AD NUMBER

ADA800276

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; FEB 1951. Other requests shall be referred to US Air Force School of Aviation Medicine, Randolph AFB, TX.

AUTHORITY

AFSAM ltr, 12 May 1970

ADA 800276

C 5028

110893

"...hen Government or other drawings, specifications or ...e used for any purpose other than in connection with ...lated Government procurement operation, the U.S. ...ereby incurs no responsibility, nor any obligation ...nd the fact that the Government may have formulated, ...in any way supplied the said drawings, specifications ...is not to be regarded by implication or otherwise as ...icensing the holder or any other person or corpora- ...ying any rights or permission to manufacture, use or ...ted invention that may in any way be related thereto."

UNCLASSIFIED

ADA800276

# DISCRIMINATORY ANALYSIS

## NONPARAMETRIC DISCRIMINATION: CONSISTENCY PROPERTIES


Evelyn Fix, Ph.D.
J.L. Hodges, Jr., Ph.D.
University of California, Berkeley



PROJECT NUMBER 21-49-004
REPORT NUMBER 4*

*(Prepared at the University of California under Contract No. AF41(128)-31)


USAF SCHOOL OF AVIATION MEDICINE
RANDOLPH FIELD, TEXAS
FEBRUARY 1951

# NONPARAMETRIC DISCRIMINATION: CONSISTENCY PROPERTIES

## 1. Introduction

The discrimination problem (two population case) may be defined as follows: a random variable Z, of observed value z, is distributed over some space (say, p-dimensional) either according to distribution F, or according to distribution G. The problem is to decide, on the basis of z, which of the two distributions Z has.

The problem may be classified in various ways into subproblems. One pertinent method of classification is according to the amount of information assumed to be available about F and G. We may distinguish three stages:

(i) F and G are completely known

(ii) F and G are known except for the values of one or more parameters

(iii) F and G are completely unknown, except possibly for assumptions about existence of densities, etc.

Subproblem (i) has been, in a sense, completely solved. The solution is implicit in the Neyman-Pearson lemma [1], and was made explicit by Welch [2]. We may without loss of generality assume the existence of density functions, say f and g, corresponding to F and G, since F and G are absolutely continuous with respect to F + G. If f and g

are known, the discrimination should depend only on $\frac{f(z)}{g(z)}$.
An appropriate (positive) constant c is chosen, and the following rule is observed:

If $\frac{f(z)}{g(z)} > c$, we decide in favor of F

If $\frac{f(z)}{g(z)} < c$, we decide in favor of G

If $\frac{f(z)}{g(z)} = c$, the decision may be made in an arbitrary manner

These procedures are known to have optimum properties with regard to control of probability of misclassification (probability of wrong decision). We shall refer to this as the "likelihood ratio procedure," and denote it by L(c).

For simplicity, we shall assume throughout the paper that the borderline case $f(z) = cg(z)$ can be neglected. Formally, we postulate that

$$P\{ f(Z) = cg(Z)\} = 0$$

regardless of whether Z comes from F or G. Since the classification is arbitrary when $f(z) = cg(z)$, it hardly seems worth while to introduce complications into the methods to allow for it. However, it is not difficult to extend our methods to take care of the situation which arises when

$$P\{ f(Z) = cg(Z)\} > 0.$$

The choice of c depends on considerations relating to the relative importance of the two possible errors: saying Z is distributed according to G when in fact it is distributed according to F, and conversely. Two choices of c have been widely advocated:

(a) Take $c = 1$

(b) Choose $c$ so that the two probabilities of error are equal.

Choice (a) has been called "logical"; choice (b) yields the minimax procedure. In this paper we shall not concern ourselves with the choice of $c$, but shall assume that a given positive $c$ is a datum of the problem.

The usual approach to subproblem (ii) is as follows. We assume there are available samples from the two distributions, say

$X_1, X_2, \ldots, X_m$ : sample from $F$

$Y_1, Y_2, \ldots, Y_n$ : sample from $G$.

We assume further that $F$ and $G$ are known in form; that is, that we know them except for the values of some real parameters, which may be denoted collectively by $\theta$. We may denote the distributions corresponding to a given $\theta$ by $F_\theta$, $G_\theta$. The procedure currently employed is to use the $X$'s and $Y$'s to estimate $\theta$, by, say, $\hat{\theta}$, and then to proceed as under (i), using the distributions $F_{\hat{\theta}}$, $G_{\hat{\theta}}$ as though they were known to be correct.

The most familiar example of this process is the linear discriminant function [3]. There, it is (tacitly) assumed that $F$ and $G$ are $p$-variate normal distributions having the same (unknown) covariance matrix , and unknown expectation vectors. The two expectation vectors and the covariance matrix are estimated from the samples, and the likelihood ratio procedure is then employed, using the estimated values as though they were known to be correct.

Not much is known about the desirability of the usual method of attack on (ii). We give in Section 3 a theorem concerning asymptotic properties of the method. Undoubtedly, this procedure is reasonable provided the assumed parametric form is correct. But the validity of the use of the linear discriminant function with data obviously not normal or, if normal, with obviously unequal covariance matrices has been of general concern. Presumably, very bad results may ensue if a procedure is used, based on certain assumptions about parametric form, when those assumptions are not even approximately correct.

There seems to be a need for discrimination procedures whose validity does not require the amount of knowledge implied by the normality assumption, the homoscedastic assumption, or any assumption of parametric form. The present paper is, as far as the authors are aware, the first one to attack subproblem (iii): can reasonable discrimination procedures be found which will work even if no parametric form can be assumed?

It is not to be expected that any procedure can be guaranteed to give good results without any restriction whatsoever on the distributions F and G. To clarify this point, we need to state a precise meaning for "good results." This is done in Section 2, with the introduction of the concept of "consistency." We then proceed in Section 4 to prove, under weak restrictions on the densities f and g, the consistency of a class of nonparametric procedures there proposed. A modification of these procedures is then considered in Section 5.

It may be noted that all of the methods and results of this paper can be extended without difficulty to the situation in which there are more than two populations to be discriminated.

The authors are engaged in further work along the lines here laid down. Specifically, some sampling experiments are being conducted, intended to throw some light on the performance of the procedures for moderate sample sizes; and asymptotic properties of a class of sequential nonparametric discriminatory procedures is being investigated. It is intended to prepare further reports setting forth the results.

## 2. The notion of consistency.

In setting out to define an optimum property in statistical inference, it is useful to have in mind the limit of excellance beyond which it is not possible to go. The procedures L(c) described in Section 1 provide such a limit in the case of nonparametric discrimination: we cannot, with any nonparametric classification procedure, expect to do better than the best which is possible when the densities themselves are assumed to be known. This fact is intuitively obvious, but if desired an exact proof is easily given. When f and g are known, Z is sufficient for the classification, with respect to $(Z; X_1, X_2, \ldots, X_m; Y_1, Y_2, \ldots, Y_n)$.

and we may (by using randomization) exactly duplicate (with a procedure based on Z) the performance characteristic of any procedure based on $(Z; X_1, X_2, \ldots, X_m; Y_1, Y_2, \ldots, Y_n)$.

Thus, no nonparametric procedure can have probabilities of error less than those of a likelihood ratio procedure. On the other hand, we shall propose in Sections 4 and 5 classes of (sequences of) nonparametric procedures which, in the limit as m and n tend to infinity, have the same probabilities of error as the procedures L(c). We may therefore reasonably say that our procedures are <u>consistent with</u> the likelihood ratio procedures.

There are two different notions of consistency for sequences of statistical decision functions, and it may be worth while to distinguish them. Suppose that the decision space is finite (as is the case in discriminatory analysis when there are finitely many populations). Let the possible decisions be denoted by $\delta_1, \delta_2, \ldots, \delta_r$. Now suppose we are considering two sequences of decision functions, say $\{\Delta_n'\}$ and $\{\Delta_n''\}$. How should we define the notion that these two sequences tend to agree with each other, or be consistent with each other, as $n \to \infty$? On the one hand, we might require that in the limit there should be close agreement between the probabilities of decision; on the other hand we might require that in the limit there be high probability of agreement of decision. The former requirement relates to the performance characteristics of the decision functions; the latter requirement relates to the decision functions themselves. We have then two definitions:

<u>Definition</u> 1. We shall say that the sequences $\{\Delta_n'\}$ and $\{\Delta_n''\}$ <u>are consistent in the sense of performance characteristics</u> if, whatever be the true distributions, and whatever be $\epsilon > 0$, there exists a number $N$ such that whenever $m > N$ and $n > N$,

$$\left| P\{\Delta_m' = \delta_i\} - P\{\Delta_n'' = \delta_i\} \right| < \epsilon$$

for every decision $\delta_i$.

<u>Definition</u> 2. We shall say that the sequences $\{\Delta_n'\}$ and $\{\Delta_n''\}$ <u>are consistent in the sense of decision functions</u> if, whatever be the true distributions, and whatever be $\epsilon > 0$, there exists a number $N$ such that whenever $m > N$ and $n > N$,

$$P\{\Delta_m' = \Delta_n''\} > 1 - \epsilon.$$

We observe that consistency in the second sense implies that in the first, since $P(\Delta_m' \neq \Delta_n'')$ is not less than each of the quantities $P(\Delta_m' = \delta_i \text{ and } \Delta_n'' \neq \delta_i) \geq P(\Delta_m' = \delta_i) - P(\Delta_n'' = \delta_i)$.

The definitions are not equivalent however, as the following trivial example shows. If $\Delta'$ and $\Delta''$ each denotes (for any $m,n$) the process of choosing between two alternatives $\delta_1$ and $\delta_2$ by tossing a coin, then $P(\Delta' = \Delta'') = \frac{1}{2}$, while

$$P(\Delta' = \delta_i) = P(\Delta'' = \delta_i) = \frac{1}{2} \quad \text{for } i = 1, 2.$$

Inasmuch as it is customary to evaluate decision functions solely in terms of their performance characteristics, Definition 1 is the more natural. However all proofs of consistency given in this paper provide consistency in the stronger sense of the second definition, and consequently we shall adopt it.

Since our procedures are based on two samples, we must consider a double limit process as both $m$ and $n$ tend to infinity. To avoid difficulties which would otherwise arise in Section 5, we shall assume throughout that $m$ and $n$ approach infinity at the same speed. Precisely, we assume $\frac{m}{n}$ and $\frac{n}{m}$ are both bounded away from $0$ as $n,m \longrightarrow \infty$. Whenever we write "$m,n \longrightarrow \infty$" this restriction should be understood. Our restriction has the effect of reducing the limiting process from a double to a single one.

In the sequel we shall be comparing certain discriminatory procedures with procedures of the type $L(c)$. It is convenient to introduce

Definition 3. A sequence $\{\Delta_{m,n}\}$ of discriminatory procedures, based on $Z$ and on samples $X_1, X_2, \cdots, X_m$ from $F$ and $Y_1, Y_2, \cdots, Y_n$ from $G$, is said to be consistent with $L(c)$ if, whatever be the distributions $F$ and $G$, regardless of whether $Z$ is distributed according to $F$ or according to $G$, and whatever be $\varepsilon > 0$, we can assure

$$P\{\Delta_{m,n} \text{ and } L(c) \text{ yield the same classification of } Z\} > 1 - \varepsilon$$

provided only that $m$ and $n$ are sufficiently large.

We may also define a corresponding notion of uniform consistency. If, in Definition 3, the bound on probability of agreement can be assured for all $F$ and $G$ with a single size specification on $m$ and $n$, we say that $\{\Delta_{m,n}\}$ is uniformly consistent with $L(e)$.

3. Consistency for the parametric case.

We shall now demonstrate that the analogy of the notion of consistency just introduced with the like-named notion in point estimation, is more than formal. Consider the problem of parametric discrimination (subproblem (ii)) of Section 1.

We shall from time to time have occasion to consider probabilities computed under the assumption that $Z$ is distributed according to $F$, or according to $G$. It is convenient to let $P_1$ and $P_2$ denote probabilities computed under these respective assumptions.

Let $\mathfrak{F}$ and $\mathfrak{G}$ be classes of densities parametrized by parameters denoted collectively by $\theta$. Let there be a notion of convergence introduced in the space $\Theta$ of parameter values. Suppose there is given a sequence $\{\hat{\theta}_{m,n}\}$ of estimates for $\theta$, $\hat{\theta}_{m,n}$ being a function of $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$.

<u>Theorem</u> 1. <u>If</u>

(a) <u>the estimates $\{\hat{\theta}_{m,n}\}$ are consistent,</u>

(b) <u>for every $\theta$, $f_\theta(z)$ and $g_\theta(z)$ are continuous functions of $\theta$ for every $z$ except perhaps for $z \in Z_0$, where</u>

$P_i(Z_\theta) = 0$, $i = 1, 2$, <u>then the sequence of discrimination procedures</u> $\{\hat{L}_{m,n}(c)\}$ <u>obtained by applying the likelihood ratio principle with critical value</u> $c > 0$ <u>to</u> $f_{\hat{\theta}_{m,n}}(z)$ and $g_{\hat{\theta}_{m,n}}(z)$ <u>is consistent with</u> $L(c)$.

<u>Proof</u>. The idea of the proof is very simple: Since $\hat{\theta}_{m,n}$ is consistent, $\hat{\theta}_{m,n}$ will probably be near $\theta$ if $m$ and $n$ are large. But since $f_\theta$ and $g_\theta$ are continuous, this means that $f_{\hat{\theta}_{m,n}}$ will probably be near $f_\theta$, and $cg_{\hat{\theta}_{m,n}}$ will probably be near $cg_\theta$. Therefore, it is not likely to make much difference whether we compare $f_\theta$ and $cg_\theta$ or $f_{\hat{\theta}_{m,n}}$ and $cg_{\hat{\theta}_{m,n}}$.

Fix $c > 0$, $\varepsilon > 0$, and $\theta \varepsilon \Theta$. Find $\delta > 0$ so small that $P_i\{|f_\theta(Z) - cg_\theta(Z)| \leqq \delta\} < \frac{1}{2}\varepsilon$, $i = 1, 2$. (This is possible since $P_i\{|f_\theta(Z) - cg_\theta(Z)| \leqq u\}$ is the cumulative function of the random variable $|f_\theta(Z) - cg_\theta(Z)|$ and hence is continuous on the right, and by assumption takes on the value $0$ when $u = 0$). We now assume that $z$ does not lie in $Z_\theta$, thus excluding an event of zero probability. Since $f_\theta(z)$ is a continuous function of $\theta$ for all $z$, we can associate with every $z$ a quantity $\eta_1(z) > 0$ such that

$$|f_{\hat{\theta}}(z) - f_\theta(z)| < \frac{\delta}{2} \quad \text{whenever} \quad |\hat{\theta} - \theta| < \eta_1(z).$$

A like function $\eta_2(z)$ arises if $f$ is replaced by $cg$. Let

$\eta(z) = \min\{\eta_1(z), \eta_2(z)\}$ and find $\eta > 0$ such that $P_i(\eta(Z) < \eta) < \frac{1}{4}\varepsilon$, $i = 1, 2$.

Using finally the consistency of the estimates, choose $M$ and $N$ large enough so that whenever $m > M$ and $n > N$, $P\{|\hat{\theta}_{m,n} - \theta| > \eta\} < \frac{1}{4}\varepsilon$. Combining the above, a disagreement between $L(c)$ and $\hat{L}_{m,n}(c)$ will arise with probability less than $\varepsilon$.

Remarks. (1) The dependence of the discontinuity sets $Z_\theta$ on $\theta$ is important. Were we to demand the stronger property that $f_\theta(z)$ and $g_\theta(z)$ be continuous in $\theta$ for all $z \notin Z$, $Z$ a fixed set, $P_i(Z) = 0$, $i = 1, 2$, we should exclude many cases which are included under the theorem as given.

(2) Two notions of convergence in $\Theta$ are involved: that with respect to which the estimates are consistent, and that with respect to which the densities are continuous. These need not be the same, provided the former implies the latter.

(3) If uniformity is added to the hypotheses of theorem 1, it may also be added to the conclusions. Specifically, if the estimates $\hat{\theta}_{m,n}$ are uniformly consistent, if the densities $f$ and $g$ are uniformly continuous functions of $\theta$, uniformly in $z$, and if the $\delta$ of the proof of theorem 1 may be fixed independently of $\theta$, then that proof goes through for all $\theta$ using the same value of $\varepsilon$. We can then conclude the uniform consistency of $\{\hat{L}_{m,n}(c)\}$.

## 4. Nonparametric discrimination and its consistency.

Let us next consider the discrimination problem of the third kind delineated in Section 1. We admit the possibility that the densities $f$ for $X$ and $g$ for $Y$ may be any in certain classes $\mathfrak{F}$ and $\mathfrak{G}$ of densities which are too large to be characterized by a finite number of parameters. Thus, $\mathfrak{F}$ and $\mathfrak{G}$ may consist of all uniformly continuous densities, or of all continuous densities, or of all densities continuous save at most at countably many points. Can we have any discrimination procedures which are reasonable to use when so little is assumed about the populations being discriminated?

Recall that, once $c$ has been selected and $Z$ has been observed to have the value $z$, the only information needed to carry out the procedure $L(c)$ are the two real numbers $f(z)$ and $g(z)$. In the procedure $\hat{L}_{m,n}(c)$, we employed the estimate for $\theta$ as a means of obtaining estimates for $f_\theta(z)$ and $g_\theta(z)$. In the nonparametric case there is no $\theta$ to be estimated, but we may instead proceed to estimate the numbers $f(z)$ and $g(z)$ directly. Once estimates have been obtained, we may apply the procedure $L(c)$, using these estimates instead of $f(z)$ and $g(z)$. We shall designate such procedures by $L^{\#}(c, \hat{f}, \hat{g})$, where $\hat{f}$ and $\hat{g}$ are the estimates for $f$ and $g$.

Before considering the problem of estimating the densities, let us note the properties which such estimates should have if we are to be able to prove the consistency of $L^{\#}(c, \hat{f}, \hat{g})$ with $L(c)$.

Theorem 2. If $\hat{f}_{m,n}(z)$ and $\hat{g}_{m,n}(z)$ are consistent estimates for $f(z)$ and $g(z)$ for all $z$ except possibly $z \in Z_{f,g}$

where $P_i(Z_{f,g}) = 0$, $i = 1, 2$, then $\{L^{\#}_{m,n}(c, \hat{f}, \hat{g})\}$ is consistent with $L(c)$.

The proof follows lines similar to that of theorem 1, and will be omitted.

Our problem is now to find consistent estimates for $f(z)$ and $g(z)$. We shall for brevity consider $f(z)$ only, as analogous remarks apply to $g(z)$. We fix $z$, since the argument is the same for each value. Our basic idea is this: the proportion of the $m$ X's which fall in a stated (small) neighborhood of $z$ may be used to estimate the X-probability in that neighborhood. The ratio of this estimated probability to the measure of the neighborhood is then an estimate of the average value of $f(x)$ near $z$. This is in turn an estimate of $f(z)$ itself if we make some assumption about the smoothness of $f$. To obtain consistency, we may let the neighborhood shrink down to $z$ as $m \to \infty$, so that the average of $f(x)$ over the neighborhood will approach $f(z)$; but we will take care to have the neighborhood shrink slowly enough so that the proportion of the X's therein will have a positive expectation. This will assure that the proportion of X's in the neighborhood is a consistent estimate of the probability.

It is obvious that we cannot hope to estimate $f(z)$ from $X_1, X_2, \cdots, X_m$ unless some continuity assumption is made. For, otherwise we could alter $f(z)$ arbitrarily without in any way changing the distribution of $X_1, X_2, \cdots$, and thus without changing the distribution of any sequence of estimates based on $X_1, X_2, \cdots$.

Now let $\mu$ denote Lebesgue measure in our (p-dimensional) sample space, and let $|x - y|$ denote the (Euclidean) distance between points $x$ and $y$ of this space.

__Lemma__ 3. __If__ $f(x)$ __is continuous at__ $x = z$, __and if__ $\{\Delta_m\}$ __is a sequence of sets such that__

$$\lim_{m \to \infty} \sup_{d \in \Delta_m} |z-d| = 0,$$

__and__ $\lim_{m \to \infty} m \cdot \mu(\Delta_m) = \infty$, __and if__ $M$ __is the number of__ $X_1, X_2, \cdots, X_m$ __which lie in__ $\Delta_m$, __then__ $\{\frac{M}{m\mu(\Delta_m)}\}$ __is a consistent estimate for__ $f(z)$.

__Proof__. Observe that $\frac{P(\Delta_m)}{\mu(\Delta_m)} \to f(z)$ as $m \to \infty$. If $f(z) > 0$, $m P(\Delta_m) \to \infty$. Since $\mu(\Delta_m) \to 0$, $P(\Delta_m) \to 0$ and we conclude $\frac{M}{mP(\Delta_m)} \xrightarrow[P]{} 1$. Combining $\frac{M}{m\mu(\Delta_m)} \xrightarrow[P]{} f(z)$ as was to be shown. If $f(z) = 0$, $E\left(\frac{M}{m\mu(\Delta_m)}\right) = \frac{P(\Delta_m)}{\mu(\Delta_m)} \to 0$ and the Markoff lemma completes the proof.

We have in lemma 3 a class of estimates, any of which, by virtue of theorem 2, will provide consistent discrimination of any (nonparametric) classes $\mathfrak{F}$ and $\mathfrak{G}$ whose members are continuous (except possibly for a set of values of zero measure).

5. Alternative procedures.

While the procedures $L^*\left(c, \frac{M}{m\mu(\Delta_m)}, \frac{N}{n\mu(\Delta_m)}\right)$ of the last

Section provide consistent discrimination, the question of their applicability when $m$ and $n$ are not large remains open. (Like criticism may of course be applied to any asymptotic theorem.) We shall in the present section suggest some alternative estimates for $f(z)$ and $g(z)$, which seem on intuitive grounds more likely to give good results than the estimates proposed before. The former estimates are the natural ones when thinking of the simplicity of consistency proofs, but need not be desirable in practice.

The main practical difficulty in using the former estimates lies in the choice of the regions $\{\Delta_m\}$, (and the corresponding regions for $g$, say $\{\Lambda_n\}$). If these regions are made too small, the numbers $M$ and $N$ of sample points falling into them will be too small, so that the proportions $\frac{M}{m}$ and $\frac{N}{n}$ will not be accurate estimates for the corresponding probabilities $P_1(\Delta_m)$, $P_2(\Lambda_n)$. On the other hand, if the regions are made too large, these probabilities will not be good approximations for $f(z)\mu(\Delta_m)$ and $g(z)\mu(\Lambda_n)$. We are between twin perils and must steer a middle course. We might, for example, decide the smallest values of $M$ and $N$ we could tolerate, and choose $\Delta_m$ and $\Lambda_n$ just big enough to include the chosen number of points. But to do so alters the probabilistic properties; now $M$ and $N$ are fixed and $\Delta$ and $\Lambda$ are random. Are the results of lemma 3 still valid?

Even if they are we may still be in difficulties. It may happen that near $z$ there are numerous X's, but few Y's; but by going a little further we find the situation reversed. The indication is clearly for $\pi_1$, but if we take separate $\Delta$ and $\Lambda$ the estimated f and g may be close. To avoid this difficulty the following idea is suggested: Choose a number k, and take in the neighborhood of z a single region, $\Delta_{M,N}$, containing a total of k points of either sample. Intuitively this procedure seems sound, but since $M + N = k$ we have introduced dependence of our estimates and further altered the probabilistic properties. The question which now arises is whether or not estimates for $f(z)$ and $g(z)$ based on M and N, when so determined, are still consistent.

As a first step in answering these questions, observe that we may by means of a preliminary transformation reduce our space from p dimensions to one. Let $\rho(x,y)$ denote a non-negative real valued function of pairs $(x,y)$ of points in the sample space. Suppose $\rho$ is so constructed that when $x_n \longrightarrow x$, $\rho(x_n, x) \longrightarrow 0$, and suppose further that for each z, except perhaps for $z \in Z_{f,g}$ where $P_i(Z_{f,g}) = 0$, $i = 1, 2$; $\rho(X,z)$ and $\rho(Y,z)$ are random variables possessing densities, say $f_z(x)$ and $g_z(x)$, continuous and not both 0 at 0. (These properties are satisfied, for example, by $\rho(x,y) = \sqrt[p]{|x-y|}$). We now replace the problem of deciding whether $f(z)$ or $cg(z)$ is the larger, by the problem of deciding whether $f_z(0)$ or $cg_z(0)$ is larger; and further replace the samples $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ by $\rho(X_1,z), \rho(X_2,z), \cdots, \rho(X_m,z)$ and $\rho(Y_1,z), \rho(Y_2,z), \cdots, \rho(Y_n,z)$, respectively. We may now, without real loss of generality, assume that $f$ and $g$ are densities of non-negative univariate random variables, and that $z = 0$.

Theorem 4. Let $X$ and $Y$ be non-negative. Let $f$ and $g$ be positive and continuous at $0$. Let $k(m,n)$ be a positive-integer-valued function such that $k(m,n) \longrightarrow \infty$, $\frac{1}{m} k(m,n) \longrightarrow 0$, and $\frac{1}{n} k(m,n) \longrightarrow 0$, as $m$ and $n \longrightarrow \infty$. (This tendency being restricted so that $\frac{m}{n}$ is bounded away from $0$ and $\infty$). Define

$U = k^{th}$ smallest value of combined samples of $X$'s and $Y$'s,

$M =$ number of $X$'s $\leqq U$,

$N =$ number of $Y$'s $\leqq U$.

Then $\frac{M}{mU}$ is a consistent estimate for $f(0)$ and $\frac{N}{nU}$ is a consistent estimate for $g(0)$.

Proof. Fix $\epsilon > 0$ and $\delta > 0$. Define $k_1(m,n)$ and $k_2(m,n)$ by $k_1(m,n) + k_2(m,n) = k(m,n)$ and $\frac{k_1(m,n)}{k_2(m,n)} = \frac{mf(0)}{ng(0)}$.

Define $v(m,n) = \frac{k_1(m,n)}{mf(0)(1+\delta)^2}$ and $w(m,n) = \frac{k_1(m,n)}{mf(0)(1-\delta)^2}$.

Observe $v(m,n) = \frac{k_2(m,n)}{ng(0)(1+\delta)^2}$ and $w(m,n) = \frac{k_2(m,n)}{ng(0)(1-\delta)^2}$.

Define

$M^v_{m,n} =$ number of $X$'s $< v(m,n)$,

$M^w_{m,n} =$ number of $X$'s $< w(m,n)$,

$N^v_{m,n} =$ number of $Y$'s $< v(m,n)$.

$$N^{w}_{m,n} = \text{number of } Y\text{'s} < w(m,n).$$

Using the continuity and positiveness of $f$ and $g$ at 0, find $q > 0$ so small that when $0 \leq x \leq q$, $\left|\frac{f(x)}{f(0)} - 1\right| < \delta$ and $\left|\frac{g(x)}{g(0)} - 1\right| < \delta$. Find $m_1$, $n_1$ such that when $m > m_1$ and $n > n_1$, $w(m,n) < q$, and make these restrictions. Observe

$$E(M^{v}_{m,n}) = m \int_{0}^{v(m,n)} f(x)dx \quad \text{and hence}$$

$$mf(0)\ v(m,n)(1-\delta) < E(M^{v}_{m,n}) < mf(0)\ v(m,n)(1+\delta).$$

Similarly observe

$$mf(0)\ w(m,n)(1-\delta) < E(M^{w}_{m,n}) < mf(0)\ w(m,n)(1+\delta)$$

$$ng(0)\ v(m,n)(1-\delta) < E(N^{v}_{m,n}) < ng(0)\ v(m,n)(1+\delta)$$

$$ng(0)\ w(m,n)(1-\delta) < E(N^{w}_{m,n}) < ng(0)\ w(m,n)(1+\delta)$$

Thus, $E(M^{v}_{m,n}) < \dfrac{k_1(m,n)}{1+\delta}$, $\qquad E(M^{w}_{m,n}) > \dfrac{k_1(m,n)}{1-\delta}$,

$E(N^{v}_{m,n}) < \dfrac{k_2(m,n)}{1+\delta}$, $\qquad E(N^{w}_{m,n}) > \dfrac{k_2(m,n)}{1-\delta}$.

The random variables involved are binomials, whose expectations tend to $\infty$, but more slowly than the numbers of trials, as $m,n \longrightarrow \infty$. Therefore, if we take $m_2$, $n_2$ large enough, we can assure

$$P\left(M^{v}_{m,n} < k_1(m,n)\right) > 1 - \epsilon$$

$$P\left(N^{V}_{m,n} < k_2(m,n)\right) > 1 - \varepsilon$$

$$P\left(M^{W}_{m,n} > k_1(m,n)\right) > 1 - \varepsilon$$

$$P\left(N^{W}_{m,n} > k_2(m,n)\right) > 1 - \varepsilon$$

as soon as $m > m_2$ and $n > n_2$, which restriction we now make. Combining, using the fact that $U$ will exceed $v(m,n)$ if $M^{V}_{m,n} + N^{V}_{m,n} < k(m,n)$, we have

$$P\left(U > v(m,n)\right) > 1 - 2\varepsilon,$$

$$P\left(U < w(m,n)\right) > 1 - 2\varepsilon.$$

The event $U > v(m,n)$ implies the event that all $X$'s $< v(m,n)$ are among the first $k$ $X$'s and $Y$'s and hence the event $M^{V}_{m,n} \leqq M$. Therefore, $P(M^{V}_{m,n} \leqq M) \geqq P\left(U > v(m,n)\right) > 1 - 2\varepsilon$. Similarly, $P(M^{W}_{m,n} \geqq M) > 1 - 2\varepsilon$. Restricting $m > m_3$, $n > n_3$, we can further assure

$$P\left(M^{V}_{m,n} > mf(0)\ v(m,n)(1-\delta)^2\right) > 1 - \varepsilon,$$

$$P\left(M^{W}_{m,n} < mf(0)\ w(m,n)(1+\delta)^2\right) > 1 - \varepsilon.$$

Combining,

$$P\left(\frac{M}{m} < f(0)\ w(m,n)(1+\delta)^2\right) > 1 - 3\varepsilon,$$

$$P\left(\frac{M}{m} > f(0)\ v(m,n)(1-\delta)^2\right) > 1 - 3\varepsilon.$$

Hence $$P\left\{f(0) \cdot \frac{v(m,n)}{w(m,n)}(1-\delta)^2 < \frac{M}{mU} < f(0) \cdot \frac{w(m,n)}{v(m,n)}(1+\delta)^2\right\} > 1 - 10\varepsilon.$$

Since $\frac{v(m,n)}{w(m,n)} = \frac{(1-\delta)^2}{(1+\delta)^2}$, the conclusion $\frac{M}{mU} \xrightarrow{p} f(0)$ is at hand.

A similar argument shows $\frac{N}{nU} \xrightarrow{p} g(0)$.

A situation in which one of the densities is 0 at 0 can be dealt with by a corresponding but simpler argument which we omit. The effect of theorem 4 is to assure us of satisfactory large sample results if we employ procedures of the following kind:

Choose k, a positive integer which is large but small compared to the sample sizes. Specify a metric in the sample space, for example ordinary Euclidean distance. Pool the two samples and find, of the k values in the pooled samples which are nearest to z, the number M which are X's. Let $N = k-M$ be the number which are Y's. Proceed with the likelihood ratio discrimination, using however $\frac{M}{m}$ in place of $f(z)$ and $\frac{N}{n}$ in place of $g(z)$. That is, assign Z to F if and only if

$$\frac{M}{m} < c \frac{N}{n} .$$

UNCLASSIFIED

ATI 110 633 (COPIES OBTAINABLE FROM CADO)

UNIVERSITY OF CALIFORNIA, BERKELEY

DISCRIMINATORY ANALYSIS - NONPARAMETRIC DISCRIMINATION: CONSISTENCY PROPERTIES - PROJECT REPORT

EVELYN FIX; J.L. HODGES, JR. FEB'51 21PP

USAF SCHOOL OF AVIATION MEDICINE, RANDOLPH AIR FORCE BASE, TEX., USAF CONTR. NO. AF-41(128)-31 (REPORT NO. 4)

STATISTICAL ANALYSIS SCIENCES, GENERAL (33) MATHEMATICS (3)

UNCLASSIFIED

# REFERENCES


1. J. NEYMAN and E. S. PEARSON, "Contributions to the theory of testing statistical hypotheses," Stat. Res. Memoirs, Vol. 1 (1936), pp. 1-37.
2. B. L. WELCH, "Note on discriminant functions," Biometrika, Vol. 31 (1939), pp. 218-220.
3. R. A. FISHER, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, Vol. 7 (1936), pp. 179-188.