DISTINGUISHABILITY OF SETS OF DISTRIBUTIONS

(The case of independent and identically distributed chance variables)

By Wassily Hoeffding[1] and J. Wolfowitz[2]

University of North Carolina and Cornell University

1. Introduction. Suppose it is desired to make one of two decisions, d_1 and d_2, on the basis of independent observations on a chance variable whose distribution F is known to belong to a set $\mathfrak{F}$. There are given two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{F}$ such that decision $d_1(d_2)$ is strongly preferred if F is in $\mathcal{G}$ ($\mathcal{H}$). Then it is reasonable to look for a test (decision rule) which makes the probability of an erroneous decision small when F belongs to $\mathcal{G}$ or $\mathcal{H}$, and at the same time exercises some control over the number of observations required to reach a decision when F is in $\mathfrak{F}$ (not only in $\mathcal{G}$ or $\mathcal{H}$).

This paper is concerned with criteria that enable us to decide whether, for given sets $\mathfrak{F}$, $\mathcal{G}$, and $\mathcal{H}$, there exists a test of the described type. More precisely, we shall consider several classes of tests, such as the class of all fixed sample size tests, or the class of all tests which terminate with probability one whenever F is in $\mathfrak{F}$. Thus the restriction to tests in one of these classes is equivalent to imposing some sort of control, of a purely qualitative nature, on the sample size. We then shall try to find necessary and/or sufficient conditions for the existence of a test in a given class which makes the maximum error probability in $\mathcal{G} \cup \mathcal{H}$ less than any preassigned positive number.

If such a test exists, we shall say that the sets $\mathcal{G}$ and $\mathcal{H}$ are distinguishable[3] in the given class $\mathfrak{J}$ of tests. If $\mathfrak{J}$ is the class of all fixed sample size tests, the distinguishability of $\mathcal{G}$ and $\mathcal{H}$ in $\mathfrak{J}$ is equivalent to the existence of what has been called a uniformly consistent sequence of tests for testing $F \varepsilon \mathcal{G}$ against $F \varepsilon \mathcal{H}$.

The sets $\mathcal{G}$ and $\mathcal{H}$ will be called indistinguishable in $\mathfrak{J}$ if for any test in $\mathfrak{J}$ the sum of the maximum error probability in $\mathcal{G}$ and the maximum error probability in $\mathcal{H}$ is at least one. (There always exists a trivial test for which this sum is equal to one.) In section 2 it will be shown that, with the present restriction to sequences of independent and identically distributed chance variables, two sets are either distinguishable or indistinguishable in any of the classes $\mathfrak{J}$ which we shall consider.

Received May 20, 1957.

[1] The research of this author was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under contract No. AF 18(600)-458. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] The research of this author was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under contract No. AF 18(600)-685. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[3] In [3] the term distinguishable was used in another sense.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

Since we confine ourselves to tests based on a sequence $X_1, X_2, \cdots$ of independent, identically distributed chance variables, we may restrict ourselves to sequential tests. A sequential test is determined by the sample size function N and the terminal decision function ϕ, and will be denoted by (N, ϕ). Here N is a chance variable whose values are the non-negative integers and $+\infty$, and whose conditional distribution, given any sequence $\mathbf{x} = (x_1, x_2, \cdots)$ of possible values of $X_1, X_2, \cdots$, is such that the probability of $N \leqq n$ does not depend on $x_{n+1}, x_{n+2}, \cdots$, for all non-negative integers n. The function ϕ is a function of $(x_1, \cdots, x_N)$ whose values range from 0 to 1. The test (N, ϕ) consists in taking one observation on each of the first N chance variables of the sequence, finding the corresponding value of ϕ, and making decision d_1 or d_2 with respective probabilities $1 - \phi$ and ϕ. The function ϕ and the conditional distribution function of N given $\mathbf{x}$ are always understood to be measurable on the appropriate σ-field. The sample size function N and the terminal decision function ϕ are said to be non-randomized if the respective functions $P[N \leqq n \mid \mathbf{x}]$ and $\phi(\mathbf{x})$ take on the values 0 and 1 only. A test (N, ϕ) will be called non-randomized if both N and ϕ are non-randomized.

We use the term distribution synonymously with probability measure. The set $\mathfrak{F}$ consists of distributions on a fixed σ-field $\mathfrak{A}$ of subsets of a space $\mathfrak{X}$. Unless we state otherwise, we shall assume that $\mathfrak{X}$ is a k-dimensional Euclidean space and $\mathfrak{A}$ the k-dimensional Borel field. A distribution on $\mathfrak{A}$ will then be called a k-dimensional or k-variate distribution. If F is a distribution on $\mathfrak{A}$, we denote by $F[A]$ the probability of the set $A \varepsilon \mathfrak{A}$ and by $F(x) = F(x^{(1)}, \cdots, x^{(k)})$, $x \varepsilon \mathfrak{X}$, the associated distribution function, that is, $F(x) = F[\{y \mid y^{(1)} \leqq x^{(1)}, \cdots, y^{(k)} \leqq x^{(k)}\}]$. With the usual definition (see [5])[4] of the distribution of a sequence $\mathbf{X} = (X_1, X_2, \cdots)$ of independent chance variables with identical marginal distribution F, we denote by $P_F[B]$ the probability of a measurable set B in the range of $\mathbf{X}$, and write $E_F\psi$ for the expected value of a function ψ of $\mathbf{X}$.

According to our definitions, the probability of an erroneous decision when test (N, ϕ) is used is equal to $E_F\phi$ if $F \varepsilon \mathcal{G}$, and to $E_F(1 - \phi)$ if $F \varepsilon \mathcal{H}$. Thus the sets $\mathcal{G}$ and $\mathcal{H}$ are distinguishable in a class $\mathfrak{J}$ of tests if and only if for every $\epsilon > 0$ there exists a test (N, ϕ) in $\mathfrak{J}$ such that $E_F\phi < \epsilon$ for $F \varepsilon \mathcal{G}$ and $E_F(1 - \phi) < \epsilon$ for all $F \varepsilon \mathcal{H}$.

2. Modes of distinguishability. We shall be concerned with the distinguishability of two sets of distributions in various classes $\mathfrak{J}$ of tests, which are defined in terms of properties of the distribution of the sample size function N. Some classes of particular interest are the following.

$\mathfrak{J}_0$: $P_F[N < \infty] = 1$ if $F \varepsilon \mathfrak{F}$

$\mathfrak{J}_1(r)$: $E_F N^r < \infty$ if $F \varepsilon \mathfrak{F}$ $(r > 0)$.

$\mathfrak{J}_1$: $E_F N^r < \infty$ for all $r > 0$ if $F \varepsilon \mathfrak{F}$.

$\mathfrak{J}_2$: $E_F e^{tN} < \infty$ for some $t = t(F) > 0$ if $F \varepsilon \mathfrak{F}$.

$\mathfrak{J}_3$: $\max(N) < \infty$.

[4] The numbers in square brackets refer to the bibliography listed at the end.

It will be noted that each of the successive classes contains the one following. Some classes of obvious interest have been omitted because, for the purposes of our investigation, they are equivalent to some of the classes listed above. Thus if two sets are distinguishable in one of the classes $\mathfrak{J}_0, \cdots, \mathfrak{J}_3$, they are also distinguishable in the corresponding subclass which contains only the non-randomized tests; this follows from Theorem 2.1 below. If two sets are distinguishable in $\mathfrak{J}_3$ (the class of "truncated" sequential tests), they are clearly distinguishable in the class of all fixed sample size tests; for if (N, ϕ) is any test in $\mathfrak{J}_3$, and we put $N' = \max(N), \phi' = E[\phi \mid \mathbf{x}]$, then (N', ϕ') is a fixed sample size test such that $E_F\phi' = E_F\phi$ for all F.

In view of the importance of the two extreme classes, $\mathfrak{J}_0$ and $\mathfrak{J}_3$, we shall use the following terms. If two sets of distributions are distinguishable (indistinguishable) in $\mathfrak{J}_0$, they will be called distinguishable ($\mathfrak{F}$)(indistinguishable ($\mathfrak{F}$)). If two sets are distinguishable (indistinguishable) in $\mathfrak{J}_3$, we shall say that they are finitely distinguishable (finitely indistinguishable).

The classes $\mathfrak{J}_i$ have been defined in terms of the set $\mathfrak{F}$ to which the distribution of X_j is assumed to belong (without displaying $\mathfrak{F}$ in the notation). It may be of interest to consider also the corresponding classes where $\mathfrak{F}$ is replaced by some subset of $\mathfrak{F}$ (compare Lemma 4.1 in section 4). It will be clear that Theorems 2.1 and 4.1 below can be immediately extended to such classes.

Our list does not contain the subclass of $\mathfrak{J}_1(r)$ where $E_F N^r$ is bounded for $F \varepsilon \mathfrak{F}$, nor the subclass of $\mathfrak{J}_0$ where $P_F[N > n] \to 0$ as $n \to \infty$, uniformly for $F \varepsilon \mathfrak{F}$. The reason for this omission is that two sets $\mathcal{G}$ and $\mathcal{H}$ which are distinguishable in one of these classes are finitely distinguishable. This follows from the following fact: If (N, ϕ) is a test such that $P_F[N > n] \to 0$ as $n \to \infty$, uniformly for $F \varepsilon \mathcal{G} \cup \mathcal{H}$, then for every $\epsilon > 0$ there exists a test (N', ϕ') such that $\max(N') < \infty$ and $| E_F\phi' - E_F\phi | < \epsilon$ for all $F \varepsilon \mathcal{G} \cup \mathcal{H}$. This is so since, by our assumption, we can choose an integer $n = n(\epsilon)$ such that $P_F[N > n] < 2\epsilon$ for all $F \varepsilon \mathcal{G} \cup \mathcal{H}$, and the test (N', ϕ') defined by

$$\phi' = \phi, \qquad N' = N \text{ if } N \leqq n; \qquad \phi' = \tfrac{1}{2}, \qquad N' = n \text{ if } N > n$$

has the stated property.

Let $\mathfrak{J}$ be any class of tests. If $\Phi = \Phi(\mathfrak{J})$ denotes the class of all terminal decision functions ϕ of the tests in $\mathfrak{J}$, the statement that $\mathcal{G}$ and $\mathcal{H}$ are distinguishable in $\mathfrak{J}$ can be expressed by the equation

$$\sup_{\phi \varepsilon \Phi} \inf_{G \varepsilon \mathcal{G}, H \varepsilon \mathcal{H}} (E_H \phi - E_G \phi) = 1. \tag{2.1}$$

Whenever $\mathfrak{J}$ contains a trivial test such that $\phi = \text{const}$, the left side of (2.1) is at least zero. Let us say that a test in $\mathfrak{J}$ is nontrivial for distinguishing between $\mathcal{G}$ and $\mathcal{H}$ if $\sup_{G \varepsilon \mathcal{G}} E_G\phi < \inf_{H \varepsilon \mathcal{H}} E_H\phi$. Thus the left side of (2.1) is positive if and only if $\mathfrak{J}$ contains a nontrivial test for distinguishing between $\mathcal{G}$ and $\mathcal{H}$. The following theorem shows that if $\mathfrak{J}$ is one of the classes $\mathfrak{J}_0, \cdots, \mathfrak{J}_3$ (or one of the "equivalent" classes mentioned above), then the existence in $\mathfrak{J}$ of a nontrivial

test for distinguishing between $\mathcal{G}$ and $\mathcal{H}$ is sufficient for $\mathcal{G}$ and $\mathcal{H}$ to be distinguishable in $\mathfrak{J}$, and even in the class $\mathfrak{J}'$ which consists of the non-randomized tests in $\mathfrak{J}$. The special case of the theorem where $\mathfrak{J}$ is the class of all non-randomized fixed sample size tests is contained in a lemma of Berger [1] (which is there attributed to Bernoulli).

We denote by Φ and Φ' the classes of the terminal decision functions of the tests in $\mathfrak{J}$ and $\mathfrak{J}'$, respectively.

THEOREM 2.1. *If $\mathfrak{J}$ is one of the classes $\mathfrak{J}_0, \cdots, \mathfrak{J}_3$, then*

$$\sup_{\phi \varepsilon \Phi} \inf_{G \varepsilon \mathcal{G}, H \varepsilon \mathcal{H}} (E_H \phi - E_G \phi) > 0 \tag{2.2}$$

implies

$$\sup_{\phi \varepsilon \Phi'} \inf_{G \varepsilon \mathcal{G}, H \varepsilon \mathcal{H}} (E_H \phi - E_G \phi) = 1 \tag{2.3}$$

Hence

$$\sup_{\phi \varepsilon \Phi} \inf_{G \varepsilon \mathcal{G}, H \varepsilon \mathcal{H}} (E_H \phi - E_G \phi) = 0 \text{ or } 1. \tag{2.4}$$

For the proof of Theorem 2.1 we require the following

LEMMA. *If $\mathfrak{J}$ is one of the classes $\mathfrak{J}_0, \cdots, \mathfrak{J}_3$, and (N, ϕ) is in $\mathfrak{J}$, then for every $\epsilon > 0$ there is a test (N', ϕ') in $\mathfrak{J}$ such that N' is non-randomized and $|\phi' - \phi| < \epsilon$.*

PROOF. Let N' be the least integer $n \geqq 1$ such that

$$P[N > n \mid \mathbf{x}] < \epsilon.$$

Define ϕ' by

$$\phi' = E[\phi \mid N \leqq n, \mathbf{x}] \text{ if } N' = n, \qquad n = 1, 2, \cdots.$$

Thus (N', ϕ') is a test, and N' is non-randomized.

We have for every $n \geqq 1$

$$\begin{aligned} P[N' > n] = P\{P[N > n \mid \mathbf{x}] \geqq \epsilon\} &\leqq \epsilon^{-1} E P[N > n \mid \mathbf{x}] \\ &= \epsilon^{-1} P[N > n]. \end{aligned}$$

Since for any increasing function h on the nonnegative integers

$$Eh(N) = h(0) + \sum_{n=0}^{\infty} [h(n+1) - h(n)] P[N > n],$$

it follows that if N satisfies the condition for any of the classes $\mathfrak{J}_0, \cdots, \mathfrak{J}_3$, so does N'. Hence (N', ϕ') is in $\mathfrak{J}$.

Now if $N' = n$, we have from the definition of ϕ'

$$\begin{aligned} \phi - \phi' &= P[N \leqq n \mid \mathbf{x}] E[\phi \mid N \leqq n, \mathbf{x}] + P[N > n \mid \mathbf{x}] E[\phi \mid N > n, \mathbf{x}] - \phi' \\ &= P[N > n \mid \mathbf{x}](E[\phi \mid N > n, \mathbf{x}] - \phi'). \end{aligned}$$

Thus $|\phi - \phi'| \leqq P[N > n \mid \mathbf{x}]$ if $N' = n$. But $N' = n$ implies $P[N > n \mid \mathbf{x}] < \epsilon$, for all n. This completes the proof of the lemma.

PROOF OF THEOREM 2.1. If condition (2.2) is satisfied, $\mathfrak{J}$ contains a test (N, ϕ) such that

$$\alpha = \sup_{G \varepsilon \mathcal{G}} E_G \phi < \inf_{H \varepsilon \mathcal{H}} E_H \phi = \beta.$$

By the preceding lemma we may and shall assume that N is non-randomized.

Let ϵ be any positive number. The theorem will be proved by showing that there is a non-randomized test (N', ϕ') in $\mathfrak{J}$ such that

$$\inf_{H \varepsilon \mathcal{H}} E_H \phi' - \sup_{G \varepsilon \mathcal{G}} E_G \phi' > 1 - \epsilon. \tag{2.5}$$

Choose a positive integer m which satisfies the inequality

$$\left(\frac{2}{\beta - \alpha}\right)^2 \frac{1}{m} < \frac{\epsilon}{2}.$$

Define the test (N', ϕ') as follows. First apply test (N, ϕ), and denote the resulting values of N and ϕ by N_1 and ϕ_1. Then apply the same test to a new independent sequence of observations and note the values N_2 and ϕ_2 of N and ϕ. Continue in this way until m independent sequences of observations have been taken. The total sample size is $N' = N_1 + \cdots + N_m$. Since N is non-randomized, so is N'. Now put

$$\bar{\phi} = \frac{1}{m} \sum_{i=1}^{m} \phi_i,$$

$$\phi' = \begin{cases} 1 & \text{if} \quad \bar{\phi} > \dfrac{\alpha + \beta}{2} \\ 0 & \text{if} \quad \bar{\phi} \leqq \dfrac{\alpha + \beta}{2}. \end{cases}$$

Thus (N', ϕ') is a non-randomized test.

The chance variables $\phi_1, \cdots, \phi_m$ are independent, and each has the same distribution as ϕ. Hence $E\bar{\phi} = E\,\phi$, and the variance of $\bar{\phi}$ is less than $1/m$.

If $G \varepsilon \mathcal{G}$, then $E_G\phi \leqq \alpha$, so that

$$\begin{aligned} E_G \phi' &= P_G\left[\bar{\phi} - E_G\bar{\phi} > \frac{\alpha + \beta}{2} - E_G\phi\right] \\ &\leqq P_G\left[\bar{\phi} - E_G\bar{\phi} > \frac{\beta - \alpha}{2}\right] \\ &\leqq \frac{1}{m}\left(\frac{2}{\beta - \alpha}\right)^2 \end{aligned}$$

by Chebyshev's inequality. Hence

$$\sup_{G \varepsilon \mathcal{G}} E_G \phi' < \frac{\epsilon}{2}.$$

In a similar way it is seen that

$$\sup_{H \varepsilon \mathcal{H}} E_H(1 - \phi') < \frac{\epsilon}{2},$$

so that inequality (2.5) is satisfied.

We now show that the test (N', ϕ') is in $\mathfrak{J}$. For $\mathfrak{J} = \mathfrak{J}_0$ and $\mathfrak{J} = \mathfrak{J}_3$ this is obvious. Since for $r > 0$,

$$(N')^r = \left(\sum_{i=1}^m N_i\right)^r \leqq (m \max_{i=1,\cdots,m} N_i)^r = m^r \max_{i=1,\cdots,m} (N_i^r) \leqq m^r \sum_{i=1}^m N_i^r$$

and each N_i has the same distribution as N, we have $E(N')^r < \infty$ whenever $EN^r < \infty$. This proves the statement for $\mathfrak{J} = \mathfrak{J}_1(r)$ and $\mathfrak{J} = \mathfrak{J}_1$.

Finally, if $Ee^{tN} < \infty$, where $t > 0$, put $t' = t/m$. Since $N_1, \cdots, N_m$ are independent and distributed as N, $Ee^{t'N'} = Ee^{tN} < \infty$.

Thus (N', ϕ') is in $\mathfrak{J}$ in every case. The proof is complete.

It should be noted that if $X_1, X_2, \cdots$ are not independent and identically distributed, the analog of Theorem 2.1 is not true in general.

3. Sufficient conditions for distinguishability. Let $\mathcal{K}$ be a set of distributions on $\mathfrak{A}$. A distance in $\mathcal{K}$ is a nonnegative function δ of the pairs (G, H) of distributions in $\mathcal{K}$ such that $\delta(G, G) = 0$, $\delta(G, H) = \delta(H, G)$, and $\delta(G, H) \leqq \delta(G, K) + \delta(H, K)$, for all G, H, and K in $\mathcal{K}$. (We do not require that $\delta(G, H) = 0$ imply $G = H$.) We write $\delta(G, \mathcal{H})$ for $\inf_{H \varepsilon \mathcal{H}} \delta(G, H)$, and $\delta(\mathcal{G}, \mathcal{H})$ for $\inf_{G \varepsilon \mathcal{G}} \delta(G, \mathcal{H})$.

Let F_n denote the empiric distribution of the first n members, $X_1, \cdots, X_n$ of a sequence of independent chance variables with the common distribution $F \varepsilon \mathfrak{F}$; that is, $nF_n[A]$ is the number of indices $i \leqq n$ for which $X_i \varepsilon A$. We assume throughout that the set $\mathcal{K}$ in which a distance δ is defined contains $\mathfrak{F}$ and all empiric distributions.

We shall say that a distance δ is *consistent* in $\mathfrak{F}$ if for every $\epsilon > 0$

$$\lim_{n \to \infty} P_F[\delta(F_n, F) > \epsilon] = 0 \tag{3.1}$$

whenever $F \varepsilon \mathfrak{F}$. The distance δ will be called *uniformly consistent* in $\mathfrak{F}$ if the convergence in (3.1) is uniform for $F \varepsilon \mathfrak{F}$.

In this section we derive sufficient conditions for distinguishability in terms of uniformly consistent distances. We first mention a few examples of such distances. If $\mathfrak{F}$ is the set of all distributions on the k-dimensional Borel field $\mathfrak{A}$, and $\mathfrak{X}$ denotes the k-dimensional Euclidean space, the distance

$$D(G, H) = \sup_{x \varepsilon \mathfrak{X}} | G(x) - H(x) | \tag{3.2}$$

is known to be uniformly consistent in $\mathfrak{F}$ (see, for example, [4]). So is the distance

$$\left(\int_{\mathfrak{X}} | G(x) - H(x) |^r \, dK\right)^{1/r},$$

where $r \geqq 1$ and K is a fixed distribution on $\mathfrak{A}$, since it is bounded above by $D(G, H)$. A further example of a uniformly consistent distance is

$$D_\omega(G, H) = D(G_\omega, H_\omega), \tag{3.3}$$

where G_ω and H_ω are the distributions, according to G and H, of a fixed, real- or vector-valued measurable function ω on $\mathfrak{X}$. If $\mu(F)$ denotes the mean of a one-dimensional distribution F, the distance $|\mu(G) - \mu(H)|$ is uniformly consistent in any class of distributions with bounded variances.

A sufficient condition for finite distinguishability is the following. If the distance δ is uniformly consistent in $\mathcal{G} \cup \mathcal{H}$ and

$$\delta(\mathcal{G}, \mathcal{H}) > 0, \tag{3.4}$$

then the sets $\mathcal{G}$ and $\mathcal{H}$ are finitely distinguishable.

This can be seen by using the test with $N = n$ fixed and $\phi = 1$ or 0 according as $\delta(F_n, \mathcal{G}) - \delta(F_n, \mathcal{H}) \geqq 0$ or < 0. If $F \varepsilon \mathcal{G}$, then $\delta(F_n, \mathcal{G}) \leqq \delta(F_n, F)$ and $\delta(F_n, \mathcal{H}) \geqq \delta(F, \mathcal{H}) - \delta(F_n, F) \geqq \delta(\mathcal{G}, \mathcal{H}) - \delta(F_n, F)$. Hence $E_F\phi$ does not exceed

$$\sup_{F \varepsilon \mathcal{G} \cup \mathcal{H}} P_F[\delta(F_n, F) \geqq \tfrac{1}{2}\delta(\mathcal{G}, \mathcal{H})].$$

We obtain the same upper bound for $E_F(1 - \phi)$, $F \varepsilon \mathcal{H}$. Our assumptions imply that the bound tends to 0 as $n \to \infty$.

In the proof of the next theorem we shall make use of a test defined as follows. Let δ be a distance, $\{c_i\}$, $i = 1, 2, \cdots$, a sequence of positive numbers, and $\{n_i\}$, $i = 1, 2, \cdots$, an increasing sequence of positive integers. Put

$$\delta_i = \max\,[\delta(F_{n_i}, \mathcal{G}), \delta(F_{n_i}, \mathcal{H})].$$

Take successive independent samples of sizes n_1, $n_2 - n_1$, $n_3 - n_2$, $\cdots$. Continue sampling as long as $\delta_i < c_i$. Stop sampling as soon as $\delta_i \geqq c_i$, and apply the terminal decision function

$$\phi = \begin{cases} 1 \text{ if } \delta(F_{n_i}, \mathcal{G}) \geqq \delta(F_{n_i}, \mathcal{H}) \\ 0 \text{ if } \delta(F_{n_i}, \mathcal{G}) < \delta(F_{n_i}, \mathcal{H}). \end{cases}$$

Thus $N = n_i$, where i is the least integer for which $\delta_i \geqq c_i$.

We shall refer to this test as the test $T(\delta, \{c_i\}, \{n_i\})$.

THEOREM 3.1. (a) *If the distance δ is uniformly consistent in $\mathfrak{F}$, then any two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{F}$ for which*

$$\max\,[\delta(F, \mathcal{G}), \delta(F, \mathcal{H})] > 0 \textit{ if } F \varepsilon \mathfrak{F} \tag{3.6}$$

are distinguishable ($\mathfrak{F}$).

(b) *If, for every $c > 0$, there exist two positive numbers $A(c)$ and $B(c)$ such that for all integers $n > 0$ and all $F \varepsilon \mathfrak{F}$*

$$P_F[\delta(F_n, F) \geqq c] \leqq A(c)e^{-B(c)n}, \tag{3.7}$$

then any two subsets $\mathcal{G}$ *and* $\mathcal{H}$ *of* $\mathcal{F}$ *which satisfy* (3.6) *are distinguishable in the class of tests* (N, ϕ) *such that* $E_F e^{tN} < \infty$ *for some* $t = t(F) > 0$ if $F \varepsilon \mathcal{F}$.

PROOF. Let α be a positive number. Part (a) will be proved by showing that the sequences $\{c_i\}$ and $\{n_i\}$ can be so chosen that the test $(N, \phi) = T(\delta, \{c_i\}, \{n_i\})$ satisfies the conditions

$$E_F\phi \leqq \alpha \text{ if } F \varepsilon \mathcal{G}, \qquad E_F(1 - \phi) \leqq \alpha \text{ if } F \varepsilon \mathcal{H} \tag{3.8}$$

and

$$P_F[N < \infty] = 1 \text{ if } F \varepsilon \mathcal{F}. \tag{3.9}$$

Let $\{c_i\}$ be a sequence of positive numbers such that

$$\lim_{i\to\infty} c_i = 0. \tag{3.10}$$

Choose the positive numbers $\alpha_1, \alpha_2, \cdots$ so that

$$\sum_{i=1}^{\infty} \alpha_i \leqq \alpha. \tag{3.11}$$

Since δ is uniformly consistent in $\mathcal{F}$, we can choose the integers $n_1 < n_2 < \cdots$ in such a way that

$$P_F[\delta(F_{n_i}, F) \geqq c_i] \leqq \alpha_i, \qquad i = 1, 2, \cdots \tag{3.12}$$

for all $F \varepsilon \mathcal{F}$.

If $F \varepsilon \mathcal{G}$,

$$\begin{aligned} E_F\phi &= \sum_{j=1}^{\infty} P_F[\delta_i < c_i \quad \text{for} \quad i < j, \delta_j \geqq c_j, \delta(F_{n_j}, \mathcal{G}) \geqq \delta(F_{n_j}, \mathcal{H})] \\ &\leqq \sum_{j=1}^{\infty} P_F[\delta(F_{n_j}, \mathcal{G}) \geqq c_j] \\ &\leqq \sum_{j=1}^{\infty} P_F[\delta(F_{n_j}, F) \geqq c_j]. \end{aligned}$$

It now follows from (3.12) and (3.11) that $E_F\phi \leqq \alpha$ if $F \varepsilon \mathcal{G}$. In a similar way it is seen that $E_F(1 - \phi) \leqq \alpha$ if $F \varepsilon \mathcal{H}$. Thus the conditions (3.8) are satisfied.

The terminal sample size N takes on the values $n_1, n_2 \cdots$, and we have

$$P_F[N > n_j] = P_F[\delta_i < c_i, i = 1, \cdots, j] \leqq P_F[\delta_j < c_j].$$

By the triangle inequality,

$$\delta_j \geqq \delta^* - \delta(F_{n_j}, F),$$

where

$$\delta^* = \max\,[\delta(F, \mathcal{G}), \delta(F, \mathcal{H})].$$

By assumption, $\delta^* > 0$ for all $F \varepsilon \mathfrak{F}$.

Hence if $F \varepsilon \mathfrak{F}$,

$$P_F[N > n_j] \leqq P_F[\delta(F_{n_j}, F) > \delta^* - c_j]. \tag{3.13}$$

Since $c_j \to 0$, we have $\delta^* - c_j > c_j$ for j sufficiently large, and then the right side of (3.13) is $\leqq \alpha_j$. By (3.11), $\alpha_j \to 0$ as $j \to \infty$. Thus $P_F[N > n_j] \to 0$ as $j \to \infty$, which implies (3.9). This completes the proof of part (a).

Now suppose that the assumption of part (b) is satisfied. The sequences $\{c_i\}$ and $\{n_i\}$ can be so chosen that, in addition to $\lim c_i = 0$ and $n_i < n_{i+1}$,

$$\liminf_{i \to \infty} i^{-1}(2n_i - n_{i+1}) > 0 \tag{3.14}$$

and

$$\sum_{i=1}^{\infty} A(c_i) e^{-B(c_i)n_i} \leqq \alpha. \tag{3.15}$$

(For instance, put $M(c) = \max [A(c), 1/B(c)]$; choose $c_1, c_2, \cdots$ so that $c_i > 0$, $\lim c_i = 0$ and $M(c_i) \leqq mi^{1/2}$, $i = 1, 2, \cdots$, with a suitable number $m > 0$; and put $n_i = ni$, where n is so large that

$$\sum_{i=1}^{\infty} mi^{1/2} e^{-nm^{-1}i^{1/2}} \leqq \alpha).$$

The inequalities (3.7) and (3.15) imply that conditions (3.11) and (3.12) are fulfilled. Hence the conditions (3.8) are satisfied.

For a fixed $F \varepsilon \mathfrak{F}$, choose the integer h so that $c_i \leqq \delta^*/2$ for $i > h$. Then for $i > h$, due to (3.13) and (3.7),

$$P_F[N > n_i] \leqq P_F[\delta(F_{n_i}, F) > \delta^*/2] \leqq ae^{-bn_i},$$

where $a = A(\delta^*/2)$ and $b = B(\delta^*/2)$ are positive numbers.

Now for any real t,

$$\begin{aligned} E_F e^{tN} &= \sum_{j=1}^{\infty} e^{tn_j} P_F[N = n_j] \\ &\leqq e^{tn_1} + \sum_{i=1}^{\infty} e^{tn_{i+1}} P_F[N > n_i]. \end{aligned}$$

Thus $E_F e^{tN} < \infty$ if the series

$$\sum_i e^{tn_{i+1} - bn_i}$$

converges. If $t \leqq b/2$,

$$tn_{i+1} - bn_i \leqq -\frac{b}{2}(2n_i - n_{i+1}),$$

so that the series converges due to (3.14). The proof is complete.

The assumption of Theorem 3.1, part (b) is satisfied if $\mathfrak{F}$ is any set of k-dimen-

sional distributions ($k \geqq 1$) and $\delta = D$, the distance defined by (3.2). This is implied by the following theorem due to Kiefer and one of the authors [4]: For every integer $k \geqq 1$ there exist two positive numbers a and b such that for all $c > 0$, all integers $n > 0$, and all k-dimensional distributions F

$$P_F[D(F_n, F) \geqq c] \leqq ae^{-bc^2 n}. \tag{3.16}$$

(For $k = 1$ the inequality (3.16), with $b = 2$, was proved by Dvoretzky, Kiefer and one of the authors [2].) Hence we can state the following corollary.

COROLLARY 3.1. *If $\mathfrak{F}$ is any set of k-dimensional distributions ($k \geqq 1$), then any two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{F}$ for which*

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

are distinguishable in the class of tests (N, ϕ) such that $E_F e^{tN} < \infty$ for some $t = t(F) > 0$ if $F \varepsilon \mathfrak{F}$.

4. Necessary conditions for distinguishability. Let P and Q be two distributions on a σ-field $\mathcal{B}$ of subsets of an arbitrary space $\mathcal{Y}$, and let Ψ be the class of all measurable functions on $\mathcal{Y}$ with values ranging from 0 to 1. We denote by d the distance defined by

$$d(P, Q) = \sup_{\psi \varepsilon \Psi} | E_P \psi - E_Q \psi |. \tag{4.1}$$

We note some alternative expressions for d. Let ν be any σ-finite measure with respect to which P and Q are absolutely continuous (for instance, $\nu = P + Q$), and denote by p and q densities (Radon-Nikodym derivatives) of P and Q with respect to ν. Then

$$d(P, Q) = \int_{\{p>q\}} (p - q)\, d\nu = \frac{1}{2} \int | p - q |\, d\nu = 1 - \int \min (p, q)\, d\nu. \tag{4.2}$$

(Here and in what follows, an integral whose domain of integration is not indicated is extended over the entire space.) Also

$$d(P, Q) = \sup_{B \varepsilon \mathcal{B}} | P[B] - Q[B] |. \tag{4.3}$$

For any distribution G on $\mathcal{A}$ we denote by $G^{(n)}$ the distribution of n independent chance variables each of which has the distribution G. We write $\mathcal{G}^{(n)}$ for the set of all $G^{(n)}$ such that $G \varepsilon \mathcal{G}$.

It is easily seen from (4.1) that

$$d(G, H) \leqq d(G^{(n)}, H^{(n)}) \leqq d(G^{(n+1)}, H^{(n+1)}), \qquad n = 1, 2, \cdots \tag{4.4}$$

and from the last expression in (4.2), using the inequality $\min (ab, cd) \geqq \min (a, c) \min (b, d)$, where a, b, c, d are all positive, that

$$d(G^{(n)}, H^{(n)}) \leqq 1 - (1 - d(G, H))^n \leqq n\, d(G, H). \tag{4.5}$$

(See also Kruskal [6], p. 29.)

The convex hull, $C\mathcal{P}$, of a set $\mathcal{P}$ of distributions on a common σ-field is defined

as the set of all distributions $\lambda_1 P_1 + \cdots + \lambda_r P_r$, where r is any positive integer, $P_1, \cdots, P_r$ are in $\mathcal{P}$, and $\lambda_1, \cdots, \lambda_r$ are positive numbers whose sum is 1.

In order that two sets $\mathcal{G}$ and $\mathcal{H}$ be finitely distinguishable it is necessary that

$$d(C\mathcal{G}^{(n)}, C\mathcal{H}^{(n)}) > 0 \tag{4.6}$$

for some n or, equivalently,

$$\lim_{n\to\infty} d(C\mathcal{G}^{(n)}, C\mathcal{H}^{(n)}) = 1. \tag{4.7}$$

This is known and follows easily from the definition (4.1) and Theorem 2.1.

If the set $\mathcal{G} \cup \mathcal{H}$ is dominated, that is to say, if the distributions in $\mathcal{G} \cup \mathcal{H}$ are absolutely continuous with respect to a fixed σ-finite measure, then condition (4.7) is also sufficient for $\mathcal{G}$ and $\mathcal{H}$ to be finitely distinguishable. This is contained in Theorem 6 of Kraft [7] and follows from a theorem of LeCam (Theorem 5 of Kraft [7]) which is equivalent to the statement that if the set $\mathcal{P}_1 \cup \mathcal{P}_2$ is dominated, then

$$\max_{\phi\varepsilon\Phi} \inf_{P_1\varepsilon\mathcal{P}_1, P_2\varepsilon\mathcal{P}_2} (E_{P_2}\phi - E_{P_1}\phi) = d(C\mathcal{P}_1, C\mathcal{P}_2), \tag{4.8}$$

where Φ denotes the set of all measurable functions ϕ such that $0 \leqq \phi \leqq 1$.

If condition (4.6) is satisfied, then

$$d(\mathcal{G}, \mathcal{H}) > 0. \tag{4.9}$$

In fact, $d(C\mathcal{G}^{(n)}, C\mathcal{H}^{(n)}) \leqq d(\mathcal{G}^{(n)}, \mathcal{H}^{(n)}) \leqq n\, d(\mathcal{G}, \mathcal{H})$ by (4.5). This weaker but much simpler necessary condition for finite distinguishability will be shown in section 5 to be also sufficient under certain assumptions.

To obtain necessary conditions for non-finite distinguishability we first prove the following lemma.

LEMMA 4.1. *If*

$$d(F_0^{(n)}, C\mathcal{G}^{(n)}) = d(F_0^{(n)}, C\mathcal{H}^{(n)}) = 0 \tag{4.10}$$

for all n, then the sets $\mathcal{G}$ and $\mathcal{H}$ are indistinguishable in the class of tests (N, ϕ) with $P_{F_0}[N < \infty] = 1$.

PROOF. Let (N, ϕ) be any test such that $P_{F_0}[N < \infty] = 1$. Define $\phi_n = \phi$ if $N \leqq n$, $\phi_n = 0$ if $N > n$. Thus ϕ_n is a function of $x_1, \cdots, x_n$ only, and $\phi_n \leqq \phi$. Let K be a member of $C\mathcal{G}^{(n)}$, so that $K = \lambda_1 G_r^{(n)} + \cdots + \lambda_r G_r^{(n)}$, $G_i \varepsilon \mathcal{G}$, $\lambda_i > 0$, $\Sigma\lambda_i = 1$. Then[5]

$$E_K \phi_n = \Sigma\lambda_i E_{G_i}\phi_n \leqq \Sigma\lambda_i E_{G_i}\phi \leqq \sup_{G\varepsilon\mathcal{G}} E_G\phi.$$

Hence

$$E_{F_0}\phi_n - \sup_{G\varepsilon\mathcal{G}} E_G\phi \leqq E_{F_0}\phi_n - E_K\phi_n \leqq d(F_0^{(n)}, K)$$

[5] Here $E_K\phi_n$ denotes the expected value of ϕ_n when the joint distribution of $(X_1, \cdots, X_n)$ is K. We keep the notation $E_G\phi_n$ when $X_1, \cdots, X_n$ are independent and each X_i has the distribution G.

for all $K \varepsilon C\mathcal{G}^{(n)}$. Therefore

$$E_{F_0}\phi_n - \sup_{G\varepsilon\mathcal{G}} E_G\phi \leqq d(F_0^{(n)}, C\mathcal{G}^{(n)}) = 0.$$

Since $P_{F_0}(N > n) \to 0$ as $n \to \infty$, $E_{F_0}\phi_n$ converges to $E_{F_0}\phi$. Hence

$$E_{F_0}\phi \leqq \sup_{G\varepsilon\mathcal{G}} E_G\phi. \tag{4.11}$$

In a similar way, if we use, instead of ϕ_n, the function $\phi_n' = \phi$ if $N \leqq n$, $\phi_n' = 1$ if $N > n$, we find that

$$E_{F_0}\phi \geqq \inf_{H\varepsilon\mathcal{H}} E_H\phi. \tag{4.12}$$

Inequalities (4.11) and (4.12) imply the Lemma.

THEOREM 4.1. *In order that the sets $\mathcal{G}$ and $\mathcal{H}$ be distinguishable* ($\mathfrak{F}$) *it is necessary that*

$$\max\,[d(F^{(n)}, C\mathcal{G}^{(n)}), d(F^{(n)}, C\mathcal{H}^{(n)})] > 0 \tag{4.13}$$

for some n if $F \varepsilon \mathfrak{F}$ and hence that

$$\max\,[d(F, \mathcal{G}), d(F, \mathcal{H})] > 0 \text{ if } F \varepsilon \mathfrak{F}. \tag{4.14}$$

PROOF. The necessity of (4.14) follows immediately from Lemma 4.1. That (4.13) implies (4.14) follows from inequality (4.5).

That the condition (4.13) can be violated when inequality (4.14) is satisfied can be seen from an example given by Kraft ([7], p. 132) to show the non-equivalence of two conditions equivalent to (4.6) and (4.9). Nevertheless the simple necessary condition (4.14) is also sufficient under certain restrictions on the set of distributions, as will be seen in section 5.

We conclude this section by showing that a known necessary condition for distinguishability is implied by condition (4.14) of Theorem 4.1.

For any two distributions F and G on $\mathcal{A}$ and any set $\mathcal{G}$ of distributions on $\mathcal{A}$ define

$$\tau(F, G) = \int f \log (f/g)\, d\nu, \qquad \tau(F, \mathcal{G}) = \inf_{G\varepsilon\mathcal{G}} \tau(F, G),$$

where ν denotes a σ-finite measure with respect to which F and G are absolutely continuous, with densities f and g. Note that $0 \leqq \tau\,(F, G) \leqq \infty$. It has been shown in [3] that if $\tau(F, \mathcal{G}) = 0$, then F and $\mathcal{G}$ are indistinguishable in the class of tests with $E_F N < \infty$. Now

$$-\tfrac{1}{2}\tau(F, G) = \int f \log (g/f)^{1/2}\, d\nu$$

$$\leqq \log \int f(g/f)^{1/2}\, d\nu = \log \int (fg)^{1/2}\, d\nu$$

and (see Kraft [7], Lemma 1)

$$d^2(F, G) \leqq 1 - \left\{\int (fg)^{1/2}\, d\nu\right\}^2.$$

Hence $\tau(F, \mathcal{G}) = 0$ implies $d(F, \mathcal{G}) = 0$. Thus, by Lemma 4.1 (with $F_0 = F$ and $\mathcal{H}$ consisting only of F) F and $\mathcal{G}$ are even indistinguishable in the class of tests with $P_F[N < \infty] = 1$. It is easy to construct examples where $d(F, \mathcal{G}) = 0$ and $\tau(F, \mathcal{G}) > 0$, so that condition (4.14) is actually better than the corresponding condition with d replaced by τ.

5. Necessary and sufficient conditions for distinguishability. In this section we shall show that the necessary conditions of section 4 are also sufficient for distinguishability under certain restrictions on the sets of distributions. Most of our results will be such that if the necessary condition is satisfied, the sets are not only distinguishable ($\mathfrak{F}$), but even distinguishable in a stronger sense.

If $\mathcal{G}$ consists of a single distribution G, then, by Theorem 4.1, $\mathcal{G}$ and $\mathcal{H}$ are distinguishable ($\mathcal{G} \cup \mathcal{H}$) only if $d(G^{(n)}, C\mathcal{H}^{(n)}) > 0$ for some n. If $\mathcal{H}$ is dominated, this condition is sufficient for $\mathcal{G}$ and $\mathcal{H}$ to be finitely distinguishable, by the Le Cam-Kraft theorem mentioned in section 4. More generally, we can state the following.

If $\mathcal{G}$ is finite and $\mathcal{H}$ is dominated, then $\mathcal{G}$ and $\mathcal{H}$ are either finitely distinguishable or are indistinguishable ($\mathcal{G} \cup \mathcal{H}$), depending on whether the condition

$$\max\,[d(F^{(n)}, C\mathcal{G}^{(n)}), d(F^{(n)}, C\mathcal{H}^{(n)})] > 0 \tag{5.1}$$

for some n if $F \varepsilon \mathcal{G} \cup \mathcal{H}$ is or is not satisfied. Condition (5.1) is equivalent to

$$d(G^{(n)}, C\mathcal{H}^{(n)}) > 0 \tag{5.2}$$

for some n if $G \varepsilon \mathcal{G}$.

That condition (5.1) is necessary for $\mathcal{G}$ and $\mathcal{H}$ to be distinguishable ($\mathcal{G} \cup \mathcal{H}$) follows from Theorem 4.1. On the other hand, if (5.1) is satisfied, so is (5.2). Hence if the distributions in $\mathcal{G}$ are denoted by $G_1, \cdots, G_r$, then, by Le Cam's theorem, G_i and $\mathcal{H}$ are finitely distinguishable, for each i. Thus, given $\epsilon > 0$, there exists an integer n and tests (n, ϕ_i) such that $E_{G_i}\phi_i < \epsilon$ and

$$\sup_{H \varepsilon \mathcal{H}} E_H(1 - \phi_i) < \epsilon, i = 1, \cdots, r.$$

Put $\phi = \phi_1\phi_2 \cdots \phi_r$. Then $\phi \leqq \phi_i$ and $1 - \phi \leqq \sum_{i=1}^{r} (1 - \phi_i)$. Hence $E_{G_i}\phi < \epsilon$ for all i and $E_H(1 - \phi) < r\epsilon$ if $H \varepsilon \mathcal{H}$. Therefore $\mathcal{G}$ and $\mathcal{H}$ are finitely distinguishable, and condition (5.2) is equivalent to (5.1).

If both $\mathcal{G}$ and $\mathcal{H}$ are countably infinite sets, it is no longer true that $\mathcal{G}$ and $\mathcal{H}$ are either finitely distinguishable or indistinguishable. To see this, let $\mathcal{G} = \{G_i\}$ and $\mathcal{H} = \{H_i\}$, $i = 1, 2, \cdots$, where G_i and H_i are univariate normal distributions with respective means a and b ($a \neq b$) and common variance σ_i^2, such that $\lim \sigma_i^2 = \infty$. It follows from a result of Stein [8] (or from Theorem 3.1) that $\mathcal{G}$ and $\mathcal{H}$ are distinguishable ($\mathfrak{N}$), where $\mathfrak{N}$ denotes the class of univariate normal distributions. But one readily verifies that $\lim_i d(G_i, H_i) = 0$, so that the sets are finitely indistinguishable.

In what follows it will be shown that the simple condition

$$\max\,[d(F, \mathcal{G}), d(F, \mathcal{H})] > 0 \text{ if } F \varepsilon \mathfrak{F} \tag{5.3}$$

which, by Theorem 4.1, is necessary for $\mathcal{G}$ and $\mathcal{H}$ to be distinguishable ($\mathfrak{F}$), is also

sufficient under rather general assumptions. Under somewhat more stringent assumptions the necessary condition $d(\mathcal{G}, \mathcal{H}) > 0$ for finite distinguishability (see (4.9)) will also be shown to be sufficient.

A comparison of the results of sections 3 and 4 shows that if δ is any uniformly consistent distance in a set $\mathfrak{F}$, then $d(\mathcal{G}, \mathcal{H}) = 0$ implies $\delta(\mathcal{G}, \mathcal{H}) = 0$ whenever $\mathcal{G} \subset \mathfrak{F}$ and $\mathcal{H} \subset \mathfrak{F}$. Theorems 3.1 and 4.1 also show that if the set $\mathfrak{F}$ has the property that there exists a uniformly consistent δ such that, for all $F \varepsilon \mathfrak{F}$ and all $\mathcal{G} \subset \mathfrak{F}$, $\delta(F, \mathcal{G}) = 0$ implies (and hence is equivalent to) $d(F, \mathcal{G}) = 0$, then the necessary condition (4.14) is also sufficient for two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{F}$ to be distinguishable ($\mathfrak{F}$). Similarly, if for all $\mathcal{G} \subset \mathfrak{F}$ and all $\mathcal{H} \subset \mathfrak{F}$, $\delta(\mathcal{G}, \mathcal{H}) = 0$ implies $d(\mathcal{G}, \mathcal{H}) = 0$, then any two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{F}$ are finitely distinguishable if and only if $d(\mathcal{G}, \mathcal{H}) > 0$.

We first consider conditions which ensure that $D(F, \mathcal{G}) = 0$ implies $d(F, \mathcal{G}) = 0$. Let F and G be two k-dimensional distributions and ϵ a nonnegative number. Suppose that there is an integer J with the following property. There exist J non-overlapping k-dimensional intervals $I_1, \cdots, I_j$ such that (i) $F - G$ is monotone[6] in each I_j, and (ii) if V denotes the complement of $\bigcup_{j=1}^{J} I_j$, then min $(F[V], G[V]) \leqq \epsilon$. Write $J(F, G; \epsilon)$ for the least integer J having this property. If such a finite J does not exist, define $J(F, G; \epsilon) = \infty$.

Note that if $F - G$ is monotone in a set C, the difference of the densities, $f - g$, is of constant sign in C except in a subset of probability 0 according to both F and G.

LEMMA 5.1. *If F and G are two k-dimensional distributions,*

$$d(F, G) \leqq 2^k J(F, G; \epsilon)\, D(F, G) + \epsilon. \tag{5.4}$$

PROOF. We may assume that $J = J(F, G; \epsilon)$ is finite. Then there exist J non-overlapping intervals $I_1, \cdots, I_J$ which satisfy the conditions (i) and (ii). We have

$$2d(F, G) = \sum_{j=1}^{J} \int_{I_j} |f - g|\, d\nu + \int_V |f - g|\, d\nu.$$

Now

$$\int_V |f - g|\, d\nu \leqq \int_V (f + g)\, d\nu = 2 \int_V f d\nu + \int_V (g - f)\, d\nu$$

$$= 2F[V] + \sum_{j=1}^{J} \int_{I_j} (f - g)\, d\nu \leqq 2F[V] + \sum_{j=1}^{J} \left| \int_{I_j} (f - g)\, d\nu \right|.$$

Also,

$$\int_{I_j} |f - g|\, d\nu = \left| \int_{I_j} (f - g)\, d\nu \right| \leqq 2^k D\,(F, G).$$

[6] An additive function L on $\mathcal{R}$ is monotone in a set C is either $L \leqq [A]\ L\ [B]$ whenever $A \subset B \subset C$ or $L\ [A] \geqq L\ [B]$ whenever $A \subset B \subset C$.

Hence

$$2d(F, G) \leqq 2J \cdot 2^k D(F, G) + 2F[V].$$

By symmetry, the term $2F[V]$ can be replaced by $2G[V]$, and hence also by 2ϵ. This implies (5.4).

THEOREM 5.1. *Let $\mathfrak{F}$ be a set of k-dimensional distributions, $k \geqq 1$.* (a) *If*

$$\sup_{G \varepsilon \mathfrak{F}} J(F, G; \epsilon) < \infty \tag{5.5}$$

for all $F \varepsilon \mathfrak{F}$ and all $\epsilon > 0$, then two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{F}$ are distinguishable ($\mathfrak{F}$) *if and only if*

$$\max [d(F, \mathcal{G}), d(F, \mathcal{H})] > 0 \tag{5.6}$$

for all $F \varepsilon \mathfrak{F}$. Moreover, if condition (5.6) *is satisfied, then $\mathcal{G}$ and $\mathcal{H}$ are distinguishable in the class of tests (N, ϕ) such that $E_F e^{tN} < \infty$ for some $t = t(F) > 0$ if $F \varepsilon \mathfrak{F}$.* (b) *If*

$$\sup_{F \varepsilon \mathfrak{F}, G \varepsilon \mathfrak{F}} J(F, G; \epsilon) < \infty \tag{5.7}$$

for all $\epsilon > 0$, then two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{F}$ are finitely distinguishable if and only if

$$d(\mathcal{G}, \mathcal{H}) > 0. \tag{5.8}$$

PROOF. The necessity of conditions (5.6) and (5.8) has been proved in section 4. If condition (5.5) is satisfied, then, by Lemma 5.1, $D(F, \mathcal{G}) = 0$ implies $d(F, \mathcal{G}) = 0$ for all $F \varepsilon \mathfrak{F}$ and all $\mathcal{G} \subset \mathfrak{F}$. Hence if (5.6) is satisfied, the assumption of Corollary 3.1 is fulfilled, which implies part (a). The proof of part (b) is similar, referring to (3.4) with $\delta = D$.

The assumption of Theorem 5.1, part (b) (and hence that of part (a)) is satisfied for most parametric sets of *univariate* distributions which are commonly used as models in statistics. In such sets $\mathfrak{F}$ the minimum number of intervals in which $f - g$ is of constant sign is usually bounded, and then even $\sup_{F \varepsilon \mathfrak{F}, G \varepsilon \mathfrak{F}} J(F, G; 0)$ is finite. For example, if F and G are any two univariate normal distributions, then $J(F, G; 0) \leqq 3$. This is also true if the singular normal distributions (with zero variance) are included.

The assumption of part (a) is satisfied if $\mathfrak{F}$ is any subclass of the class of all distributions on the subsets of a fixed countable set S. Since the points of S can be arranged in a sequence, we may assume that S is the set of the positive integers. If $F \varepsilon \mathfrak{F}$ and $\epsilon > 0$, choose the integer M so that $F[x > M] < \epsilon$. Since we can choose M intervals each of which contains exactly one positive integer $\leqq M$, we have $J(F, G; \epsilon) \leqq M$ for all $G \varepsilon \mathfrak{F}$, so that condition (5.5) is satisfied.

Actual statistical observations are either integer-valued or integer multiples of a fixed unit of measurement. In this sense it can be said that the assumption of part (a) is satisfied for all classes of distributions which actually occur in statistics.

If $\mathcal{G}$ and $\mathcal{H}$ are two arbitrary sets of distributions over a fixed countable set,

then $\mathcal{G}$ and $\mathcal{H}$ can be *finitely* indistinguishable even when $d(\mathcal{G}, \mathcal{H}) > 0$. This is shown by the following example. For $r = 1, 2, \cdots$ and $k = 1, \cdots, r$ define the sets

$$A_r = \{i2^{-r} \mid i = 1, 2, \cdots, 2^r\},$$

$$A_{r,k} = \{(j2^{r-k+1} + i)2^{-r} \mid i = 1, 2, \cdots, 2^{r-k}; j = 0, 1, \cdots, 2^{k-1} - 1\}.$$

Let G_r and $H_{r,k}$ be the discrete distributions whose elementary probability functions are

$$g_r(x) = 2^{-r}\chi(x; A_r), \qquad h_{r,k}(x) = 2^{-r+1}\chi(x; A_{r,k}),$$

where $\chi(x; A) = 1$ or 0 according as $x \varepsilon A$ or $x \not\varepsilon A$. Let $\mathcal{G} = \{G_r\}, r = 1, 2, \cdots$, and $\mathcal{H} = \{H_{r,k}\}, k = 1, \cdots, r; r = 1, 2, \cdots$. The reader can verify that

$$d(G_r, H_{s,k}) \geqq \tfrac{1}{2}$$

for all r, s, and k, so that $d(\mathcal{G}, \mathcal{H}) > 0$.

Now denote by $G_r^{(n)}$ and $H_{r,k}^{(n)}$ the distributions of n independent chance variables each of which has the distribution G_r and $H_{r,k}$, respectively, and by $g_r^{(n)}$ and $h_{r,k}^{(n)}$ their elementary probability functions. Let $H_r^{(n)}$ denote the distribution in $C\mathcal{H}^{(n)}$ whose elementary probability function is

$$h_r^{(n)} = r^{-1} \sum_{k=1}^{r} h_{r,k}^{(n)}.$$

Writing $g_r^{(n)}$, g_r, etc. for the chance variables $g_r^{(n)}(X_1, \cdots, X_n)$, $g_r(X_1)$, etc., and E for the expected value when the distribution of X_i is G_r, we have

$$2\, d(G_r^{(n)}, H_r^{(n)}) = E \mid (h_r^{(n)}/g_r^{(n)}) - 1 \mid \leqq (E[(h_r^{(n)}/g_r^{(n)}) - 1]^2)^{1/2}$$
$$= (E(h_r^{(n)}/g_r^{(n)})^2 - 1)^{1/2}$$

We calculate

$$E(h_r^{(n)}/g_r^{(n)})^2 = r^{-2} \sum_{j=1}^{r} \sum_{k=1}^{r} (E(h_{rj}\, h_{rk}/g_r^2))^n = 1 + (2^n - 1)r^{-1}.$$

It follows that $\lim_{r\to\infty} d(G_r^{(n)}, H_r^{(n)}) = 0$ for every n. Therefore $d(\mathcal{G}^{(n)}, C\mathcal{H}^{(n)}) = 0$ for all n, so that the sets $\mathcal{G}$ and $\mathcal{H}$ are finitely indistinguishable. Note, however, that since $d(\mathcal{G}, \mathcal{H}) > 0$, the sets are distinguishable in the sense of part (a) of Theorem 5.1 with $\mathcal{F} = \mathcal{G} \cup \mathcal{H}$ and, more generally, with $\mathcal{F}$ denoting any class of distributions on the subsets of $\bigcup_{r=1}^{\infty} A_r$ such that condition (5.6) is satisfied.

We shall see that all conclusions of Theorem 5.1 are true also for arbitrary sets of k-dimensional normal distributions, for any $k \geqq 1$. However, for $k > 1$ this cannot be deduced from Theorem 5.1 since the multidimensional D distance does not have the properties required by the theorem. It can be shown that if $\mathcal{F}$ is any set of non-singular bivariate normal distributions, the assumption of part (a) is satisfied. But for arbitrary sets of bivariate (possibly singular) normal distributions, $D(F, \mathcal{G}) = 0$ does not imply $d(F, \mathcal{G}) = 0$. (For instance, if F_c

denotes the bivariate normal distribution with means $(c, -c)$, unit variances and correlation coefficient 1, and $\mathcal{G} = \{F_c \mid c > 0\}$, then $D(F_0, \mathcal{G}) = 0$ but $d(F_0, \mathcal{G}) = 1$.) Moreover, $D(\mathcal{G}, \mathcal{H}) = 0$ does not imply $d(\mathcal{G}, \mathcal{H}) = 0$ even for sets of non-singular bivariate normal distributions. (Thus if G_c denotes the bivariate normal distribution with means $(c, -c)$, unit variances, and correlation coefficient $(1 + c^2)^{-1}$, if $\mathcal{G} = \{G_c \mid c < 0\}$ and $\mathcal{H} = \{G_c \mid c > 0\}$, then $D(\mathcal{G}, \mathcal{H}) = 0$ but $d(\mathcal{G}, \mathcal{H}) > 0$.)

For a fixed $k \geqq 1$ let $\mathfrak{N}$ denote the set of all k-dimensional normal distributions. To prove the statement at the beginning of the preceding paragraph it is sufficient to display a distance δ such that $\delta(\mathcal{G}, \mathcal{H}) = 0$ implies $d(\mathcal{G}, \mathcal{H}) = 0$ whenever $\mathcal{G} \subset \mathfrak{N}$ and $\mathcal{H} \subset \mathfrak{N}$, and δ satisfies assumption (3.7) of Theorem 3.1 with $\mathfrak{F} = \mathfrak{N}$. We shall show this to be true for the distance δ^* defined as follows.

For any k-dimensional distribution F with finite moments of the second order define $\theta(F) = (\mu(F), \Sigma(F))$, where $\mu(F)$ denotes the vector of the means and $\Sigma(F)$ the covariance matrix of F. Denote by Θ the range of $\theta(F)$. Define the function $d^*(\theta_1, \theta_2)$, $\theta_1, \theta_2 \varepsilon \Theta$ by

$$d^*(\theta_1, \theta_2) = d(F_1, F_2) \text{ if } F_i \varepsilon \mathfrak{N} \quad \text{and} \quad \theta(F_i) = \theta_i, \qquad i = 1, 2.$$

Now define δ^* by

$$\delta^*(F_1, F_2) = d^*(\theta(F_1), \theta(F_2))$$

for any two k-dimensional distributions F_1 and F_2 with finite moments of the second order.

The function δ^* is a distance[7] in the set of distributions for which it is defined. Obviously $\delta^*(\mathcal{G}, \mathcal{H}) = 0$ if and only if $d(\mathcal{G}, \mathcal{H}) = 0$ for $\mathcal{G} \subset \mathfrak{N}$ and $\mathcal{H} \subset \mathfrak{N}$.

Now let F_n be the empiric distribution of n independent chance variables $X_1, \cdots, X_n$, each of which has the distribution $F \varepsilon \mathfrak{N}$. Put $\theta(F) = \theta = (\mu, \Sigma)$ and $\theta(F_n) = \hat{\theta} = (\hat{\mu}, \hat{\Sigma})$. Thus $\hat{\mu}$ is the sample mean vector and $\hat{\Sigma}$ the sample covariance matrix. We have

$$\delta^*(F_n, F) = d^*(\hat{\theta}, \theta).$$

It follows from the definition of d^* that the distribution of $d^*(\hat{\theta}, \theta)$ does not change if each X_i is subjected to the same non-singular linear transformation. Hence the distribution of $d^*(\hat{\theta}, \theta)$ depends only on the rank r of Σ. If $r = k$, we may assume that $\theta = (0, I) = \theta_0$ (say), where 0 denotes the zero vector with k components and I the $k \times k$ unit matrix. If $1 \leqq r < k$, the distribution of $d^*(\hat{\theta}, \theta)$ is the same, only with k replaced by r. If $r = 0$, $d^*(\hat{\theta}, \theta) = 0$ with probability one. Thus we may confine ourselves to the case $r = k$, $\theta = \theta_0$. We have only to show that for every $c > 0$ there exist numbers $A(c)$ and $B(c)$ such that for all integers $n > 0$,

$$P[d^*(\hat{\theta}, \theta_0) > c] < A(c)e^{-B(c)n}. \tag{5.9}$$

[7] Recall that $\delta^*(F_1, F_2) = 0$ need not imply $F_1 = F_2$.

Now the function $d^*(\theta, \theta_0)$ is continuous at $\theta = \theta_0$ in the usual sense. Hence it is easily seen that (5.9) is satisfied if for every $\epsilon > 0$ the probability of each of the inequalities

$$|\hat{\mu}_i| > \epsilon, \qquad |\hat{\sigma}_{ii} - 1| > \epsilon, \qquad |\hat{\rho}_{ij}| > \epsilon,$$
$$i \neq j, i, j = 1, \cdots, k,$$

where $\hat{\rho}_{ij} = \hat{\sigma}_{ij} (\hat{\sigma}_{ii} \hat{\sigma}_{jj})^{-1/2}$, and $\hat{\mu}_i$ and $\hat{\sigma}_{ij}$ are the components of $\hat{\mu}$ and $\hat{\Sigma}$, does not exceed a bound of the form $A(\epsilon) \exp(-B(\epsilon)n)$ with $B(\epsilon) > 0$. That the latter is true is seen by considering the well-known distributions of $\hat{\mu}_i$, $\hat{\sigma}_{ii}$, and $\hat{\rho}_{ij}$. This completes the proof.

In the proof we could have equally well used, instead of d, the distance

$$d_1(F, G) = \left\{\int (f^{1/2} - g^{1/2})^2 \, d\nu\right\}^{1/2} = 2^{1/2}\{1 - \rho(F, G)\}^{1/2}$$

where ν denotes a measure such that F and G have densities, f and g, with respect to ν, and

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

For we have (see, for instance, Kraft [7], Lemma 1)

$$1 - \rho(F, G) \leqq d(F, G) \leqq (1 - \rho^2(F, G))^{1/2},$$

so that the distances d and d_1 are equivalent for our purposes.

Define $d_1^*(\theta_1, \theta_2)$ and $\delta_1^*(F, G)$ in terms of d_1 just like d^* and δ^* were defined in terms of d. We shall write $\rho(\theta_1, \theta_2)$ for $\rho(F_1, F_2)$ if $F_i \varepsilon \mathfrak{N}$, $\theta(F_i) = \theta_i$. Thus $d_1^*(\theta_1, \theta_2) = 2^{1/2}(1 - \rho(\theta_1, \theta_2))^{1/2}$. If Σ_1 and Σ_2 are nonsingular,

$$\rho(\theta_1, \theta_2) = |\Sigma_1|^{1/4} |\textstyle\sum_2|^{1/4} \left|\dfrac{\Sigma_1 + \Sigma_2}{2}\right|^{-1/2} \times \exp\left\{-\frac{1}{4}(\mu_1 - \mu_2)' (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)\right\}, \tag{5.10}$$

where μ_1 and μ_2 are regarded as column vectors and the prime denotes the transpose. (Compare Kraft [7], p. 129, where there are some misprints.) If Σ_1 has rank r, $1 \leqq r < k$, then $\rho(\theta_1, \theta_2) = 0$ unless Σ_2 also has rank r and the normal distributions with $\theta = \theta_1$ and $\theta = \theta_2$ assign probability one to the same r-dimensional plane, H; in this case $\rho(\theta_1, \theta_2)$ is equal to an expression like (5.10), with μ_i and Σ_i now denoting the means and covariances, in a common coordinate system, of the corresponding r-dimensional normal distributions on H. If the rank of Σ_1 is 0, then $\rho(\theta_1, \theta_2) = 0$ or 1 according as $\theta_1 \neq \theta_2$ or $\theta_1 = \theta_2$.

If Γ and Δ are subsets of Θ, write $\rho(\theta, \Delta)$ for $\sup_{\theta' \varepsilon \Delta} \rho(\theta, \theta')$ and $\rho(\Gamma, \Delta)$ for $\sup_{\theta \varepsilon \Gamma} \rho(\theta, \Delta)$. If $\mathcal{G} \subset \mathfrak{N}$, define $\theta(\mathcal{G}) = \{\theta(F) \mid F \varepsilon \mathcal{G}\}$.

Expressing the conditions (5.6) and (5.8) of Theorem 5.1 in terms of ρ, we can summarize the foregoing as follows.

THEOREM 5.2. *Let $\mathfrak{N}$ be the set of all k-dimensional normal distributions, $k \geqq 1$.*

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

$$\min [\rho(\theta(F), \theta(\mathcal{G})), \rho(\theta(F), \theta(\mathcal{H}))] < 1 \tag{5.11}$$

for all $F \varepsilon \mathfrak{F}$. Moreover, if condition (5.11) is satisfied, $\mathcal{G}$ and $\mathcal{H}$ are distinguishable in the class of tests (N, ϕ) such that $E_F e^{tN} < \infty$ for some $t = t(F) > 0$ if $F \varepsilon \mathfrak{F}$.

(b) *Two subsets $\mathcal{G}$ and $\mathcal{H}$ of $\mathfrak{N}$ are finitely distinguishable if and only if*

$$\rho(\theta(\mathcal{G}), \theta(\mathcal{H})) < 1. \tag{5.12}$$

We observe that condition (5.11) can be expressed in an alternative form. Note that $\rho(\theta_1, \theta_2) = 1$ if and only if $\theta_1 = \theta_2$. If $\theta = (\mu, \Sigma) \varepsilon \Theta$, where Σ is nonsingular, and $\Delta \subset \Theta$, then $\rho(\theta, \Delta) = 1$ if and only if there is a sequence $\{\theta_i\}$ in Δ such that each of the real components of θ_i converges to the corresponding component of θ(in the ordinary sense). If Σ is singular of rank r, the same is true, but with the additional condition that the normal distributions with parameters θ_i and θ assign probability one to the same r-dimensional plane. Thus, for instance, if $\mathfrak{F}$ is a set of non-singular distributions, condition (5.11) is equivalent to the statement that, for every $F \varepsilon \mathfrak{F}$, the Euclidean distance of $\theta(F)$ from $\theta(\mathcal{G})$ or from $\theta(\mathcal{H})$ is positive.

Condition (5.12) does not seem to have an equally simple interpretation.

By way of illustration, let $\mathcal{G}$ and $\mathcal{H}$ denote two sets of univariate normal distributions with positive variances such that $\mu < 0$ if $(\mu, \sigma^2) \varepsilon \theta(\mathcal{G})$ and $\theta(\mathcal{H}) = \{(\mu, \sigma^2) \mid (-\mu, \sigma^2) \varepsilon \theta(\mathcal{G})\}$. Then $\mathcal{G}$ and $\mathcal{H}$ are finitely distinguishable if and only if μ/σ is bounded away from 0 in $\theta(\mathcal{H})$. They are always distinguishable ($\mathcal{G} \cup \mathcal{H}$). If $\mathfrak{F}$ denotes a set of normal distributions with positive variances which contains $\mathcal{G} \cup \mathcal{H}$, then $\mathcal{G}$ and $\mathcal{H}$ are distinguishable ($\mathfrak{F}$) if and only if the distance of every point $(0, \sigma^2) \varepsilon \theta(\mathfrak{F})$ from $\theta(\mathcal{H})$ is positive.

REFERENCES

[1] A. BERGER, "On uniformly consistent tests", *Ann. Math. Stat.*, Vol. 22 (1951), pp. 289–293.

[2] A. DVORETZKY, J. KIEFER, AND J. WOLFOWITZ, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator", *Ann. Math. Stat.*, Vol. 27 (1956), pp. 642–669.

[3] W. HOEFFDING, "The role of assumptions in statistical decisions", *Proc. Third Berkeley Symposium on Math. Stat. and Prob.*, University of California Press (1956), pp. 105–116.

[4] J. KIEFER AND J. WOLFOWITZ, "On the deviations of the empiric distribution of vector chance variables", to appear in *Trans. Amer. Math. Soc.*

[5] A. KOLMOGOROV, *Foundations of the Theory of Probability* (English translation), Chelsea, 1950.

[6] W. H. KRUSKAL, "On the problem of nonnormality in relation to hypothesis testing". (Dittoed)

[7] C. KRAFT, "Some conditions for consistency and uniform consistency of statistical procedures". University of California Publications in Statistics, Vol. 2 (1955), pp. 125–142.

[8] C. STEIN, "A two-sample test for a linear hypothesis whose power is independent of the variance", *Ann. Math. Stat.*, Vol. 16 (1945), pp. 243–258.