ON SOME PROBLEMS IN CLASSIFICATION: CLASSIFIABILITY, ASYMPTOTIC RELATIVE EFFICIENCY, AND A COMPLETE CLASS THEOREM

by

Albert J. Kinderman*

Technical Report No. 178

May 1972

University of Minnesota
Minneapolis, Minnesota

*Submitted as a Thesis to the Faculty of the Graduate School of the University of Minnesota in partial fulfillment of the requirements for the degree of Doctor of Philosophy. This research was supported in part by an NDEA title IV fellowship and U.S. Army Grant DA-ARO-D-31-124-70-G102.

Albert John Kinderman
448 words

ABSTRACT

On Some Problems in Classification: Classifiability, Asymptotic Relative Efficiency, and a Complete Class Theorem

The problem of classifying one population into one of several alternative populations on the basis of random samples is viewed from various perspectives. Firstly, the relationship between classification and the works of W. Hoeffding and J. Wolfowitz (Ann. Math. Statist. 29 (1958) 700-718) on the distinguishability of sets of distributions is explored. A classification problem is specified by the set of possible distributions of the random variables sampled from the alternative distributions. The problem is said to be classifiable (finitely classifiable) if, for every positive ϵ, there exists a sequential classification rule (a finite sample size classification rule) such that the maximum probability of incorrectly classifying the given population is less than ϵ. The equivalence of the classifiability of a classification problem and the distinguishability of two particular sets of distributions generated by the alternatives is established. The results of Hoeffding and Wolfowitz are then applied to obtain necessary and sufficient conditions on the set of alternative distributions for classifiability and finite classifiability. A class of distance functions on the space of distributions is used to simplify these conditions. In particular, the Kolmogorov distance is used when the random variables take values in a Euclidean space to prove a problem is classifiable if, and only if, no two of the alternative distributions are identical. For finite classifiability, a sufficient condition is that the alternative distributions are uniformly separated.

Albert John Kinderman
448 words

Several examples of sequential classification rules are given. A class of minimum distance rules is described, as well as a class of rules based on the idea of sequential tests of power one as considered by H. Robbins (Ann. Math. Statist. 41 (1970) 1397-1409). Under certain conditions, each class contains rules sufficient to keep the probabilities of misclassification less than any prespecified limit.

Secondly, finite classification rules are compared through the computation of asymptotic relative efficiencies. In one example, univariate normal distributions are considered and a standard classification rule is shown to compare favorably with a two sample test for the equality of means. In a second example, classification rules based on linear rank statistics are shown to have the same asymptotic Pitman efficiency (relative to analogous rules based on sample means) as the corresponding two sample rank test for equality (relative to Student's t test).

Finally, classification is viewed from a decision theoretic perspective and a complete class theorem is proved when the distributions are univariate normal with a common known variance. An adaptation of the techniques of T. Matthes and D. Truax (Ann. Math. Statist. 38 (1967) 681-697) and M. L. Eaton (Ann. Math. Statist. 41 (1970) 1884-1888) is used to find an essentially complete class of rules relative to the class of all classification rules invariant under translations and change of sign.

Approved
Somesh Das Gupta
4 May, 1972

I dedicate this thesis

to

Susie

Acknowledgement

I gratefully acknowledge Professor Somesh Das Gupta for his guidance and encouragement. I also wish to thank Professors William Sudderth and Michael Perlman for many helpful discussions, and Mr. Gerald DuChaine for his skilled typing.

TABLE OF CONTENTS

Introduction

The problem of classification involves trying to correctly associate one population with one of several alternative populations. The mathematical model for classification is based on the assumption that the experimental data from each population are independent observations of some random variable. Thus, corresponding to each population, there is a random variable whose distribution may or may not be known. Usually, a correct classification is made when the two random variables assigned to the associated populations have the same distribution. In some cases, however, correctness of classification is based on the equality of means, variances, or some other feature of the distributions.

Various cases have been considered in the literature. For two alternatives, both of whose distributions are completely specified, B. Welch [29] found the Bayes rules. The case of several completely specified alternatives were considered by R. von Mises [22], who minimizes the maximum probability of error. More recently, J. Yao [31] set this same problem in a game-theoretic framework, found the Bayes rules, and proved the game has a value. In the case of two normal alternatives with unknown parameters, A. Wald [27] suggested substituting estimates for the actual parameters in any of several test statistics taken from the case of known parameters. Much work has since been done on finding the distributions of these statistics. A unified account of most of this material can be found (with references) in T. W. Anderson [2].

Other writers have considered the existence of minimax rules, admissible rules, and complete classes of rules. T. W. Anderson and

R. Bahadur [3] studied the case of different but known covariance structures for normal alternatives and found the minimal complete class within the class of linear rules. B. Ellison [12] generated a class of admissible rules for the case of normal alternatives with unknown means and known covariance structures. P. K. Bhattacharya and S. Das Gupta [4] considered univariate exponential families and generated a class of admissible rules. They also found the minimax rule in the normal case, both for known and unknown variance. For the multivariate normal case, S. Das Gupta [10] found the minimax invariant rule and showed it to be admissible. Further admissibility results were obtained by J. Kiefer and R. Schwartz [17], who showed that their techniques could be applied to the classification problem, in particular, to prove the admissibility of the likelihood ratio criterion (see Anderson [2]).

In another direction, S. Das Gupta [9] considered nonparametric classification rules. He proved the consistency of two minimum distance rules, one based on the Wilcoxon rank sum statistic and the other based on the Kolmogorov distance. M. Woinsky and L. Kurz [30] studied a rule based on the Wilcoxon statistic which first sequentially estimates a shift parameter between alternative distributions and then takes a finite sample on which the classification is based. Unfortunately, their error estimates appear to be unfounded.

Another formulation of the classification problem has been considered by T. Cacoullos [6] and M. S. Srivastava [26]. They consider finding which of several alternative populations is "closest" to a specified population. In particular, they restrict themselves to

normal distributions, where closeness can be defined in terms of the Mahalanobis distance. Cacoullos found some admissible invariant rules, while M. S. Srivastava used the techniques of Kiefer and Schwartz to prove the admissibility of some other rules.

A related problem is that of distinguishability. It involves deciding to which of several sets of distributions, the distribution of a random variable X belongs, based on independent observations on X. Note that if each of the sets consists of a single known distribution, then the problem is equivalent to classification with specified alternatives. W. Hoeffding and J. Wolfowitz [16] consider the case of two sets of alternatives and find conditions under which there exists a test with maximum error less than any pre-specified limit. This work has been extended by D. Freedman [15] and L. Fisher and J. Van Ness [14] to the case of a countable number of alternatives.

Our own work consists of several parts. The first takes a new view of classification. Following Hoeffding and Wolfowitz, we ask what are necessary and sufficient conditions on the alternative distributions to insure the existence of a classification rule with the maximum probability of error less than any pre-specified limit. In Chapter I, the theory is developed, using the results of Hoeffding and Wolfowitz. Our principal result is that, in any practical situation (observations in Euclidean space), there exists a sequential classification rule with arbitrarily small probabilities of error, as long as no two of the alternatives are equal. For the existence of a rule based on some finite sample, the alternatives must be uniformly separated. In Chapter II, examples are given of different types of

sequential classification rules. In particular, we show the effectiveness of minimum distance rules in controlling the probabilities of error. Also, we use the idea of a test of power one (Robbins [24]) to generate sequential classification rules.

The second part of our work (Chapter III) is a first attempt to compare finite sample size classification rules through the computation of relative efficiencies. First, we compare a standard classification rule with a two-sample test for the equality of means when the underlying distributions are univariate normal with a common unit variance. We show that the classification rule asymptotically uses 9/8 as many observations as the two-sample test when the maximum probability of error goes to 0. Next, we introduce a family of classification rules based on ranks. For the particular case of one sided location-shift alternatives, the classification rules based on ranks have the same asymptotic Pitman efficiency (relative to the analogous rules based on sample means) as the corresponding two sample rank tests for equality (relative to Student's t test).

Finally, in Chapter IV, we prove a complete class theorem for univariate normal distributions with a common unit variance. Restricting ourselves to rules invariant under translations and change of sign, we first consider Bayes rules with respect to priors with finite support. Then, using the techniques of T. Matthes and D. Truax [21] and M. Eaton [11], we find an essentially complete class of rules in the class of all invariant rules.

CHAPTER I

Theory of Classifiability

1.1 Introduction and Summary.

In general, the classification problem involves $k + 1$ populations, $\pi_0, \pi_1, \ldots, \pi_k$, of experimental units. We wish to match π_0 with one of $\pi_1, \ldots, \pi_k$. Observations are available from each of the units, and those from π_i are independent and identically distributed as X_i, a random variable taking values in the measurable space $(\mathcal{X}, \mathcal{A})$ for each i. Also, since each X_i is based on a different population, they are assumed independent. π_0 and π_i are matched only if X_0 and X_i have the same distribution. It is assumed that π_0 can be matched with one of the π_i. We are interested in controlling the probabilities of mismatching or misclassifying π_0. In other words, we are interested in finding a class of rules based on sequences of observations from the different populations such that, no matter how small ϵ, there is a rule for which the probabilities of error are less than ϵ.

Several specific cases have previously been considered in the literature, usually in a different framework. If F_i, the distribution of X_i, is completely specified for each $i = 1, \ldots, k$, then the classification problem is identical with testing k simple hypotheses, $F_0 = F_i$. Thus, for each ϵ, there is a test based on a finite number of observations from π_0 with all the error probabilities less than ϵ if and only if the F_i's are distinct. If each F_i is not completely specified, but is known to be in $\mathcal{F}_i$, some collection of distributions, then we can test the k composite hypotheses, $F_0 \in \mathcal{F}_i$ based on

observations from π_0. This is exactly the problem considered by Kraft [19] and Hoeffding and Wolfowitz [16]. They showed that to achieve arbitrarily small probabilities of error, in some circumstances one could use a finite sample size rule or a sequential rule and in other circumstances there does not exist any rule. The results of Hoeffding and Wolfowitz are so important to our work that many of them are listed in Section 1.3.

Our work is based on different assumptions. We assume that F, the distribution of the vector $(X_1, \ldots, X_k)$, is known to be in Ω, a collection of distributions. Note that, since the X_i's are independent, $F = F_1 \times \ldots \times F_k$ is a k-fold product distribution. Also, since we can take Ω to be $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$, our framework includes that of Hoeffding and Wolfowitz. On the other hand, our assumptions about sampling are different. We assume sampling is done on the vector $(X_0, X_1, \ldots, X_k)$; that is, we sample from all the populations, not just from π_0. Our problem thus becomes to test the k composite hypothesis $G \in \mathcal{G}_i$, where

$$\mathcal{G}_i = \{F_0 \times \ldots \times F_k : F_0 = F_i, F_1 \times \ldots \times F_k \in \Omega\},$$

based on samples from $(X_0, \ldots, X_k)$ with distribution G. Seen in this light, classification fits exactly into the framework of Hoeffding and Wolfowitz. Our results are based on applying their results and then using the special structure of the $\mathcal{G}_i$ to get simplified results for classification.

Since the basic constraints on the distributions derive from Ω, we will call the classification problem with $F_1 \times \ldots \times F_k$, restricted to be in Ω the classification problem based on Ω, or $C(\Omega)$. Our

major results about $C(\Omega)$ require much notation to state precisely, but we can indicate their content. If the problem is set in a Euclidean space, then we can always find a sequential test with maximum error less than ε, as long as no two X_i have the same distribution. In order to be able to use a finite sample size test, the distributions F_i must be uniformly separated, an idea to be made more precise with distance functions in Section 1.5.

For simplicity of notation, the bulk of the chapter concerns the case $k = 2$. Section 1.2 gives many of the necessary definitions and notations, and Section 1.3 lists the needed results of Hoeffding and Wolfowitz. A special class of distance functions is considered in Section 1.4, and the main results are proved in Section 1.5. In Section 1.6, the case $k \geq 2$ is considered and in Section 1.7 the case of normal distributions is considered with special attention to the existence of rules which do not sample all the populations. In Section 1.8, the theory is extended in another direction by generalizing the definition of a match. The last section is a discussion of some of the assumptions and implications of the theory.

1.2 Definitions and Notations.

Restricting ourselves to the case $k = 2$, our first convention is to assume that there is only one correct classification for each problem. Thus, even if $F_0 = F_1 = F_2$, we assume that π_0 should be classified with just one of π_1 and π_2. To indicate which choice is correct, we will write $(0\text{-}j)$ when π_0 should be classified with π_j. Of course, by assumption, if $(0\text{-}j)$ obtains, then necessarily $F_0 = F_j$.

In general, $\mathcal{J}$, $\mathcal{G}$, and $\mathcal{H}$ will be arbitrary collections of distributions. $\mathcal{G} \times \mathcal{H}$ will indicate the collection of product distributions $G \times H$, where $G \in \mathcal{G}$ and $H \in \mathcal{H}$. For any Ω, a collection of product distributions, the classification problem based on Ω, $C(\Omega)$, asks us to classify π_0 into π_1 or π_2 when the distribution of (X_1, X_2), $F_1 \times F_2$, is known to be in Ω. The following definition describes the property of $C(\Omega)$ in which we are interested.

Definition 1.1.

$C(\Omega)$ is classifiable in the class of rules $\mathcal{J}$ if, for every $\epsilon > 0$, there is a rule in $\mathcal{J}$ such that, no matter what $F_1 \times F_2 \in \Omega$,

under (0-1), the probability of classifying π_0 with π_2 is less than ϵ, and

under (0-2), the probability of classifying π_0 with π_1 is less than ϵ.

Our purpose is to find conditions on Ω that insure $C(\Omega)$ is classifiable in $\mathcal{J}$, for different specific $\mathcal{J}$'s.

Let $y_1, y_2, \ldots$ be a sequence of observations from a random variable Y. A rule based on observations from Y can be represented by (N, φ), where N is a stopping time and φ is a terminal decision rule. A stopping time is a random variable that assumes positive integer values (or plus infinity), and whose conditional distribution, given $y_1, y_2, \ldots$, is such that the probability of $N \leq n$ is not dependent on $y_{n+1}, y_{n+2}, \ldots$. N is said to be non-randomized if $P[N \leq n | y_1, \ldots, y_n]$ is either 0 or 1. The value $\varphi(y_1, \ldots, y_N)$ of the terminal decision rule φ will be the conditional probability of classifying π_0 with π_2, given N and $y_1, \ldots, y_N$. φ is non-

randomized if $\varphi(y_1, \ldots, y_N)$ is either 0 or 1. When the test (N, φ) is based on observations from Y and Y has distribution F, we will use the subscript F in probability statements about N and φ to indicate the underlying probability structure.

We can now rewrite the definition of classifiability. Let $Y = (X_0, X_1, X_2)$ and recall that under (0-j), $F_0 = F_j$.

Definition 1.1'.

$C(\Omega)$ is classifiable in $\mathcal{J}$ if, for every $\epsilon > 0$, there is a rule (N, φ) in $\mathcal{J}$ such that, for all $G \times H \in \Omega$,

$E(\varphi) < \epsilon$ when Y has distribution $G \times G \times H$, and

$E(1-\varphi) < \epsilon$ when Y has distribution $H \times G \times H$.

The different classes of rules of interest can now be defined in terms of their stopping times.

Definition 1.2.

For any $\mathcal{F}$,

$$\mathcal{J}_0(\mathcal{F}) = \{(N, \varphi): P_F[N < \infty] = 1 \text{ for all } F \in \mathcal{F}\}, \text{ and}$$

$$\mathcal{J}_1(\mathcal{F}) = \{(N, \varphi): \text{there exists } M \ni N \leq M \text{ for all } F \in \mathcal{F}\}.$$

We will use the terms classifiable $(\mathcal{F})$ and finitely classifiable $(\mathcal{F})$ when $C(\Omega)$ is classifiable in $\mathcal{J}_0(\mathcal{F})$ and $\mathcal{J}_1(\mathcal{F})$, respectively. In fact, we often drop the $\mathcal{F}$ when the choice is clear.

The work of Hoeffding and Wolfowitz (hereafter referred to as H-W) is based on the following definition of the distinguishability of two arbitrary sets of distributions, $\mathcal{G}$ and $\mathcal{H}$, when observations are taken from Y with distribution F.

Definition 1.3.

$\mathcal{G}$ and $\mathcal{H}$ are distinguishable in $\mathcal{J}$ if, for every $\varepsilon > 0$, there exists a rule $(N, \varphi) \in \mathcal{J}$ such that

$$E_F(\varphi) < \varepsilon \quad \text{for all } F \in \mathcal{G}, \text{ and}$$

$$E_F(1-\varphi) < \varepsilon \quad \text{for all } F \in \mathcal{H}.$$

An equivalent definition which is sometimes useful is the following:

Definition 1.3'.

$\mathcal{G}$ and $\mathcal{H}$ are distinguishable in $\mathcal{J}$ if

$$\sup_{\varphi} \inf_{G \in \mathcal{G}, H \in \mathcal{H}} (E_H(\varphi) - E_G(\varphi)) = 1 .$$

We use the terms distinguishable and finitely distinguishable when $\mathcal{G} \cup \mathcal{H} \subset \mathcal{F}$ and $\mathcal{G}$ and $\mathcal{H}$ are distinguishable in $\mathcal{J}_0(\mathcal{F})$ and $\mathcal{J}_1(\mathcal{F})$, respectively.

The analysis of distinguishability is based on distances between distributions. The term distribution denotes a probability measure on some measurable space. Let $\mathcal{K}$ be a collection of distributions and δ be a function of two distributions.

Definition 1.4.

δ is a distance in $\mathcal{K}$ if, for all G, H, and K in $\mathcal{K}$,

a) $\delta(G, G) = 0$,

b) $\delta(G, H) = \delta(H, G)$, and

c) $\delta(G, H) \leq \delta(G, K) + \delta(K, H)$.

Several specific distances which are essential to the theory of H-W are the Kolmogorov distance D and the total variation distance d defined below. They will be used frequently without further reference to their definitions.

Definition 1.5.

For $\mathcal{H}$ the set of distributions on some Euclidean space $\mathcal{X}$,

$$D(F, G) = \sup_{x \in \mathcal{X}} |F(x) - G(x)|,$$

where $F(x)$ and $G(x)$ are the cumulative distribution functions of the distributions (probability measure) F and G.

Definition 1.6.

For $\mathcal{H}$ any set of distributions on a measurable space $(\mathcal{X}, \mathcal{G})$, define

$$d(F, G) = \sup_{A \in \mathcal{G}} |F[A] - G[A]|$$

(where $F[A]$ and $G[A]$ are the measures of the set A under the distributions F and G). Several equivalent definitions can be given for d (see H-W). In particular, if F and G are absolutely continuous with respect to a σ-finite measure ν with densities (Radon-Nikodym derivatives) f and g, then

$$d(F, G) = F[B] - G[B],$$

where

$$B = \{x: f(x) > g(x)\}.$$

Let $F^{(n)}$ be the empiric distribution based on $y_1, \ldots, y_n$, a sequence of observations on Y with distribution $F \in \mathcal{F}$; that is, $nF^{(n)}[A]$ is the number of indices $i \leq n$ for which $y_i \in A$. We assume that the set $\mathcal{H}$ on which a distance δ is defined contains $\mathcal{F}$ and all the empiric distributions based on random variables Y with distribution $F \in \mathcal{F}$.

Definition 1.7.

A distance δ is consistent in $\mathcal{F}$ if, for every $\epsilon > 0$ and all $F \in \mathcal{F}$,

$$\lim_{n\to\infty} P_F[\delta(F^{(n)}, F) > \epsilon] = 0. \tag{1.1}$$

Definition 1.8.

A distance δ is uniformly consistent in $\mathfrak{F}$ if, for every $\epsilon > 0$, the convergence in (1.1) is uniform in $\mathfrak{F}$.

An example of a uniformly consistent distance in the Kolmogorov distance D. If $\mathcal{X}$ is a k-dimensional Euclidean space, then Kiefer and Wolfowitz [18] showed there exists two positive numbers a and b such that, for all $\epsilon > 0$, all $n > 0$, and all k-dimensional distributions F,

$$P_F[D(F^{(n)}, F) > \epsilon] \leq ae^{-b\epsilon^2 n}. \tag{1.2}$$

Then, as n increases, the left hand side of (1.2) converges to 0 uniformly in $\mathfrak{F}$, the collection of all distributions on $(\mathcal{X}, \mathcal{A})$.

If $\mathcal{N}_k$ is the collection of all k-variate normal distributions, we will use another distance d_1.

Definition 1.9.

For $F, G \in \mathcal{N}_k$, define

$$d_1(F, G) = 2^{1/2}\{1 - \rho(F, G)\}^{1/2},$$

where the measure of affinity $\rho(F, G)$ is given by

$$\rho(F, G) = \int (fg)^{1/2} d\nu,$$

f and g being the densities of F and G with respect to some σ-finite measure ν. If F and G have non-singular covariance matrices Σ and ψ and means μ and ν respectively, then

$$\rho(F, G) = |\Sigma|^{1/4}|\psi|^{1/4}\left|\frac{\Sigma+\psi}{2}\right|^{-1/2}\exp\{-\tfrac{1}{4}(\mu-\nu)'(\Sigma+\psi)^{-1}(\mu-\nu)\}$$

(see Kraft [19] and H-W).

For any distance δ, we will write $\delta(G, \mathcal{H})$ for the infimum over $\mathcal{H}$ of $\delta(G, H)$ and $\delta(\mathcal{G}, \mathcal{H})$ for the infimum over $\mathcal{G}$ of $\delta(G, \mathcal{H})$. Similarly we will write $\rho(G, \mathcal{H})$ for the supremum over $\mathcal{H}$ of $\rho(G, H)$ and $\rho(\mathcal{G}, \mathcal{H})$ for the supremum over $\mathcal{G}$ of $\rho(G, \mathcal{H})$ when $\mathcal{G}$ and $\mathcal{H}$ are collections of normal distributions.

1.3 Results of Hoeffding and Wolfowitz on Distinguishability of Distributions.

In this section $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ are arbitrary collections of distributions, where $\mathcal{G}$ and $\mathcal{H}$ are both subsets of $\mathcal{F}$. The following theorems are results of Hoeffding and Wolfowitz.

Theorem 1.1 (Theorem 3.1 of H-W).

If the distance δ is uniformly consistent in $\mathcal{F}$ and, for all $F \in \mathcal{F}$,

$$\max[\delta(F, \mathcal{G}), \delta(F, \mathcal{H})] > 0,$$

then $\mathcal{G}$ and $\mathcal{H}$ are distinguishable $(\mathcal{F})$.

Corollary.

If $\mathcal{F}$ is a collection of distributions on a k-dimensional Euclidean space and, for all F in $\mathcal{F}$,

$$\max[D(F, \mathcal{G}), D(F, \mathcal{H})] > 0,$$

then $\mathcal{G}$ and $\mathcal{H}$ are distinguishable $(\mathcal{F})$.

Theorem 1.2.

If the distance δ is uniformly consistent in $\mathcal{F}$ and

$$\delta(\mathcal{G}, \mathcal{H}) > 0,$$

then $\mathcal{G}$ and $\mathcal{H}$ are finitely distinguishable $(\mathcal{F})$.

Theorem 1.3.

In order that $\mathcal{G}$ and $\mathcal{H}$ be finitely distinguishable $(\mathcal{F})$, it is necessary that

$$d(\mathcal{G}, \mathcal{H}) > 0.$$

Theorem 1.4 (Theorem 4.1 of H-W).

In order that $\mathcal{G}$ and $\mathcal{H}$ be distinguishable ($\mathcal{F}$), it is necessary that

$$\max[d(F, \mathcal{G}), d(F, \mathcal{H})] > 0 \quad \text{for all} \quad F \quad \text{in} \quad \mathcal{F}.$$

Theorem 1.5.

If δ is a uniformly consistent distance in $\mathcal{F}$, and

$$\delta(\mathcal{G}, \mathcal{H}) = 0 \quad \text{implies} \quad d(\mathcal{G}, \mathcal{H}) = 0,$$

then $\mathcal{G}$ and $\mathcal{H}$ are finitely distinguishable ($\mathcal{F}$) if, and only if, $\delta(\mathcal{G}, \mathcal{H}) > 0$.

Theorem 1.6.

If δ is a uniformly consistent distance in $\mathcal{F}$, and, for all F in $\mathcal{F}$,

$$\delta(F, \mathcal{G}) = 0 \quad \text{implies} \quad d(F, \mathcal{G}) = 0 \quad \text{and}$$

$$\delta(F, \mathcal{H}) = 0 \quad \text{implies} \quad d(F, \mathcal{H}) = 0,$$

then $\mathcal{G}$ and $\mathcal{H}$ are distinguishable ($\mathcal{F}$) if, and only if,

$$\max[\delta(F, \mathcal{G}), \delta(F, \mathcal{H})] > 0 \quad \text{for all} \quad F \quad \text{in} \quad \mathcal{F}.$$

Theorem 1.7 (Theorem 5.2 of H-W).

Let $\mathcal{N}_k$ be the set of all k-dimensional normal distributions.

a) If $\mathcal{F} \subset \mathcal{N}_k$, then two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathcal{F}$ are distinguishable ($\mathcal{F}$) if, and only if, for all $F \in \mathcal{F}$,

$$\min[\rho(F, \mathcal{G}), \rho(F, \mathcal{H})] < 1.$$

b) Two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathcal{F}$ are finitely distinguishable ($\mathcal{F}$) if, and only if,

$$\rho(\mathcal{G}, \mathcal{H}) < 1.$$

Remarks:

The corollary to Theorem 1.1 follows from the uniform consistency of D, mentioned in Section 1.2. Theorems 1.2 and 1.3 are contained in the text of H-W on pages 706 and 710. Theorems 1.5 and 1.6 follow from comparisons of Theorems 1.2 and 1.1 with Theorems 1.3 and 1.4, respectively.

One further result of H-W that will be useful shows that we can restrict ourselves to non-randomized rules. $\mathcal{J}$ will be either $\mathcal{J}_0$ or $\mathcal{J}_1$, and $\mathcal{J}'$ will be that subset of $\mathcal{J}$ consisting exclusively of rules with non-randomized stopping times and non-randomized terminal decision rules.

Theorem 1.8 (Theorem 2.1 of H-W).

If $\mathcal{J}$ is either $\mathcal{J}_0$ or $\mathcal{J}_1$, then

$$\sup_{\mathcal{J}} \inf_{G,H} (E_H(\varphi) - E_G(\varphi)) > 0$$

implies

$$\sup_{\mathcal{J}'} \inf_{G,H} (E_H(\varphi) - E_G(\varphi)) = 1.$$

1.4 On Distance Functions and Their Properties.

Since the classification problem is based on observations on a $k+1$ vector $Y = (X_0, \ldots, X_k)$ with distribution $F = F_0 \times F_1 \times \ldots \times F_k$, we will be interested in distances defined on sets of product distributions. If $\mathcal{K}$ is a collection of distributions on $(\mathcal{X}, \mathcal{A})$, for each n we consider a distance δ_n defined on $\mathcal{K}^n$, the class of all n-fold product distributions on $(\mathcal{X}^n, \mathcal{A}^n)$, $F = F_1 \times \ldots \times F_n$, $F_i \in \mathcal{K}$.

Definition 1.10.

A family of distances $\{\delta_n\}$ is said to satisfy condition A in $\mathcal{K}$ if, for any n and for any F_j, G_j $(j = 1,\ldots, n)$ in $\mathcal{K}$,

$$\delta_n(F_1 \times \ldots \times F_n, G_1 \times \ldots \times G_n)$$

$$= \delta_{n-1}(F_1 \times \ldots \times F_{i-1} \times F_{i+1} \times \ldots \times F_n,$$

$$G_1 \times \ldots G_{i-1} \times G_{i+1} \times \ldots \times G_n)$$

whenever $F_i = G_i$, for any $i = 1,\ldots, n$.

Definition 1.11.

A family of distances $\{\delta_n\}$ is said to satisfy condition B in $\mathcal{K}$ if, for any n and any F_j, G_j $(j = 1,\ldots, n)$ in $\mathcal{K}$,

$$\delta_n(F_1 \times \ldots \times F_n, G_1 \times \ldots \times G_n)$$

$$\geq \delta_n(F_1 \times \ldots \times F_n, G_1 \times \ldots \times G_{i-1} \times F_i \times G_{i+1} \times \ldots \times G_n),$$

for any $i = 1,\ldots, n$.

Many common distances used in statistics satisfy both conditions A and B. Three important examples are D, d, and a distance which is analogous to ordinary Euclidean distance. The following lemma is immediate from the definitions 1.10 and 1.11.

Lemma 1.1.

If δ is a distance in $\mathcal{K}$, and δ_n is defined for each n on $\mathcal{K}^n$ by

$$\delta_n(F_1 \times \ldots \times F_n, G_1 \times \ldots \times G_n)$$

$$= [(\delta(F_1, G_1))^2 + \ldots + (\delta(F_n, G_n))^2]^{\frac{1}{2}},$$

then the family $\{\delta_n\}$ satisfies conditions A and B.

Lemma 1.2.

If $\mathcal{X}$ is a m-dimensional Euclidean space, $\mathcal{A}$ is the Borel field of subsets of $\mathcal{X}$, and D_n is defined for each n in the set of all distributions on $(\mathcal{X}^n, \mathcal{A}^n)$ by

$$D_n(F, G) = \sup_{x \in \mathcal{X}^n} |F(x) - G(x)|,$$

then the family $\{D_n\}$ satisfies conditions A and B.

Proof:

The verification of conditions A and B will be demonstrated in the case $n = 2$, as the general demonstration is analogous, but lengthy.

Recall that the cumulative distribution function of the product distribution $F \times G$ is $F(x)G(y)$. To verify that D_n satisfies condition A, note that for any distributions F, G, H, and K on $(\mathcal{X}, \mathcal{A})$,

$$\begin{aligned} D_2(F \times H, G \times H) &= \sup_{x,y \in \mathcal{X}} |F(x)H(y) - G(x)H(y)| \\ &= \sup_{x,y \in \mathcal{X}} H(y)|F(x) - G(x)| \\ &= \sup_{x \in \mathcal{X}} |F(x) - G(x)| \\ &= D_1(F, G). \end{aligned}$$

As for condition B, note that for x fixed,

$$\begin{aligned} D_2(F \times H, G \times K) &\geq \sup_{y \in \mathcal{X}} |F(x)H(y) - G(x)K(y)| \\ &\geq \lim_{y \to \infty} |F(x)H(y) - G(x)K(y)| \\ &= |F(x) - G(x)| . \end{aligned}$$

Thus, $D_2(F \times H, G \times K) \geq \sup_{x \in \mathcal{X}} |F(x) - G(x)| = D_1(F, G)$.

Since $D_2(F \times H, G \times K) = D_2(H \times F, K \times G)$, these two demonstrations apply equally to either coordinate. □

Lemma 1.3.

If $(\mathcal{X}, \mathcal{A})$ is any measurable space and d_n is defined for each n in the set of all distributions on $(\mathcal{X}, \mathcal{A})$ by

$$d_n(F, G) = \sup_{A \in \mathcal{A}^n} |F[A] - G[A]|,$$

then the family of distances $\{d_n\}$ satisfies conditions A and B.

Proof:

As mentioned in Section 1.2 (see also Kraft [19] and H-W), d_n has the alternate form

$$d_n(F, G) = F[B] - G[B], \text{ where}$$

$$B = \{x: f(x) > g(x)\},$$

f and g being densities (Radon-Nikodym derivatives) of F and G with respect to some σ-finite measure ν. Note that if f and h are densities of F and H with respect to ν_1 then $f h$ is a density of $F \times H$ with respect to $\nu_1 \times \nu_1$. Thus if F, G, and H have densities f, g, and h with respect to ν_1,

$$d_2(F \times H, G \times H) = F \times H[C] - G \times K[C],$$

where $C = \{(x, y): f(x)h(y) > g(x)h(y)\}$. To verify condition A, let

$$D = \{x: f(x) > g(x)\} \quad \text{and}$$

$$E = \{y: h(y) > 0\}.$$

Then $D \times E = C$, and

$$d_2(F \times H, G \times H) = F \times H[D \times E] - G \times H[D \times E]$$

$$= F[D]H[E] - G[D]H[E]$$

$$= F[D] - G[D]$$

$$= d_1(F,G).$$

The equality of lines 2 and 3 above follows from $H[E] = 1$. To verify condition B, note that (from Definition 1.6 of d_n)

$$d_2(F \times H, G \times K) \geq F \times H[D \times \mathcal{Y}] - G \times K[D \times \mathcal{Y}]$$

$$= F[D]H[\mathcal{Y}] - G[D]K[\mathcal{Y}]$$

$$= F[D] - G[D]$$

$$= d_1(F, G)$$

$$= d_2(F \times H, G \times H).$$

Since $d_2(F \times H, G \times K) = d_2(H \times F, K \times G)$, these two demonstrations apply equally to either coordinate. □

The property of distances satisfying conditions A and B that we will use is demonstrated in the following lemma.

Lemma 1.4.

If the family of distances $\{\delta_n\}$ defined on $\mathcal{H}^n$ satisfies conditions A and B, then for any F_j, G_j $(j = 1,\ldots, n)$ in $\mathcal{H}$,

$$\sum_{j=1}^{n} \delta_1(F_j, G_j) \geq \delta_n(F_1 \times \ldots \times F_n, G_1 \times \ldots \times G_n)$$

$$\geq \delta_1(F_i, G_i), \quad i = 1,\ldots, n.$$

Proof:

For $\{\delta_n\}$ satisfying conditions A and B,

$$\delta_n(F_1 \times \cdots \times F_n, G_1 \times \cdots \times G_n) \geq \delta_n(F_1 \times \cdots \times F_n, G_1 \times \cdots \times G_{n-1} \times F_n)$$

$$= \delta_{n-1}(F_1 \times \cdots \times F_{n-1}, G_1 \times \cdots \times G_{n-1}).$$

Repeating this argument, we first find

$$\delta_n(F_1 \times \cdots \times F_n, G_1 \times \cdots \times G_n) \geq \delta_i(F_1 \times \cdots \times F_i, G_1 \times \cdots \times G_i),$$

and then

$$\delta_n(F_1 \times \cdots \times F_n, G_1 \times \cdots \times G_n) \geq \delta_1(F_i, G_i), \text{ for } i = 1, \ldots, n.$$

On the other hand, δ_n is a distance, so

$$\delta_n(F_1 \times \cdots \times F_n, G_1 \times \cdots \times G_n)$$

$$\leq \delta_n(F_1 \times \cdots \times F_n, G_1 \times \cdots \times G_{n-1} \times F_n)$$

$$+ \delta_n(G_1 \times \cdots \times G_{n-1} \times F_n, G_1 \times \cdots \times G_n)$$

$$= \delta_{n-1}(F_1 \times \cdots \times F_{n-1}, G_1 \times \cdots \times G_{n-1}) + \delta_1(F_n, G_n).$$

By induction, we get

$$\delta_n(F_1 \times \cdots \times F_n, G_1 \times \cdots \times G_n) \leq \sum_{j=1}^{n} \delta_1(F_j, G_j). \quad \square$$

1.5 Conditions for Classifiability.

The role that distinguishability plays in the classification problem, hinted at in Section 1.1, will now be made precise. A comparison of Definitions 1.1' and 1.3 will show that classifiability is just distinguishability in the particular circumstances of classification, as the following theorem states. Let

$$\mathcal{G}(\Omega) = \{(F \times G \times H): F = G \text{ and } G \times H \in \Omega\} \text{ and}$$

$$\mathcal{H}(\Omega) = \{(F \times G \times H): F = H \text{ and } G \times H \in \Omega\}.$$

Theorem 1.9.

$C(\Omega)$ is classifiable in $\mathcal{T}$ if and only if $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ are distinguishable in $\mathcal{T}$.

Proof:

$C(\Omega)$ is classifiable in $\mathcal{T}$, if, and only if, for every $\epsilon > 0$, there is a rule (N, φ) in $\mathcal{T}$ (based on $Y = (X_0, X_1, X_2)$ with distribution $F = F_0 \times F_1 \times F_2$) such that, for all $G \times H$ in Ω,

$$E(\varphi) < \epsilon \text{ when } F_0 \times F_1 \times F_2 = G \times G \times H, \text{ and}$$

$$E(1-\varphi) < \epsilon \text{ when } F_0 \times F_1 \times F_2 = H \times G \times H.$$

But $F_0 \times F_1 \times F_2 = G \times G \times H$ and $G \times H \in \Omega$ if, and only if, $F_0 \times F_1 \times F_2 \in \mathcal{G}(\Omega)$, and $F_0 \times F_1 \times F_2 = H \times G \times H$ and $G \times H \in \Omega$ if, and only if, $F_0 \times F_1 \times F_2 \in \mathcal{H}(\Omega)$. Thus $C(\Omega)$ is classifiable in $\mathcal{T}$ if, and only if, for every $\epsilon > 0$, there is a rule (N, φ) in $\mathcal{T}$ such that

$$E(\varphi) < \epsilon \text{ for all } F = F_0 \times F_1 \times F_2 \in \mathcal{G}(\Omega) \text{ and}$$

$$E(1-\varphi) < \epsilon \text{ for all } F = F_0 \times F_1 \times F_2 \in \mathcal{H}(\Omega),$$

which is just the definition of the distinguishability of $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ in $\mathcal{T}$ based on observations on Y with distribution F. □

Theorem 1.9 allows us to immediately restate Theorems 1.1-1.6 as theorems of necessary and sufficient conditions on $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ for $C(\Omega)$ to be classifiable. We will not do that, but instead examine the implications of the assumption of conditions A and B for the distances involved.

Lemma 1.5.

If $\{\delta_n\}$ $n = 1,2,3$ satisfy conditions A and B, then

a) for $G \times H$ in Ω,

$$\delta_3(H \times G \times H, \mathcal{C}(\Omega)) = 0 \text{ iff } \delta_1(G, H) = 0, \text{ and}$$

$$\delta_2(G \times G \times H, \mathcal{H}(\Omega)) = 0 \text{ iff } \delta_1(G, H) = 0.$$

b) $\delta_3(\mathcal{C}(\Omega), \mathcal{H}(\Omega)) = 0$ iff there exists a sequence $\{G_j \times H_j\}$ in Ω such that $\lim_{j \to \infty} \delta_1(G_j, H_j) = 0$.

Proof:

a) If $\delta_3(H \times G \times H, \mathcal{C}(\Omega)) = 0$, then there is a sequence $G_j \times G_j \times H_j \in \mathcal{C}(\Omega)$ such that $\delta_3(H \times G \times H, G_j \times G_j \times H_j)$ converges to 0. But, under conditions A and B,

$$\delta_1(G, H) = \delta_1(H, G) \leq \delta_1(H, G_j) + \delta_1(G_j, G)$$

$$\leq 2\, \delta_3(H \times G \times H, G_j \times G_j \times H_j) \text{ for all } j,$$

so $\delta_1(G, H) = 0$.

If $\delta_1(G, H) = 0$, then

$$\delta_3(H \times G \times H, G \times G \times H) \leq \delta_1(H, G) + \delta_1(G, G) + \delta_1(H, H) = 0.$$

Since $G \times H \in \Omega$ implies $G \times G \times H \in \mathcal{C}(\Omega)$, we get

$$\delta_3(H \times G \times H, \mathcal{C}(\Omega)) \leq \delta_3(H \times G \times H, G \times G \times H) = 0.$$

The second statement under a) is proved in the same way, with the roles of G and H reversed.

b) If $\delta_3(\mathcal{C}(\Omega), \mathcal{H}(\Omega)) = 0$, then there exist sequences $\{G_j \times H_j\}$ and $\{F_j \times K_j\}$ in Ω such that $\delta_3(G_j \times G_j \times H_j, K_j \times F_j \times K_j)$ converges to 0. Under conditions A and B,

$$\delta_1(G_j, H_j) \leq \delta_1(G_j, K_j) + \delta_1(K_j, H_j)$$

$$\leq 2\delta_3(G_j \times G_j \times H_j, K_j \times F_j \times K_j),$$

so that $\delta_1(G_j, H_j)$ converges to 0.

Conversely, if there is a sequence $\{G_j \times H_j\}$ in Ω such that $\delta_1(G_j, H_j)$ converges to 0, then

$$\delta_3(\mathcal{G}(\Omega), \mathcal{H}(\Omega)) \leq \delta_3(G_j \times G_j \times H_j, H_j \times G_j \times H_j)$$

$$\leq \delta_1(G_j, H_j) + \delta_1(G_j, G_j) + \delta_1(H_j, H_j)$$

$$= \delta_1(G_j, H_j) \quad \text{for all } j,$$

so that $\delta_3(\mathcal{G}(\Omega), \mathcal{H}(\Omega)) = 0$. □

Lemma 1.5 and Theorem 1.9 together give simple conditions for classifiability. The following notations will be useful in what follows. For $\{\delta_i\}$ $i = 1,2,3$, a family of distances satisfying conditions A and B, let

$$\Delta(\Omega) = \inf_{G \times H \in \Omega} \delta_1(G, H),$$

$$\Omega_\delta = \{G \times H \in \Omega : \delta_1(G, H) = 0\}, \text{ and}$$

$$\Omega_0 = \{G \times H \in \Omega : G = H\}.$$

Theorem 1.10.

If $\{\delta_i\}$ $i = 1,2,3$ is a family of distances satisfying conditions A and B and δ_3 is uniformly consistent in $\mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$, then

$$\Delta(\Omega) > 0$$

implies $C(\Omega)$ is finitely classifiable.

Proof:

By Lemma 1.5 b), $\Delta(\Omega) > 0$ implies $\delta_3(\mathcal{G}(\Omega), \mathcal{H}(\Omega)) > 0$, which, in turn, by Theorem 1.2, implies $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ are finitely distinguishable. By the equivalence of Theorem 1.9, $C(\Omega)$ is finitely classifiable. □

Theorem 1.11.

If $\{\delta_i\}$ $i = 1,2,3$ is a family of distances satisfying conditions A and B and δ_3 is uniformly consistent in $\mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$, then

$$\Omega_\delta = \emptyset$$

implies $C(\Omega)$ is classifiable.

Proof:

$\Omega_\delta = \emptyset$ implies $\delta_1(G, H) > 0$ for all $G \times H \in \Omega$. By Lemma 5 b) $\delta_1(G, H) > 0$ implies $\delta_3(G \times G \times H, \mathcal{H}(\Omega)) > 0$ and $\delta_3(H \times G \times H, \mathcal{G}(\Omega)) > 0$. Thus

$$\max[\delta_3(F, \mathcal{G}(\Omega)), \delta_3(F, \mathcal{H}(\Omega))] > 0 \quad \text{for all} \quad F = F_0 \times F_1 \times F_2$$

in $\mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$. By Theorem 1.1, $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ are distinguishable. Thus, by Theorem 1.9, $C(\Omega)$ is classifiable. □

In order to find necessary and sufficient conditions for classifiability, we need the following lemma.

Lemma 1.6.

If $C(\Omega)$ is classifiable, then $\Omega_0 = \emptyset$.

Proof:

If $\Omega_0 \neq \emptyset$, then there exists $G \times H \in \Omega$ such that $G = H$. Thus, $G \times G \times H = H \times G \times H$, so that if Y has distribution $F = F_0 \times F_1 \times F_2$ and $\epsilon < \frac{1}{2}$,

$$E(\varphi) < \epsilon \quad \text{when} \quad F = G \times G \times H \quad \text{implies}$$

$$E(1-\varphi) > 1 - \epsilon > \epsilon \quad \text{when} \quad F = H \times G \times H, \text{ for any test } (N, \varphi).$$

By definition, $C(\Omega)$ cannot be classifiable under such circumstances, so, by contraposition, Ω_0 must be empty. □

Theorem 1.12.

If $\{\delta_i\}$ $i = 1,2,3$ is a family of distances satisfying conditions A and B, δ_3 is uniformly consistent in $\mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$, and $\delta_1(G, H) = 0$ implies $G = H$, then $C(\Omega)$ is classifiable if,and only if, $\Omega_0 = \emptyset$.

Proof:

Since in this case $\Omega_\delta = \Omega_0$, Theorem 1.11 says $\Omega_0 = \emptyset$ is sufficient. By Lemma 1.6, $\Omega_0 = \emptyset$ is necessary. □

Corollary.

If Ω is a collection of 2-fold product distributions on $2m$ dimensional Euclidean space, then $C(\Omega)$ is classifiable if and only if $\Omega_0 = \emptyset$.

Proof:

By Lemma 1.2, $\{D_n\}$ satisfies conditions A and B. As mentioned in Section 1.2, D_3 is uniformly consistent for any collection of distributions on $3m$ dimensional Euclidean space. Also, $D_1(F, G) = 0$ implies $F = G$. Thus $\{D_n\}$ satisfies the conditions of Theorem 1.12 and $C(\Omega)$ is classifiable if, and only if, $\Omega_0 = \emptyset$. □

The following theorem shows that conditions A and B, although satisfied by several important distances, are not essential to the simple necessary and sufficient conditions of Theorem 1.12.

Theorem 1.13.

If δ_3 is uniformly consistent in $\mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$,

$\delta_3(G \times G \times H, \mathcal{H}(\Omega)) = 0$ implies $G = H$, and

$\delta_3(H \times G \times H, \mathcal{G}(\Omega)) = 0$ implies $G = H$,

then $C(\Omega)$ is classifiable if and only if $\Omega_0 = \emptyset$.

Proof:

By Lemma 6, $\Omega_0 = \emptyset$ is necessary. If $\Omega_0 = \emptyset$, then for all $G \times H \in \Omega$, $G \neq H$. By contraposition, the assumptions imply

$$\delta_3(H \times G \times H, \mathcal{G}(\Omega)) > 0 \quad \text{and}$$

$$\delta_3(G \times G \times H, \mathcal{H}(\Omega)) > 0 \quad \text{for all} \quad G \times H \in \Omega.$$

Thus, for any $F \in \mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$

$$\max[\delta_3(F, \mathcal{G}(\Omega)), \delta_3(F, \mathcal{H}(\Omega))] > 0,$$

and by Theorem 1.1 $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ are distinguishable. Therefore, by Theorem 1.9, $C(\Omega)$ is classifiable. □

Note that Theorem 1.13 could have been proved by appealing to Theorem 1.6. Also, Theorem 1.12 could be stated as a corollary to Theorem 1.13, since the assumptions of the former imply the assumptions of the latter. A theorem similar to Theorem 1.13 based on Theorem 1.5 could be proved for finite classifiability.

1.6 Classification Into One of Several Populations.

In the case of $k + 1$ populations, $\pi_0, \pi_1, \ldots, \pi_k$, the results of the previous sections can be extended by similar methods of analysis. In this section, we point out the extensions, supplying proofs only when the analogy is not obvious.

Let $\Omega = \{G_1 \times \ldots \times G_k\}$ be some collection of possible distributions of the vector of observations $(X_1, \ldots, X_k)$ taken from $\pi_1, \ldots, \pi_k$. There are k possible decisions corresponding to the possible matches of population π_0 with one of π_j, $j = 1, \ldots, k$. Again we assume there is only one correct match, indicated by $(0\text{-}j)$, and that if $(0\text{-}j)$ holds, then $F_0 = F_j$, where F_j is the distribution of X_j. Let

$$\mathcal{G}_j(\Omega) = \{G_j \times G_1 \times \ldots \times G_k : G_1 \times \ldots \times G_k \in \Omega\},\ j = 1, \ldots, k,$$

and let any rule be represented by $(N, \varphi_1, \ldots, \varphi_k)$, where N is a stopping time and φ_j is the conditional probability of classifying π_0 with π_j, given N and $y_1, \ldots, y_N$.

Definition 1.12.

$C(\Omega)$ is classifiable in $\mathcal{T}$ (based on $Y = (X_0, \ldots, X_k)$ with distribution F) if, for every $\varepsilon > 0$, there is a test $(N, \varphi_1, \ldots, \varphi_k) \in \mathcal{T}$ such that

$$E(1 - \varphi_j) \leq \varepsilon \quad \text{for all } F \text{ in } \mathcal{G}_j(\Omega),\ j = 1, \ldots, k.$$

Similarly, if $\mathcal{G}_1, \ldots, \mathcal{G}_k$ are any k collections of distributions, we can define the distinguishability of $\mathcal{G}_1, \ldots, \mathcal{G}_k$ based on Y with distribution F.

Definition 1.13.

$\mathcal{G}_1, \ldots, \mathcal{G}_k$ are <u>distinguishable</u> in $\mathcal{T}$ if, for every $\varepsilon > 0$, there is a test $(N, \varphi_1, \ldots, \varphi_k) \in \mathcal{T}$ such that

$$E(1 - \varphi_j) \leq \varepsilon \quad \text{for all } F \text{ in } \mathcal{G}_j,\ j = 1, \ldots, k.$$

It is clear with these definitions that $C(\Omega)$ is (finitely) classifiable if, and only if, $\mathcal{G}_1(\Omega), \ldots, \mathcal{G}_k(\Omega)$ are (finitely) distinguishable, in analogy with Theorem 1.9.

Results for the classifiability of $C(\Omega)$ depend on the following characterization of the distinguishability of $\mathcal{G}_1, \ldots, \mathcal{G}_k$.

Theorem 1.14.

$\mathcal{G}_1, \ldots, \mathcal{G}_k$ are (finitely) distinguishable if, and only if, $\mathcal{G}_1, \ldots, \mathcal{G}_k$ are pairwise (finitely) distinguishable, i.e., if, and only if, $\mathcal{G}_i$ and $\mathcal{G}_j$ are (finitely) distinguishable for any $1 \leq i \neq j \leq k$.

Proof:

Assume $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are (finitely) distinguishable. Then, for any $\epsilon > 0$, there is a rule $(N, \varphi_1, \ldots, \varphi_k)$ such that

$$E(1 - \varphi_j) \leq \epsilon \quad \text{for all} \quad F \quad \text{in} \quad \mathcal{C}_j, \; j = 1, \ldots, k.$$

For distinguishing between $\mathcal{C}_i$ and $\mathcal{C}_j$, let $\psi_i = \varphi_i$ and $\psi_j = 1 - \psi_i = 1 - \varphi_i$. Then, since $\varphi_1 + \ldots + \varphi_k = 1$,

$$E(1 - \psi_i) = E(1 - \varphi_i) \leq \epsilon \quad \text{for all} \quad F \quad \text{in} \quad \mathcal{C}_i, \text{ and}$$

$$E(1 - \psi_j) = E(\varphi_i) \leq E(1 - \varphi_j) \leq \epsilon \quad \text{for all} \quad F \quad \text{in} \quad \mathcal{C}_j,$$

and $\mathcal{C}_i$ and $\mathcal{C}_j$ are (finitely) distinguishable.

On the other hand, assume $\mathcal{C}_1, \ldots, \mathcal{C}_k$ are pairwise (finitely) distinguishable. By Theorem 1.8, for each $\epsilon > 0$, and for each pair (i, j), $1 \leq i < j \leq k$, there is a rule $(N(i,j), \varphi_i(j), \varphi_j(i))$ where $N(i, j)$, $\varphi_i(j)$, and $\varphi_j(i)$ are non-randomized, such that

$$E(1 - \varphi_i(j)) < \epsilon \quad \text{for all} \quad F \quad \text{in} \quad \mathcal{C}_i, \text{ and}$$

$$E(1 - \varphi_j(i)) < \epsilon \quad \text{for all} \quad F \quad \text{in} \quad \mathcal{C}_j,$$

where $\varphi_i(j)$ is the conditional probability of deciding $F \in \mathcal{C}_i$ given $y_1, \ldots, y_{N(i,j)}$, when $\mathcal{C}_j$ is the competitor. Let $N = \max\{N(i, j)\}$ and

$$\Delta_i = \sum_{j \neq i} \varphi_i(j).$$

Δ_i is just the sum of the conditional probabilities of deciding $F \in \mathcal{C}_i$ over all competitors, where the dependence on $N(i, j)$ has been suppressed. Note that if for some i $\Delta_i = k - 1$, then $\varphi_i(j) = 1$ for all $j \neq i$, which implies $\varphi_j(i) = 0$ and $\Delta_j \leq k - 2$ for all

$j \neq i$. Define a decision rule $\psi = (\psi_1, \ldots, \psi_k)$ such that

$$\psi_i(y_1, \ldots, y_N) = 0 \quad \text{whenever} \quad \Delta_i < \max_j \Delta_j.$$

Thus $\Delta_i = k - 1$ implies $\psi_i = 1$, since $\Delta_j \leq k - 2$ and $\psi_j = 0$ for all $j \neq i$. Then, for $F \in \mathcal{G}_i$,

$$E(1 - \psi_i) \leq P[\psi_i < 1] \leq P[\Delta_i < k - 1].$$

Since $\varphi_i(j)$ is non-randomized,

$$E(1 - \psi_i) \leq P[\Delta_i < k - 1] = P[\varphi_i(j) = 0 \text{ for some } j \neq i]$$

$$\leq \sum_{j \neq i} P[\varphi_i(j) = 0]$$

$$= \sum_{j \neq i} E(1 - \varphi_i(j))$$

$$\leq (k-1)\epsilon \quad \text{when} \quad F \in \mathcal{G}_i, \; i = 1, \ldots, k.$$

Thus $\mathcal{G}_1, \ldots, \mathcal{G}_k$ are (finitely) distinguishable. □

Corollary.

$C(\Omega)$ is (finitely) classifiable if and only if $\mathcal{G}_1(\Omega), \ldots, \mathcal{G}_k(\Omega)$ are pairwise (finitely) distinguishable.

The analogue of Lemma 1.5 gives the properties of distances satisfying conditions A and B necessary to obtain the analogues of Theorems 1.10-1.11.

Lemma 1.7.

If the family $\{\delta_n\}$ $n = 1, \ldots, k + 1$ satisfies conditions A and B, then

a) for $G_1 \times \ldots \times G_k \in \Omega$,

$$\delta_{k+1}(G_j \times G_1 \times \ldots \times G_k, \mathcal{G}_i(\Omega)) = 0 \quad \text{iff} \quad \delta_1(G_j, G_i) = 0,$$

b) $\delta_{k+1}(\mathcal{G}_j, \mathcal{G}_i) = 0$ iff there exists a sequence $G_1^{(n)} \times \ldots \times \mathcal{G}_k^{(n)}$ in Ω such that $\lim \delta_1(G_j^{(n)}, G_i^{(n)}) = 0$.

The proof of this lemma is analogous to the proof of Lemma 1.5, which is just the case $k = 2$.

To complete the analogy to Section 1.5, we need

$$\Delta(i, j) = \inf_{\Omega}\{\delta_1(G_i, G_j)\}, \quad \Delta(\Omega) = \inf_{i \neq j} \{\Delta(i, j)\},$$

$$\Omega_\delta(i, j) = \{G_1 \times \ldots \times G_k \in \Omega: \delta_1(G_i, G_j) = 0\}, \quad \Omega_\delta = \bigcup_{i \neq j} \Omega_\delta(i, j),$$

$$\Omega_0(i, j) = \{G_1 \times \ldots \times G_k \in \Omega: G_i = G_j\}, \text{ and } \Omega_0 = \bigcup_{i \neq j} \Omega_0(i, j).$$

Although other theorems could be proved, the principal results are the following.

Theorem 1.15.

If $\{\delta_i\}$ $i = 1, \ldots, k + 1$ satisfy conditions A and B, and if δ_{k+1} is uniformly consistent in $\mathcal{G}_1 \cup \ldots \cup \mathcal{G}_k$, then $\Omega_\delta = \emptyset$ or $\Delta(\Omega) > 0$ are sufficient for $C(\Omega)$ to be classifiable or finitely classifiable, respectively.

Proof:

If $\Delta(\Omega) > 0$, then $\Delta(i, j) > 0$ for all $i \neq j$. By Lemma 1.7b), $\Delta(i, j) > 0$ implies $\delta_{k+1}(\mathcal{G}_i(\Omega), \mathcal{G}_j(\Omega)) > 0$, so by Theorem 1.2, $\mathcal{G}_i(\Omega)$ and $\mathcal{G}_j(\Omega)$ are finitely distinguishable, for all $i \neq j$. By Theorem 1.14, $\mathcal{G}_i(\Omega), \ldots, \mathcal{G}_k(\Omega)$ are finitely distinguishable, thus $C(\Omega)$ is finitely classifiable.

If $\Omega_\delta = \emptyset$, then $\Omega_\delta(i, j) = \emptyset$, for all $i \neq j$. By Lemma 1.7a), $\Omega_\delta(i, j) = \emptyset$ implies, for all F in $\mathcal{G}_i(\Omega) \cup \mathcal{G}_j(\Omega)$,

$$\max[\delta_{k+1}(F, \mathcal{G}_i(\Omega)), \delta_{k+1}(F, \mathcal{G}_j(\Omega))] > 0,$$

so by Theorem 1.1, $\mathcal{G}_i(\Omega)$ and $\mathcal{G}_j(\Omega)$ are distinguishable. By Theorem 1.14, $\mathcal{G}_1(\Omega),\ldots,\mathcal{G}_k(\Omega)$ are distinguishable, thus $C(\Omega)$ is classifiable. □

Theorem 1.16.

If $\{\delta_i\}$ $i = 1,\ldots, k+1$ satisfy conditions A and B, δ_{k+1} is uniformly consistent in $\mathcal{G}_1(\Omega) \cup \ldots \cup \mathcal{G}_k(\Omega)$, and $\delta_1(G_i, G_j) = 0$ implies $G_i = G_j$, then $C(\Omega)$ is classifiable if and only if $\Omega_0 = \emptyset$.

Proof:

Combine Theorem 1.15 with a lemma similar to Lemma 1.6. □

Corollary.

If $\Omega \subset \mathcal{F}^k$, where $\mathcal{F}$ is a collection of distributions on a Euclidean space, then $C(\Omega)$ is classifiable if and only if $\Omega_0 = \emptyset$.

Proof:

Note that $\{D_n\}$ satisfies the conditions of Theorem 1.16. □

1.7 Equivalence Relations.

If the experimenter is willing to match population π_0 with population π_j when something less stringent than $F_0 = F_j$ obtains, the preceding theory must be modified. In this section, we assume a match cannot occur unless $F_0 \sim F_j$, where $\sim$ is an equivalence relation, and that there is only one correct match.

Definition 1.14.

$\sim$ is an equivalence relation on a class $\mathcal{F}$, if for all F, G, and H in $\mathcal{F}$,

a) $F \sim F$,

b) $F \sim G$ implies $G \sim F$, and

c) $F \sim G$ and $G \sim H$ implies $F \sim H$.

Obviously, equality is an equivalence relation. If $\mathcal{F}$ is the collection of distributions on the real line with finite means, then

$F \sim G$ if and only if $\left|\int x dF - \int x dG\right| = 0$

defines an equivalence relation on $\mathfrak{F}$.

Definition 1.15.

A distance δ is compatible with an equivalence relation on $\mathfrak{F}$ if, for any F and G in $\mathfrak{F}$, $F \sim G$ implies $\delta(F, G) = 0$.

Definition 1.16.

A distance δ is congruent with an equivalence relation on $\mathfrak{F}$ if, for any F and G in $\mathfrak{F}$,

$F \sim G$ if, and only if, $\delta(F, G) = 0$.

Note that any distance is compatible with equality by definition, and distances congruent with equality were used in Theorems 1.12 and 1.16 to obtain the simple necessary and sufficient conditions for classifiability.

We will also use the following modified definitions. Let

$$\tilde{\mathcal{G}}(\Omega) = \{F \times G \times H: F \sim G \text{ and } G \times H \in \Omega\} \text{ and}$$

$$\tilde{\mathcal{H}}(\Omega) = \{F \times G \times H: F \sim H \text{ and } G \times H \in \Omega\}.$$

Now, we can define $C(\Omega)$ to be classifiable if and only if $\tilde{\mathcal{G}}(\Omega)$ and $\tilde{\mathcal{H}}(\Omega)$ are distinguishable.

An examination of the proof of Lemma 1.5, shows that, since the proof is entirely in terms of distances, the only way equality is used is through $\delta_1(G, G) = 0$. Thus, using an equivalence relation and a compatible distance, we get the following analogue of Lemma 1.5.

Lemma 1.8.

If $\{\delta_i\}$ $i = 1,2,3$ satisfy conditions A and B and δ_1 is compatible with $\sim$, then

a) for $F \times G \times H \in \tilde{\mathcal{G}}(\Omega)$,

$$\delta_3(F \times G \times H, \tilde{\mathcal{H}}(\Omega)) = 0 \text{ iff } \delta_1(G, H) = 0, \text{ and}$$

for $F \times G \times H \in \tilde{\mathcal{H}}(\Omega)$,

$$\delta_3(F \times G \times H, \tilde{\mathcal{G}}(\Omega)) = 0 \text{ iff } \delta_1(G, H) = 0.$$

b) $\delta_3(\tilde{\mathcal{G}}(\Omega), \tilde{\mathcal{H}}(\Omega)) = 0$ iff there exists a sequence $\{G_j \times H_j\}$ in Ω such that $\lim \delta_1(G_j, H_j) = 0$.

Proof:

The proof of Lemma 1.5 can be repeated with minor revisions. □

Again we define

$$\Delta(\Omega) = \inf_{\Omega} \delta_1(G, H),$$

$$\Omega_\delta = \{G \times H \in \Omega: \delta_1(G, H) = 0\}, \text{ and}$$

$$\Omega_\sim = \{G \times H \in \Omega: G \sim H\}.$$

The following theorems, stated without proof, are the analogues of the important theorems of previous sections.

Theorem 1.17.

If $\{\delta_i\}$ $i = 1,2,3$ satisfy conditions A and B, δ_1 is compatible with $\sim$, and δ_3 is uniformly consistent in $\tilde{\mathcal{G}}(\Omega) \cup \tilde{\mathcal{H}}(\Omega)$, then

$$\Omega_\delta = \emptyset \text{ or } \Delta(\Omega) > 0$$

are sufficient for $C(\Omega)$ to be classifiable or finitely classifiable, respectively.

Theorem 1.18.

If $\{\delta_i\}$ $i = 1,2,3$ satisfy conditions A and B, δ_1 is congruent with $\sim$, and δ_3 is uniformly consistent in $\tilde{\mathcal{G}}(\Omega) \cup \tilde{\mathcal{H}}(\Omega)$, then $C(\Omega)$ is classifiable iff $\Omega_\sim = \emptyset$.

An example of the use of these theorems concerns the collection $\mathfrak{F}$ of all distributions on the real line with finite means and variances bounded by M. Writing $\mu(F)$ for the mean of F, we will say $F \sim G$ if, and only if, $\mu(F) = \mu(G)$. For $F = F_1 \times \ldots \times F_n$ and $G = G_1 \times \ldots \times G_n$ in $\mathfrak{F}^n$ define

$$\delta_n(F, G) = \left(\sum_{i=1}^{n} (\mu(F_i) - \mu(G_i))^2\right)^{\frac{1}{2}}.$$

Such a family satisfies conditions A and B, and δ_1 is congruent with $\sim$. We can extend the definition of δ_n to all distributions on n-dimensional Euclidean space by interpreting the F_i and G_i in the definition as the marginal distributions. Thus, for any distributions F and G on R^3,

$$\begin{aligned} P[\delta_3(F, G) > \epsilon] &= P[\sum_{1}^{3} (\mu(F_i) - \mu(G_i))^2 > \epsilon^2] \\ &\leq \sum_{1}^{3} P[(\mu(F_i) - \mu(G_i))^2 > \epsilon^2/3] \\ &= \sum_{1}^{3} P[|\mu(F_i) - \mu(G_i)| > \epsilon/\sqrt{3}]. \end{aligned}$$

In particular, if G is the sample distribution based on m observations of (X_1, X_2, X_3),

$$P[\delta_3(F, G) > \epsilon] \leq \sum_{1}^{3} P[|\mu(F_i) - \bar{X}_i| > \epsilon/\sqrt{3}] \leq \frac{9M}{m\epsilon^2},$$

so δ_3 is uniformly consistent in $\tilde{\mathcal{G}}(\Omega) \cup \tilde{\mathcal{H}}(\Omega)$. Thus, by Theorem 1.18, $C(\Omega)$ is classifiable if, and only if, Ω contains no pair of alternatives with the same mean. Similarly, by Theorem 1.17, if we assume the alternatives always have a difference between their means greater than $\Delta > 0$, $C(\Omega)$ is finitely classifiable.

1.8 Classification of Normal Distributions: Sampling Fewer Than Three Populations.

In certain situations, samples are not needed from all three populations in order to construct rules with arbitrarily small error probabilities. We consider the special case of normal populations, first, when Ω is such that only observations on X_0 are needed, and then when Ω is such that only observations on X_0 and X_1 are needed.

If only observations from X_0 are available, the following theorem completely characterizes the type of Ω for which $C(\Omega)$ is still classifiable.

Theorem 1.20.

$C(\Omega)$ is(finitely) classifiable on the basis of observations on X_0 if, and only if, there exist $\mathcal{G}$ and $\mathcal{H}$ such that $\mathcal{G}$ and $\mathcal{H}$ are (finitely) distinguishable and $\Omega \subset \mathcal{G} \times \mathcal{H}$.

Proof:

If $C(\Omega)$ is classifiable on the basis of observations on X_0 alone, define $\mathcal{G}_1$ and $\mathcal{H}_1$ by $\mathcal{G}_1 = \{G: (G \times H) \in \Omega \text{ for some } H\}$ and $\mathcal{H}_1 = \{H: (G \times H) \in \Omega \text{ for some } G\}$. Note that if F is the distribution of X_0 and $F \in \mathcal{G}_1$, then there is a $(G \times H) \in \Omega$ such that $F \times G \times H \in \mathcal{G}(\Omega)$. Thus if (N, φ) is a classification rule depending on X_0 alone for which

$$E(\varphi) \leq \varepsilon \text{ for all } F \times G \times H \in \mathcal{G}(\Omega),$$

$$E(1-\varphi) \leq \varepsilon \text{ for all } F \times G \times H \in \mathcal{H}(\Omega), \text{ and}$$

$$P[N < \infty] = 1 \text{ for all } F \times G \times H \in \mathcal{G}(\Omega) \cup \mathcal{H}(\Omega),$$

then, since $E(\varphi)$ depends only on F (the distribution of X_0), we

have

$$P[N < \infty] = 1 \quad \text{and} \quad E(\varphi) \leq \epsilon \quad \text{for all} \quad F \in \mathcal{G}_1.$$

Similarly, we obtain

$$P[N < \infty] = 1 \quad \text{and} \quad E(1-\varphi) \leq \epsilon \quad \text{for all} \quad F \in \mathcal{H}_1.$$

Thus, if $C(\Omega)$ is classifiable on the basis of X_0 alone, then $\mathcal{G}_1$ and $\mathcal{H}_1$ are distinguishable and $\Omega \subset \mathcal{G}_1 \times \mathcal{H}_1$. In the finitely classifiable case, an analogous argument holds. To complete the proof, note that if $\Omega \subset \mathcal{G} \times \mathcal{H}$, where $\mathcal{G}$ and $\mathcal{H}$ are (finitely) distinguishable, then a test of $F \in \mathcal{G}$ versus $F \in \mathcal{H}$ serves to classify $C(\Omega)$ based only on observations from X_0. □

For the particular case of normal distributions, Theorem 1.20 and Theorem 1.7 together yield necessary and sufficient conditions for (finite) classifiability based on X_0 in terms of the measure of affinity ρ (Section 1.2). To find conditions for classifiability based only on X_0 and X_1 when X_0 and X_1 have normal distributions, we need the following lemma.

Lemma 1.9.

Let F_i and G_i be in $\mathcal{P}_k$, with means μ_i and ν_i and positive definite covariance matrices Σ_i and Ψ_i. Then $\lim_{i \to \infty} \rho(F_i, G_i) = 1$ if and only if

a) all the characteristic roots of $\Sigma_i^{-1} \Psi_i$ converge to 1 and

b) $\lim_{i \to \infty} (\mu_i - \nu_i)'(\Sigma_i + \Psi_i)^{-1}(\mu_i - \nu_i) = 0$.

Proof:

For any $F, G \in \mathcal{P}_k$ with means μ and ν and positive definite covariance matrices Σ and Ψ, note that $\rho(F, G)$ can be considered as the product of two factors,

$$\rho_1(\Sigma, \psi) = |\Sigma|^{1/4}|\psi|^{1/4}\left|\frac{\Sigma + \psi}{2}\right|^{-1/2}, \text{ and}$$

$$\rho_2(\mu, \nu, \Sigma, \psi) = \exp\{-\tfrac{1}{4}(\mu-\nu)'(\Sigma + \psi)^{-1}(\mu-\nu)\}.$$

Also, $0 \leq \rho_i \leq 1$, so that $\rho(F_i, G_i)$ converges to 1 if, and only if, ρ_1 and ρ_2 converge to 1. Thus, we need only show that ρ_1 converges to 1 iff a) obtains and ρ_1 converges to 1 iff b) obtains.

First, note that ρ_1 can be rewritten as

$$\rho_1(\Sigma, \psi) = |\Sigma|^{1/4}|\psi|^{1/4}|\Sigma|^{-1/2}\left|\frac{I + \Sigma^{-1}\psi}{2}\right|^{-1/2}$$

$$= |\Sigma^{-1}\psi|^{1/4}\left|\frac{I + \Sigma^{-1}\psi}{2}\right|^{-1/2}.$$

Thus, if λ_j, $j = 1,\ldots, k$ are the characteristic roots of $\Sigma^{-1}\psi$, then

$$\rho_1(\Sigma, \psi) = \prod_{j=1}^{k} \lambda_j^{1/4}\left(\frac{1+\lambda_j}{2}\right)^{-1/2} = \prod_{j=1}^{k}\left[\frac{4\lambda_j}{(1+\lambda_j)^2}\right]^{1/4}$$

$$= \prod_{j=1}^{k}\left[\frac{(1+\lambda_j)^2 - (1-\lambda_j)^2}{(1+\lambda_j)^2}\right]^{1/4} = \prod_{j=1}^{k}\left[1 - \left(\frac{1-\lambda_j}{1+\lambda_j}\right)^2\right]^{1/4}.$$

Thus, it is immediate that

$\lim \rho_1(\Sigma_i, \psi_i) = 1$ if, and only if, a) obtains,

i.e., all the characteristic roots of $\Sigma_i^{-1}\psi_i$ converge to 1. Also, since ρ_2 is e to a negative power,

$\lim \rho_2(\mu_i, \nu_i, \Sigma_i, \psi_i) = 1$ if,and only if, b) obtains,

i.e., $\lim(\mu_i - \nu_i)'(\Sigma_i + \psi_i)^{-1}(\mu_i - \nu_i) = 0$. By the comments of the paragraph above, this completes the proof. □

If we ignore X_2 (and H) and try to classify $C(\Omega)$ on the basis of X_0 and X_1, alone, then, equivalently, we are trying to distinguish

$$\mathcal{G}_2(\Omega) = \{F \times G: F = G \text{ for some } G \times H \in \Omega\} \text{ and}$$

$$\mathcal{H}_2(\Omega) = \{F \times G: F = H \text{ for some } G \times H \in \Omega\}.$$

Note that $F \times G \in \mathcal{H}_2(\Omega)$ if, and only if, $G \times F \in \Omega$.

In the particular case of univariate normal distributions with common unknown variance, $\mathcal{G}_2(\Omega)$ and $\mathcal{H}_2(\Omega)$ become

$$\mathcal{G}_2 = \{F \times G: F, G, H \in \mathcal{H}_1, \eta \neq \mu = \nu, \theta^2 = \sigma^2 = \tau^2 > 0 \text{ for some } G \times H \in \Omega\} \text{ and}$$

$$\mathcal{H}_2 = \{F \times G: F, G, H \in \mathcal{H}_1, \eta = \mu \neq \nu, \theta^2 = \sigma^2 = \tau^2 > 0 \text{ for some } G \times H \in \Omega\},$$

where μ, ν, η, σ^2, θ^2, and τ^2 are the means and variances of F, G, and H, the distributions of X_0, X_1 and X_2. Implicit in the definitions of $\mathcal{G}_2$ and $\mathcal{H}_2$ is the assumption that Ω contains no $G \times H$ for which $G = H$.

Theorem 1.21.

a) If

$$\inf_{\Omega} \frac{|\nu-\eta|}{\tau} > 0,$$

then $C(\Omega)$ is finitely classifiable on the basis of X_0 and X_1.

b) If

$$\inf_{\Omega} |\nu-\eta| > 0,$$

then $C(\Omega)$ is classifiable on the basis of X_0 and X_1.

Proof:

a) By Lemma 1.9, $\rho(\mathcal{G}_2, \mathcal{H}_2) = 1$ if, and only if, there exist $F_i \times G_i \in \mathcal{G}_2$ with means μ_i and variances σ_i^2 and $H_i \times K_i \in \mathcal{H}_2$ with means ν_i and η_i and variances τ_i^2 such that

$$\lim_{i \to \infty} (\sigma_i^2/\tau_i^2) = 1 \text{ and} \tag{1.3}$$

(1.4) $$\lim_{i\to\infty} ((\mu_i - \nu_i)^2 + (\mu_i - \eta_i)^2)/(\sigma_i^2 + \tau_i^2) = 0 .$$

If $|\nu_i - \eta_i| / \tau_i \geq \epsilon$, then

$$(x-a)^2 + (x-b)^2 \geq (a-b)^2/2$$

implies

(1.5) $$((\mu_i - \nu_i)^2 + (\mu_i - \eta_i)^2)/(\sigma_i^2 + \tau_i^2) \geq (\mu_i - \eta_i)^2/2\tau_i^2(\frac{\sigma_i^2}{\tau_i^2} + 1)$$

$$\geq \epsilon^2/2\ (\frac{\sigma_i^2}{\tau_i^2} + 1).$$

Thus, since

$$\inf_{\Omega} \frac{|\nu-\eta|}{\tau} > 0 \Rightarrow \inf_{\mathcal{H}_2} \frac{|\nu-\eta|}{\tau} > 0,$$

(1.5) implies that (1.3) and (1.4) cannot occur simultaneously and hence that $\rho(\mathcal{C}_2, \mathcal{H}_2) < 1$. By Theorem 1.7, $\mathcal{C}_2$ and $\mathcal{H}_2$ are finitely distinguishable ($C(\Omega)$ is finitely classifiable) on the basis of X_0 and X_1.

b) By Theorem 1.7, to show $C(\Omega)$ is classifiable on the basis of X_0 and X_1, it is sufficient to verify that, for all $F \times G \in \mathcal{C}_2 \cup \mathcal{H}_2$,

$$\min[\rho(F \times G, \mathcal{C}_2), \rho(F \times G, \mathcal{H}_2)] < 1.$$

If $F \times G \in \mathcal{C}_2$, both with means μ and variances σ^2, then $\rho(F \times G, \mathcal{H}_2) = 1$ if, and only if, there exist $F_i \times G_i$ in $\mathcal{H}_2$ with means ν_i and η_i and common variances τ_i^2 such that

$$\lim_{i\to\infty} \tau_i^2 = \sigma^2$$

$$\lim_{i\to\infty} ((\mu-\nu_i)^2 + (\mu-\eta_i)^2)/(\sigma^2 + \tau_i^2) = 0.$$

Now $|\nu_i - \eta_i| \geq \epsilon$ implies

$$((\mu-\nu_i)^2 + (\mu-\eta_i)^2)/(\sigma^2+ \tau_i^2) \geq \varepsilon^2/2(\sigma^2 + \tau_i^2),$$

so

$$\inf_\Omega |\nu-\eta| > 0 \Rightarrow \rho(F \times G, \mathcal{H}_2) < 1.$$

Similarly if $F \times G \in \mathcal{H}_2$, then $\rho(F \times G) < 1$. Thus $\inf_\Omega|\nu-\eta| > 0$ implies $\mathcal{C}_2$ and $\mathcal{H}_2$ are distinguishable ($C(\Omega)$ is classifiable) on the basis of X_0 and X_1. □

If X_0, X_1, and X_2 have a common known variance σ_0^2, then the conditions for finite and sequential classifiability of $C(\Omega)$ are weaker. Note that in this case,

$$\inf_\Omega \frac{|\nu-\eta|}{\sigma} = \sigma_0^{-1} \inf_\Omega |\nu-\eta|$$

so when $\inf_\Omega |\nu-\eta| > 0$, Theorem 1.21 a) implies $C(\Omega)$ is finitely classifiable. The condition for sequential classifiability is not so simple. If $F_0 \times G_0 \in \mathcal{C}_2$ with equal means μ_0, then $\rho(F_0 \times G_0, \mathcal{H}_2) = 1$ if, and only if, there exist $F_i \times G_i \in \mathcal{H}_i$ with means ν_i and η_i such that

$$\lim_{i\to\infty} [(\mu_i- \mu_0)^2 + (\nu_i- \mu_0)^2] = 0.$$

A condition necessary and sufficient to insure $\rho(F_0 \times G_0, \mathcal{H}_2) < 1$ is thus

(1.6) $\quad \forall\, G_0 \times H_0 \in \Omega$ (G_0 with mean ν_0),

there exists $\varepsilon > 0$ such that

$$\inf_{\Omega, |\nu-\nu_0|\leq\varepsilon} |\nu-\eta| > 0,$$

since this insures that ν_i and η_i cannot both converge to the same finite limit ν_0. Condition (1.6) is actually a necessary and

sufficient condition for the classifiability of $C(\Omega)$ on the basis of X_0 and X_1 alone, since $\rho(F \times G, G_2)$ is always less than one if F and G have different means.

Further results, similar to the above, can be found when we do not assume that X_1 and X_2 have equal variances, but sufficient conditions are no longer as easy to state in terms of Ω. In all cases involving normal distributions, it is Theorem 1.7 which gives necessary and sufficient conditions for distinguishability in terms of the measure of affinity ρ, and only by applying Theorem 1.7 do we get a reduction to more natural conditions for classification in terms of restrictions on Ω.

1.9 Discussion.

There are several points of the preceding theory which should be clarified. Among these are the basic assumptions and the sampling method.

The first assumption is that the distribution F_0 of observations from π_0 is equal (or at least equivalent) to some F_j. In this regard, our work is distinct from that of Cacoullos [6], who considered finding the population closest to π_0. Our theory also does not apply to the classification problem where π_0 is assumed to be at a distance less than ϵ from the closest population.

The second assumption is that there is only one correct match, even if two alternative distributions are equal. This implies that even if populations π_1 and π_2 cannot be distinguished on the basis of X_1 and X_2, there are other characteristics which do allow one (theoretically) to tell π_1 from π_2. Using those same character-

istics, π_0 can be matched with exactly one of π_1 and π_2. However, if $F_1 = F_2$, the experimenter is evidently measuring the wrong characteristic and can do no better than guessing. If such an assumption were not made, we could replace it by the assumption that π_0 is correctly matched with π_j if, and only if, $F_0 = F_j$. Thus, if $F_1 = F_2$, any match would be correct, but there may not exist stopping rules for which N is finite with probability one. If we knew $F_1 = F_2$, we would do no sampling and match π_0 with any π_j, while if we knew $F_1 \neq F_2$, we would apply all of the theory developed above. Unfortunately, deciding whether $F_1 = F_2$ or $F_1 \neq F_2$ may not be possible. For example, if all F are univariate normal with variance 1, then we must decide (with error probabilities less than ϵ) whether $\mu_2 - \mu_1 = 0$ or $\mu_2 - \mu_1 \neq 0$. This is equivalent to testing the mean of $Y = X_2 - X_1$, which, using Theorem 1.7, we can show cannot be done with $P[N < \infty] = 1$.

Our other important assumptions involve sampling. Except as noted in Section 1.8, our tests are based on repeated sampling of the vector $Y = (X_0, X_1, X_2)$. This is equivalent to requiring equal sample sizes at each stage of experimentation. From a theoretical point of view, this is not much of a restriction. For example, if $C(\Omega)$ is classifiable and any sampling scheme is used such that $\min(n_1, n_2, n_3)$ increases without limit (where n_i is the number of samples from X_i), then, by properly ignoring observations, we can consider the rest of the observations to be taken a vector at a time and use the rules we know to exist. Similarly, if $C(\Omega)$ is finitely classifiable and $n(\epsilon)$ observations on Y are needed to have maximum error less than ϵ, then for any sampling scheme with $\min(n_i) \geq n(\epsilon)$, there will be a test based on that sample with maximum error less than ϵ.

CHAPTER II

Some Sequential Classification Rules

2.1 Introduction and Summary.

The theory of the previous chapter gives conditions under which $C(\Omega)$ is classifiable. In this chapter, we will describe several classes of sequential rules for actually achieving maximum error probabilities less than ϵ.

Two of these classes, described in Section 2.2, are based on uniformly consistent distances between distributions. The first is based on a class of sequential rules given by Hoeffding and Wolfowitz [16] for the problem of distinguishing two sets of distributions $\mathcal{G}$ and $\mathcal{H}$. These rules can be applied directly to the sets of distributions $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ associated with the classification problem $C(\Omega)$. The second class of rules, which we call minimum distance rules, are based on computing the distances between the sample distribution based on observations from X_0 and the sample distributions based on X_1 and X_2. Sampling on $Y = (X_0, X_1, X_2)$ continues until one of the sample distances is large, at which point the decision is made that X_0 has the same distribution as that X_i corresponding to the smaller sample distance. Theorem 2.2 verifies that such a rule can be chosen with arbitrarily small maximum error probabilities.

A third class of classification rules is based on sequential tests of power one. As explicated by H. Robbins [24], tests of power one for H versus K exist if there is a stopping rule such that

$$P[N < \infty] \leq \epsilon \quad \text{under} \quad H \quad \text{and}$$

$$P[N < \infty] = 1 \quad \text{under} \quad K.$$

We exploit the existence of such stopping rules in simultaneously testing $F = G$ and $F = H$, where F, G, and H are the distributions of X_0, X_1, and X_2. These rules are described in Section 2.3, where we prove that they have maximum error probabilities less than ϵ (Theorem 2.3) and give examples of the stopping rules of interest.

2.2 Minimum Distance Rules.

For the problem of distinguishing two sets of distributions $\mathcal{G}$ and $\mathcal{H}$ based on observations on Y, Hoeffding and Wolfowitz described a class of rules with arbitrarily small error probabilities. We first present that class and then make use of it. Let δ be a distance $\{c(i)\}$ a sequence of positive numbers, and $\{n(i)\}$ an increasing sequence of positive integers. To define a rule, let

$$(2.1) \qquad \delta(i) = \max[\delta(F_{n(i)}, \mathcal{G}), \delta(F_{n(i)}, \mathcal{H})],$$

where $F_{n(i)}$ is the sample distribution based on $n(i)$ independent observations on Y, and then take successive samples of sizes $n(1)$, $n(2) - n(1)$, $n(3) - n(2)$,... until $\delta(i) \geq c(i)$. If i is the least integer for which $\delta(i) \geq c(i)$, then let $N = n(i)$ and apply the terminal decision rule

$$\varphi = \begin{cases} 1 \text{ , if } \delta(F_N, \mathcal{G}) \geq \delta(F_N, \mathcal{H}) \text{ and} \\ 0 \text{ , otherwise .} \end{cases}$$

This rule (N, φ) is also denoted by $T(\delta, c(i), n(i))$.

Hoeffding and Wolfowitz use these rules in proving their Theorem 3.1 (our Theorem 1.1) which we restate here in terms of $T(\delta, c(i), n(i))$.

Theorem 2.1.

If δ is uniformly consistent in $\mathfrak{F}$, $\mathcal{G}$ and $\mathcal{H}$ are subsets of $\mathfrak{F}$, and

$$\max[\delta(F, \mathcal{G}), \delta(F, \mathcal{H})] > 0 \quad \text{for all} \quad F \in \mathcal{F},$$

then, for all $\epsilon > 0$, there exist sequences $\{c(i)\}$ and $\{n(i)\}$ such that the rule $(N, \varphi) = T(\delta, c(i), n(i))$ based on observations on Y with distribution F satisfies

$$E_F(\varphi) \leq \epsilon \quad \text{if} \quad F \in \mathcal{G}, \; E_F(1-\varphi) \leq \epsilon \quad \text{if} \quad F \in \mathcal{H}, \text{ and}$$

$$P_F[N < \infty] = 1 \quad \text{if} \quad F \in \mathcal{F}.$$

Note that the conclusions of Theorem 2.1 imply by definition that $\mathcal{G}$ and $\mathcal{H}$ are distinguishable. We will not give a proof of Theorem 2.1, as it can be found in H-W, but we will display a modified version of it in the proof of Theorem 2.2 below.

If we are interested in the classification problem $C(\Omega)$ then a direct comparison with the problem of distinguishing $\mathcal{G}(\Omega)$ and $\mathcal{H}(\Omega)$ (as was done in the proof of Theorem 1.9) shows that they are in fact identical problems. Thus a rule (N, φ) for $C(\Omega)$ with maximum error probabilities less than ϵ can be found among the class of rules $T(\delta, c(i), n(i))$ based on $Y = (X_0, X_1, X_2)$ if δ is uniformly consistent in $\mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$. In fact, an exact prescription for $c(i)$ and $n(i)$ can be easily given. Let $c(i)$ be any positive sequence converging to zero and let $\alpha(i)$ be any positive constants such that $\Sigma\alpha(i) \leq \epsilon$. Then choose $n(1) < n(2) < n(3) \ldots$ such that

$$P_F[\delta(F_{n(i)}, F) \geq c(i)] \leq \alpha(i), \; i = 1,2,3,\ldots$$

for all $F \in \mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$. This is possible since δ is uniformly consistent in $\mathcal{G}(\Omega) \cup \mathcal{H}(\Omega)$.

One difficulty with the rule $T(\delta, c(i), n(i))$ is the calculation of $\delta(i)$ as defined in (2.1). It involves first computing $F_{n(i)}$,

the sample distribution of $Y = (X_0, X_1, X_2)$, and then computing $\inf \delta(F_{n(i)}, G)$, $G \in \mathcal{G}(\Omega)$, and $\inf \delta(F_{n(i)}, H)$, $H \in \mathcal{H}(\Omega)$. For example, in the simplest case when each X_i is a real valued random variable, Y has a trivariate distribution, and, although each $G \in \mathcal{G}(\Omega)$ will be a product distribution, $F_{n(i)}$ in general will not be, so that finding the infimum of $\delta(F_{n(i)}, G)$ can be difficult.

These difficulties are avoided in the following class of rules, called minimum distance rules. Let F, G, and H denote the distribution of X_0, X_1, and X_2 respectively and define

$$\Delta(i) = \max[\delta(F_{n(i)}, G_{n(i)}), \delta(F_{n(i)}, H_{n(i)})],$$

where $F_{n(i)}$, $G_{n(i)}$, $H_{n(i)}$ are the sample distributions based on $n(i)$ observations on X_0, X_1, X_2 respectively. The minimum distance rule consists of taking samples of size $n(1)$, $n(2) - n(1)$,... until $\Delta(i) \geq c(i)$, then setting $N = n(i)$ and applying the terminal decision rule,

$$\varphi = \begin{cases} 1 & \text{if } \delta(F_N, G_N) \geq \delta(F_N, H_N) \text{ and} \\ 0 & \text{otherwise .} \end{cases}$$

This rule will be denoted by $M(\delta, c(i), n(i))$. The use of minimum distance rules is justified by the following theorem.

Theorem 2.2.

If Ω is a subset of $\mathcal{F} \times \mathcal{F}$ where δ is uniformly consistent in $\mathcal{F}$ and

$$\Omega_\delta = \{G \times H \in \Omega: \delta(G, H) = 0\} = \emptyset,$$

then,

(2.2) $\forall \epsilon > 0$, there exists a minimum distance rule $M(\delta, n(i), c(i))$ such that

$$E(\varphi) \leq \epsilon \quad \text{if} \quad F \times G \times H \in \mathcal{G}(\Omega),$$

$$E(1-\varphi) \leq \epsilon \quad \text{if} \quad F \times G \times H \in \mathcal{H}(\Omega), \text{ and}$$

$$P(N < \infty) = 1 \text{ if} \quad F \times G \times H \in \mathcal{G}(\Omega) \cup \mathcal{H}(\Omega).$$

Proof:

Let $\alpha(j) > 0$ be such that $\Sigma\alpha(j) \leq \epsilon$ and let $c(j)$ be a sequence of positive constants converging to 0. Then, since δ is uniformly consistent, there exist an increasing sequence of positive integers $n(j)$ such that

$$(2.3) \qquad P_F[\delta(K_{n(j)}, K) > \frac{c(j)}{2}\ ] \leq \frac{\alpha(j)}{2} \quad \text{for all} \quad K \in \mathcal{F}.$$

Then for the minimum distance rule $M(\delta, n(i), c(i))$, if $F = G$ and $G \times H \in \Omega$,

$$E[\varphi] = \sum_j P[\Delta(i) < c(i),\ i < j;\ \Delta(j) \geq c(j);\ \delta(F_{n(j)}, G_{n(j)}) \geq \delta(F_{n(j)}, H_{n(j)})]$$

$$\leq \sum_j P[\delta(F_{n(j)}, G_{n(j)}) \geq c(j)].$$

Since $\delta(F_{n(j)}, G_{n(j)}) \leq \delta(F_{n(j)}, F) + \delta(F, G) + \delta(G, G_{n(j)})$ and $F = G$ implies $\delta(F, G) = 0$,

$$E[\varphi] \leq \sum_j [\delta(F_{n(j)}, F) + \delta(G_{n(j)}, G) \geq c(j)].$$

Now,

$$P[\delta(F_{n(j)}, F) + \delta(G_{n(j)}, G) \geq c(j)]$$

$$\leq P[\delta(F_{n(j)}, F) \geq c(j)/2] + P[\delta(G_{n(j)}, G) \geq c(j)/2],$$

which by (2.3) yields

$$(2.4) \qquad P[\delta(F_{n(j)}, F) + \delta(G_{n(j)}, G) \geq c(j)] \leq \alpha(j).$$

Thus, for $F \times G \times H$ in $\mathcal{G}(\Omega)$,

$$E[\varphi] \leq \Sigma\alpha(j) \leq \epsilon.$$

Similarly, for $F \times G \times H$ in $\mathcal{H}(\Omega)$ $(F = H)$, $E[1-\varphi] \leq \epsilon$.

To prove $P[N < \infty] = 1$, note that

$$P[N > n(j)] \leq P[\Delta(j) < c(j)]$$

$$= P[\delta(F, F_{n(j)}) + \Delta(j) + \delta(H, H_{n(j)}) < \delta(F, F_{n(j)}) + c(j) + \delta(H, H_{n(j)})].$$

But

$$\delta(F, H) \leq \delta(F, F_{n(j)}) + \delta(F_{n(j)}, H_{n(j)}) + \delta(H_{n(j)}, H)$$

$$\leq \delta(F, F_{n(j)}) + \Delta(j) + \delta(H_{n(j)}, H)$$

so

$$P[N > n(j)] \leq P[\delta(F, H) < \delta(F, F_{n(j)}) + c(j) + \delta(H, H_{n(j)})]$$

$$= P[\delta(F, F_{n(j)}) + \delta(H, H_{n(j)}) > \delta(F, H) - c(j)].$$

Similarly, replacing H by G in the above argument,

$$P[N > n(j)] \leq P[\delta(F, F_{n(j)}) + \delta(G, G_{n(j)}) > \delta(F, G) - c(j)]. \tag{2.5}$$

Because Ω_δ is empty, $\delta(G, H)$ must be positive. If $F = H$, then $\delta(F, G) = \delta(H, G) > 0$. Since $c(j)$ converges to zero, for large j we have $c(j) < \delta(F, G) - c(j)$, which together with (2.5) implies

$$P[N > n(j)] \leq P[\delta(F, F_{n(j)}) + \delta(G, G_{n(j)}) > c(j)].$$

By the inequality (2.4),

$$P[N > n(j)] \leq \alpha(j).$$

Similarly, if $F = G$, then $\delta(F, H) > 0$ and, for large j,

$$P[N > n(j)] \leq P[\delta(F, F_{n(j)}) + \delta(H, H_{n(j)}) > c(j)] \leq \alpha(j).$$

Since $\Sigma\alpha(j) \leq \epsilon$ and $\alpha(j)$ converges to zero, we have

$$P[N < \infty] = \lim_{j \to \infty} P[N \leq n(j)] = 1,$$

for all $F \times G \times H$ in $\mathcal{C}(\Omega) \cup \mathcal{H}(\Omega)$. □

Corollary.

If Ω is a collection of pairs of distributions on a Euclidean space, then

$\Omega_0 = \emptyset$ if, and only if, (2.2) obtains.

Proof:

$\Omega_0 = \emptyset$ is necessary for $C(\Omega)$ to be classifiable, which would be true if (2.2) obtained. Since the Kolmogorov distance D is uniformly consistent on all distributions on a Euclidean space (see Section 1.2, equation (1.2)) and $D(F, G) = 0$ if, and only if, $F = G$, Theorem 2.2 applied to D implies that $\Omega_0 = \Omega_D = \emptyset$ is sufficient for (2.2) to obtain. □

Note that the above is nearly equivalent to the corollary to Theorem 1.12. However, the above corollary specifies the particular rules one can use to obtain arbitrarily small probabilities of error, i.e., the minimum distance rules based on D.

2.3 Classification Rules Adapting an Idea of Robbins.

Another class of classification rules can be constructed using a concept developed by H. Robbins and others in a series of papers ([7], [8], [24]). Using their concept of sequential tests of power one, we simultaneously test

(2.6) $H_{01}: F = G$ against $K_{01}: F \neq G$, and

$H_{02}: F = H$ against $K_{02}: F \neq H$,

where F, G, and H are the distributions of X_0, X_1, and X_2 respectively. The resulting procedure will have a prescribed maximum probability of error ε under the conditions of Theorem 2.3 below.

Specifically, we will consider collections of distributions $\mathfrak{F}$ satisfying the following condition:

(2.7) there exists a stopping time N (based on observations on X and Y with distributions G and H) for which G and H in $\mathfrak{F}$ implies

$$P[N < \infty] \leq \varepsilon \quad \text{if} \quad G = H, \text{ and}$$

$$P[N < \infty] = 1 \quad \text{if} \quad G \neq H.$$

For such an $\mathfrak{F}$, Robbins' test of power one of H_0: $G = H$ rejects H_0 if N is finite. To apply this concept to classification, we consider two tests of power one, one of H_{01}:$F = G$ based on N_{01}, and one of H_{02}: $F = H$ based on N_{02}; we sample until one of the hypotheses is rejected, at which point we accept the other. Formally, we define a rule (N, φ), denoted by $\underline{R(N_{01}, N_{02})}$, by

(2.8) $$N = \min\{N_{01}, N_{02}\}, \text{ and}$$

$$\varphi = \begin{cases} 1 & \text{if } N_{01} \leq N_{02}, \text{ and} \\ 0 & \text{if } N_{01} > N_{02}. \end{cases}$$

Theorem 2.3.

If Ω is a subset of $\mathfrak{F} \times \mathfrak{F}$, where $\mathfrak{F}$ satisfies (2.7), and if $\Omega_0 = \emptyset$, then there exists a rule $R(N_{01}, N_{02})$ such that

$$E(\varphi) \leq \varepsilon \quad \text{for} \quad F \times G \times H \in \mathcal{G}(\Omega),$$

$$E(1-\varphi) \leq \varepsilon \quad \text{for} \quad F \times G \times H \in \mathcal{H}(\Omega), \text{ and}$$

$$P[N < \infty] = 1 \quad \text{for} \quad F \times G \times H \in \mathcal{G}(\Omega) \cup \mathcal{H}(\Omega).$$

Proof:

First, $\Omega \subset \mathfrak{F} \times \mathfrak{F}$ implies both G and H are in $\mathfrak{F}$. Since either $F = G$ or $F = H$ obtains, F is also in $\mathfrak{F}$. Thus there exist two stopping times, N_{01}(based on X_0 and X_1) and N_{02} (based on X_0 and X_2), such that

$$P[N_{01} < \infty] \leq \epsilon \quad \text{if} \quad H_{01}: F = G,$$

$$P[N_{01} < \infty] = 1 \quad \text{if} \quad K_{02}: F \neq G,$$

$$P[N_{02} < \infty] \leq \epsilon \quad \text{if} \quad H_{02}: F = H, \text{ and}$$

$$P[N_{02} < \infty] = 1 \quad \text{if} \quad K_{02}: F \neq H.$$

For $F \times G \times H \in \mathcal{G}(\Omega)$, $\Omega_0 = \emptyset$ implies $G = F \neq H$. Thus for the rule $R(N_{01}, N_{02})$ defined by (2.8),

$$E(\varphi) = P[N_{01} \leq N_{02}] \leq P[N_{01} < \infty] + P[N_{02} = \infty] \leq \epsilon, \text{ and}$$

$$P[N < \infty] \geq P[N_{02} < \infty] = 1.$$

Similarly for $F \times G \times H \in \mathcal{H}(\Omega)$, $\Omega_0 = \emptyset$ implies $G \neq F = H$, and thus

$$E(1-\varphi) = P[N_{01} > N_{02}] \leq P[N_{02} < \infty] \leq \epsilon, \text{ and}$$

$$P[N < \infty] \geq P[N_{01} < \infty] = 1. \square$$

Theorem 2.3 can easily be extended to the case of k alternatives $G_1, \ldots, G_k$. If $\Omega \subset \mathfrak{F}^k$, $\mathfrak{F}$ satisfying (2.7), then there are k stopping times, $N_{01}, \ldots, N_{0k}$, (N_{0j} based on X_0 and X_j) such that

(2.9)
$$P[N_{0j} < \infty] \leq \epsilon, \text{ if } F = G_j,$$

$$P[N_{0j} < \epsilon] = 1, \text{ if } F \neq G_j.$$

Thus we can define $R(N_{01}, \ldots, N_{0k})$ by letting N be the least n such that $N_{0j} \leq n$ for $k - 1$ different j, and deciding $F = G_j$

for that j with $N_{0j} > N = n$. In other words, sample until $k - 1$ of the hypotheses $H_{0j}: F = G_j$ have been rejected and then accept the one hypothesis left. That all the probabilities of error are less than ϵ follows from (2.9), since, if $F = G_j$,

$$P[\text{reject } H_{0j}] \leq P[N_{0j} < \infty] \leq \epsilon.$$

In the rest of this section, we consider examples of $\mathfrak{F}$ and N satisfying (2.7). They are taken from Robbins [24]. In each case, there exists a sequence of statistics $T_i(X, Y)$ (based on i observations from X and Y with distributions G and H), such that, if $G = H$,

$$P[T_i(X, Y) \geq c(i) \text{ for some } i \geq k] \leq \epsilon, \tag{2.10}$$

while if $G \neq H$,

$$T_i(X, Y) \text{ converges (almost surely) to } b \neq 0.$$

If $c(i)$ can be chosen so that (2.10) obtains and $c(i)$ converges to zero, then the stopping rule N defined by

$$N = \begin{cases} \text{least } i \geq k \text{ for which } T_i(X, Y) \geq c(i) \\ \infty, \text{ if no such } i \text{ exists,} \end{cases}$$

satisfies

$$P[N < \infty] \leq \epsilon \quad \text{when } G = H \text{ and}$$

$$P[N < \infty] = 1 \quad \text{when } G \neq H.$$

This is just (2.7), the desired condition on N.

In the following examples, we will list $\mathfrak{F}$, $T_i(X, Y)$, k, and $c(i)$. Recall that for the classification problem, two stopping rules, N_{01}, based on $T_i(X_0, X_1)$ and N_{02}, based on $T_i(X_0, X_2)$, are used to define the classification rule $R(N_{01}, N_{02})$ given by (2.8).

Examples.

a) If $\mathfrak{F}$ is the class of univariate normal distributions with unit variances, then let

$$T_i(X, Y) = \bar{X} - \bar{Y},$$

$$k = 1, \text{ and}$$

$$c(i) = [(i+m)(a^2 + \log(i/(m+1))]^{\frac{1}{2}}/i,$$

where m is any positive constant and

$$a = (-2 \log 2\varepsilon)^{\frac{1}{2}}.$$

b) For the same $\mathfrak{F}$ and $T_i(X, Y)$ as above, we can let k be any positive integer and choose

$$c(i) = [(ka^2 + \log i)/ki]^{\frac{1}{2}}, \text{ and}$$

a to be the solution to

$$\varepsilon = 2(1 - \Phi(a) + a\varphi(a)),$$

where Φ and φ are, respectively, the cumulative distribution function and density function of a standard normal variable.

c) If $\mathfrak{F}$ is a collection of univariate normal distributions with a common unknown variance σ^2, let

$$T_i(X, Y) = |\bar{X} - \bar{Y}|/[i^{-1}\sum_{j=1}^{i} (X_j - Y_j - \bar{X} + \bar{Y})^2]^{\frac{1}{2}},$$

k be any positive integer, and

$$c(i) = [(ti)^{1/i} - 1]^{\frac{1}{2}}.$$

Here, $t = (1 + a^2/(m-1))^m/m$, where m is any positive constant and a is a solution to

$$\varepsilon = 2(1 - F_{k-1}(a) + a\, f_{k-1}(a)),$$

F_{k-1} and f_{k-1} being the cumulative distribution function and density function of Student's t distribution with $k - 1$ degrees of freedom.

d) If $\mathfrak{F}$ is any collection of distributions on some Euclidean space, let

$$T_i(X, Y) = D(G_i, H_i),$$

where D is the Kolmogorov distance and G_i and H_i are the sample distributions based on i observations on X and Y respectively. Also, let

$$c(i) = [(i+1)(\log 4 + 2 \log i)]^{\frac{1}{2}}/i$$

and choose k so that

$$\varepsilon \geq 2 \sum_{i=k}^{\infty} \exp[-i^2c^2(i)/(i+1)].$$

CHAPTER III

Asymptotic Relative Efficiency of Some Classification Rules

3.1 Introduction and Summary.

When we know $C(\Omega)$ to be classifiable or finitely classifiable, then we have to decide which of several classification rules we should use. In this chapter, we will consider two examples where $C(\Omega)$ is finitely classifiable and compare rules on the basis of the number of observations required to achieve a fixed maximum on the probabilities of error. In both cases, we cannot compute exactly the number of necessary observations, but, for two competing rules, we will compute the limit of the ratio of the numbers required by each rule.

The first example concerns univariate normal alternatives with a common unit variance. We compare a well known classification rule (Anderson [2]) with a two sample test for the equality of means. Letting the maximum of the probabilities of error approach 0, we find (Theorem 3.1) that the classification rule requires 9/8 as many observations as the two-sample test.

The second example concerns classification rules based on linear rank statistics. Considering one-sided location-shift alternatives, we compute the Pitman asymptotic efficiency of the non-parametric rules relative to their parametric analogues. An exposition of the concept of Pitman efficiency is given in Puri and Sen [23].

3.2 Two Finite Sample Size Rules: The Normal Case.

When X_0, X_1, and X_2 are univariate normal with means μ, ν, and η and a common known variance (taken to be 1) and if we assume Ω is such that

$$\Delta = \inf_{\Omega} |\nu - \eta| > 0,$$

then there exist classification rules based on a finite number of observations on X_0 and X_1 with arbitrarily small probabilities of error (see Section 1.8). Of course, in the same circumstances, there will also exist rules based on a finite number of observations on X_0, X_1, and X_2 with arbitrarily small probabilities of error. We consider two rules, one based only on X_0 and X_1 and one making use of X_0, X_1, and X_2, and compare the sample sizes necessary to achieve the same probabilities of error.

The first rule is a two sample test for the equality of the means of X_0 and X_1, μ and ν. The second is the appropriate univariate version of a common classification rule considered by Anderson [2] and others. Let $\bar{X}_{im}$ $i = 0,1, 2$ be the sample mean of m observations on X_i and let

$$W_m = \bar{X}_{0m} - \bar{X}_{1m} \quad \text{and}$$

$$T_m = (2\bar{X}_{0m} - \bar{X}_{1m} - \bar{X}_{2m})(\bar{X}_{2m} - \bar{X}_{1m}).$$

We define the two rules, denoted by $\underline{W(c, m)}$ and $\underline{T(m)}$ as follows:

$W(c, m)$:
$$\psi_W = \begin{cases} 1 & \text{if } |W_m| > c, \text{ and} \\ 0 & \text{if } |W_m| \leq c, \end{cases}$$

$T(m)$:
$$\psi_T = \begin{cases} 1 & \text{if } T_m > 0, \text{ and} \\ 0 & \text{if } T_m \leq 0. \end{cases}$$

ψ is the conditional probability of deciding $\mu = \eta$ (X_0 and X_2 have the same distribution).

To compare W and T, we first assume a fixed maximum probability of error α and then consider the smallest n and m for which $W(c,n)$ and $T(m)$ have maximum probabilities of error less than α. Denoting these sample sizes as $n(\alpha)$ and $m(\alpha)$, we will examine the ratio $n(\alpha)/m(\alpha)$ as α approaches 0. We will show that

$$\lim_{\alpha \to 0} n(\alpha)/m(\alpha) = 4/3,$$

but first, we must study the limiting behavior of several different functions of α. This is done in Lemmas 3.1-3.5.

Let Φ and φ, respectively, be the cumulative distribution function and density function of the standard normal random variable. The first lemma is a classical result (see, for example, Feller [13]).

Lemma 3.1.

$$\lim_{y \to \infty} \varphi(y)/y\Phi(-y) = 1.$$

Lemma 3.2.

For any constant b,

$$\lim_{y \to \infty} \Phi(-by)/\Phi(-y) = \begin{cases} 0 & \text{if } b > 1, \\ 1 & \text{if } b = 1, \text{ and} \\ \infty & \text{if } b < 1. \end{cases}$$

Proof:

Consider the identity

$$\frac{\Phi(-by)}{\Phi(-y)} = \left(\frac{by\ \Phi(-by)}{\varphi(by)}\right)\left(\frac{\varphi(by)}{b\varphi(y)}\right)\left(\frac{\varphi(y)}{y\Phi(-y)}\right). \tag{3.1}$$

For $b > 0$, $y \to \infty$ implies $by \to \infty$. By Lemma 3.1, the first and third terms on the right hand side of (3.1) converge to 1. Thus the limiting behavior of the left hand side of (3.1) depends on

$$\varphi(by)/b\varphi(y) = b^{-1}\exp\{-(b^2-1)y^2/2\},$$

which yields the result for $b > 0$. For $b \leq 0$, $\Phi(-by) \geq 1/2$ for $y \geq 0$, while $\Phi(-y)$ converges to 0 as y increases. □

Lemma 3.3.

For any function $b(y)$, the existence of $0 < c < \infty$ such that

$$\lim_{y \to \infty} \Phi(-b(y)y)/\Phi(-y) = c$$

implies

$$\lim_{y \to \infty} b(y) = 1.$$

Proof:

Assume

$$\lim_{y \to \infty} \sup\,(b(y) - 1) > 0.$$

Then there must exist a sequence y_n such that $y_n \to \infty$ and $b(y_n) \geq 1 + \epsilon$ for some $\epsilon > 0$. For that sequence,

$$\Phi(-b(y_n)y_n)/\Phi(-y_n) \leq \Phi(-(1+\epsilon)y_n)/\Phi(-y_n),$$

where the right hand side converges to 0 by Lemma 3.2. Thus the left hand side must also converge to 0, contradicting the assumption that $c > 0$. Similarly, assuming

$$\lim_{y \to \infty} \inf(b(y)-1) < 0$$

contradicts $c < \infty$. □

We will now compare the rates at which $n(\alpha)$ and $m(\alpha)$ approach ∞ as α approaches 0.

a) Consider first $W(c, n)$. If we define $\delta = \mu - \nu$ and $\theta(n) = (\frac{n}{2})^{\frac{1}{2}}$, then $W_n = X_{0n} - X_{1n}$ is normal with mean δ and variance $(\theta(n))^{-2}$. The two probabilities of error

of the rule $W(c, n)$ are thus the probability of falsely classifying X_0 with X_2 when $\mu = \nu(\delta=0)$,

$$(3.2) \qquad P_0[|W_n| > c] = 2\Phi(-c\theta(n)),$$

and the probability of falsely classifying X_0 with X_1 when $\mu = \eta$ and $|\delta| \geq \Delta$,

$$(3.3) \qquad P_\delta[|W_n| \leq c] = \Phi(-(\delta-c)\theta(n)) - \Phi(-(\delta+c)\theta(n)).$$

$n(\alpha)$ is thus the least integer such that, for some c, (3.2) is less than α and (3.3) is less than α for all $|\delta| \geq \Delta$. Since we are interested in asymptotic behavior, we will ignore the fact that n must be an integer.

Because (3.2) is decreasing in n, the solution(for fixed c) of

$$2\Phi(-c\theta(n)) = \alpha$$

will be the least n sufficient to insure that (3.2) is less than α. Writing $y(\alpha) = \Phi^{-1}(1-\alpha/2)$, the above equation is equivalent to

$$(3.4) \qquad c\theta(n) = y(\alpha).$$

Moreover, differentiation of (3.3) with respect to δ or an appeal to a stronger (multivariate) result of Anderson [1] shows that the value of 3.3 is monotonically decreasing as $|\delta|$ increases, so it is sufficient to find n and c such that

$$\Phi(-(\Delta-c)\theta(n)) - \Phi(-(\Delta+c)\theta(n)) = \alpha.$$

Substituting (3.4) and $a = \Delta/c$ into the above yields

$$(3.5) \qquad \Phi(-(a-1)y(\alpha)) - \Phi(-(a+1)y(\alpha)) = \alpha.$$

If $a(\alpha)$ is the solution of (3.5) then $c(\alpha) = \Delta/a(\alpha)$ and

$$n(\alpha) = 2[y(\alpha)a(\alpha)/\Delta]^2.$$

We will now consider the limiting behavior of $a(\alpha)$ and hence, implicitly, that of $n(\alpha)$. In what follows, we will write $f(\alpha) \sim g(\alpha)$ if

$$\lim_{\alpha \to 0} f(\alpha)/g(\alpha) = 1.$$

Lemma 3.4.

For $a(\alpha)$ and $y(\alpha)$ as defined above,

$$\alpha \sim \Phi(-(a(\alpha) - 1)y(\alpha)) \quad \text{and}$$

$$\lim_{\alpha \to 0} a(\alpha) = 2.$$

Proof:

Consider the value of

$$\Phi(-(a-1)y(\alpha)) - \Phi(-(a+1)y(\alpha)) \tag{3.6}$$

for a fixed value of a as $\alpha \to 0$ $(y(\alpha) \to \infty)$. Using Lemma 3.2, we find that for $a = 1$, (3.6) converges to $\Phi(0) = 1/2$ while for $a = 2$, (3.6) is always less than $\Phi(-y(\alpha)) = \alpha/2$. Since (3.6) is a continuous function of α, the solution $a(\alpha)$ of (3.5) (3.6 equated to α) must lie between 1 and 2, for sufficiently small values of α. Thus

$$\frac{\Phi(-(a(\alpha) + 1)y(\alpha))}{\Phi(-(a(\alpha) - 1)y(\alpha))} \leq \frac{\Phi(-2y(\alpha))}{\Phi(- y(\alpha))},$$

which, by Lemma 3.2, converges to 0.

Suppressing the dependencies on α, (3.5) can be written as

$$\frac{\Phi(-(a-1)y)}{\alpha}\left[1 - \frac{\Phi(-(a+1)y)}{\Phi(-(a-1)y}\right] = 1.$$

By the discussion above, the expression within brackets converges to 1, and thus,

$$\alpha \sim \Phi(-(a-1)y).$$

Since $\alpha = 2\Phi(-y)$, Lemma 3.3 implies that $(a-1)$ converges to 1. □

b) Consider now the rule $T(m)$. The relevant statistic,

$$T_m = (2\bar{X}_{0m} - \bar{X}_{1m} - \bar{X}_{2m})(\bar{X}_{2m} - \bar{X}_{1m}),$$

is the product of two independent normal random variables with means $2\mu - \nu - \eta$ and $\eta - \nu$ and variances $6/m$ and $2/m$, respectively. Writing $\delta = \eta - \nu$ and $s(m) = (m/6)^{\frac{1}{2}}$, note that when X_0 and X_1 have the same distribution, $\mu = \nu$ and $2\mu - \nu - \eta = -\delta$ and the probability of error is

$$\begin{aligned}(3.7)\qquad P_\delta[T_m > 0] &= P[2\bar{X}_{0m} - \bar{X}_{1m} - \bar{X}_{2m} > 0,\ \bar{X}_{2m} - \bar{X}_{1m} > 0] \\ &\quad + P[2\bar{X}_{0m} - \bar{X}_{1m} - \bar{X}_{2m} < 0,\ \bar{X}_{2m} - X_{1m} < 0] \\ &= [1 - \Phi(\delta s(m))][1 - \Phi(-3^{\frac{1}{2}}\delta s(m))] \\ &\quad + \Phi(\delta s(m))\Phi(-3^{\frac{1}{2}}\delta s(m)) \\ &= [1 - \Phi(\delta s(m))]\Phi(3^{\frac{1}{2}}\delta s(m)) + \Phi(\delta s(m))[1 - \Phi(3^{\frac{1}{2}}\delta s(m))] \\ &= \Phi(\delta s(m)) + \Phi(3^{\frac{1}{2}}\delta s(m)) - 2\Phi(\delta s(m))\Phi(3^{\frac{1}{2}}\delta s(m)).\end{aligned}$$

Similarly, when X_0 and X_1 have the same distribution, $2\mu - \nu - \eta = \delta$, and the probability of error is

$$(3.8)\qquad P_\delta[T_m \leq 0] = \Phi(-\delta s(m)) + \Phi(-3^{\frac{1}{2}}\delta s(m)) - 2\Phi(-\delta s(m))\Phi(3^{\frac{1}{2}}\delta s(m)).$$

Further manipulations show that the values of (3.7) and (3.8) are equal, so that the probabilities of error are equal and depend only on $|\delta|$ and m.

We will now show that the probabilities of error are decreasing as $|\delta|$ increases. Differentiation of (3.7) with respect to δ yields (suppressing m),

$$s\varphi(\delta s) + 3^{\frac{1}{2}}s\varphi(3^{\frac{1}{2}}\delta s) - 2[s\varphi(\delta s)\Phi(3^{\frac{1}{2}}\delta s) + 3^{\frac{1}{2}}s\varphi(3^{\frac{1}{2}}\delta s)\Phi(\delta s)]$$

$$= s\varphi(\delta s)[1 - 2\Phi(3^{\frac{1}{2}}\delta s)] + 3^{\frac{1}{2}}s\varphi(3^{\frac{1}{2}}\delta s)[1 - 2\Phi(\delta s)],$$

which is negative for all $\delta > 0$. Since the probabilities of error depend only on $|\delta|$, they are decreasing as $|\delta|$ increases. Thus, to find the least m for which both probabilities of error are less than α for all $|\delta| \geq \Delta$, it is sufficient to solve

(3.9) $$\alpha = \Phi(\Delta s(m)) + \Phi(3^{\frac{1}{2}}\Delta s(m)) - 2\Phi(\Delta s(m))\Phi(3^{\frac{1}{2}}\Delta s(m)) \quad \text{for } m(\alpha).$$

Lemma 3.5.

For $m(\alpha)$, the solution to (3.9), $\alpha \sim \Phi(-\Delta s(m(\alpha)))$.

Proof:

Using the equality of (3.7) and (3.8) and suppressing $m(\alpha)$, (3.9) yields

$$1 = \frac{\alpha}{\alpha} = \frac{\Phi(-\Delta s)}{\alpha}\left[1 + \frac{\Phi(-3^{\frac{1}{2}}\Delta s)}{\Phi(-\Delta s)} - 2\Phi(-3^{\frac{1}{2}}\Delta s)\right].$$

Since $\alpha \to 0$ implies $m(\alpha) \to \infty$ and $\Delta s \to \infty$, Lemma 3.2 implies that the expression within brackets converges to 1. □

We are now in a position to calculate the limit of $m(\alpha)/m(\alpha)$.

Lemma 3.6.

For $m(\alpha)$ and $n(\alpha)$ as considered above,

$$\lim_{\alpha \to 0} n(\alpha)/m(\alpha) = 4/3.$$

Proof:

Combining Lemma 3.4 and Lemma 3.5, we find

$$\Phi(-(a(\alpha) - 1)y(\alpha)) \sim \Phi(-\Delta s(m(\alpha))).$$

Equivalently (using Lemma 3.1 and suppressing α and m),

$$\lim_{\alpha \to 0}\left(\frac{(a-1)y}{\varphi((a-1)y)}\right)\left(\frac{\varphi(\Delta s)}{\Delta s}\right) = 1.$$

Taking logarithms, we obtain

$$\lim_{\alpha \to 0} [\log(a-1)y - \log \Delta s - \Delta^2 s^2/2 + (a-1)^2 y^2/2] = 0. \tag{3.10}$$

Substituting $(n/2)^{\frac{1}{2}} \Delta/a$ for y (see (3.4)) and $(m/6)^{\frac{1}{2}}$ for s, we find

$$\log(a-1)y - \log \Delta s = \tfrac{1}{2} \log(a-1)^2 y^2 - \tfrac{1}{2} \log \Delta^2 s^2$$

$$= \tfrac{1}{2} \log(a-1)^2 (n\Delta^2/2a^2) - \tfrac{1}{2} \log(m\Delta^2/6)$$

$$= \tfrac{1}{2} \log(3(a-1)^2 n/(ma^2)),$$

and

$$-\Delta^2 s^2/2 + (a-1)^2 y^2/2 = \Delta^2[-m/12 + (a-1)^2 n/4a^2]$$

$$= \frac{\Delta^2 m}{12} [-1 + 3(a-1)^2 n/(ma^2)].$$

Using these equalities and defining

$$g(\alpha) = \log(3(a-1)^2 n/(ma^2)),$$

(3.10) can be rewritten as

$$\lim_{\alpha \to 0} [(\tfrac{1}{2}) g(\alpha) + m(\alpha)\Delta^2 (e^{g(\alpha)} - 1)/12] = 0. \tag{3.11}$$

By Lemma 3.4, a converges to 2, so $3(a-1)^2/a^2$ converges to $3/4$ and thus, to show n/m converges to $4/3$, it is sufficient to show $g(\alpha)$ converges to 0. Now, for any sequence α_k converging to 0 ($m(\alpha) \to \infty$), assuming that $g(\alpha)$ converges to some $c \neq 0$ contradicts (3.11), since then,

$$m(\alpha)\Delta^2 (e^{g(\alpha)} - 1)$$

goes to ∞ or $-\infty$ depending on whether $c > 0$ or $c < 0$. Thus, $\lim_{\alpha \to 0} g(\alpha) = 0$ and $\lim_{\alpha \to 0} n(\alpha)/m(\alpha) = 4/3$. □

To properly interpret Lemma 3.6, recall that $W(c, n)$ is based on $\bar{X}_{0n}$ and $\bar{X}_{1n}$ while $T(m)$ is based on $\bar{X}_{0m}$, $\bar{X}_{1m}$ and $\bar{X}_{2m}$. Therefore, $W(c, n)$ requires a total of $2n$ observations and $T(m)$ requires a total of $3m$ observations. Let $O_W(\alpha) = 2n(\alpha)$ and $O_T(\alpha) = 3m(\alpha)$ where $O_W(\alpha)$ and $O_T(\alpha)$ are respectively, the minimum number of observations required for W and T to have maximum error probabilities less than α. If we define the asymptotic relative efficiency of $T(m)$ relative to $W(n, c)$, $e_{T,W}$, as the limit of the ratio $O_W(\alpha)/O_T(\alpha)$, then Lemma 3.6 immediately implies the following theorem.

Theorem 3.1.

For the tests $T(m)$ and $W(c, n)$, where c is chosen minimize n, the asymptotic relative efficiency of $T(m)$ relative to $W(c, n)$ is

$$e_{T,W} = 8/9.$$

It should be noted that $e_{T,W}$ is computed assuming that Δ is known and the optimal c is used in $W(c, n)$. Consider now the situation when Δ is equal to 1 but the experimenter thinks it is actually $\tilde{\Delta}$. Wishing to have a maximum error of α, an experimenter using W would choose $c = \tilde{\Delta}/a(\alpha) = \tilde{\Delta}c(\alpha)$ and $n = 2\left(\frac{y(\alpha)a(\alpha)}{\tilde{\Delta}}\right)^2 = n(\alpha)/\tilde{\Delta}^2$ ($c(\alpha)$ and $n(\alpha)$ being the correct choices for $\Delta = 1$). If $\tilde{\Delta} < 1$, the maximum probability of error of $W(c, n)$ is still

$$2\Phi(-c\theta(n)) = 2\Phi(-c(\alpha)\theta(n(\alpha))) = 2\Phi(-y(\alpha)) = \alpha.$$

On the other hand, an experimenter using T would choose $m = m(\alpha)/\tilde{\Delta}^2$ and the actual maximum probability of error of $T(m)$ would be

$$\Phi(-\tilde{\Delta}s) + \Phi(-3^{\frac{1}{2}}\tilde{\Delta}s) - 2\Phi(-\tilde{\Delta}s)\Phi(-3^{\frac{1}{2}}\tilde{\Delta}s),$$

where $s = s(m(\alpha))$. As in the previous analysis, it is the leading term in the above expression which dominates as α goes to 0, so the ratio of the two errors ($W(c, n)$ to $T(m)$) is asymptotically equivalent to

$$\frac{\alpha}{\Phi(-\tilde{\Delta}s)} \sim \frac{\Phi(-s)}{\Phi(-\tilde{\Delta}s)}$$

which, since $\tilde{\Delta} < 1$, increases to ∞ as α goes to 0.

Similar conclusions can be drawn when $\tilde{\Delta} > 1$. In fact, if $\tilde{\Delta} > 2$, the maximum error of $W(c, n)$ is asymptotically equivalent to

$$\Phi(-(a(\alpha)\tilde{\Delta}^{-1} - 1)\, y(\alpha)),$$

which is greater than $1/2$ since $a(\alpha) < 2$ implies $(a\tilde{\Delta}^{-1} - 1) < 0$. On the other hand, no matter what $\tilde{\Delta}$ is, the maximum probability of error of the test $T(m)$ converges to 0 as α goes to 0.

An analysis similar to that of this section could be done if we assumed that the sampling scheme did not take equal numbers of observations from each population, but instead, picked out the respective samples in some fixed ratio.

3.3 Some Nonparametric Classification Rules.

As illustrated in Chapter II, there are distribution-free sequential rules based on the Kolmogorov-distance D which have arbitrarily small error under the assumption that X_1 and X_2 do not have the same distribution. In this section, other distribution free rules are studied in the fixed sample size case, in an attempt to find rules which make good use of a given sample. In particular, linear rank statistics are formed from the combined rank order of all three samples and used to define classification rules, which are then compared on the basis of asymptotic relative efficiency.

If X_0, X_1, and X_2 are univariate random variables with continuous distributions F_0, F_1, and F_2, consider taking n observations on each population and forming three linear rank statistics as follows:

$$T_{Nj} = n^{-1} \sum_{i=1}^{N} E_{Ni} Z_{ji} \quad , \quad j = 0, 1, 2,$$

where E_{Ni}, $i = 1,\ldots, N$, is a sequence of constants (called scores) and Z_{ji} is 1 if the i^{th} smallest observation in the combined ordering of all $N = 3n$ observations is from population π_j, and 0 otherwise. If we define a function J_N on $(0, N/(N+1))$ by

$$J_N(x) = E_{Ni} \quad \text{if} \quad (i-1) < (N+1)x \leq i, \; i = 1,\ldots, N,$$

and write $F_{nj}(x)$ for the sample cumulative distribution function based on n observations on X_j, then T_{Nj} has an equivalent representation,

$$T_{Nj} = \int_{-\infty}^{\infty} J_N[NK_N(x)/(N+1)]dF_{nj}(x),$$

where $K_N(x) = (F_{n0}(x) + F_{n1}(x) + F_{n2}(x))/3$. These statistics can be used to form a number of classification statistics; for example, an analogue of the statistic T_N of section 3.2 would be $(2T_{N0} - T_{N1} - T_{N2})(T_{N2} - T_{N1})$.

Consider the sequence of problems

$$\Omega_N(\theta) = \{F_1 \times F_2 : F_2(x) = F_1(x + N^{-\frac{1}{2}}\theta) \text{ for some continuous } F_1\},$$

where θ is a fixed positive constant. Thus $F_2(x)$ is a translation of $F_1(x)$ to the left by $N^{-\frac{1}{2}}\theta$. Let $H_1(N)$ and $H_2(N)$ stand for $F_0 \times F_1 \times F_2 \in \mathcal{G}(\Omega_N)$ and $F_0 \times F_1 \times F_2 \in \mathcal{H}(\Omega_N)$, respectively. Since Ω_N consists of one-sided location shift alternatives, the appropriate rule V would be

decide $F_0 = F_1$ if $V_N > 0$, and

decide $F_0 = F_2$ if $V_N \leq 0$,

where

$$V_N = 2T_{N0} - T_{N1} - T_{N2} .$$

To study the limiting power of this sequence of rules for this sequence of problems, we need the asymptotic distributions of the T_{Nj}. They are found by an application of Theorem 5.6.1 of Puri and Sen.

Lemma 3.7.

If $J(x) = \lim J_N(x)$ exists and is not constant, $0 < x < 1$, then, under the usual Chernoff-Savage regularity conditions on J_N, J, and F_j, the vector with components $N^{\frac{1}{2}}(T_{Nj} - \mu_{Nj})$, $j = 0,1,2$, has a limiting normal distribution with mean vector $(0, 0, 0)$ and covariance matrix $\Sigma = (\sigma_{ij})$ under either of the sequences of hypotheses $H_i(N)$, $i = 1, 2$. Here,

$$\mu_{Nj} = \int_{-\infty}^{\infty} J(K(x))dF_j,$$

$$\sigma_{ii} = 2A^2, \ i = 0,1,2,$$

$$\sigma_{ij} = -A^2, \ 0 \leq i \neq j \leq 2,$$

where

$$K(x) = (F_0(x) + F_1(x) + F_2(x))/3, \text{ and}$$
$$A^2 = \int_0^1 J^2(u)du - (\int_0^1 J(u)du)^2.$$

Note that $K(x)$, and hence μ_{Nj}, actually depends on the hypothesis $H_i(N)$, since, under $H_i(N)$, $F_1 \times F_2$ is in $\Omega_N(\theta)$ and $F_0 = F_i$.

It follows from the above that the asymptotic distribution of $N^{\frac{1}{2}}(V_N - \mu_N)$ is normal with mean 0 and variance $18A^2$, where

$\mu_N = 2\mu_{N0} - \mu_{N1} - \mu_{N2}$. We wish to find the limiting power, or, equivalently, the limit of $P[V_N > 0]$ under $H_i(N)$. Since $V_N > 0$ is equivalent to

$$N^{\frac{1}{2}}(V_N - \mu_N)/\sqrt{18}\,A > -N^{\frac{1}{2}}\mu_N/\sqrt{18}\,A,$$

we need only find the limit of $N^{\frac{1}{2}}\mu_N$ under $H_i(N)$ to be able to apply the limiting normal theory.

Lemma 3.8.

Under certain regularity conditions,

$$\lim N^{\frac{1}{2}}(\mu_{N1} - \mu_{N2}) = \theta \int_{-\infty}^{\infty} \frac{d}{dx} J(F_1(x))dF_1(x) \tag{3.12}$$

under either $H_1(N)$ or $H_2(N)$.

Proof:

To establish (3.12), consider

$$K_t(x) = \frac{2}{3}F(x) + \frac{1}{3}F(x+t) \quad \text{and}$$

$$f(t) = \int_{-\infty}^{\infty} J(K_t(x))dF(x).$$

If $f''(t)$ exists and is finite at $t = 0$, then the Taylor expansion about 0 is

$$f(t) = f(0) + tf'(0) + O(t^2).$$

Noting that $K_0(x) = F(x)$, the above can be rewritten as

$$f(t) = \int_{-\infty}^{\infty} J(F(x))dF(x) + t\,\frac{d}{dt}\left[\int_{-\infty}^{\infty} J(K_t(x))dx\right]_{t=0} + O(t^2). \tag{3.13}$$

If the differentiation in the above expression can be taken under the integral, the second term on the right hand side of (3.13) can be written as

$$t\int_{-\infty}^{\infty} \frac{d}{dt}\left[J(K_t(x))\right]_{t=0} dx. \tag{3.14}$$

The integrand in (3.14) can be evaluated by the chain rule as

$$\left[\frac{d}{du} [J(u)]_{u=K_t(x)} \quad \frac{d}{dt} K_t(x) \right]_{t=0} =$$

$$\left[\frac{d}{du} [J(u)]_{u=K_t(x)} \quad \frac{1}{3} \frac{d}{dy} [F(y)]_{y=x+t} \right]_{t=0} =$$

$$\frac{d}{du} [J(u)]_{u=F(x)} \quad \frac{1}{3} \frac{d}{dy} [F(y)]_{y=x} =$$

$$\frac{1}{3} \frac{d}{dx} J(F(x)).$$

Thus (3.13) becomes

$$f(t) = \int_{-\infty}^{\infty} J(F(x))dF(x) + t \int_{-\infty}^{\infty} \frac{1}{3} \frac{d}{dx} J(F(x))dF(x) + O(t^2).$$

Under $H_1(N)$, $F_2(x) = F_1(x+N^{-\frac{1}{2}}\theta) = F_0(x+N^{-\frac{1}{2}}\theta)$, so $\mu_{N1} = f(N^{-\frac{1}{2}}\theta)$, or

$$\mu_{N1} = \int_{-\infty}^{\infty} J(F_1(x))dF_1(x) + \frac{N^{-\frac{1}{2}}\theta}{3} \int_{-\infty}^{\infty} \frac{d}{dx} J(F_1(x))dF_1(x) + O(N^{-1}).$$

A similar analysis shows that

$$\mu_{N2} = \int_{-\infty}^{\infty} J(F_1(x))dF_1(x) - \frac{2N^{-\frac{1}{2}}\theta}{3} \int_{-\infty}^{\infty} \frac{d}{dx} J(F_1(x))dF_1(x) + O(N^{-1})$$

and thus, under $H_1(N)$,

$$\mu_{N1} - \mu_{N2} = N^{-\frac{1}{2}}\theta \int_{-\infty}^{\infty} \frac{d}{dx} J(F_1(x))dF_1(x) + O(N^{-1}).$$

(3.12) now follows from

$$N^{\frac{1}{2}}(\mu_{N1} - \mu_{N2}) = \theta \int_{-\infty}^{\infty} \frac{d}{dx} J(F_1(x))dF_1(x) + O(N^{-\frac{1}{2}}).$$

The analysis for $H_2(N)$ is similar. □

Note that under $H_1(N)$, $\mu_{N0} = \mu_{N1}$ and $\mu_N = \mu_{N1} - \mu_{N2}$, while under $H_2(N)$, $\mu_{N0} = \mu_{N2}$ and $\mu_N = -(\mu_{N1} - \mu_{N2})$. Writing

$$c = \int_{-\infty}^{\infty} \frac{d}{dx} J(F_1(x))dF_1(x),$$

we see that $N^{\frac{1}{2}}\mu_N$ converges to θc or $-\theta c$ depending on whether $H_1(N)$ or $H_2(N)$ obtains. Thus, under $H_1(N)$, $P[V_N \leq 0]$ converges to $\Phi(-\theta c/\sqrt{18}\,A)$, while, under $H_2(N)$, $P[V_N > 0]$ converges to $\Phi(-\theta c/\sqrt{18}\,A)$, the same value. Note that we have implicitly assumed $c > 0$; if it were not, the appropriate test would just reverse the actions of V.

An alternate rule would use the sample means in an analogous fashion. Let $\Omega_M(\eta)$ be defined in the same manner as $\Omega_N(\theta)$. We consider the rule U which

$$\text{decides } F_0 = F_1, \text{ if } U_M > 0, \text{ and}$$

$$\text{decides } F_0 = F_2, \text{ if } U_M \leq 0,$$

where

$$U_M = 2\bar{X}_{0M} - \bar{X}_{1M} - \bar{X}_{2M},$$

$\bar{X}_{jM}$ being the sample mean based on M observations on X_j. Elementary calculations and a standard central limit theorem yield

$$M^{\frac{1}{2}}(U_M - M^{-\frac{1}{2}}\eta)/\sqrt{6}\,\sigma \xrightarrow{\mathcal{L}} \Phi \text{ under } H_1(N), \text{ and}$$

$$M^{\frac{1}{2}}(U_M + M^{-\frac{1}{2}}\eta)/\sqrt{6}\,\sigma \xrightarrow{\mathcal{L}} \Phi \text{ under } H_2(N),$$

where

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 dF_1(x) - \left(\int_{-\infty}^{\infty} x dF_1(x)\right)^2.$$

Using the limiting distributions, we find the limiting probabilities of error to be $\Phi(-\eta/\sqrt{6}\,\sigma)$.

To compare the rules U and V, suppose sequences of N and M are chosen such that $N^{-\frac{1}{2}}\theta \sim M^{-\frac{1}{2}}\eta$. In that case, $\Omega_N(\theta)$ and $\Omega_M(\eta)$ are asymptotically identical sequences of problems. If the limiting probabilities of error are also to be equal, then we need

$$-\theta c/\sqrt{18}\ A = -\eta/\sqrt{6}\ \sigma.$$

Equivalently,

$$\frac{M}{N} \sim \frac{\eta^2}{\theta^2} = \sigma^2 c^2/3A^2.$$

Since the rule based on V_N requires N observations and the rule based on U_M requires $3M$ observations, an appropriate definition of the Pitman asymptotic relative efficiency of V relative to U, $e_{V,U}$, is the limit of $3M/N$. We have thus proved the following theorem.

Theorem 3.2.

Under regularity conditions,

$$e_{v,u} = \sigma^2[\int_{-\infty}^{\infty} \frac{d}{dx} J(F_1(x))dF_1(x)]^2/A^2.$$

The above expression for $e_{v,u}$ is exactly the expression Puri and Sen obtain for the asymptotic efficiency of the two sample rank test of $F_0 = F_1$ relative to the two sample t-test for equality of means. They describe several special cases.

a) If $J(u) = u$ (Wilcoxon rank sums),

$$e_{v,u} \geq .864 \quad \text{for all} \quad F_1.$$

In particular, if F_1 is standard normal, $e_{v,u} = 3/\pi$.

b) If $J(u) = \Phi^{-1}(u)$ (normal scores),

$$e_{v,u} = \sigma^2[\int_{-\infty}^{\infty} \frac{f_1^2(x)}{\varphi[\Phi^{-1}(F_1(x))]} dx\ ]^2 \geq 1 \quad \text{for all} \quad F_1.$$

CHAPTER IV

A Complete Class of Invariant Classification Rules

4.1 Introduction and Summary.

In this chapter, we consider the problem of classification into one of two univariate normal distributions with unknown means and unit variances. Let X_0, X_1, and X_2 be independent normal random variables with means μ_0, μ_1, and μ_2, respectively, and a common unit variance. The problem is to test $\mu_0 = \mu_1$ versus $\mu_0 = \mu_2$ based on n observations on X_0 and m observations on each of X_1 and X_2. We restrict ourselves to the class of rules which are invariant under translation and change of sign. The distribution of a maximal invariant in the space of sufficient statistics is first obtained. Then, assuming the loss function to be zero-one, the structure of invariant Bayes rules with respect to prior distributions having finite support is studied. Our main result is the characterization of an essentially complete class of invariant rules. The proof is based on the decision theory of Abraham Wald [28] and some techniques suggested by Matthes and Truax [21] and Eaton [11].

4.2 Invariance.

Let Y_0, Y_1, and Y_2 be the respective sample means of the observations on X_0, X_1 and X_2. By sufficiency, we can restrict our attention to rules based on Y_0, Y_1, and Y_2. Consider the following groups of transformations on the space of Y's: the group of translations,

$$G_1 = \{g: g(y_0, y_1, y_2) = (y_0 + g, y_1 + g, y_2 + g),\ -\infty < g < \infty\}$$

and the group of sign changes,

$$G_2 = \{h: h(y_0, y_1, y_2) = (hy_0, hy_1, hy_2),\ h = \pm 1\}.$$

We are interested in G, the composition of G_2 with G_1. It is clear that the classification problem under consideration remains invariant under any transformation in G. Recall (Lehmann [20], Chapter 6) that any invariant rule is a function of a maximal invariant.

Lemma 4.1.

A maximal invariant in the space of Y_0, Y_1, and Y_2 under G is almost everywhere equivalent to

$$(UV, |V|),$$

where

$$U = 2Y_0 - Y_1 - Y_2, \text{ and}$$
$$V = Y_2 - Y_1.$$

Proof:

It is well known that (U, V) is a maximal invariant under G_1, the group of translations. Since U and V are non-zero with probability one, we restrict our attention to this case. Note that the transformations induced by G_2, the group of sign changes, on (U, V) is

$$h(U, V) = (hU, hV), \; h = \pm 1.$$

Clearly, $(UV, |V|)$ is invariant under G_2. Suppose now, for some U^* and V^*,

$$UV = U^*V^* \quad \text{and} \quad |V| = |V^*|.$$

Then $V^* = hV$, where $h = \pm 1$. This implies $UV = U^*(hV) = (hU^*)V$, which, in turn, implies $U = hU^*$. □

We will consider an equivalent maximal invariant, (R, S), given by

$$R = cUV/|V| \quad \text{and} \quad S = d|V|,$$

for some non-zero constants c and d.

Lemma 4.2.

For the choices $c = [(4/n) + (2/m)]^{-\frac{1}{2}}$ and $d = (2/m)^{-\frac{1}{2}}$, the random vector (R, S) has a joint density function given by $f_1(r, s;\gamma)$ when $\mu_0 = \mu_1$ and $f_2(r, s;\gamma)$ when $\mu_0 = \mu_2$, where

$$f_1(r,s;\gamma) = \begin{cases} \eta(r, s)h(\gamma) \cosh[(cr-ds)\gamma], & \text{if } s > 0, \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_2(r,s;\gamma) = \begin{cases} \eta(r, s)h(\gamma) \cosh[(cr+ds)\gamma], & \text{if } s > 0, \\ 0 & \text{otherwise .} \end{cases}$$

Here,

$$\gamma = \mu_2 - \mu_1,$$
$$\eta(r,s) = \pi^{-1} \exp[-(r^2 + s^2)/2], \text{ and}$$
$$h(\gamma) = \exp[-\gamma^2(c^2 + d^2)/2].$$

Proof:

Since Y_0 is normal with mean μ_0 and variance $1/n$, while Y_1 and Y_2 are normal with means μ_1 and μ_2 and variances $1/m$, U and V are independent normal random variables with means $2\mu_0 - \mu_1 - \mu_2$ and $\mu_2 - \mu_1 = \gamma$ and variances $(4/n) + (2/m)$ and $(2/m)$, respectively. Consider first the case when $\mu_0 = \mu_1$. Then, $2\mu_0 - \mu_1 - \mu_2 = -(\mu_2 - \mu_1) = -\gamma$, and with c and d chosen as above, cU and dV are independent normal random variables with means $-c\gamma$ and $d\gamma$ and unit variances.

The joint distribution function of (R, S) is as follows:

$$F(r,s;\gamma) = P_\gamma[R \leq r, S \leq s] = 0 \quad \text{if} \quad s \leq 0,$$

while if $s > 0$,

$$F(r,s;\gamma) = P_\gamma[R \le r,\ S \le s]$$

$$= P_\gamma[cU \le r,\ 0 < dV \le s] + P_\gamma[-cU \le r,\ 0 < -dV \le s],$$

since for $dV > 0$, $S = dV$ and $R = cUV/V = cU$, while for $dV < 0$, $S = -dV$ and $R = cUV/(-V) = -cU$. By independence, for $s > 0$,

$$F(r,s;\gamma) = P_\gamma[cU \le r]P_\gamma[0 < dV \le s] + P_\gamma[-cU \le r]P_\gamma[0 < -dV \le s].$$

Thus, since $\mu_0 = \mu_1$ and $s > 0$,

$$F(r,s;\gamma) = \Phi(r+c\gamma)[\Phi(s-d\gamma) - \Phi(-d\gamma)] + \Phi(r-c\gamma)[\Phi(s+d\gamma) - \Phi(d\gamma)],$$

where Φ is the standard normal cumulative distribution function.

The joint density of (R, S) when $\mu_0 = \mu_1$ is found by differentiation.

$$f_1(r,s;\gamma) = \begin{cases} \varphi(r+c\gamma)\varphi(s-d\gamma) + \varphi(r-c\gamma)\varphi(s+d\gamma), & \text{if } s > 0, \\ 0 & \text{, otherwise.} \end{cases}$$

Substituting $\varphi(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)$ and $\cosh x$ for $(e^x + e^{-x})/2$, we get

$$f_1(r,s;\gamma) = \begin{cases} \eta(r,s)h(\gamma)\ \cosh[(cr-ds)\gamma], & \text{if } s > 0, \text{ and} \\ 0 & \text{, otherwise,} \end{cases}$$

where $\eta(r,s)$ and $h(\gamma)$ are defined as before.

Similarly, when $\mu_0 = \mu_1$, cU and dV are independent normal random variables with means $c\gamma$ and $d\gamma$ and unit variances, and the above analysis yields $f_2(f,s;\gamma)$ when c is substituted for $-c$. □

It should be noted that the joint distributions of (R, S), f_1 and f_2, depend on μ_1 and μ_2 only through $|\mu_2 - \mu_1| = |\gamma|$. For that reason, γ will henceforth denote $|\mu_2 - \mu_1|$.

4.3 Invariant Bayes Rules.

Let $\psi(r, s)$ denote an invariant rule, $\psi(r, s)$ being the conditional probability of deciding $\mu_0 = \mu_2$, given the value (r, s) of the maximal invariant (R, S). The risk function of ψ depends on μ_0, μ_1, and μ_2 only through $\gamma = |\mu_2 - \mu_1|$ and whether $\mu_0 = \mu_1$ or $\mu_0 = \mu_2$. Let ξ be a prior distribution on μ_0, μ_1, and μ_2. ξ induces a distribution ξ_1 on γ in the plane $\mu_0 = \mu_1$ and a distribution ξ_2 on γ in the plane $\mu_0 = \mu_2$. The Bayes risk of ψ with respect to ξ is thus given by

$$r(\psi,\xi) = \int E_\gamma[\psi]d\xi_1 + \int E_\gamma[1-\psi]d\xi_2.$$

In this section, we shall only consider priors ξ with finite support. If ξ_1 assigns probability p_i to $\gamma = a_i$ $i = 1,\ldots, k$ and ξ_2 assigns probability q_i to $\gamma = b_i$ $i = 1,\ldots, \ell$, then

$$r(\psi, \xi) = \sum_{i=1}^{k} p_i \int_{s>0}\int \psi(r, s)f_1(r,s; a_i)drds + \sum_{i=1}^{\ell} q_i$$

$$- \sum_{i=1}^{\ell} q_i \int_{s>0}\int \psi(r, s)f_2(r,s; b_i)drds$$

$$= \Sigma q_i + \int_{s>0}\int \psi(r, s)[\Sigma p_i f_1(r, s; a_i) - \Sigma q_i f_2(r, s; b_i)]drds.$$

Using our expression for f_1 and f_2 (Lemma 4.2) and the following definitions,

$$(4.1) \qquad g_1(x) = \sum_{1}^{k} p_i h(a_i)\cosh(a_i x) \quad \text{and}$$

$$g_2(x) = \sum_{1}^{\ell} q_i h(b_i)\cosh(b_i x),$$

we can express $r(\psi, \xi)$ as

$$(4.2) \qquad \Sigma q_i + \int\int_{s>0} \psi(r, s)\eta(r, s)[g_1(cr-ds) - g_2(cr+ds)]drds,$$

where η and h are as in Lemma 4.2.

We say ψ is a Bayes invariant rule with respect to a prior ξ with finite support if ψ is invariant and ψ minimizes (4.2) among the class of all invariant rules. It is clear that any Bayes invariant rule with respect to ξ must satisfy (almost everywhere), for $s > 0$,

$$\psi(r, s) = \begin{cases} 1 & \text{if } g_1(cr-ds) < g_2(cr+ds), \text{ and} \\ 0 & \text{if } g_1(cr-ds) > g_2(cr+ds). \end{cases}$$

Note that any Bayes invariant rule with respect to a prior with finite support is actually an invariant Bayes rule, since the prior ξ can be chosen by assigning probability 1 to the set of μ vectors such that $\mu_0 + \mu_1 + \mu_2 = 0$, and then assigning the following probabilities:

$$\begin{aligned} &p_i/2 \text{ to } \mu_0 = \mu_1,\ \mu_2 - \mu_1 = a_i, \\ &p_i/2 \text{ to } \mu_0 = \mu_1,\ \mu_2 - \mu_1 = -a_i, \\ &q_i/2 \text{ to } \mu_0 = \mu_2,\ \mu_2 - \mu_1 = b_i, \text{ and} \\ &q_i/2 \text{ to } \mu_0 = \mu_2,\ \mu_2 - \mu_1 = -b_i. \end{aligned}$$

With this assignment of probability, the Bayes risk of any rule is equal to an expression equivalent to (4.2).

To simplify notation, we make the substitutions, $x = cr$ and $y = ds$. Then any Bayes invariant rule will satisfy

$$(4.3) \qquad \psi(x, y) = \begin{cases} 1 & \text{if } g_1(x-y) < g_2(x+y), \text{ and} \\ 0 & \text{if } g_1(x-y) > g_2(x+y). \end{cases}$$

Since $\cosh(x) = \cosh(-x)$, both $g_1(x)$ and $g_2(x)$ will be symmetric about 0. Thus, for any ψ for which (4.3) obtains, we have,

(4.4) $\psi(x, y) = \psi(y, x) = \psi(-y, -x) = \psi(-x, -y)$.

We will say that ψ satisfies the symmetry property if (4.4) obtains (a.e.). Although technically, ψ need only be defined for $y = ds > 0$, we can extend the definition to all y by requiring ψ to satisfy the symmetry property.

Let

$$Q = \{(x, y): 0 \leq |y| \leq x\}.$$

Note that Q is a cone and just the positive quadrant rotated by 45°. Since, for any (x, y), one of the points, (x, y), (y, x), $(-y, -x)$, and $(-x, -y)$, must lie in Q, any ψ satisfying the symmetry property and hence, any Bayes invariant rule, is uniquely (a.e.) defined by its values in Q.

Given a prior ξ with finite support, and g_1 and g_2 as defined in (4.1), let

$$g(x, y) = g_2(x+y) - g_1(x-y).$$

Note that $g(x, y)$ is continuous since the hyperbolic cosine and hence, each g_i, is continuous. For $x \geq 0$, define

(4.5) $$f(x) = \begin{cases} \inf\{y: (x, y) \in Q \text{ and } g(x, y) > 0\}, \text{ or} \\ x, \text{ if the above set is empty.} \end{cases}$$

Lemma 4.3.

If $f(x)$ is defined as above for a prior distribution ξ with finite support, then

(4.6) $$|f(r) - f(x)| \leq |r-x| \quad \text{for} \quad r, x \geq 0.$$

Proof:

Assume (t, v) and (x, y) are any two points in Q, that is, $0 \leq |v| \leq t$ and $0 \leq |y| \leq x$. These inequalities imply that $t-v$, $t+v$, $x-y$, and $x+y$ are all positive. Without loss of generality, take $t \geq x$. If $(v-y) > (t-x)$, then $x-y > t-v$ and $x+y < 2x-t+v \leq t+v$.

Since a_i and b_i are both positive and cosh x is strictly increasing in $|x|$, we get

$$\cosh a_i(x-y) > \cosh a_i(t-v) \text{ and}$$
$$\cosh b_i(x+y) < \cosh b_i(t+v).$$

But as g_1 and g_2 are weighted sums (with positive coefficients) of the above terms, this implies

$$g_1(x-y) > g_1(t-v) \text{ and } g_2(x+y) < g_2(t+v).$$

Thus, for $t \geq x$,

(4.7) $\quad v-y \geq t-x$ implies $g(x, y) < g(t, v)$.

In particular, if $t = x$, then

$$g(x, y) < g(x, v) \text{ if } v > y.$$

Assume now that $(x, y) = (x, f(x)) \in Q^o$, the interior of Q. From the definition of f and the above remarks, it follows that $g(x, f(x)) = 0$. Also, if (t, v), $t \geq x$, is any point in Q with $v - f(x) > t-x \geq 0$, then (4.7) implies $g(t, v) > g(x, f(x)) = 0$. A similar analysis shows that if $-(v - f(x)) > t-x$, then $g(t, v) < g(x, f(x)) = 0$. Thus, for any fixed $t \geq x$, $g(t, v) = 0$ will have a solution (equal to $f(t)$) between $f(x) + (t-x)$ and $f(x) - (t-x)$; equivalently,

(4.8) $\quad |f(t) - f(x)| \leq |t-x|$.

If x_0 is the infimum of all $x \geq 0$ such that $(x, f(x)) \in Q^o$, then the continuity of g implies $g(x_0, f(x_0)) = 0$ and either $x_0 = 0$, $x_0 = f(x_0)$ or $x_0 = -f(x_0)$. If $x_0 = 0$, then (4.6) follows from (4.8) since $(x, f(x)) \in Q^o$ for all $x > 0$. If $x_0 = f(x_0)$, then for all $0 \leq x \leq x_0$, $(x, y) \in Q^o$ implies $f(x_0) - y \geq x_0 - x$, which in turn implies $g(x, y) < g(x_0, f(x_0)) = 0$. Thus $0 \leq x \leq x_0$ implies $f(x) = x$. If $0 \leq x \leq x_0 \leq r$, then

$$|f(r) - f(x)| \leq |f(r) - f(x_0)| + |f(x_0) - f(x)|$$

$$\leq |f(r) - f(x_0)| + (x_0 - x),$$

which, since (4.8) applies to all $r \geq x_0$,

$$\leq (r - x_0) + (x_0 - x) = |r - x|.$$

If $0 \leq x \leq r \leq x_0$, then

$$|f(r) - f(x)| = |r - x|,$$

while if $0 \leq x_0 \leq x \leq r$, (4.8) applies and

$$|f(r) - f(x)| \leq |r - x|.$$

The case where $x_0 = -f(x_0)$ is handled similarly. □

Remark.

For all $x \geq x_0$, the graph $(x, f(x))$ is just the graph of the solutions to $g(x, y) = 0$, i.e., the boundary between the acceptance and rejection regions of the Bayes invariant rule ψ with respect to ξ. For $x \leq x_0$, the boundary passes out of Q, so $f(x)$ just connects the point $(x_0, f(x_0))$ to the origin by a straight line.

Note that for any $(x, y) \in Q$, $y < f(x)$ implies $g(x, y) < 0$ and $\psi(x, y) = 0$, while $y > f(x)$ implies $g(x, y) > 0$ and $\psi(x, y) = 1$. Thus the Bayes invariant rule ψ satisfies

$$(4.9) \qquad \psi(x, y) = \begin{cases} 1 & \text{if } y > f(x), \text{ and} \\ 0 & \text{if } y < f(x), \end{cases}$$

for any $(x, y) \in Q$, as well as the symmetry property (4.4).

4.4 A Complete Class Theorem.

To find an essentially complete class of invariant rules, we will use the decision theory of Abraham Wald [28]. According to Wald's theorem 3.18, if B is the class of all Bayes rules with respect to priors with finite support, then the closure of B is an essentially complete class. The closure of B is taken in the following sense of convergence: the sequence $\{\psi_n\}$ defined in R^n is said to converge to ψ (in the sense of Wald's regular convergence) if

$$\lim \int_K \psi_n(x)dx = \int_K \psi(x)dx$$

for all compact sets $K \subset R^n$ (see Wald, page 134).

Since finding the closure of B can be difficult, a useful technique is to find a closed class of rules C which contains B. Then, as the closure of B must be contained in C, C will be an essentially complete class. This technique has been employed by A. Birnbaum [5], T. Matthes and D. Truax [21] and M. Eaton [11] in proving complete class theorems in various testing situations.

Definition 4.1.

E is the class of all functions of a real variable defined on $[0, \infty]$ which satisfy

a) $f(0) = 0$

b) $|f(x) - f(y)| \leq |x-y|$ for $x, y \geq 0$.

Note that all the functions f determined through (4.5) by priors with finite support are in E (see Lemma 4.3). Note also that

condition b) implies that each f in E is continuous; in fact, E is an equi-continuous family.

Lemma 4.4.

If $< f_n >$ is a sequence of functions in E, then there exists $f \in E$ and a subsequence $< f_{n_j} >$ which converges to f uniformly on compact sets.

Proof:

Since E is an equi-continuous family and a) and b) of Definition 4.1 imply $|f(x)| \leq x$ for all $x \geq 0$, an application of the Ascoli theorem (see, for example, Royden [25], page 155) immediately yields the existence of f and the subsequence $< f_{n_j} >$ which converges to f uniformly on compact sets. It remains to show $f \in E$.

Consider condition a). Since $f_{n_j}(0) = 0$ for all j, $f(0)$ must be 0. As for condition b), if $x, y \geq 0$, then

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_{n_j}(x)| + |f_{n_j}(x) - f_{n_j}(y)| \\ &\quad + |f_{n_j}(y) - f(y)| \\ &\leq |f(x) - f_{n_j}(x)| + |x-y| + |f_{n_j}(y) - f(y)|, \end{aligned}$$

for all j. But both the first and third terms on the right converge to zero, so $|f(x) - f(y)| \leq |x-y|$. □

We now define C and show that it both contains B and is closed in Wald's regular convergence.

Definition 4.2.

A test ψ is in C if and only if

a) ψ is a function of (R, S), the maximal invariant,

b) for some f in E and for all $(x, y) \in Q$,

$$\psi(x, y) = \begin{cases} 1 \text{ (a.e.)} & \text{if } y > f(x) \\ 0 \text{ (a.e.)} & \text{if } y < f(x) \end{cases}$$

where $x = cr$ and $y = ds$, and

c) $\psi(x, y) = \psi(-x, -y) = \psi(-y, -x) = \psi(y, x)$ (a.e.).

It is clear from (4.4) and (4.9) that B is contained in C. Note also, that if f is any function in E, then a), b), and c) can be used to define a test ψ uniquely (a.e.) in C.

Lemma 4.5.

C is closed in Wald's sense of regular convergence.

Proof:

To show C is closed, consider a sequence of rules $\psi_n \in C$ such that ψ_n converges to ψ in the sense of Wald's regular convergence. We wish to show that ψ is in C.

Let f_n be the functions in E associated with each ψ_n. As stated in Lemma 4.4, there is a subsequence f_{n_j} and a function f in E such that f_{n_j} converges to f uniformly on compact sets. Let φ be the test in C associated with f. To prove that ψ is in C, we will show that $\psi = \varphi$ (a.e.), first on Q and then, through symmetry, on the whole (x, y) plane.

Restricting our attention to Q, let

$$P_1 = \{(x, y) \in Q: y < f(x)\},$$

and assume K is any compact subset of P_1. Since K is compact, there exists an N such that $(x, y) \in K$ implies $x < N$. Also, since $f(x) - y$ is continuous and strictly positive for all $(x, y) \in K$, $\alpha = \inf\{f(x) - y: (x, y) \in K\}$ is achieved for some $(x_0, y_0) \in K$ and is strictly positive. Thus, since f_{n_j} converges uniformly to

f on [0, N], for j sufficiently large, $(x, y) \in K$ implies $|f_{n_j}(x) - f(x)| < \alpha$ and therefore, $y < f_{n_j}(x)$. Since $\psi_{n_j} \in C$, this last inequality implies $\psi_{n_j}(x, y) = 0$ (a.e.) on K. Therefore,

$$\int_K \psi = \lim_{j \to \infty} \int_K \psi_{n_j} = 0,$$

which implies $\psi(x, y) = 0$ (a.e.) on K.

Similarly, if $K \subset P_2$ where

$$P_2 = \{(x, y) \in Q: y > f(x)\},$$

then $\psi(x, y) = 1$ (a.e.) on K.

But both P_1 and P_2 can be written as the union of a countable number of compact sets so, in Q, $\psi(x, y) = 1$ (a.e.) if $y > f(x)$ and $\psi(x, y) = 0$ (a.e.) if $y < f(x)$.

If $\tilde{K}$ is any of the symmetric reflections of $K \subset P_i$, then since ψ_{n_j} satisfies the symmetry condition c) of Definition 4.2,

$$\int_{\tilde{K}} \psi = \lim_{j \to \infty} \int_{\tilde{K}} \psi_{n_j} = \lim_{j \to \infty} \int_K \psi_{n_j} = \int_K \psi.$$

Thus, ψ must satisfy c) as well as a) and b) of Definition 4.2. □

Since C is closed in Wald's sense of regular convergence and contains B, the class of all Bayes invariant rules with respect to priors with finite support, we have the following theorem.

Theorem 4.1.

For the classification problem as stated in Section 4.1 and the group of transformations defined in Section 4.2. C is an essentially complete class in the class of all invariant decision rules.

4.5 A Counterexample.

Because our work is based on techniques suggested by Matthes and Truax [21] and Eaton [11], it might be expected that our results

would be similar. For normal distributions in particular, they found complete classes of rules among the class of rules with convex acceptance regions for testing that the mean is the zero vector against certain restricted alternatives. Therefore, it might be suspected that the invariant Bayes rules for our problem have either convex acceptance or rejection regions in Q. We give a counter-example to this conjecture.

Let g_1 and g_2 be defined as in (4.1). As noted in the proof of Lemma 4.3, if $(x, f(x)) \in Q_0$, then $g_1(x-f(x)) = g_2(x+f(x))$. In fact, for all $x \geq x_0 = \inf\{x:(x, f(x)) \in Q_0\}$, $(x, f(x))$ is the graph of the boundary between the acceptance and rejection regions. If we make the rotation $s = x - y$ and $t = x + y$, then Q becomes the positive quadrant in the (s, t) plane and the boundary becomes $\{(s, t): g_1(s) = g_2(t)\}$. Writing $h(s) = x + f(x)$ and $s = x - f(x)$, this becomes $(s, h(s))$ for all $s \geq s_0$, the infimum of all s such that $(s, h(s))$ is in the interior of the positive quadrant. Since $\cosh x$ has positive continuous derivatives of all orders for $x \geq 0$, so has g_i. In particular, $t = g_2(s)$ has an increasing inverse g_2^{-1} for $t \geq 1$. Thus

$$h(s) = g_2^{-1}(g_1(s))$$

(for $s \geq s_0$) has a positive continuous derivative. Also, since $x = (s + h(s))/2$, we can write

$$x = c(s) = (s+h(s))/2$$

which also has a positive continuous derivative. Then $s = c^{-1}(x)$, and

$$f(x) = [h(c^{-1}(x)) - c^{-1}(x)]/2.$$

Now, because

$$\frac{d}{dx}\, c^{-1}(x) = [\frac{dc(s)}{ds}\,|_{s=c^{-1}(x)}]^{-1} = 2[1 + h'(c^{-1}(x))]^{-1},$$

we find,

$$f'(x) = \frac{h'(c^{-1}(x)) - 1}{h'(c^{-1}(x)) + 1} \quad , x \geq x_0 \ .$$

Note that since $h' > 0$,

$$|f'(x)| < 1$$

(cf. Lemma 4.3).

After some algebra, we also find

$$f''(x) = \frac{4h''(c^{-1}(x))}{[h'(c^{-1}(x)) + 1]^3} \ .$$

We will now obtain $h''(s)$ for a particular example and show that h'' changes sign. Since $h' > 0$, this implies that $f''(x)$ changes sign and neither acceptance or rejection region can be convex.

Suppose now $n = m = 6$, so $(c^2 + d^2)/2 = 2$, and ξ_1 puts probability $(3/40)$ on $\gamma = 0.2$ and $(9/40)$ on $\gamma = 10$ while ξ_2 puts probability $(28/40) = (7/10)$ on $\gamma = 1$. Then,

$$g_1(s) = \frac{3}{40}\,[e^{-.08}\cosh\,(.2s) + 3e^{-200}\,\cosh(10s)] \text{ and}$$

$$g_2(t) = \frac{7}{10}\,e^{-2}\,\cosh\,(t).$$

Thus $h(s) = \cosh^{-1}[\ell(s)]$, where

$$\ell(s) = \frac{3}{28}[e^{1.92}\,\cosh\,(.2s) + 3e^{-198}\,\cosh\,(10s)].$$

Now, after some calculation, we obtain

$$h'(s) = \ell'(s)[\sinh\,(\cosh^{-1}(\ell(s)))]^{-1} \text{ and}$$

$$h''(s) = \frac{[\sinh(\cosh^{-1}(\ell(s)))]^2\ell''(s) - \ell(s)[\ell'(s)]^2}{[\sinh(\cosh^{-1}(\ell(s)))]^3} \ .$$

If $s \geq s_0$, then $\ell(s) \geq 1$ and the denominator of the above expression is positive. Thus we need only find two values of s for which the numerator changes sign. In particular, if $s = 5$, we have computed that $\ell(5) = 1.128$ and the value of the numerator is -0.0210, while if $s = 20$, $\ell(20) = 21.08$ and the value of the numerator is about 4800.

REFERENCES

[1] Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. Proc. Amer. Math. Soc. 6 170-176.

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

[3] Anderson, T. W. and Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. Ann. Math. Statist. 33 420-431.

[4] Bhattacharya, P. K. and Das Gupta, S. (1964). Classification between univariate exponential populations. Sankhyā 26 17-24.

[5] Birnbaum, A. (1955). Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests. Ann. Math. Statist. 26 21-36.

[6] Cacoullos, T. (1965). Comparing Mahalanobis distances I, II. Sankhyā 27 Series A 1-32.

[7] Darling, D. A. and Robbins, H. (1968). Some further remarks on inequalities for sample sums. Proc. Nat. Acad. Sci. 60 1175-1182.

[8] Darling, D. A. and Robbins, H. (1968). Some nonparametric sequential tests with power 1. Proc. Nat. Acad. Sci. 61 804-809.

[9] Das Gupta, S. (1964). Nonparametric classification rules. Sankhyā 26 25-30.

[10] Das Gupta, S. (1965). Optimum classification rules for classification into two multivariate normal populations. Ann. Math. Statist. 36 1174-1184.

[11] Eaton, M. (1970). A complete class theorem for multidimensional one-sided alternatives. Ann. Math. Statist. 41 1884-1888.

[12] Ellison, B. E. (1962). A classification problem in which information about alternative distributions is based on samples. Ann. Math. Statist. 33 213-223.

[13] Feller, W. (1968). An Introduction to Probability Theory and Its Applications. 1 (3rd ed.). Wiley, New York.

[14] Fisher, L. and Van Ness, J. W. (1969). Distinguishability of probability measures. Ann. Math. Statist. 40 381-392.

[15] Freedman, D. A. (1967). A remark on sequential discrimination. Ann. Math. Statist. 38 1666-1670.

[16] Hoeffding, W. and Wolfowitz, J. (1958). Distinguishability of sets of distributions. Ann. Math. Statist. 29 700-718.

[17] Kiefer, J. and Schwartz, R. (1965). Admissible Bayes character of T^2, R^2, and other fully invariant tests for classical multivariate normal problems. Ann. Math. Statist. 36 747-770.

[18] Kiefer, J. and Wolfowitz, J. (1958). On the deviations of the empiric distribution functions of vector chance variables. Trans. Amer. Math. Soc. 87 173-186.

[19] Kraft, C. (1955). Some conditions for consistency and uniform consistency of statistical procedures. University of California Publications in Statistics, 2 125-142.

[20] Lehmann, E. L. (1959). Testing Statistical Hypothesis. Wiley, New York.

[21] Matthes, T.K. and Truax, D. R. (1967). Tests of composite hypotheses for the multivariate exponential family. Ann. Math. Statist. 38 681-697.

[22] von Mises, R. (1945). On the classification of observation data into distinct groups. Ann. Math. Statist. 16 68-73.

[23] Puri, M. L. and Sen, P. K. (1971). Nonparametric Methods in Multivariate Analysis. Wiley, New York.

[24] Robbins, H. (1970). Statistical methods relating to the law of the iterated logarithm. Ann. Math. Statist. 41 1397-1409.

[25] Royden, H. L. (1963). Real Analysis. Macmillan, New York.

[26] Srivastava, M. S. (1967). Comparing distances between multivariate populations--the problem of minimum distances. Ann. Math. Statist. 38 550-556.

[27] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Statist. 15 145-162.

[28] Wald, A. (1950). Statistical Decision Functions. Wiley, New York.

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

[30] Woinsky, M. N. and Kurz, L. (1969). Sequential nonparametric two-way classification with prescribed maximum asymptotic error probability. Ann. Math. Statist. 40 445-455.

[31] Yao, J. S. (1971). Optimal solutions for the problem of classification into one of several populations. Tamkang Jour. Math. 2 23-28.