2 which are used to construct $\Delta(c_1,c_2)$ can be based on x. Let $\alpha = \beta = .1$; in this case, the constants are $c_1 = 1.718$ and $c_2 = 1.282$. The procedure $\Delta(c_1,c_2)$ is defined by the sets

$$S_1 = \{x;\ x \geq 1.718\} ,$$
$$S_2 = \{x;\ x \leq 1.282\} ,$$
$$S_0 = \{x;\ 1.282 < x < 1.718\} .$$

See Figure 2.2 (a).

This is an example of a situation when $c_1 > c_2$. For reasonably small α and β this situation occurs only if the distributions are quite distant. See Figure 2.2 (a).

The curves representing f_1 and f_2 intersect at x = 1.5, which is contained in the interval (1.282, 1.718). Consider a discrimination procedure $\Delta'(S_1',S_2')$ where

$$S_1' = \{x;\ 1.5 < x < 4.838\} ,$$
$$S_2' = \{x;\ -1.838 < x < 1.5\} ,$$
$$S_0' = \{x;\ x < -1.838\} \cup \{x;\ x > 4.838\} .$$

The procedure $\Delta'(S_1',S_2')$ is of exact size-(.067,.067). The

overall probability of no classification under $\Delta(c_1,c_2)$ is $\pi_1 P_1(S_0) + \pi_2 P_2(S_0) = .057\pi_1 + .057\pi_2 = .057$. The overall probability of no classification under $\Delta'(S_1',S_2')$ is $\pi_1 P_1(S_0') + \pi_2 P_2(S_0') = .033\pi_1 + .033\pi_2 = .033$. Thus $\Delta(c_1,c_2)$ is not most powerful for discrimination at level-(.1,.1). See Figure 2.2 (b).

However, a discrimination procedure can be constructed which has power one for discrimination and exact size-(.067,.067) by setting $c_1 = c_2 = 1.5$. Then the sets $S_1^* = \{x;\ f_1(x) > 1.5\ f_2(x)\}$ and $S_2^* = \{x;\ f_1(x) > 1.5\ f_2(x)\}$ define a discrimination procedure which has the desired properties. See Figure 2.2 (c).

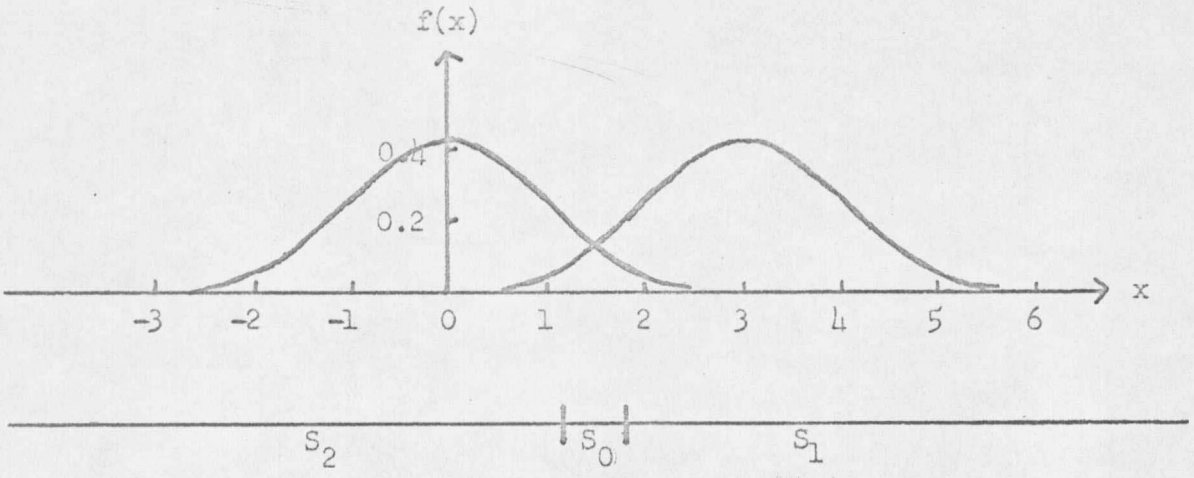

Figure 2.2 (a): The size-(.1,.1) discrimination procedure $\Delta(c_1,c_2)$ for P_1: N(3,1) and P_2: N(0,1).

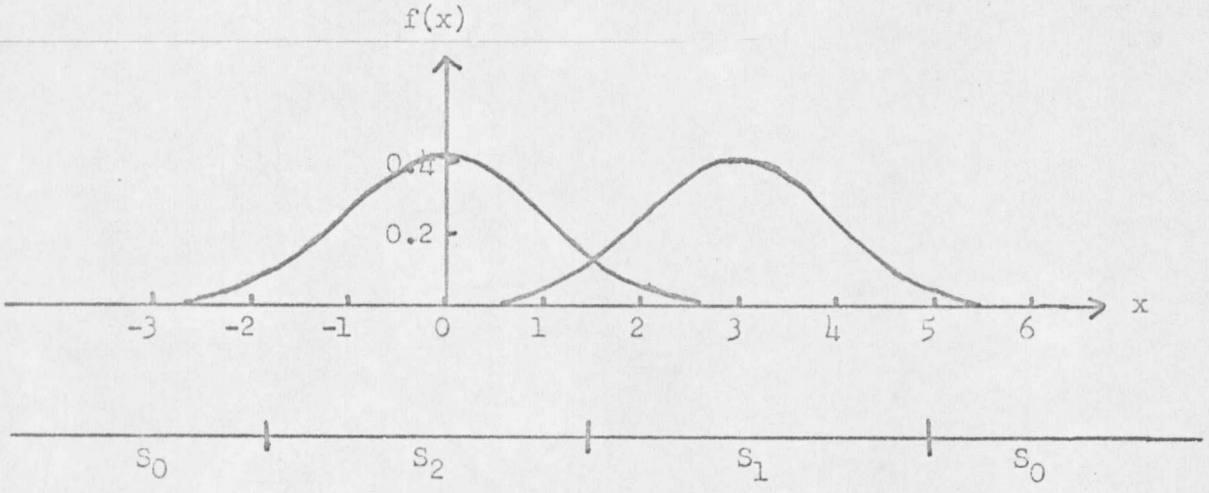

Figure 2.2 (b): A discrimination procedure of size-(.1,.1) with power for discrimination greater than the procedure in Figure 2.2 (a).

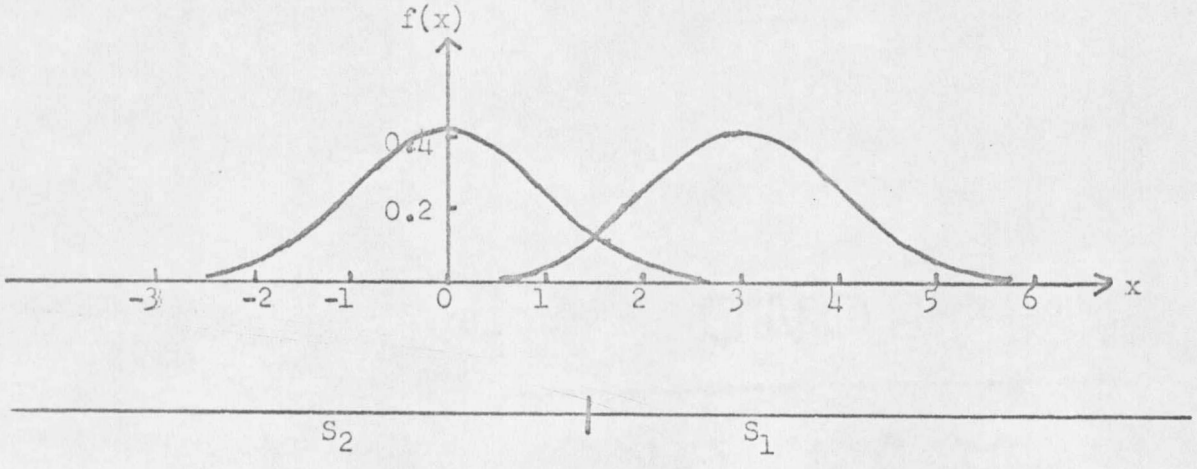

Figure 2.2 (c): A discrimination procedure with power one for discrimination and size-(.1,.1).

Example 2.2: Let $(P_\theta; \theta \in \Omega)$ be the family of univariate normal distributions where θ is the vector (μ, σ^2). Let $P_{\theta_1} = P_1$ and $P_{\theta_2} = P_2$ be any two elements of this family with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively. The sets A_1 and A_2 defined in Theorem 2.1 (i) are determined by the values of the ratio $f_1(x)/f_2(x)$ where f_1 and f_2 are the probability density functions of P_1 and P_2, respectively. For this family, the ratio has the form

$$f_1(x)/f_2(x) = k \exp\{ax^2 + bx + c\}.$$

This ratio is strictly monotone in the exponent $ax^2 + bx + c$, which is a parabola; thus the ratio f_1/f_2 is strictly monotone from $-\infty$ to the value of the vertex for this parabola and strictly monotone in the opposite sense from this value to $+\infty$. The vertex of the parabola is at the point

$$x_0 = -\frac{b}{2a} = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2}{\sigma_1^2 - \sigma_2^2}$$

If $a = \frac{\sigma_1^2 - \sigma_2^2}{2\sigma_1^2\sigma_2^2}$ is negative, the value of the ratio increases strictly from $-\infty$ to x_0 and decreases strictly from x_0 to $+\infty$. Let A_1 be the complement of a closed interval centered at x_0 and having probability $1 - \alpha$ under P_1; while the set A_2 is an open interval centered at x_0 and having probability β under P_2. If a is positive, the value of the ratio decreases strictly from $-\infty$ to x_0 and increases strictly from x_0 to $+\infty$. Let the set A_1 be an open interval centered at x_0 and having probability α under P_1; while the set A_2 is the complement of a closed interval centered at x_0 and having probability $1 - \beta$ under P_2. In either case form the discrimination procedure $\Delta(c_1, c_2)$ constructed from

A_1 and A_2 as in (2). If $a = 0$, the problem reduces to that already discussed in Example 2.1. These notions will be illustrated with an explicit example:

Let P_1 be a univariate normal distribution with mean 0 and variance 1; let P_2 be a univariate normal distribution with mean 2 and variance 25. Let $\alpha = \beta = .1$. In this case, the ratio of densities is

$$f_1(x)/f_2(x) = 5 \exp\left\{-\frac{2}{25}(6x^2 + x - 1)\right\} .$$

This ratio increases strictly from $-\infty$ to -0.083 and decreases strictly from -0.083 to $+\infty$. In order to find A_1, we take -0.083 as center of an interval with probability .9 under P_1 and A_1 will be the complement of this interval. The set A_2 will be the interval centered at -0.083 which has the required probability under P_2. These sets are as follows:

$$A_1 = \{x;\ x < -1.733\} \cup \{x;\ x > 1.567\} ,$$

$$A_2 = \{x;\ -0.768 < x < 0.602\} .$$

The discrimination procedure $\Delta(c_1, c_2)$ defines the following sets:

$$S_1 = \{x;\ -0.768 < x < 0.602\} ,$$

$$S_2 = \{x;\ x < -1.733\} \cup \{x;\ x > 1.567\} ,$$

$$S_0 = \{x;\ -1.733 \leq x \leq -0.768\} \cup \{x;\ 0.602 \leq x \leq 1.567\} .$$

In this case $c_1 < c_2$ and this procedure is most powerful for discrimination at level-(.1,.1). Any procedure which is most powerful for discrimination at level-(.1,.1) will have exact size-(.1,.1) and will agree with $\Delta(c_1, c_2)$ a.e. λ. See Figure 2.3.

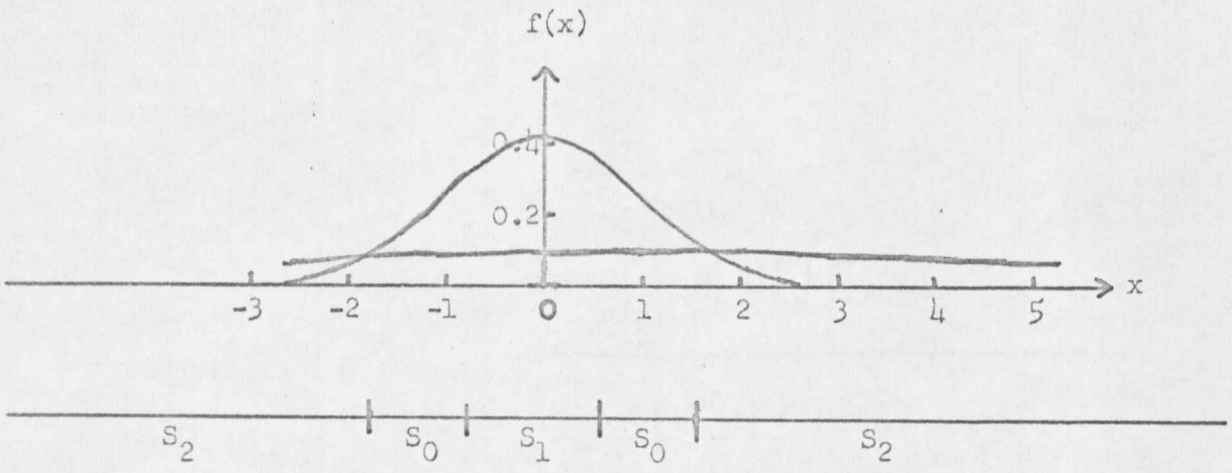

Figure 2.3: The size-(.1,.1) discrimination procedure $\Delta(c_1,c_2)$ for P_1: N(0,1) and P_2: N(2,25).

Example 2.3: Let P_1 and P_2 be a k-variate normal probability distributions with the common dispersion matrix Σ and the distinct mean vectors $\mu^{(1)}$ and $\mu^{(2)}$, respectively. The ratio of densities in this case reduces to a linear function of the k-variate observation z. This random variable U(z) has a univariate normal distribution whether z is distributed as a P_1 random variable or as a P_2 random variable, and this problem reduces to that of Example 2.1. Details can be found in Anderson (1958).

CHAPTER III

A NONPARAMETRIC DISCRIMINATION PROBLEM

3.1 A Nonparametric Discrimination Model

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two distinct populations. A random variable X is defined on the k-dimensional Euclidean space, R^k, by measurements made on elements of $\mathcal{P}_1$, and a random variable Y is defined on R^k by the corresponding measurements made on elements of $\mathcal{P}_2$. The random variables X and Y are distributed over R^k according to distinct absolutely continuous probability distributions P_1 and P_2, respectively. With P_1 and P_2 are associated absolutely continuous distribution functions F_1 and F_2, respectively; also the assumption of absolute continuity guarantees the existence of probability density functions f_1 and f_2, respectively, with respect to Lebsgue measure λ on R^k. Nothing is assumed about P_1 or P_2 except their absolute continuity and the resulting existence of probability density functions.

Let $x_1, x_2, \ldots, x_m$ be a simple random sample of observations from the distribution of the random variable X and $y_1, y_2, \ldots, y_n$ a simple random sample of observations from the distribution of the random variable Y. An individual from a population consisting of elements from $\mathcal{P}_1$ and $\mathcal{P}_2$ is observed. The k-dimensional measurements are made on the individual, which then becomes an observation z on either the random variable X or the random variable Y. On the basis of the observation z a decision is to be made whether to classify z as a P_1 random variable, to classify z as a P_2 random variable, or to reserve judgment, i.e. to make no classification. A discrimination procedure Δ will be a function from the space of the observations, namely R^k, into the decision space

$D = (d_1, d_2, d_0)$, which was defined in Chapter II. A procedure will be identified by the notation $\Delta(S_1, S_2)$, where S_1 and S_2 are the sets defined in (1).

If a discrimination procedure $\Delta(S_1, S_2)$ is used, the overall probability of misclassification would be $\pi_1 P_1(S_2) + \pi_2 P_2(S_1)$, where π_1 and π_2 are the a priori probabilities that an observation comes from P_1 and P_2, respectively. Let the size of a discrimination procedure be defined as in Definition 2.1. If Definitions 2.3, 2.4, and 2.5 are accepted for this model and arbitrary but fixed values $\alpha \in (0,1)$ and $\beta \in (0,1)$ are chosen, the discrimination problem is defined as follows: from the class of all size-(α, β) discrimination procedures find a procedure, if it exists, which is most powerful for discrimination at level-(α, β). If the probability density functions satisfy the hypotheses of Theorem 2.1, such a discrimination procedure is known to exist and it will be assumed in the following discussion that such a procedure does exist.

Under the nonparametric model the statistic $(x_1, \ldots, x_m; z; y_1, \ldots, y_n)$ is sufficient for the discrimination. If the probability distributions and their probability density functions are completely known, a discrimination procedure which is most powerful for discrimination at a given level based on this knowledge would be at least as powerful for discrimination at this level as a procedure chosen only on the basis of the statistic. Although this is intuitively obvious, it is quite simple to prove, if a proof is desired (cf. Fix and Hodges (1951)). If the probability distributions are completely known, the observation z is sufficient for the discrimi-

nation. The performance characteristics of any procedure based on $(x_1,\ldots,x_m;z;y_1,\ldots,y_n)$ can (by using randomization, if required) be exactly duplicated with a procedure based on z. Thus one can not find a discrimination procedure based on the statistic $(x_1,\ldots,x_m;z;y_1,\ldots,y_n)$ which is more powerful for discrimination at a given level than a procedure which is most powerful for discrimination based on z alone. If the probability distributions are completely known and a discrimination procedure based on z is found which is most powerful for discrimination at a given level, then it is optimal in this sense. If one does not know the probability distributions completely, then an attempt should be made to arrive at a discrimination procedure based on the statistic $(x_1,\ldots,x_m;z;y_1,\ldots,y_n)$ which approximates a procedure which is most powerful for discrimination at the desired level when both distributions are completely known.

In Theorem 2.1 the form of a discrimination procedure which is most powerful for discrimination at a given level is given. In order to construct this procedure, sets A_1 and A_2 were found such that $P_1(A_1) = \alpha$ and $P_2(A_2) = \beta$ where $A_1 = \{ x;\ f_1(x) < c_1 f_2(x)\}$ and $A_2 = \{ x;\ f_1(x) > c_2 f_2(x)\}$ where c_1 and c_2 are positive constants. The first step in constructing from the statistic a discrimination procedure of size-(α, β) is to look for a method of constructing measurable sets with approximately the required probabilities under P_1 and P_2. As long as the distributions are completely known, constructing such sets presents only computational difficulties. Under the nonparametric model, however,

constructing such sets has in addition an element of uncertainty due to lack of knowledge concerning the probability distributions. The sets, based on the statistic $(x_1,\ldots,x_m;z;y_1,\ldots,y_n)$, are random quantities and their probabilities are random variables. It is no longer possible to speak about sets whose exact probabilities are known; thus only discrimination procedures whose size approximates (α, β) can be hoped for. If sets A_1 and A_2 could be constructed in such a way that as m and n grow large without bound the probability of A_1 under P_1 approaches α and the probability of A_2 under P_2 approaches β, this would give a method with strong appeal for the construction of approximate size-(α, β) discrimination procedures when large samples are available.

The actual construction of such sets which will be suggested in this paper is based on the theory of coverages. In the following section the definitions and theorems necessary for this construction will be given.

3.2 Some Results Concerning Order Statistics and the Theory of Coverages

In this section only the definitions and results from the theory of coverages necessary for the construction will be given. For detailed treatment and proofs the reader will be referred to Wilks (1962), Tukey (1947), and Kemperman (1956).

Let $x_1,x_2,\ldots,x_m$ be a sample from a univariate probability distribution with continuous distribution function $F(x)$. If the sample is ordered by numerical value, we obtain the order statistics $\{x_{(1)}, x_{(2)},\ldots,x_{(m)}\}$ of the sample, where $x_{(1)} \leq \ldots \leq x_{(m)}$. The assumption of continuity then gives a partition of R^1 into sets $(-\infty, x_{(1)}]$,

$(x_{(1)},x_{(2)}]$,..., $(x_{(m)},+\infty)$ with probability one. These sets will be called sample blocks and will be denoted by $B_1^{(i)}$ for $i = 1,2,\ldots,m+1$ where the subscript 1 means the random variable is one-dimensional. The probability of the block $B_1^{(i)}$ will be called a coverage and denoted by c_i, i.e. $c_1 = Pr(B_1^{(1)}) = F(x_{(1)})$, $c_2 = Pr(B_1^{(2)}) = F(x_{(2)}) - F(x_{(1)})$,..., $c_{m+1} = Pr(B_1^{(m+1)}) = 1 - F(x_{(m)})$. Now these coverages are random variables and the distribution of a sum of any r of the m + 1 coverages, which is a result of the theory of the Dirichlet distribution, is given by the following theorem:

Theorem 3.1: The sum of any r of the coverages $c_1,\ldots,c_{m+1}$ has the beta distribution with parameters r and m - r + 1.

Note that a sum of r coverages is just the probability of the union of the sample blocks which determine the coverages. For a proof of this theorem, see Wilks (1962).

The theory of coverages for one-dimensional random variables also leads to some results in large sample distribution theory for the order statistics which will be used later in this chapter. The pertinent results are the following:

Theorem 3.2: If $\{x_{(1)},\ldots,x_{(m)}\}$ are the order statistics of a sample from a continuous distribution function F(x), and np_n is an integer such that $p_n = p + o(1/n)$, where $0 < p < 1$, then for large n,

$F(x_{(np_n)})$ is asymptotically distributed according to $N(p, p(1 - p)/n)$.

Corollary 3.1: If x_p^* is a unique solution of $F(x) = p$, then $x_{(np_n)}$ converges in probability to x_p^*.

For proofs of these results, see Wilks (1962).

The theory of coverages generalizes directly to multi-dimensional random variables and gives a generalization of Theorem 3.1. Let $x_1, x_2, \ldots, x_m$ be a sample from a k-dimensional probability distribution with continuous distribution function. Let $\varphi_1, \varphi_2, \ldots, \varphi_m$ be a sequence of real-valued functions defined in R^k such that $\varphi_1(x)$, $\varphi_2(x), \ldots, \varphi_m(x)$ have a continuous joint distribution function. The sample $x_1, \ldots, x_m$ and the functions $\varphi_1, \ldots, \varphi_m$ define sample blocks as follows:

$$B_k^{(1)} = \{x;\ \varphi_1(x) > a_1\}$$

where $a_1 = \max_i \varphi_1(x_i) = \varphi_1(x_{i(1)})$.

$$B_k^{(2)} = \{x;\ \varphi_1(x) < a_1,\ \varphi_2(x) > a_2\}$$

where $a_2 = \max_{i \neq i(1)} \varphi_2(x_i) = \varphi_2(x_{i(2)})$, $i(2) \neq i(1)$.

And in general, for $2 \leq p \leq m$,

$$B_k^{(p)} = \{x;\ \varphi_1(x) < a_1, \ldots, \varphi_{p-1}(x) < a_{p-1},\ \varphi_p(x) > a_p\}$$

where $a_p = \max \varphi_p(x_i) = \varphi_p(x_{i(p)})$, the maximum being taken over all i except $i(1), i(2), \ldots, i(p - 1)$; and $i(p)$ being

chosen distinct from all $i(j)$, $j < p$.

$$B_k^{(m+1)} = \{x;\ \varphi_1(x) \quad a_1, \ldots,\ \varphi_m(x) < a_m\}.$$

The coverages $c_1, \ldots, c_{m+1}$ are defined to be the probabilities of these blocks, i.e. $c_1 = \Pr(B_k^{(1)}), \ldots, c_{m+1} = \Pr(B_k^{(m+1)})$. From the assumption of continuity we know that R^k is partitioned into $m + 1$ blocks by the above procedure with probability one. The generalization of Theorem 3.1 is as follows:

Theorem 3.3: If blocks and coverages are defined as above, the sum of any r of the k-dimensional coverages $c_1, \ldots, c_{m+1}$ has the beta distribution with parameters r and $m - r + 1$.

The proof of this theorem can be found in Tukey (1947).

It has been shown by J. H. B. Kemperman (1956) that it is not necessary to have a fixed sequence of functions, since it is possible, without changing the results of Theorem 3.3, to have the choice of function at any point in the sequence depend on the observed values at the cuts of the functions already used, i.e. the values $\varphi_p(x_{i(p)})$ for $p = 1, \ldots, m$.

3.3 A Construction Based on the Theory of Coverages for Sets A_1 and A_2

With the theory of coverages in hand it is possible to construct sets whose probability under P_1 converges in probability to α. A similar approach can be used to obtain sets whose probability under P_2 converges in probability to β.

Let $a_1 = [\alpha(m+1)]$, an integer; since

$$\frac{a_1}{m+1} = \alpha - \frac{\alpha(m+1) - [\alpha(m+1)]}{m+1} = \alpha - o(1/m),$$

it can be seen that as $m \to \infty$, $a_1/(m+1) \to \alpha$.

For $x_1, x_2, \ldots, x_m$ a sample from the distribution of the random variable X and $\varphi_1, \ldots, \varphi_m$ a sequence of ordering functions defined on R^k with continuous joint distribution function, the sample blocks and the corresponding coverages $c_1, \ldots, c_{m+1}$ can be formed as in the preceding section. If a sum of a_1 of these coverages is taken, this sum is distributed as a beta random variable with a_1 and $m - a_1 + 1$ degrees of freedom. Let A_1 be the union of sample blocks associated with these coverages. Then $P_1(A_1)$ is a random variable with a beta distribution having a_1 and $m - a_1 + 1$ degrees of freedom. The mean of this distribution is $a_1/(m+1)$, which has been shown to converge to α. It follows from a simple application of Chebyshev's inequality that $P_1(A_1)$ converges in probability to α. For small m we can use the beta distribution to set an upper bound on $P_1(A_1)$ with given probability.

Let A_1 be a union of a_1 of the sample blocks. Let A_2 be a union of $a_2 = [\beta(n+1)]$ sample blocks formed from the sample $y_1, \ldots, y_n$. Asymptotically these choices will give us sets whose probabilities approximate those that are required under each distribution. Note that any choices of A_1 and A_2 constructed in this manner would give an approximate size-(α, β) discrimination procedure for large m and n. But some choices give more powerful procedures than others. This is obvious, since choices made without consideration of the other distri-

bution are not likely to give a procedure which is most powerful for discrimination. In Section 3.1 it was suggested that an attempt should be made to approximate an optimal procedure by procedures based on the sufficient statistic. As a means for this approximation, the notion of the consistency of sequences of discrimination procedures will be introduced in the next section.

3.4 Consistency of Sequences of Discrimination Procedures

Let $\Delta_{m,n}$ be a discrimination procedure based on the statistic $(x_1,\ldots,x_m;z;y_1,\ldots,y_n)$. For $m = 1,2,\ldots$ and $n = 1,2,\ldots$ let $(\Delta_{m,n})$ be a sequence of discrimination procedures. If $(\Delta^*_{m,n})$ is another sequence of discrimination procedures, the question arises of how the corresponding members of these sequences agree or are consistent with each other. One requirement could be that as m and n grow large without bound the probability that $\Delta_{m,n}(z) \neq \Delta^*_{m,n}(z)$ should converge in probability to zero under both distributions. This requirement will form the definition of consistency of the two sequences which will be used in this paper.

Definition 3.1: Two sequences $(\Delta_{m,n})$ and $(\Delta^*_{m,n})$ of discrimination procedures are said to be consistent for the discrimination problem if as $m\to\infty$ and $n\to\infty$, $\Pr\{\Delta_{m,n} \neq \Delta^*_{m,n}\}$ converges in probability to zero under both P_1 and P_2.

It is the consistency of a sequence of discrimination procedures with the procedure $\Delta(c_1,c_2)$ which is of particular interest.

Definition 3.2: A sequence of discrimination procedures $(\Delta_{m,n})$ is said to be consistent with $\Delta(c_1,c_2)$ for the discrimination problem if as $m\to\infty$ and $n\to\infty$, $\Pr\{\Delta_{m,n} \neq \Delta(c_1,c_2)\}$ converges in probability to zero under both P_1 and P_2.

The problem of finding discrimination procedures which are consistent with $\Delta(c_1,c_2)$ for the nonparametric model introduced in Section 3.1 is not determinate. In the following section some restricted models will be considered and sequences of discrimination procedures which are consistent with $\Delta(c_1,c_2)$ will be found for these models. Where this does not appear to be possible, discrimination procedures based on the theory of coverages which have some intuitive appeal will be suggested.

3.5 Examples of Discrimination Procedures Constructed by Means of the Theory of Coverages

In Examples 3.1 and 3.2 restricted models for the discrimination problem will be considered, and for these models a sequence of procedures consistent with $\Delta(c_1,c_2)$ will be found.

Example 3.1: For θ a real-valued parameter, let $(P_\theta;\theta\in\Omega)$ be a family of absolutely continuous, univariate probability distributions which has strictly monotone (increasing) density ratio in a continuous real-valued function $T(x)$. Suppose $\theta>\theta'$ and denote f_θ and $f_{\theta'}$ by f_1 and f_2, respectively. If the density ratio f_1/f_2 satisfies the hypotheses of Theorem 2.1, there is a discrimination procedure which is most powerful for discrimination between P_1 and P_2 at level-(α,β) for arbitrary

$\alpha \in (0,1)$ and $\beta \in (0,1)$; and in the following discussion the existence of such a procedure is assumed.

From Example 2.1 it is known that a procedure equivalent to $\Delta(c_1,c_2)$ is defined by the sets $A_1^* = \{t;\ t < t_\alpha^*\}$ and $A_2^* = \{t;\ t > t_{1-\beta}^*\}$ where $P_1^*(A_1^*) = \alpha$ and $P_2^*(A_2^*) = \beta$. It was also noted that t_α^* was the α^{th} percentile of the probability distribution P_1^* and $t_{1-\beta}^*$ was the $(1-\beta)^{\text{th}}$ percentile of the probability distribution P_2^*.

Let $x_1, x_2, \ldots, x_m$ be a sample of observations on the random variable X and $t_1, t_2, \ldots, t_m$ the corresponding sample on the random variable T(x). The order statistics of the latter sample are $t_{(1)}, t_{(2)}, \ldots, t_{(m)}$.

If $A_1^* = \{t;\ t < t_\alpha^*\}$ and an attempt to approximate the procedure $\Delta(c_1,c_2)$ is to be made, then an obvious choice of the $a_1 = [\alpha(m+1)]$ blocks required to form the set A_1 would be the union of those blocks determined by the a_1 least order statistics, i.e. let

$$A_1 = \bigcup_{i=1}^{a_1} {}^{X}B_1^{(i)} = \{t;\ t \leq t_{(a_1)}\}.$$

Since $A_2^* = \{t;\ t > t_{1-\beta}^*\}$, the corresponding choice of A_2 would be the $a_2 = [\beta(n+1)]$ blocks determined by the a_2 largest order statistics from the second sample on T(x), i.e. let

$$A_2 = \bigcup_{i=n-a_2+1}^{n+1} {}^{Y}B_1^{(i)} = \{t;\ t > t_{(n-a_2+1)}\}.$$

For each m and n these choices define a discrimination procedure which

will be denoted by $\Delta_{m,n}$. The sequence $(\Delta_{m,n})$ of discrimination procedures will be shown to be consistent with $\Delta(c_1,c_2)$.

From Definition 3.2, it must be shown that $\Pr\{\Delta_{m,n} \neq \Delta(c_1,c_2)\}$ converges in probability to zero under both P_1 and P_2 as $m \to \infty$ and $n \to \infty$. The sets which define the discrimination procedure $\Delta_{m,n}$ depend on the values $t_{(a_1)}$ and $t_{(n - a_2 + 1)}$; the sets which define the discrimination procedure $\Delta(c_1,c_2)$ depend on the values t^*_{α} and $t^*_{1-\beta}$. The differences in the values of the two procedures arise from any difference between t^*_{α} and $t_{(a_1)}$ and any difference between $t^*_{1-\beta}$ and $t_{(n - a_2 + 1)}$. An examination of all the possibilities shows that the probability under P_1 that $\Delta_{m,n} \neq \Delta(c_1,c_2)$ is at most

$$U = \left| F_1^*(t_{(n - a_2 + 1)}) - F_1^*(t_{1-\beta}) \right| + \left| F_1^*(t_{(a_1)}) - F_1^*(t^*_{\alpha}) \right|,$$

where F_1^* is the distribution function of $T(x)$ induced by the probability distribution P_1. Note that F_1^* is continuous since $T(x)$ and $F_1(x)$ are continuous. From Theorem 3.2, $t_{(n - a_2 + 1)}$ converges in probability to $t^*_{1-\beta}$ as $n \to \infty$. Since F_1^* is a continuous function, by a well-known result $F_1^*(t_{(n - a_2 + 1)})$ converges in probability to $F_1^*(t^*_{1-\beta})$ as $n \to \infty$. Similarly, $F_1^*(t_{(a_1)})$ converges in probability to $F_1^*(t^*_{\alpha})$ as $m \to \infty$.

Therefore, U converges in probability to zero as $m \to \infty$ and $n \to \infty$. Since $0 \leq \Pr\{\Delta_{m,n} \neq \Delta(c_1,c_2)\} \leq U$, the $\Pr\{\Delta_{m,n} \neq \Delta(c_1,c_2)\}$ converges in

probability to zero as $m \to \infty$ and $n \to \infty$ under P_1. Similarly, this probability converges to zero as $m \to \infty$ and $n \to \infty$ under P_2. By definition, $(\Delta_{m,n})$ is consistent with $\Delta(c_1, c_2)$.

If $c_1 \leq c_2$ or equivalently $t^*_{\alpha} \leq t^*_{1-\beta}$, the sequence of discrimination procedures which has just been defined is consistent with a procedure which is most powerful for discrimination at level-(α, β). The estimates $t_{(a_1)}$ and $t_{(n - a_2 + 1)}$ are consistent for t^*_{α} and $t^*_{1-\beta}$ respectively. If $t_{(a_1)} \leq t_{(n - a_2 + 1)}$, then the procedure just defined should be used. When $t_{(a_1)} > t_{(n - a_2 + 1)}$, select a value $t^* \in \left[t_{(n - a_2 + 1)}, t_{(a_1)} \right]$ and define the procedure with power one for discrimination based on the value t^*. The actual choice of the value t^* will depend on the desires of the investigator.

Example 3.2: Let $(P_\theta; \theta \in \Omega)$ be the family of all univariate normal distributions, where θ is the vector (μ, σ^2). Let $P_{\theta_1} = P_1$ and $P_{\theta_2} = P_2$ be any two elements of this family with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively. The arbitrary levels $\alpha \in (0,1)$ and $\beta \in (0,1)$ are given.

A sample $x_1, x_2, \ldots, x_m$ of observations from the distribution P_1 of the random variable X and a sample $y_1, y_2, \ldots, y_n$ of observations from the distribution P_2 of the random variable Y are in hand. The sample blocks and coverages can be used, as outlined in this chapter, to construct discrimination procedures. For each ordered pair of integers (m,n) it is desired to choose a discrimination procedure $\Delta_{m,n}$ in

such a way that the sequence $(\Delta_{m,n})$ is consistent with $\Delta(c_1,c_2)$.

When the probability distributions are completely known, Example 2.2 shows that the choices of the sets A_1 and A_2 to construct $\Delta(c_1,c_2)$ depend on

(a) the sign of $a = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}$,

(b) the value of $x_0 = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2}{\sigma_1^2 - \sigma_2^2}$,

(c) the arbitrary levels $\alpha\in(0,1)$ and $\beta\in(0,1)$.

The levels have been fixed. The sample means and variances $\bar{x}_1, \bar{x}_2$, s_1^2, and s_2^2 converge in probability to the parameters μ_1, μ_2, σ_1^2, and σ_2^2, respectively, as $m\to\infty$ and $n\to\infty$. Thus, the values of x_0 and a have estimators $\hat{x}_0$ and $\hat{a}$ which converge in probability to these parameters as $m\to\infty$ and $n\to\infty$.

Let m and n be arbitrary but fixed positive integers. The sample blocks and their coverages are formed as in Section 3.2. Suppose $\hat{a} > 0$. Then take B_1 to be the interval $(x_{(r_1)}, x_{(r_2)}]$, i.e. the interval determined by the r_1^{st} and r_2^{nd} order statistics of the first sample subject to the following conditions:

(d) $r_2 - r_1 = a_1$,

(e) the quantity $\left|(x_{(r_2)} - \hat{x}_0) - (\hat{x}_0 - x_{(r_1)})\right|$ is a minimum.

That is, take B_1 to be the union of the a_1 consecutive sample blocks for which $\left|x_{(r_2)} + x_{(r_1)} - 2\hat{x}_0\right|$ is a minimum. If $x_{(1)} < \hat{x}_0 < x_{(m)}$, this requires that $(x_{(r_1)}, x_{(r_2)}]$ be the interval containing $\hat{x}_0$ for which

the difference of the distances of the end-points from $\hat{x}_0$ is a minimum. Let B_2 be the complement of an interval $\overline{B}_2 = (y_{(r_3)}, y_{(r_4)}]$, where $\overline{B}_2$ is a set of $n - a_2 + 1$ sample blocks and $y_{(r_3)}$ and $y_{(r_4)}$ are order statistics which satisfy (e) for the sample $y_1, y_2, \ldots, y_n$.

The sets B_1 and B_2 can be used to construct a discrimination procedure as in (2); let this procedure be denoted by $\Delta_{m,n}$. The sequence $(\Delta_{m,n})$ of discrimination procedures will be shown to be consistent with $\Delta(c_1, c_2)$.

If $a > 0$, the set A_1 defined in Theorem 2.1 (i) is an open interval centered at x_0 and having probability α under P_1; let $A_1 = (x_1, x_2)$. Since P_1 is a univariate normal distribution, the set A_1 is unique and characterized by the following properties:

(f) $F_1(x_2) - F_1(x_1) = \alpha$,

(g) $x_1 + x_2 = 2x_0$.

The set A_2 is the complement of a closed interval centered at x_0 and having probability $1 - \beta$ under P_2; let $A_2 = (-\infty, y_1) \cup (y_2, \infty)$. This set is unique and $\overline{A}_2$ is characterized by properties similar to (f) and (g).

From Definition 3.2, in order to show the required consistency, it must be demonstrated that $\Pr\{\Delta_{m,n} \neq \Delta(c_1, c_2)\}$ converges in probability under both P_1 and P_2. It will suffice to show that the end-points of A_1 and B_1 coincide and that the end-points of A_2 and B_2

coincide. The former will be shown here, and the latter follows in a similar fashion.

Since B_1 is a union of a_1 sample blocks, it has been shown that $P_1(B_1)$ converges in probability to α as $m \to \infty$.

Set $e_j = F_1(x_j)$ and $u_j = [e_j(m+1)]$ for $j = 1,2$. Now $\alpha = e_2 - e_1$, so that $\alpha(m+1) = e_2(m+1) - e_1(m+1)$. Therefore, $u_2 = a_1 + u_1 + \omega$ for $\omega = 0$ or 1. Also, $x_{(u_j)}$ converges in probability to x_j for $j = 1,2$. Since $\hat{x}_0$ converges in probability to x_0, it can be shown that

$$x_{(u_2)} + x_{(u_1)} - 2\hat{x}_0 = (x_{(u_2)} - \hat{x}_0) - (\hat{x}_0 - x_{(u_1)})$$

converges in probability to zero. But from (e),

$$0 \leq \left| x_{(r_2)} + x_{(r_1)} - 2\hat{x}_0 \right| \leq \left| x_{(u_2)} + x_{(u_1)} - 2\hat{x}_0 \right|$$

since $u_2 - u_1$ converges to a_1 as $m \to \infty$. Therefore, $x_{(r_2)} + x_{(r_1)} - 2\hat{x}_0$ converges in probability to zero; or, equivalently, $x_{(r_2)} + x_{(r_1)}$ converges in probability to $2x_0$.

From the above discussion, the properties (f) and (g), and the properties of a univariate normal distribution, it follows that $x_{(r_j)}$ converges in probability to $x_{(j)}$ for $j = 1,2$, as was to be shown.

Since $\hat{a}$ is a consistent estimator for a, the probability of error from a difference in sign between $\hat{a}$ and a converges in probability to zero. Although throughout this discussion it was assumed that $\hat{a} > 0$,

a similar result can be obtained when $\hat{a} < 0$. If $\hat{a} = 0$, proceed as in Example 3.1. For this restricted model a sequence of discrimination procedures consistent with $\Delta(c_1, c_2)$ has been found.

Example 3.3: If no attempt is made to find a discrimination procedure consistent with an optimal procedure, then procedures with some appeal can still be constructed by the method suggested in this chapter. This is so, for instance, when differences of means are involved.

The type of region used in this example is felt to have some appeal for a class of unimodal, bivariate distributions which differ in means. The data used to illustrate these regions were drawn from two bivariate normal distributions as follows:

$$P_1\colon\ N\left((0,0),\begin{bmatrix}1 & 0\\ 0 & 4\end{bmatrix}\right),\quad P_2\colon\ N\left((3,0),\begin{bmatrix}1 & 2\\ 2 & 9\end{bmatrix}\right).$$

Sample sizes were $m = n = 81$.

It was decided to make $\overline{A}_1$ and $\overline{A}_2$ eight-sided regions in the plane; and the following sequence of ordering functions was chosen for $\underline{x} = (x_1, x_2)$ a general bivariate random variable: $\varphi_1(\underline{x}) = x_2$, $\varphi_2(\underline{x}) = x_1$, $\varphi_3(\underline{x}) = -x_2$, $\varphi_4(\underline{x}) = -x_1$, $\varphi_5(\underline{x}) = x_1 + x_2$, $\varphi_6(\underline{x}) = x_1 - x_2$, $\varphi_7(\underline{x}) = -x_1 - x_2$, $\varphi_8(\underline{x}) = -x_1 + x_2$. The sets S_1, S_2, S_0 defined by these functions and the drawn samples are sketched in Figure 3.1.

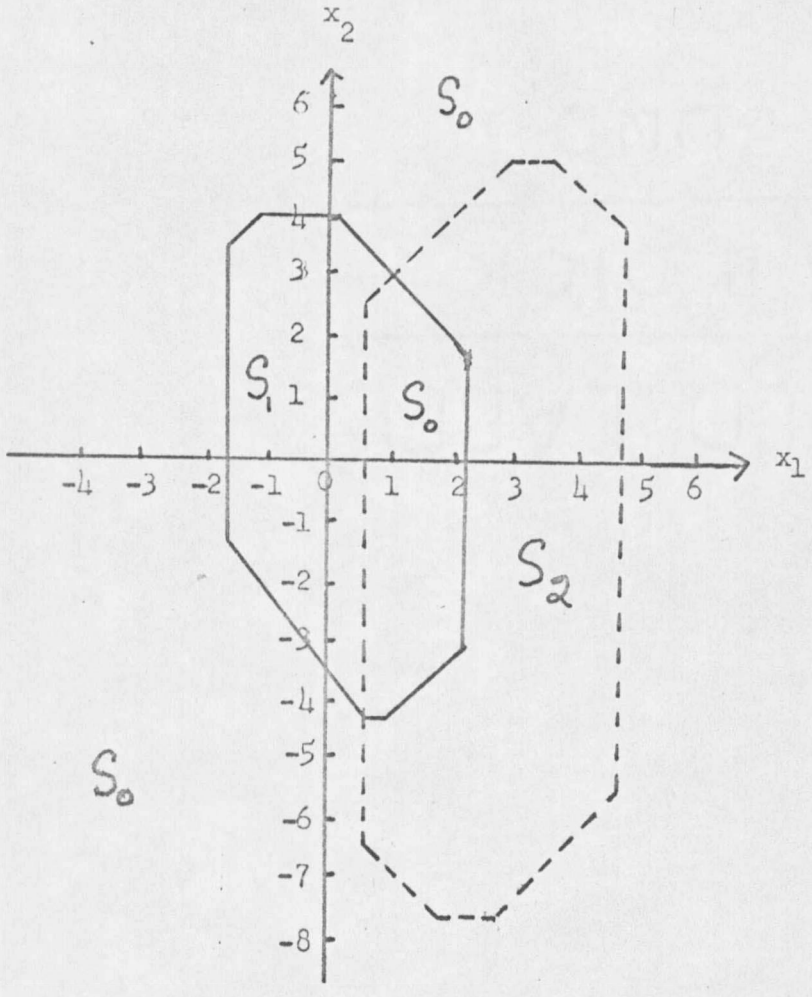

Figure 3.1: A discrimination procedure of approximate size-(.1,.1) based on samples of size 81 drawn from two bivariate normal distributions.

From the drawn samples, nineteen observations from the distribution P_1 and fourteen observations from the distribution P_2 are in the set S_0.

CHAPTER IV

SUMMARY

In Chapter I the discrimination problem for two probability distributions was introduced, and the problem was subdivided according to the knowledge in hand concerning the distributions.

The subproblem in which both distributions are completely known was considered in Chapter II. A discrimination model, in which the probability of misclassification under each distribution is controlled at a given level, was outlined. Such control on the probabilities of misclassification made it necessary to include in the sample space a point for reserving judgment, i.e. making no classification. The size of a discrimination procedure was defined in terms of the probabilities for that procedure of misclassification under each distribution. The overall probability of no classification for a procedure was defined to be the power for discrimination of that procedure. Among all discrimination procedures of a given size a procedure was sought which minimized this probability, i.e. a procedure which was most powerful for discrimination at the given level. If the ratio of the probability density functions was a continuous random variable which was defined a.e. λ, the discrimination problem under this model was solved in Theorem 2.1. Several examples were given to illustrate the concepts of this theorem.

In Chapter III a nonparametric discrimination model was outlined using the concepts of size and power as defined in Chapter II. A method, based on the theory of coverages, for constructing discrimination procedures whose size approximates a given size for large m and n was suggested. A definition of consistency for sequences of discrimi-

nation procedures was given and for restricted models a sequence of procedures consistent with an optimal procedure was found. The problem of finding such a sequence for the nonparametric problem is not determinate. When no such sequence is sought, a type of region which has strong appeal for unimodal distributions with different means was suggested.

LITERATURE CITED

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugen. 7 179.

Fix, E. and Hodges, J. L. Jr. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 4.

Kemperman, J. H. B. (1956). Generalized tolerance limits. Ann. Math. Stat., 27 180-186.

Kendall, M. G. (1966). Discrimination and classification. Multivariate Analysis: Proceedings of an International Symposium. Edited by Paruchuri R. Krishnaiah.

Tukey, J. W. (1947). Nonparametric estimation II. Statistically equivalent blocks and tolerance regions - the continuous case. Ann. Math. Stat. 18 529-539.

Welch, B. L. (1939). Note on discriminant functions. Biometrika 31 218-220.

Wilks, S. S. (1962). Mathematical Statistics. Wiley, New York.