CLASSIFIABILITY AND DESIGNS FOR SAMPLING

By

Somesh Das Gupta[1] and Albert Kinderman[2]
University of Minnesota   Carnegie-Mellon University

Technical Report No. 209

May, 1973

[1]The research was supported by U. S. Army Grant DA-ARO-D-31-124-70-G102.

[2]The research was supported partly by U. S. Army Grant DA-ARO-D-31-124-70-G102 and partly by a grant from the Sarah Mellon Scaife Foundation.

1. <u>Introduction</u>. We consider the problem of identifying a population π_0 with one of two specified populations π_1 and π_2, based on real-valued observations on sampled units from the populations. Let X_0, X_1, X_2 denote the random vector of observations on a unit drawn from π_0, π_1, and π_2, respectively. Let G_i be the distribution of X_i $(i = 0, 1, 2)$; it is assumed that $G_1 \times G_2$ belongs to a given collection Ω of distributions and G_0 equals either G_1 or G_2. The classification problem thus posed will be denoted by $C(\Omega)$. When G_1 and G_2 are known and Ω is a singleton set, rules based on indpendent observations on X_0 are considered. When Ω is not a singleton set, most authors devise rules based on observations on X_0, X_1, and X_2 without any precise reference to the nature of Ω. In this paper we consider the following basic problem of error control in classification for which we have found no reference in the existing literature. Given the set Ω, any $\epsilon > 0$, and a collection R of rules based on a sequence of independent observations on $X = (X_0, X_1, X_2)$, the problem is to investigate conditions for which there exists a rule in R having probabilities of correct classification (or identification) at least $1 - \epsilon$ uniformly for $G_1 \times G_2$ in Ω. If such a rule exists in the class of all sequential rules which terminate with probability 1 in Ω, we say $C(\Omega)$ is "classifiable" (or sequentially classifiable); $C(\Omega)$ is said to be finitely classifiable if such a rule exists in the class of all fixed sample size rules. With minor modifications of our results one may also treat the problem with unequal sample sizes from the three populations.

It turns out that the classifiability problem is equivalent to a special case of the problem of distinguishability between two sets of distributions, considered by Hoeffding and Wolfowitz (1958) (hereafter referred to as H-W). However, the special structure of our problem, with some additional assumptions, reduce the general results of H-W to simple ones in terms of the set Ω. These are obtained in Section 3 following some preliminary basic results on distance functions in Section 2. Special emphasis is given to the multivariate normal family.

A related question in the problem of controlling the probabilities of correct classification (arbitrarily and uniformly in Ω) is whether it is sufficient to consider rules based on independent observations on X_0 and X_1 or only observations on X_0. In Section 4, examples involving normal distributions are given which illustrate the differences between the various structures associated with finite sampling and sampling on X_0 and X_1 or X_0 alone.

Having raised the issue of finite sampling and sampling fewer than three populations, we must note that we have only found conditions on Ω for $C(\Omega)$ to be classifiable in each situation. There remains the problem of finding relative efficiencies and the problem of robustness against wrong specification of Ω, neither of which is treated here or in the literature.

In the last section, the problem of classification into more than 2 populations is treated from the viewpoint of error-control.

2. On Distance Functions.

A distance function δ is a pseudometric on a collection of distributions (probability measures) $\mathcal{G}$ defined on some measurable space $(\mathcal{X}, \mathcal{A})$. Let $\hat{G}_R$ be the empirical distribution based on the first R observations of a sequence independent random variables with the common distribution G. A distance δ is said to be uniformly consistent on $\mathcal{G}$ if, for every $\epsilon > 0$,

$$P_G [\delta (\hat{G}_R, G) > \epsilon] \to 0 \quad \text{as} \quad R \to \infty$$

uniformly for $G \in \mathcal{G}$. An example of a uniformly consistent distance on the set $\mathcal{G}$ of all distributions defined on m-dimensional Euclidean space $(\mathcal{X}, \mathcal{A})$ is the Kolmogorov distance given by

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

where F and G are c.d.f.'s. (Kiefer and Wolfowitz, 1958)

Since the classification problem is based on observations on a vector random variable $X = (X_0, X_1, X_2)$ with mutually independent X_i's, each X_i being defined on $(\mathcal{X}, \mathcal{A})$, we consider distances defined on collections of product distributions. Let δ_n denote a distance function defined on a collection of n-fold product distributions on $(\mathcal{X}^n, \mathcal{A}^n)$. The following condition on δ_n plays a major role in later sections.

Condition A.

Let $\mathcal{G}$ be a collection of distributions on $(\mathcal{X}, \mathcal{A})$ and $\mathcal{G}^n$ be the corresponding collection of n-fold product distributions on $(\mathcal{X}^n, \mathcal{A}^n)$. For a distance function δ_n defined on $\mathcal{G}^n \times \mathcal{G}^n$ there exists a distance function δ_1 defined on $\mathcal{G} \times \mathcal{G}$ such that for any sequences of distributions $\{F_{ik}\}_{k=1}^{\infty}$ and $\{G_{ik}\}_{k=1}^{\infty}$, $(i=1,\ldots, n)$ in $\mathcal{G}$

$$\lim_{k \to \infty} \delta_n(F_{1k} \times .. \times F_{nk} \,,\; G_{1k} \times .. \times G_{nk}) = 0$$

if, and only if

$$\lim_{k \to \infty} \delta_1 (F_{ik}, G_{ik}) = 0, \quad i = 1, \ldots, n.$$

In particular, if $G_{ik} \equiv G_i$, $i = 1, .., n$, Condition A implies that $F_{ik} \times .. \times F_{nk}$ converges to $G_1 \times .. \times G_n$ in the metric of δ_n, iff F_{ik} converges to G_i in the metric of δ_1 for each i.

The following sufficient condition for A to hold is easier to verify in some cases.

Condition A'.

There exist distances δ_k defined on $\mathcal{G}^k \times \mathcal{G}^k$, such that for any $k = 1, .., n$ and any $i = 1, .., k$,

$$\delta_k (F_1 \times .. \times F_k, G_1 \times .. \times G_k) \geq \delta_{k-1} (F_1 \times .. \times F_{i-1} \times F_{i+1} \times .. \times F_k, \; G_1 \times .. \times G_{i-1} \times G_{i+1} \times .. \times G_k)$$

with equality whenever $F_i = G_i$.

Lemma 2.1 Condition A' ⇒ Condition A.

Proof From Condition A', we get

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

$$\delta_{n-1} (F_1 \times .. \times F_{n-1}, G_1 \times .. \times G_{n-1}) \geq \delta_{n-2} (F_2 \times .. \times F_{n-1}, G_2 \times .. \times G_{n-1}).$$

Proceeding in this way and combining all such inequalities, we obtain

(2.1) $$\delta_n (F_1 \times .. \times F_n, G_1 \times .. \times G_n) \geq \delta_1 (F_i, G_i) \,, \quad i = 1, \ldots, n.$$

Also

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

By induction, we get

$$\text{(2.2)} \qquad \delta_n (F_1 \times .. \times F_n, G_1 \times .. \times G_n) \le \sum_{i=1}^{n} \delta_1 (F_i, G_i)$$

Condition A follows from (2.1) and (2.2).

We next consider two distances of special interest which satisfy A'. Let $(\mathfrak{X}, \mathfrak{G})$ be the m-dimensional Euclidean Space, and let D_n denote the Kolmogorov distance on the collection of all distributions on mn-dimensional Euclidean Space. For any sample space $(\mathfrak{X}, \mathfrak{G})$, the total variation metric d_n on the collection of all distributions on $(\mathfrak{X}^n, \mathfrak{G}^n)$ is given by

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

where $F[A]$ is the F-measure of the set A.

The proof of the following lemma is straightforward and is omitted (see Kinderman (1972) for a proof).

<u>Lemma 2.2.</u> The distance functions D_n and d_n satisfy the condition A'.

Given a distance function δ_1 on $\mathcal{G} \times \mathcal{G}$ we can construct a distance δ_n which satisfies A and is defined on the collection of distributions on $(\mathfrak{X}^n, \mathfrak{G}^n)$ whose marginal distributions are all in $\mathcal{G}$. For any distribution F on $(\mathfrak{X}^n, \mathfrak{G}^n)$, let $F_1, \ldots, F_n$ be the marginal distributions of F; i.e. $F_1[A] = F [A \times \mathfrak{X} \times .. \times \mathfrak{X}]$. For any F and G defined on $(\mathfrak{X}^n, \mathfrak{G}^n)$, define δ_n by

$$\delta_n(F, G) = \sum_{i=1}^{n} \delta_1(F_i, G_i)$$

If F and G are product distributions, then F and G must be the product of their respective marginals, (i.e. $F = F_1 \times .. \times F_n$) and δ_n thus defined will clearly satisfy the condition A.

If δ_1 is a uniformly consistent distance on $\mathcal{G}$, then δ_n as defined above, will also be uniformly consistent on the collection of all distributions whose marginals are in $\mathcal{G}$.

For two m-dimensional normal distributions $F_1 = \mathcal{N}_m(\mu_1, \Sigma_1)$ and $F_2 = \mathcal{N}_m(\mu_2, \Sigma_2)$, define a distance δ_1^* by

$$\delta_1^* (F_1, F_2) = [\int (\sqrt{f_1} - \sqrt{f_2})^2 \, d\nu]^{\frac{1}{2}},$$

where f_1 and f_2 are the p.d.f.'s of F_1 and F_2, respectively, with respect to the m-dimensional Lebesgue measure ν. An equivalent definition is

$$\delta_1^* (F_1, F_2) = \sqrt{2} \; [1 - \rho_1 (F_1, F_2)]^{\frac{1}{2}},$$

where

$$\rho_1 (F_1, F_2) = \int \sqrt{f_1 f_2} \; d\nu$$

is a measure of affinity.

For an m-dimensional distribution F (on the Euclidean space) with finite second moments, we define $N(F)$ to be the m-dimensional normal distribution with the first and second moments respectively equal to those of F. For such distributions F_1 and F_2, define

$$\delta_1^* (F_1, F_2) = \delta_1^* (N (F_1), N (F_2)).$$

The definition of δ_1^* is thus extended to all m-dimensional distributions with finite second moments.

Let $\hat{F}_n$ be the empirical distribution based on n i.i.d. observations with the common distribution F. When $F = \mathfrak{n}_m(\mu, \Sigma)$, the distribution of $\delta_1^* (\hat{F}_n, F)$ under F does not depend on μ and Σ. Thus δ_1^* is uniformly consistent in $\mathfrak{n}_m$.

For distributions $F_1, \ldots, F_n, H_1, \ldots, H_n$ in $\mathfrak{n}_m$, define

$$\rho_n (F_1 \times .. \times F_n, H_1 \times .. \times H_n) = \prod_{i=1}^{n} \rho_1 (F_i, H_i),$$

and

$$\delta_n^* (F_1 \times .. \times F_n, H_1 \times .. \times H_n) = \sqrt{2} \; [1 - \rho_n(F_1 \times .. \times F_n, H_1 \times .. \times H_n)]^{\frac{1}{2}}.$$

When F_i's and H_i's are m-dimensional distributions with finite second-order moments, define

$$\delta_1^* (F_1 \times .. \times F_n, H_1 \times .. \times H_n) = \delta_n^* (N(F_1) \times .. \times N(F_n), N(H_1) \times .. \times N(H_n)).$$

We then extend the definition of δ_n^* to the class of all mn-variate distributions with finite second-order moments by defining δ_n^* in terms of the product of n m-dimensional marginal distributions. Thus δ_n^* is uniformly consistent on the class of all mn-variate normal distributions.

Kraft (1955) has shown that

$$1 - \rho_1(F, H) \le d_1(F, H) \le [1 - \rho_1^2 (F, H)]^{\frac{1}{2}}$$

for any two m-dimensional distributions F and H with finite second moments, where ρ_1 and δ_1^* are related as above. It then follows that the convergence to 0 in d_1 is equivalent to convergence to 1 in ρ_1 which again is equivalent to convergence to 0 in δ_1^*. Note that (δ_n^*, δ_1^*) satisfies the condition A.

For two families of distributions $\mathcal{G}$ and $\mathcal{H}$ and a distance function δ we define the following notations:

$$\delta(G, \mathcal{H}) = \inf_{H \in \mathcal{H}} \delta(G, H), \quad \delta(\mathcal{G}, \mathcal{H}) = \inf_{\substack{G \in \mathcal{G} \\ H \in \mathcal{H}}} \delta(G, H).$$

3. Classifiability with samples from all three populations π_0, π_1, π_2.

Let $(\mathcal{X}, \mathcal{G})$ be the sample space of each of the random variables X_0, X_1, and X_2. We shall write $\pi_0 \sim \pi_i$, when the correct decision is to identify π_0 with π_i. A classification rule consists of a stopping variable N and a terminal decision function $\varphi = (\varphi_1, \varphi_2)$ where φ is a measurable function of N independent observations on $X = (X_0, X_1, X_2)$, and $\varphi_i(\cdot)$ is the conditional probability of making the decision $\pi_0 \sim \pi_i$ given N and the sequence of N independent observations on X. Let R be a collection of classification rules. $C(\Omega)$ is said to be classifiable in R if, for every $\epsilon > 0$ there is a rule (N, φ) in R such that

$$\pi_0 \sim \pi_i,\ G_1 \times G_2 \in \Omega \Rightarrow E\,\varphi_i \geq 1 - \epsilon$$

for $i = 1, 2$. Let

$$\Omega_0 = \{G_1 \times G_2 \ :\ G_1 \times G_2 \in \Omega,\ G_1 = G_2\}$$

$$\mathcal{G}_i(\Omega) = \{G_0 \times G_1 \times G_2 \ :\ G_0 = G_i,\ G_1 \times G_2 \in \Omega\}, \quad i = 1, 2.$$

According to H-W, $\mathcal{G}_1(\Omega)$ and $\mathcal{G}_2(\Omega)$ are distinguishable in the class of rules R, if for every $\epsilon > 0$ there is a rule (N, φ) in R such that

$$G_0 \times G_1 \times G_2 \in \mathcal{G}_i(\Omega) \Rightarrow E\,\varphi_i \geq 1 - \epsilon$$

for $i = 1, 2$.

The following two lemmas are immediate from the definitions.

Lemma 3.1. If $C(\Omega)$ is classifiable in R, then Ω_0 is empty.

Lemma 3.2. $C(\Omega)$ is classifiable in R, iff $\mathcal{G}_1(\Omega)$ and $\mathcal{G}_2(\Omega)$ are distinguishable in R.

We shall be concerned mainly with two particular classes of classification rules. Let R_s be the class of all rules (N, φ) for which $P_G[N < \infty] = 1$ whenever $G \in \mathcal{G}_1(\Omega) \cup \mathcal{G}_2(\Omega)$. We say $C(\Omega)$ is <u>sequentially classifiable</u> or (simply) classifiable when $C(\Omega)$ is classifiable in R_s. $C(\Omega)$ is said to be <u>finitely classifiable</u>, if it is classifiable in the class of all truncated sequential rules (i.e. $\max N < \infty$). As pointed out by H-W, the above definition is equivalent to the classifiability of $C(\Omega)$ in the class R_f of all fixed sample-size rules.

The main results of this Section are stated below. For a distance function δ_1 on the family of distributions on $(\mathcal{X}, \mathcal{A})$, define

$$\Omega_0(\delta_1) = \{G_1 \times G_2 \in \Omega : \ \delta_1(G_1, G_2) = 0\}$$

$$\delta_1(\Omega) = \inf_{G_1 \times G_2 \in \Omega} [\delta_1(G_1, G_2)].$$

<u>Theorem 3.1</u> (a) If $C(\Omega)$ is classifiable, then $\Omega_0 = \phi$.

(b) Suppose δ_3 is a distance function on a family of distributions on $(\mathcal{X}^3, \mathcal{A}^3)$ such that δ_3 satisfies condition A (in Section 2) and δ_3 is uniformly consistent in $\mathcal{G}_1(\Omega) \cup \mathcal{G}_2(\Omega)$. Then $C(\Omega)$ is classifiable if $\Omega_0(\delta_1) = \phi$.

(c) If the condition on δ_3 in (b) holds, and

$$\delta_1 (G_1, G_2) = 0 \Rightarrow G_1 = G_2 \qquad (\text{i.e. } \Omega_0(\delta_1) = \Omega_0),$$

then $C(\Omega)$ is classifiable, iff $\Omega_0 = \phi$.

<u>Theorem 3.2</u>. (a) If $C(\Omega)$ is finitely classifiable, then $d_1(\Omega) > 0$.

where d_1 is the total variation metric on the class of all distributions on $(\mathcal{X}, \mathcal{G})$.

(b) If the condition on δ_3 in Theorem 3.1 (b) holds, and

$$\delta_1(\Omega) > 0,$$

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

(c) If the condition on δ_3 in Theorem 3.1 (b) holds, and

$$\delta_1(\Omega) = 0 \Rightarrow d_1(\Omega) = 0,$$

then $C(\Omega)$ is finitely classifiable, iff $d_1(\Omega) > 0$.

These theorems will be proved using the following main lemma.

Lemma 3.3. Let δ_3 be a distance function on a family of distributions on $(\mathcal{X}^3, \mathcal{G}^3)$, and suppose δ_3 satisfies condition A (in Section 2)

Then

(a) for every $G_1 \times G_2 \in \Omega$,

$$\delta_3(G_1 \times G_1 \times G_2, \mathcal{G}_2(\Omega)) = 0 \Leftrightarrow \delta_1(G_1, G_2) = 0$$

$$\Leftrightarrow \delta_3(G_2 \times G_1 \times G_2, \mathcal{G}_1(\Omega)) = 0,$$

and

(b)

$$\delta_3(\mathcal{G}_1(\Omega), \mathcal{G}_2(\Omega)) = 0 \Leftrightarrow \delta_1(\Omega) = 0.$$

Proof. (a) We shall only prove $\delta_3(G_2 \times G_1 \times G_2, \mathcal{G}_1(\Omega)) = 0 \Leftrightarrow \delta_1(G_1, G_2) = 0$; the other part follows the same way.

Suppose $\delta_3(G_2 \times G_1 \times G_2, \mathcal{G}_1(\Omega)) = 0$. Then there exists a sequence $\{G_{1k} \times G_{1k} \times G_{2k}\}$ in $\mathcal{G}_1(\Omega)$ such that

$$\delta_3(G_2 \times G_1 \times G_2, G_{1k} \times G_{1k} \times G_{2k}) \to 0 \quad \text{as} \quad k \to \infty.$$

From Condition A, we then have

$$\delta_1 (G_2, G_{1k}) \to 0, \quad \delta_1 (G_1, G_{1k}) \to 0, \quad \delta_1 (G_2, G_{2k}) \to 0 \quad \text{as } k \to \infty.$$

Since

$$0 \le \delta_1 (G_1, G_2) \le \delta_1 (G_1, G_1\ ) + \delta_1 (G_{1k}, G_2)$$

we have $\delta_1 (G_1, G_2) = 0$.

Suppose $\delta_1 (G_1, G_2) = 0$. From condition A, we get

$$\delta_3 (G_2 \times G_1 \times G_2, \ G_1 \times G_1 \times G_2) = 0,$$

which implies $\delta_3 (G_2 \times G_1 \times G_2, \mathcal{G}_1 (\Omega)\ ) = 0$.

(b) Suppose $\delta_3 (\mathcal{G}_1 (\Omega), \mathcal{G}_2(\Omega)\ ) = 0$. Then there exist sequences $\{G_{1j} \times G_{1j} \times G_{2j}\}$ in $\mathcal{G}_1(\Omega)$ and $\{G_{2k} \times G_{1k} \times G_{2k}\}$ in $\mathcal{G}_2(\Omega)$ such that

$$\delta_3 (G_{1j} \times G_{1j} \times G_{2j}, \ G_{2k} \times G_{1k} \times G_{2k}) \to 0 \quad \text{as } j, k \to \infty.$$

From condition A, we then have

$$\delta_1 (G_{1j}, G_{2k}) \to 0, \quad \delta_1 (G_{1j}, G_{1k}) \to 0, \quad \delta_1 (G_{2j}, G_{2k}) \to 0.$$

Since

$$0 \le \delta_1 (G_{1k}, G_{2k}) \le \delta_1 (G_{1j}, G_{1k}) + \delta_1 (G_{1j}, G_{2k}),$$

we get $\delta_1 (\Omega) = 0$.

Suppose $\delta_1 (\Omega) = 0$. Then there exists a sequence $\{G_{1k} \times G_{2k}\}$ in Ω such that

$$\delta_1 (G_{1k}, G_{2k}) \to 0 \quad \text{as } k \to \infty.$$

By condition A, we get

$$\delta_3 (G_{1k} \times G_{1k} \times G_{2k}, G_{2k} \times G_{1k} \times G_{2k}) \to 0$$

which implies $\delta_3 (\mathcal{G}_1 (\Omega), \mathcal{G}_2 (\Omega)\ ) = 0$.

Proof of Theorem 3.1.

(a) Lemma 3.1 is repeated here for the sake of completeness.

(b) According to H-W (Theorem 3.1 (a)), $\mathcal{G}_1(\Omega)$ and $\mathcal{G}_2(\Omega)$ are distinguishable if

$$\max [\delta_3 (G_0 \times G_1 \times G_2, \mathcal{G}_1(\Omega)), \delta_3 (G_0 \times G_1 \times G_2, \mathcal{G}_2(\Omega)] > 0$$

for all $G_0 \times G_1 \times G_2 \in \mathcal{G}_1(\Omega) \cup \mathcal{G}_2(\Omega)$. However, from lemma 3.3, we get

$$\delta_3(G_2 \times G_1 \times G_2, \mathcal{G}_1(\Omega)) = 0 \Leftrightarrow \delta_1(G_1, G_2) = 0,$$

$$\delta_3(G_1 \times G_1 \times G_2, \mathcal{G}_2(\Omega)) = 0 \Leftrightarrow \delta_1(G_1, G_2) = 0$$

for any $G_1 \times G_2 \in \Omega$. Hence the above condition is equivalent to $\Omega_0(\delta_1) = \phi$. The part (c) follows from (a) and (b).

Proof of Theorem 3.2.

(a) According to H-W (equation 4.9), if $\mathcal{G}_1(\Omega)$ and $G_2(\Omega)$ are finitely distinguishable, then

$$d_3 (\mathcal{G}_1(\Omega), \mathcal{G}_2(\Omega) > 0;$$

which, by Lemma 3.3 (b), is equivalent to $d_1(\Omega) > 0$.

(b) According to H-W (see equation 3.4), a sufficient condition for finite distinguishability of $\mathcal{G}_1(\Omega)$ and $\mathcal{G}_2(\Omega)$ is

$$\delta_3 (G_1(\Omega), \mathcal{G}_2(\Omega)) > 0$$

which, by Lemma 3.3 (b), is equivalent to $\delta_1(\Omega) > 0$.

The part (c) follows from (a) and (b).

Corolloray 3.1. If $(\mathfrak{X}, \mathfrak{a})$ is the m-dimensional Euclidean space, then $C(\Omega)$ is classifiable iff $\Omega_0 = \phi$.

Proof. We apply Theorem 3.1 (c) with Kolmogorov distances D_3 and D_1 for δ_3 and δ_1, respectively. The uniform consistency of D_3 was shown by Kiefer and Wolfowitz (1958). Note also that (D_3, D_1) satisfies condition A.

Remark. In Theorem 3.1, it is assumed that δ_3 is uniformly consistent in $\mathcal{G}_1(\Omega) \cup \mathcal{G}_2(\Omega)$. This may be replaced by the following: δ_1 is uniformly consistent in $\mathfrak{F}$ and $\Omega \subset \mathfrak{F} \times \mathfrak{F}$. In that case, as noted in Section 2, a distance δ_3 is uniformly consistent on $\mathcal{G}_1(\Omega) \cup \mathcal{G}_2(\Omega) \subset \mathfrak{F} \times \mathfrak{F} \times \mathfrak{F}$ and (δ_3, δ_1) satisfies condition A so that Theorem 3.1 applies as stated.

When $(\mathfrak{X}, \mathcal{Q})$ is the m-dimensional Euclidean space, Theorem 3.2 may be applied to prove that the condition $d_1(\Omega) > 0$ (or $D_1(\Omega) > 0$) is necessary and sufficient for $C(\Omega)$ to be finitely classifiable, if $D_1(\Omega) = 0 \Rightarrow d_1(\Omega) = 0$. However, when $\Omega \subset \mathfrak{h}_m \times \mathfrak{h}_m$, $\mathfrak{h}_m$ being the collection of all m-dimensional nonsingular normal distributions $(m > 1)$, H-W shows that $D_1(\Omega) = 0$ need not imply $d_1(\Omega) = 0$. As in H-W, we shall derive a necessary and sufficient condition for $C(\Omega)$ to be finitely classifiable in terms of measure of affinity when $\Omega \subset \mathfrak{h}_m \times \mathfrak{h}_m$.

Theorem 3.3. If $\Omega \subset \mathfrak{h}_m \times \mathfrak{h}_m$, then $C(\Omega)$ is finitely classifiable iff

$$\rho_1(\Omega) \equiv \sup_{G_1 \times G_2 \in \Omega} [\rho_1(G_1, G_2)] < 1.$$

Proof. From the discussion in Section 2, it follows that

$$\rho_3(G_1(\Omega), G_2(\Omega)) = 1 \Leftrightarrow d_3(\mathcal{G}_1(\Omega), \mathcal{G}_2(\Omega)) = 0,$$

where

$$\rho_3(\mathcal{G}_1(\Omega), \mathcal{G}_2(\Omega)) = \sup [\rho_3(G_0 \times G_1 \times G_2, H_0 \times H_1 \times H_2) : G_0 \times G_1 \times G_2 \in \mathcal{G}_1(\Omega), H_0 \times H_1 \times H_2 \in \mathcal{G}_2(\Omega)].$$

From condition A, and Lemma 3.3, we get

$$d_3(\mathcal{G}_1(\Omega),\ \mathcal{G}_2(\Omega)\ ) = 0 \ \Leftrightarrow\ d_1(\Omega) = 0$$

$$\Leftrightarrow\ \rho_1(\Omega) = 1.$$

The theorem now follows from Theorem 5.2 (b) of H-W.

4. On Designs for Sampling.

For the purpose of uniformly controlling the errors of misclassification the possible sampling designs are denoted by Si or Fi $(i = 1, 2, 3)$ which respectively indicate that $C(\Omega)$ is classifiable in the class of all sequential rules or the class of all finite sample-size rules based on $(X_0, \ldots, X_{i-1})$. For example, Section 3 deals with S3- and F3-classifiability of $C(\Omega)$.

We first consider the S2- and F2- classifiability of $C(\Omega)$, which are equivalent, respectively, to the finite and sequential distinguishability of $\mathfrak{F}_1(\Omega)$ and $\mathfrak{F}_2(\Omega)$, where

$$\mathfrak{F}_1(\Omega) = \{ G_1 \times G_1 \ : \ G_1 \times G_2 \in \Omega \}$$

$$\mathfrak{F}_2(\Omega) = \{ G_2 \times G_1 \ : \ G_1 \times G_2 \in \Omega \}.$$

Theorem 4.1.

(a) A necessary condition for $C(\Omega)$ to be finitely classifiable on the basis of observations on (X_0, X_1) is

$$d_1(\Omega) > 0$$

(b) If δ_2 satisfies the condition A and is uniformly consistent on $\mathfrak{F}_1(\Omega) \cup \mathfrak{F}_2(\Omega)$, a sufficient condition for $C(\Omega)$ to be finitely classifiable on the basis of observations on (X_0, X_1) is

$$\delta_1(\Omega) > 0.$$

Proof. Since $F2 \Rightarrow F3$, the part (a) follows from Theorem 3.2(a). For the part (b) we note, as in the proof of Theorem 3.2, that the condition

$$\delta_2(\mathfrak{F}_1(\Omega), \mathfrak{F}_2(\Omega)) > 0 \qquad (4.1)$$

is a sufficient condition for the finite distinguishability of $\mathfrak{F}_1(\Omega)$ and $\mathfrak{F}_2(\Omega)$ when δ_2 is uniformly consistent on $\mathfrak{F}_1(\Omega) \cup \mathfrak{F}_2(\Omega)$, and is a necessary condition when $\delta_2 = d_2$, the total variation metric. If (4.1) does not hold, there exist sequences $\{F_{1k} \times F_{2k}\}$ and $\{G_{1k} \times G_{2k}\}$ in Ω such that

$$\lim_{k \to \infty} \delta_2 (F_{1k} \times F_{1k}, G_{2k} \times G_{1k}) = 0,$$

which, by condition A, is equivalent to

$$\lim_{k \to \infty} \delta_1 (F_{1k}, G_{2k}) = 0 = \lim_{k \to \infty} \delta_1 (F_{1k}, G_{1k}).$$

This implies

$$\lim_{k \to \infty} \delta_1 (G_{1k}, G_{2k}) = 0,$$

or $\delta_1(\Omega) = 0$.

The implication of Theorem 4.1 becomes clearer when normal distributions are considered. It follows from Sections 2 and 3 and Theorem 4.1 that

$$\delta_1^* (\Omega) > 0 \quad (\text{or}, \ \rho_1(\Omega) < 1)$$

is a necessary and sufficient condition for F2- and F3- classifiability of $C(\Omega)$. Thus, in this case, sampling from π_3 is not required in order to control errors with finite sample-size rules.

The analysis of the S2- classifiability of $C(\Omega)$ does not yield such neat results. For that reason, we only present a necessary and sufficient condition in its simplest form. As an example, the only easily expressed necessary condition is $\Omega_0 = \phi$, which, as seen from Theorem 3.1, is necessary for S3-classifiability and hence trivially necessary for S2-classifiability. One simple sufficient condition is presented below; no proof is given since the analysis is similar to that for F2-classifiability. Let

$$\Omega_1 = \{ G_1 : G_1 \times G_2 \in \Omega \}$$

and

$$\Omega_2 = \{ G_2 : G_1 \times G_2 \in \Omega \}$$

be the coordinate projections of Ω.

Theorem 4.2. If δ_2 satisfies condition A and is uniformly consistent on $\mathfrak{F}_1(\Omega) \cup \mathfrak{F}_2(\Omega)$, a sufficient condition for $C(\Omega)$ to be classifiable on the basis of observations on (X_0, X_1) is

$$\delta_1(G_1, \Omega_2) > 0 \quad \text{for all } G_1 \in \Omega_1.$$

Note now that $C(\Omega)$ is classifiable or finitely classifiable on the basis of observations on X_0 alone, iff Ω_1 and Ω_2 are distinguishable or finitely distinguishable. Necessary and sufficient conditions for S1- and F1-classifiability of $C(\Omega)$ follow directly from the results of H-W; for example, the condition $\delta_1(\Omega_1, \Omega_2) > 0$ is sufficient for F1-classifiability when δ_1 is uniformly consistent on $\Omega_1 \cup \Omega_2$. It may be noted that $C(\Omega)$ is (finitely) classifiable on the basis of observations on X_0 iff there exists $\mathcal{G}$ and $\mathcal{H}$ such that $\mathcal{G}$ and $\mathcal{H}$ are (finitely) classifiable and $\Omega \subset \mathcal{G} \times \mathcal{H}$.

The trivial relationships among Fi and Si classifiability for different i can be expressed as follows:

$$\begin{array}{ccccc} F1 & \Rightarrow & F2 & \Rightarrow & F3 \\ \Downarrow & & \Downarrow & & \Downarrow \\ S1 & \Rightarrow & S2 & \Rightarrow & S3 \end{array}$$

As noted before, $F2 \Leftrightarrow F3$ hold for normal distributions. However, none of the other possible implications hold for normal distributions as indicated by the following examples.

Example 1. $F2 \not\Rightarrow F1$, $F2 \not\Rightarrow S1$, $S2 \not\Rightarrow F1$, $S2 \not\Rightarrow S1$.

Let Ω be the collection of all products of pairs of univariate noraml distributions whose means differ by 1 and whose variances are 1. Then both Ω_1 and Ω_2 are the set of all $N(\mu, 1)$ distributions and they are clearly indistinguishable. On the other hand, $X_0 - X_1$ is normally distributed with variance 2 and mean 0 or ± 1 according as X_0 and X_1 have the same distribution or not. Thus the errors can be uniformly controlled with a finite number of observations on (X_0, X_1). This shows $F2 \not\Rightarrow S1$.

Example 2. $S1 \not\Rightarrow F2$, $S2 \not\Rightarrow F2$.

Let

$$\Omega = \{N(\mu_1, 1) \times N(\mu_2, 1) \ : \ \mu_1 > 0, \mu_2 < 0\}.$$

Then

$$\Omega_1 = \{N(\mu, 1) \ : \ \mu > 0\}, \quad \Omega_2 = \{N(\mu, 1) \ : \ \mu < 0\}.$$

Thus $C(\Omega)$ is S1 and S2 classifiable, since Ω_1 and Ω_2 are distinguishable; but $C(\Omega)$ is not F2-classifiable, since $\delta_1^*(\Omega) = 0$. This shows $S1 \not\Rightarrow F2$.

Example 3. S3 $\not\Rightarrow$ F3, S3 $\not\Rightarrow$ S2.

Let

$$\Omega = \{N(\mu_1, 1) \times N(\mu_2, 1) \ : \ \mu_1 \neq \mu_2\}.$$

Then $\Omega_0 = \phi$ and $C(\Omega)$ is S3-classifiable; but $C(\Omega)$ is not F3-classifiable since $\delta_1^*(\Omega) = 0$. In fact, $C(\Omega)$ is not S2-classifiable either, since

$$\delta_2^* (F, \mathfrak{F}_2(\Omega)) = 0 \quad \text{for every } F \in \mathfrak{F}_1(\Omega),$$

and hence

$$\max \{ \delta_2^* (F, \mathfrak{F}_1(\Omega)), \ \delta_2^* (F, \mathfrak{F}_2(\Omega)) \} = 0$$

for every $F \in \mathfrak{F}_1(\Omega)$. This follows from the fact that $\mathfrak{F}_1(\Omega)$ consists of all products of pairs with equal means while $\mathfrak{F}_2(\Omega)$ consists of all products of pairs with unequal means.

When the normal distributions are assumed to have unknown but equal variance σ^2, Ω is a subset of $\{N(\mu_1, \sigma^2) \times N(\mu_2, \sigma^2) : \sigma^2 > 0\}$. Then $C(\Omega)$ is F2-classifiable, if

$$\inf_\Omega \left|\frac{\mu_1 - \mu_2}{\sigma}\right| > 0,$$

and $C(\Omega)$ is S2-classifiable, if

$$\inf_\Omega |\mu_1 - \mu_2| > 0.$$

This follows from Theorems 4.1 and 4.2 and the definition of ρ_1.

5. Classification into one of Several Populations.

The problem here is to identify a population π_0 with one of specified populations $\pi_1, \ldots, \pi_k$. Let $X_0, X_1, \ldots, X_k$ denote the random vector of observations on a unit drawn from $\pi_0, \pi_1, \ldots, \pi_k$, respectively. Let G_i be the distribution of X_i $(i = 0, 1, \ldots, k)$ and it is assumed that G_0 equals one of $G_1, \ldots, G_k$, and $G_1 \times \ldots \times G_k$ belongs to a given set Ω of distributions. Let

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

$j = 1, \ldots, k$. A classification rule is given by $(N, \varphi_1, \ldots, \varphi_k)$ where N is a stopping variable and φ_j is the conditional probability of identifying π_0 with π_j (i.e. the decision $G_0 = G_j$) given N and N observations on $Y = (X_0, X_1, \ldots, X_k)$. Let F denote the distribution of Y.

$C(\Omega)$ will be said to be classifiable in a class of rules R if, for every $\epsilon > 0$ there is a rule $(N, \varphi_1, \ldots, \varphi_k)$ in R such that

$$F \in \mathcal{G}_j(\Omega) \;\Rightarrow\; E_F\,(1 - \varphi_j) \leq \epsilon$$

for all $j = 1, \ldots, k$.

Similarly, if $\mathcal{G}_1, \ldots, \mathcal{G}_k$ are any k collections of distributions F of Y, we can define the distinguishability of $\mathcal{G}_1, \ldots, \mathcal{G}_k$ based on Y as follows: $\mathcal{G}_1, \ldots, \mathcal{G}_k$ are distinguishable in a class of rules R if, for every $\epsilon > 0$ there is a test $(N, \varphi_1, \ldots, \varphi_k)$ in R such that

$$F \in \mathcal{G}_j \;\Rightarrow\; E_F\,(1 - \varphi_j) \leq \epsilon,$$

$j = 1, \ldots, k$. It is clear that $C(\Omega)$ is (finitely) classifiable iff $\mathcal{G}_1(\Omega), \ldots, \mathcal{G}_k(\Omega)$ are (finitely) distinguishable.

Results for the classifiability of $C(\Omega)$ depend on the following characterization of the distinguishability of $G_1,..G_k$ which is stated without proof. (For a proof, see Kinderman (1972)).

Theorem 5.1. $G_1,\ldots,G_k$ are (finitely) distinguishable iff $G_1,\ldots,G_k$ are pairwise (finitely) distinguishable.

As in Section 3, we define

$$\delta_1(\Omega) = \inf_{(G_1 x..x G_k)\varepsilon\,\Omega} [\delta_1(G_i, G_j) \ : \ i \neq j]$$

$$\Omega_0(\delta_1) = \{G_1 \, x..x \, G_k \ \varepsilon \ \Omega \ : \ \delta_1(G_i, G_j) = 0 \ \text{ for some } \ i \neq j\}$$

$$\Omega_0 = \{G_1 \, x..x \, G_k \ \varepsilon \ \Omega \ : \ G_i = G_j \ \text{ for some } \ i \neq j\}.$$

The following extensions of the results in Section 3 are stated without proofs (which are almost analogous).

Theorem 5.2 (a) If δ_{k+1} satisfies the condition A and is uniformly consistent in $G_1(\Omega) \cup \ldots \cup G_k(\Omega)$, then $\Omega_0(\delta_1) = \phi$ and $\delta_1(\Omega) > 0$ are respectively sufficient to ensure classifiability and finite classifiability of $C(\Omega)$.

(b) If δ_{k+1} satisfies the conditions stated in (a) and $\delta_1(G_i, G_j) = 0 \Rightarrow G_i = G_j$ then $C(\Omega)$ is classifiable, iff $\Omega_0 = \phi$.

Note that the Kolmogorov distances satisfies the conditions stated in Theorem 5.2 (b).

References

Hoeffding, W. and Wolfowitz, J. (1958). Distinguishability of probability measures. Ann. Math. Statist. 29 700-718.

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.

Kinderman, A. (1972). On some problems in classification: classifiability, asymptotic relative efficiency, and a complete class theorem. Technical Report # 178, School of Statistics, University of Minnesota.

Kraft, C. (1955). Some conditions for consistency and uniform consistency of Statistical procedures. University of California Publication in Statistics. 2 125-142.