PROCEEDINGS *of the* THIRD BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY

Held at the Statistical Laboratory
University of California
December, 1954
July and August, 1955

VOLUME I

CONTRIBUTIONS TO THE THEORY OF STATISTICS

EDITED BY JERZY NEYMAN

For further effective support of the Symposium thanks must be given the National Science Foundation, the United States Air Force Research and Development Command, the United States Army Office of Ordnance Research, and the United States Navy Office of Naval Research.

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
1956

THE ROLE OF ASSUMPTIONS IN STATISTICAL DECISIONS

WASSILY HOEFFDING
UNIVERSITY OF NORTH CAROLINA

1. Introduction

In order to obtain a good decision rule for some statistical problem we start by making assumptions concerning the class of distributions, the loss function, and other data of the problem. Usually these assumptions only approximate the actual conditions, either because the latter are unknown, or in order to simplify the mathematical treatment of the problem. Hence the assumptions under which a decision rule is derived are ordinarily not satisfied in a practical situation to which the rule is applied. It is therefore of interest to investigate how the performance of a decision rule is affected when the assumptions under which it was derived are replaced by another set of assumptions.

We shall confine ourselves to the consideration of assumptions concerning the class of distributions. Investigations of particular problems of this type are numerous in the literature. There are many studies of the performance of "standard" tests under "non-standard" conditions, for example [3], where further references are given. Most of them considered only the effect of deviations from the assumptions on the significance level of the test. The relatively few studies of the effect on the power function include several papers by David and Johnson, the latest of which is [6]. For some problems tests have been proposed whose significance level is little affected by certain deviations from standard assumptions, for instance R. A. Fisher's randomization tests (see section 3; see also Box and Andersen [4]). Some other relevant work will be mentioned later.

In sections 2, 3, and 4 we shall be concerned with problems of the following type. Let P denote the joint distribution of the random variables under observation. Suppose that we contemplate making the assumption that P belongs to a class $\mathcal{P}_1$, but we admit the possibility that actually P is contained in another class, $\mathcal{P}_2$. The performance of a decision rule (decision function) d is assumed to be expressed by the given risk function $r(P, d)$, defined for all $P \in \mathcal{P}_1 + \mathcal{P}_2$ and all d in D, the class of decision rules available to the statistician. Let d_i be a decision rule which is optimal in some specified sense (for example, minimax) under the assumption $P \in \mathcal{P}_i$, $i = 1, 2$. Suppose first that the optimal rule d_i is unique except for equivalence in $\mathcal{P}_1 + \mathcal{P}_2$, for $i = 1, 2$, that is, if d_i' is also optimal for $P \in \mathcal{P}_i$ then $r(P, d_i') = r(P, d_i)$ for all $P \in \mathcal{P}_1 + \mathcal{P}_2$. Then we may assess the consequences of the assumption $P \in \mathcal{P}_1$ when actually $P \in \mathcal{P}_2$ by comparing the values $r(P, d_1)$ and $r(P, d_2)$ for $P \in \mathcal{P}_2$. If the optimal rules are not unique, we may pick out from the class of rules which are optimal for $P \in \mathcal{P}_1$ a subclass of rules which come closest to optimality under the assumption $P \in \mathcal{P}_2$, and compare their performance with that of the rules which are optimal under the latter assumption. In

This research was supported by the United States Air Force through the Office of Scientific Research of the Air Research and Development Command, and by the Office of Ordnance Research, U.S. Army, under Contract DA-04-200-ORD-355.

some situations other ways of approaching the problem may be more adequate (see, for example, section 2).

In section 2 the consequences of assuming that a distribution is continuous are discussed. Problems involved in comparing assumptions of varying generality are considered in section 3. Section 4 is concerned with cases where decision rules derived under assumptions of normality retain their optimal properties when these assumptions are relaxed.

The last three sections deal with distinguishable sets of distributions, a concept related to the problem of the existence of unbiased or consistent tests under given assumptions. Criteria for the distinguishability of two sets by means of a test based on finitely many observations and by a sequential test are considered and their uses illustrated in sections 5 and 7. An example where two sets are indistinguishable by a nonrandomized test, but distinguishable by a randomized test, is discussed in section 6.

2. The assumption of a continuous distribution

The assumption that we are dealing with a class of continuous distributions is usually made when actually the observations are integer multiples of the unit of measurement h, a (small) positive constant. Suppose that a sample $x = (x_1, \cdots, x_n)$ is a point in R^n, and let $\boldsymbol{\rho}_1$ be a class of distributions (probability measures) which are absolutely continuous with respect to n-dimensional Lebesgue measure. Let S be the set of all points in R^n whose coordinates are integer multiples of h. Let us suppose that when we say that the distribution is $P_1 \in \boldsymbol{\rho}_1$, we "have in mind" that the distribution is $P_2 = f(P_1)$, where the probability measure P_2 is defined by

$$(1) \qquad P_2(\{y\}) = P_1\left(\left\{x: y_j - \frac{h}{2} < x_j \leqq y_j + \frac{h}{2},\ j = 1, \cdots, n\right\}\right)$$

for all $y = (y_1, \cdots, y_n)$ in S. Let $\boldsymbol{\rho}_2 = \{f(P): P \in \boldsymbol{\rho}_1\}$. Thus we are interested in the consequences of assuming $P \in \boldsymbol{\rho}_1$ when actually $P \in \boldsymbol{\rho}_2$.

Let d be a decision function which is optimal in some sense under the assumption $P \in \boldsymbol{\rho}_1$. Then any decision rule which differs from d only on the set S is equivalent to d for $P \in \boldsymbol{\rho}_1$. Since $P(S) = 1$ for all $P \in \boldsymbol{\rho}_2$, the mere fact that a rule is optimal for $P \in \boldsymbol{\rho}_1$ does not tell us anything about its performance when $P \in \boldsymbol{\rho}_2$; indeed, it can be as bad as we please under the latter assumption.[1] Of course, in general there are rules which are optimal under either assumption. But the main reason for making the simplifying assumption of continuity is that we do not want to bother with rules which are optimal for $P \in \boldsymbol{\rho}_2$. Now it is clear that if there is a determination d' of d which is sufficiently regular, its risk at $P_2 = f(P_1)$ will differ arbitrarily little from the risk at P_1 if h is small enough; also, d' may not be much worse than an optimal rule for $P \in \boldsymbol{\rho}_2$. We shall not investigate here under what conditions such a regular decision rule exists or how small h has to be in order that the assumption of continuity cause little harm. These questions may deserve attention. Fortunately, when a statistician applies a decision rule, he is likely to choose the most regular determination available anyway. However, the theoretical statistician might do well to be careful when he neglects sets of measure zero.

[1] The author's attention was drawn to situations of this kind by H. Robbins some years ago.

3. Assumptions of varying generality

Suppose we consider making one of two assumptions, $P \in \boldsymbol{\rho}_1$ and $P \in \boldsymbol{\rho}_2$, where $\boldsymbol{\rho}_1 \subset \boldsymbol{\rho}_2$. The second assumption is safer, but with the first assumption we may achieve a smaller risk.

The consequences of making the broader assumption when actually the narrower assumption is justified may be called serious if any decision rule which is "good" under the broader assumption is "bad" under the narrower assumption. Thus the consequences will depend on what we mean by a good decision rule. But even with a given definition of "good" or "best" the consequences may depend on the class of decision rules at our disposal. For example, suppose we require a minimax estimator of the mean μ of a normal distribution when the loss function is the squared deviation from μ, and we assume that the variance σ^2 does not exceed a given number A. If we are restricted to estimators based on a sample of fixed size, the minimax estimator is the sample mean $\bar{x}$ and does not depend on A. On the other hand, if we are permitted to choose the sample size in advance, and the cost of sampling is taken into account, the minimax estimator will depend on A. If A_2 is substantially larger than A_1, the assumption $\sigma^2 \leqq A_2$ will give us a unique minimax estimator whose performance is poor under the assumption $\sigma^2 \leqq A_1$.

Sometimes a considerable broadening of the assumption does not lead to serious consequences when the narrower assumption is justified. Thus in the standard problems concerning the variance of a normal distribution we need, when the mean is completely unknown, just one more observation to obtain the same expected loss as when the mean is known. Somewhat similar results have been obtained in certain cases where a parametric class of distributions is enlarged to a nonparametric class. Several examples can be found in [9]. For instance, consider the problem of testing whether two distributions are equal (and not otherwise specified) against the alternative that the distributions are normal with common variance and means $\mu_1 < \mu_2$. The uniformly most powerful similar test, based on two random samples of fixed size, is asymptotically as powerful in large samples (in a sense explained in [9]) as the corresponding standard test for testing the equality of the means of two normal distributions. (The former test is of the randomization type introduced by R. A. Fisher; its optimal properties were proved by Lehmann and Stein [12].) Here we assumed that the class of alternatives is the same under both assumptions. Actually the test retains its property of being uniformly most powerful similar even when the class of alternatives is enlarged to a nonparametric class of distributions of an exponential type (see Lehmann and Stein [12]). If the class is further extended, a uniformly most powerful similar test will in general not exist, and it will be necessary to specify against what types of alternatives the power of a test should be large. This can be done in many ways, and an optimal test and its performance in the class of normal distributions will depend on this specification.

4. Nonparametric justifications of assumptions of normality

Given a decision rule d which is optimal in a specified sense under the assumption that P is in a class $\boldsymbol{\rho}_1$, it is of interest to determine other classes $\boldsymbol{\rho}$ such that d is optimal (in the same or a suitably extended sense) under the assumption $P \in \boldsymbol{\rho}$. If optimal means minimax, an obvious sufficient condition for d to remain a minimax rule in $\boldsymbol{\rho} \supset \boldsymbol{\rho}_1$ is that the risk of d in $\boldsymbol{\rho}$ attain its maximum in $\boldsymbol{\rho}_1$. Situations of this type were considered by Hodges and Lehmann [8].

In certain cases we find that a decision rule derived under the assumption of a normal distribution retains its optimal character in a large, nonparametric class of distributions. One result of this type, concerning the minimax character of Markov estimators, can be found in [8]. Similar though weaker results can be obtained in certain testing problems.

As an example consider the following extension of Student's problem. Let $\mathcal{Q}$ be the class of distributions F with finite mean $\mu(F)$, positive variance $\sigma^2(F)$ and such that

$$\int_{-\infty}^{\infty} |x-\mu(F)|^3 dF(x) \leqq M\sigma^3(F), \tag{2}$$

where M is fixed. Let $\mathcal{Q}_\delta$ be the subclass of $\mathcal{Q}$ with $\sqrt{n}\,\mu(F)/\sigma(F) = \delta$. We want to test the hypothesis $F \in \mathcal{Q}_\delta$, $\delta \leqq 0$, against the alternative $F \in \mathcal{Q}_\delta$, $\delta > 0$. We restrict ourselves to the class D of tests d based on n independent observations from F, with critical region $W = W(d)$. We choose the risk function

$$r(F, d) = \begin{cases} aP(W|F) & \text{if} \quad F \in \mathcal{Q}_\delta, \delta \leqq 0, \\ b[1-P(W|F)] & \text{if} \quad F \in \mathcal{Q}_\delta, \delta \geqq \delta_1, \\ 0 & \text{elsewhere}, \end{cases} \tag{3}$$

where $P(W|F)$ denotes the probability of $(X_1, \cdots, X_n) \in W$ when the X_j are independent with the common distribution F, and a, b, and δ_1 are positive constants.

Let d_0 be the test with critical region $W_0 = \{t > c\}$, where

$$t = \frac{n^{1/2}\bar{x}}{s}, \qquad \bar{x} = n^{-1}\sum_{j=1}^{n} x_j, \qquad s^2 = (n-1)^{-1}\sum_{j=1}^{n}(x_j - \bar{x})^2, \tag{4}$$

and the constant c is determined by

$$a[1 - S_{n-1}(c, 0)] = bS_{n-1}(c, \delta_1); \tag{5}$$

here $S_{n-1}(x, \delta)$ denotes the noncentral Student distribution function with $n-1$ degrees of freedom and noncentrality parameter δ. It can be shown by standard methods that d_0 is the minimax test in the subclass $\mathcal{Q}^0$ of $\mathcal{Q}$ which consists of the normal distributions. By an inequality of Berry and Esseen (see, for example, [7]) the distribution function $F_n(y)$ of $n^{1/2}[\bar{X} - \mu(F)]/\sigma(F)$ converges to the standard normal distribution function $\Phi(y)$ uniformly for $F \in \mathcal{Q}$ (and uniformly in y) as $n \to \infty$. Also, for any $\epsilon > 0$, $P[|s/\sigma(F) - 1| < \epsilon | F] \to 1$ uniformly for $F \in \mathcal{Q}$. Hence it can be shown that for any real δ and for all $F \in \mathcal{Q}_\delta$ we have

$$|P(t \leqq y|F) - \Phi(y - \delta)| \leqq C_n(\delta), \qquad -\infty < y < \infty, \tag{6}$$

where $C_n(\delta)$ depends on n, δ, and M only and tends to 0 as $n \to \infty$, for δ fixed. It follows that

$$|P(t \leqq y|F) - S_{n-1}(y|\delta)| \leqq 2C_n(\delta), \qquad -\infty < y < \infty \tag{7}$$

for all $F \in \mathcal{Q}_\delta$.

Now if $\mathcal{Q}^*$ denotes the subclass of $\mathcal{Q}$ with $\mu(F) = 0$ and $\sigma^2(F) = 1$, we have

$$\sup_{F \in \mathcal{Q}_\delta} P(t > c|F) = \sup_{F \in \mathcal{Q}^*} P\left[\frac{n^{1/2}\bar{x} + \delta}{s} > c \,\middle|\, F\right], \tag{8}$$

which is a nondecreasing function of δ. The same is true of the infimum in $\mathcal{Q}_\delta$. Hence we obtain

$$\sup_{F\in\mathcal{F}} r(F, d_0) \leqq \sup_{F\in\mathcal{F}_0} r(F, d_0) + \epsilon, \tag{9}$$

$$\leqq \inf_{d\in D} \sup_{F\in\mathcal{F}} r(F, d) + \epsilon$$

where $\epsilon = 2 \max \{aC_n(0), bC_n(\delta_1)\}$. Thus the maximum risk in $\mathcal{Q}$ of Student's test d_0 exceeds the minimax risk in $\mathcal{Q}$ by at most ϵ, where ϵ is arbitrarily small for n sufficiently large. (Note that the minimax risk is bounded away from zero as $n \to \infty$.)

In the corresponding problem with $\sigma(F) = \sigma$ fixed we find in a similar way the stronger result that the maximum risk of the $\bar{x}$-test in $\mathcal{Q}_\delta$ [the class with $\mu(F) = \delta\sigma/\sqrt{n}$] lies within a small ϵ of its "normal" risk, uniformly in δ. The argument which was used above does not permit us to decide whether an analogous result is true when $\sigma(F)$ is unrestricted.

The explanation for the near-optimal behavior of the "normal" decision rules in these cases is, of course, the distribution-free character of the central limit theorem, combined with the fact that the class $\mathcal{Q}$ was so chosen as to make the approach to the normal distribution uniform.

5. Distinguishable sets of distributions

If we relax the assumptions more and more, the minimax risk will in general increase, and eventually we may reach a point where the maximum risk of any decision rule is not smaller than the risk of a rule which does not depend on the observations. We shall consider criteria for recognizing when this or a similar situation occurs in testing problems.

Consider a testing (or two decision) problem such that one or the other decision is definitely preferred according as the distribution P belongs to $\mathcal{P}^1$ or $\mathcal{P}^2$, two disjoint subsets of the given class $\mathcal{P}$. Unless otherwise stated we assume that each P in $\mathcal{P}$ is a probability measure on $(\mathcal{X}, \mathcal{A})$, where $\mathcal{X}$ is the space of infinite sequences $x = (x_1, x_2, \cdots)$ of real numbers and $\mathcal{A}$ is Kolmogorov's extension to $\mathcal{X}$ of the ordinary Borel field.

A test will be called finite if it depends only on a finite number of coordinates (observations) x_j. By the critical function of a finite test we mean a measurable function ψ from $\mathcal{X}$ to the interval [0, 1] such that $1 - \psi(x)$ [$\psi(x)$] is the probability of taking the decision corresponding to $P \in \mathcal{P}^1$ [$P \in \mathcal{P}^2$] when x is the sequence of observations.

Let D be any class of finite tests, and let Ψ be the class of the critical functions of tests in D. We shall say that the sets $\mathcal{P}^1$ and $\mathcal{P}^2$ are distinguishable in D if there exists a ψ in Ψ such that

$$\sup_{P\in\mathcal{P}^1} E(\psi \mid P) + \sup_{P\in\mathcal{P}^2} E(1 - \psi \mid P) < 1, \tag{10}$$

where $E(f \mid P)$ is the expected value of $f(X)$ when X has the distribution P. Otherwise $\mathcal{P}^1$ and $\mathcal{P}^2$ are said to be indistinguishable in D. [The property of ψ expressed in (10) has an obvious relation with unbiasedness.]

Let D_f denote the class of all finite tests, and let D_n, $n = 1, 2, \cdots$, be the class of all fixed sample size tests based on the observations $(x_1, \cdots, x_n)$. Two sets which are distinguishable in D_f will be called finitely distinguishable. We observe that two sets are finitely distinguishable if and only if they are distinguishable in D_n for some n.

Berger and Wald [2] gave conditions under which two sets of distributions are dis-

tinguishable in the class of all nonrandomized tests in D_n if and only if they are disjoint. (Their theorem 3.1 is stated in a slightly more special form.)

A sufficient condition for two sets to be indistinguishable in D_n can be stated as follows. Let $\mathcal{X}_n$ be the space of points $(x_1, \cdots, x_n)$, and let $\mathcal{A}_n$, $\mathcal{P}_n$ and $\mathcal{P}_n^i$ be the σ-field of subsets of $\mathcal{X}_n$ and the classes of distributions on $\mathcal{A}_n$ which are determined by $\mathcal{A}$, $\mathcal{P}$ and $\mathcal{P}^i$ in an obvious way. For any two distributions P_1 and P_2 on $\mathcal{A}_n$ we denote by ν any measure on $\mathcal{A}_n$ relative to which P_1 and P_2 are absolutely continuous, and by p_1 and p_2 the respective densities (Radon-Nikodym derivatives). With this notation, the sets $\mathcal{P}^1$ and $\mathcal{P}^2$ (or, equivalently, the sets $\mathcal{P}_n^1$ and $\mathcal{P}_n^2$) are indistinguishable in D_n if for any $\epsilon > 0$ there exist two distributions, $P_1 \in \mathcal{P}_n^1$ and $P_2 \in \mathcal{P}_n^2$, such that

$$\int |p_1 - p_2| \, d\nu < \epsilon , \tag{11}$$

where the integral extends over $\mathcal{X}_n$. This follows from the inequality

$$\inf_{P \in \mathcal{P}^2} \int \psi dP - \sup_{P \in \mathcal{P}^1} \int \psi dP \leqq \int \psi \, (p_2 - p_1) \, d\nu . \tag{12}$$

The statement of the condition remains true in the more general form where P_i is any mixture of distributions in $\mathcal{P}_n^i$ with respect to some probability measure ξ_i on a σ-field of subsets of $\mathcal{P}_n^i$, subject to an obvious measurability condition. The proof is similar and uses theorem 3 of Robbins [13]. A theorem of Le Cam (see Kraft [15]) implies that if the distributions in $\mathcal{P}_n^1$ and $\mathcal{P}_n^2$ are absolutely continuous with respect to a fixed measure, the condition expressed in (11), with P_1 and P_2 mixtures, is also necessary for the indistinguishability of $\mathcal{P}_n^1$ and $\mathcal{P}_n^2$.

With $S = \{x: p_1(x) > p_2(x)\}$ we have

$$\tfrac{1}{2} \int |p_1 - p_2| \, d\nu = \sup_{A \in \mathcal{A}_n} |P_1(A) - P_2(A)| = P_1(S) - P_2(S) . \tag{13}$$

The first equation (13) shows that condition (11) is independent of the choice of ν. The last expression in (13) is often convenient when applying this condition.

It follows from an earlier remark that two sets $\mathcal{P}^1$ and $\mathcal{P}^2$ are finitely indistinguishable if the condition expressed in (11) is satisfied for every n.

We shall say that $\mathcal{P}^1$ and $\mathcal{P}^2$ are finitely absolutely distinguishable if for any $\epsilon > 0$ there exists a finite test with critical function ψ such that

$$\sup_{P \in \mathcal{P}_n^1} E(\psi | P) + \sup_{P \in \mathcal{P}_n^2} E(1 - \psi | P) < \epsilon . \tag{14}$$

This property has also been expressed by saying that there exists a uniformly consistent sequence of tests [1].

Now suppose that each P in $\mathcal{P}$ is the distribution of a sequence of independent, identically distributed random variables. Then if two sets are finitely distinguishable, they are finitely absolutely distinguishable. This is a simple partial extension of a theorem of Berger [1]; the theorem gives a necessary and sufficient condition for the existence of a uniformly consistent sequence of nonrandomized tests. Further interesting results on the existence of a uniformly consistent sequence of tests were recently obtained by Kraft [15].

We now give three examples of finitely indistinguishable sets.

Example 5.1. If P is the distribution of independent, normal random variables with mean μ and variance σ^2, and $\mathcal{P}^i$ is the set with $\mu = \mu_i$, $0 < \sigma^2 < \infty$, then $\mathcal{P}^1$ and $\mathcal{P}^2$ are finitely indistinguishable. Condition (11) is satisfied for every n if P_i is the distribution

with $\mu = \mu_i$ and σ sufficiently large. The corresponding result for tests with constant power in $\boldsymbol{\rho}^1$ and $\boldsymbol{\rho}^2$ was proved by Dantzig [5] in 1940.

Example 5.2. If P is the distribution of independent, normal random variables with means $\mu_1, \mu_2, \cdots$ and common variance σ^2, and $\boldsymbol{\rho}^i$ the set with $\sigma = \sigma_i$, $-\infty < \mu_j < \infty$, $j = 1, 2, \cdots$, then $\boldsymbol{\rho}^1$ and $\boldsymbol{\rho}^2$ are finitely indistinguishable. Here we can apply the general form of condition (11). For if P_i is the mixture of the P in $\boldsymbol{\rho}_n^i$ according to ξ_i, where under ξ_i the means $\mu_1, \cdots, \mu_n$ are independent normal with zero mean and variance τ_i^2, such that $\sigma_1^2 + \tau_1^2 = \sigma_2^2 + \tau_2^2$, then $P_1 = P_2$.

Example 5.3. This is a further extension of Student's problem (see section 4). Let $\boldsymbol{\mathcal{Q}}^i$ be the class of all distributions F on the real line with finite mean $\mu(F)$ and positive variance $\sigma^2(F)$ such that $\mu(F)/\sigma(F) = \gamma_i$, $\gamma_1 < \gamma_2$. Let $\boldsymbol{\rho}^i$ be the class of distributions of independent random variables with common distribution $F \in \boldsymbol{\mathcal{Q}}^i$. Then $\boldsymbol{\rho}^1$ and $\boldsymbol{\rho}^2$ are finitely absolutely distinguishable if $\gamma_1 < 0 < \gamma_2$, and finitely indistinguishable if $\gamma_2 \leqq 0$ or $\gamma_1 \geqq 0$.

If $\gamma_1 < 0 < \gamma_2$, it is easy to show with the aid of Chebyshev's inequality that the tests with critical functions $\psi_n(x) = 0$ or 1 according as $\sum_1^n x_j \leqq 0$ or > 0 form a uniformly consistent sequence.

If $\gamma_1 \geqq 0$, condition (11) is satisfied for every n if P_i is the distribution with $F = F_i$, where F_i ascribes probabilities $1 - \pi_i$ and $\pi_i = (1 + t_i^2)^{-1}$ to the respective points $\gamma_2 - t_2^{-1}$ and $\gamma_2 + t_2$; here $t_2 > 0$, $t_1 = f(t_2)$ is the positive root (unique for t_2 small) of

$$(1 - \gamma_2 t_2)\, t_1^2 + \gamma_1 (1 + t_2^2)\, t_1 - \gamma_2 t_2 - t_2^2 = 0\,, \tag{15}$$

and $t_2 \to 0$. The case $\gamma_2 \leqq 0$ can be reduced to this case.

6. Sets distinguishable only by randomized tests: An example

Some results of Lehmann [11] suggest that two sets may be distinguishable in D_n but indistinguishable in the class D_n' of nonrandomized tests in D_n. We shall consider a problem where this situation occurs. We denote by $\Psi_n(\Psi_n')$ the class of critical functions of the tests in $D_n(D_n')$. Thus if $\psi \in \Psi_n'$, $\psi(x) = 0$ or 1 for all x.

Let $\boldsymbol{\mathcal{Q}}_\mu$ be a class of distributions F on the real line with mean μ and variance 1, which contains all distributions with this property which assign probability 1 to at most three points. Let $\boldsymbol{\rho}_{\mu, n}$ be the class of all distributions of n independent random variables with a common distribution in $\boldsymbol{\mathcal{Q}}_\mu$. We shall show that $\boldsymbol{\rho}_{\lambda, n}$ and $\boldsymbol{\rho}_{\mu, n}$ are distinguishable in D_n for all $\lambda \neq \mu$ and all $n = 1, 2, \cdots$, but indistinguishable in D_n' for any n unless $|\lambda - \mu|$ exceeds a positive constant (which depends on n). It is clearly sufficient to take $\lambda = -h$, $\mu = h > 0$. We denote by $E(f|F)$ the expected value of $f(X)$ when the components of X are independent with the common distribution F.

We first prove the second part of the statement in the stronger form: For any n and for any $\alpha \in (0, 1)$ the inequalities

$$\sup_{F \in F_{-h}} E(\psi \,|\, F) \leqq \alpha \leqq \inf_{F \in F_h} E(\psi \,|\, F) \tag{16}$$

cannot both be satisfied with $\psi \in \Psi_n'$ unless h exceeds a positive number which depends only on n (and is of order $n^{-1/2}$). If ψ is in Ψ_n' and satisfies the first inequality (16), we must have

$$\psi(y, \cdots, y) = 0 \quad \text{if} \quad \alpha\,[1 + (y + h)^2]^n < 1\,, \tag{17}$$

for all real y. For if $t = y + h \neq 0$, let F' be the distribution (in $\mathcal{Q}_{-h}$) which assigns the probabilities $(1 + t^2)^{-1}$ and $1 - (1 + t^2)^{-1}$ to the respective points $t - h$ and $-t^{-1} - h$. Then $a \geqq E(\psi|F') \geqq \psi(t - h, \cdots, t - h)(1 + t^2)^{-n}$. This implies (17) for $y + h \neq 0$. If $y + h = 0$, we use a similar argument with F' any distribution in $\mathcal{Q}_{-h}$ which assigns to the point $-h$ a probability arbitrarily close to 1.

Similarly, for any $\psi \in \Psi_n'$ which satisfies the second inequality (16) we must have

$$\psi(y, \cdots, y) = 1 \quad \text{if} \quad (1 - a)[1 + (y - h)^2]^n < 1, \tag{18}$$

for all real y. Taking $y = -h$ and $y = h$, we find that a $\psi \in \Psi_n'$ cannot satisfy both inequalities (16) if $[1 + (2h)^2]^n < \max [a^{-1}, (1 - a)^{-1}]$; and hence cannot satisfy them for any a if $[1 + (2h)^2]^n < 2$. [This is not the best bound which can be obtained from (17) and (18).]

We now show that for any $h > 0$, any $n \geqq 1$, and any $a \in (0, 1)$ condition (16), with at least one strict inequality, can be satisfied by a randomized test in D_n. Let $a = hn^{1/2}$, $-a < c < a$,

$$k(c) = a + c + (a - c)^{-1}, \qquad b = \frac{-k(-c)}{2}, \qquad d = \frac{k(c)}{2}, \tag{19}$$

$$\phi(y) = \begin{cases} 0 & \text{if} \quad y \leqq b, \\ \dfrac{y - b}{d - b} & \text{if} \quad b < y < d, \\ 1 & \text{if} \quad d \leqq y. \end{cases} \tag{20}$$

If we let $\psi(x) = \phi\left(n^{-1/2}\sum_1^n x_j\right)$, we have

$$\sup_{F \in \mathcal{F}_{-h}} E(\psi|F) \leqq \frac{1}{2ak(c)} < 1 - \frac{1}{2ak(-c)} \leqq \inf_{F \in \mathcal{F}_h} E(\psi|F) \tag{21}$$

for $|c| < a$. As c increases from $-a$ to a, either side of (21) decreases continuously from 1 to 0.

We sketch the proof of (21). Let $f(y)$ be any polynomial of the second degree such that $\phi(y) \leqq f(y)$ for all real y. If $g(x) = f(n^{-1/2}\sum x_j)$, then $E(\psi|F) \leqq E(g|F)$, and $E(g|F)$ is constant in $\mathcal{Q}_\mu$ for each μ. Now choose f so as to minimize $E(g|F)$, $F \in \mathcal{Q}_\mu$.

7. Sequentially distinguishable sets of distributions

We shall restrict ourselves to sequences of independent random variables with a common distribution F. Suppose that $F \in \mathcal{Q}$, and let $\mathcal{Q}^1$ and $\mathcal{Q}^2$ be two disjoint subsets of $\mathcal{Q}$. Let $D_s = D_s(\mathcal{Q})$ be the class of all sequential tests for taking one of two decisions, a_1 and a_2, which terminate with probability one for all $F \in \mathcal{Q}$. We denote by $Pr\{a_i|F, d\}$ and $E(n|F, d)$, respectively, the probability of the decision a_i and the expected number of observations required to reach a decision when the distribution is F and test d is used.

The sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$ will be called sequentially distinguishable (indistinguishable) at F if there exists (does not exist) a d in D_s such that $E(n|F, d) < \infty$ and

$$\sup_{F \in \mathcal{F}^1} Pr\{a_2|F, d\} + \sup_{F \in \mathcal{F}^2} Pr\{a_1|F, d\} < 1. \tag{22}$$

If the left side of (22) is arbitrarily small for some d in D_s with $E(n|F, d) < \infty$, then $\mathcal{Q}^1$ and $\mathcal{Q}^2$ are said to be sequentially absolutely distinguishable at F. If $\mathcal{Q}^1$ and $\mathcal{Q}^2$ are sequentially [absolutely] distinguishable (indistinguishable) at every F in a class $\mathcal{Q}^*$, then $\mathcal{Q}^1$ and $\mathcal{Q}^2$ will be said to be sequentially [absolutely] distinguishable (indistinguishable) in $\mathcal{Q}^*$.

Note that these definitions are stated in terms of the sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$ rather than in terms of the corresponding sets of distributions of sequences. Statements such as $\mathcal{Q}^1$ and $\mathcal{Q}^2$ are finitely indistinguishable will have an obvious meaning in this context.

A sufficient condition for two sets to be sequentially indistinguishable is implied by an inequality proved in [10]. Let $F_1 \in \mathcal{Q}^1$, $F_2 \in \mathcal{Q}^2$, $F \in \mathcal{Q}$, and let ν be a measure relative to which these three distributions are absolutely continuous, with respective densities f_1, f_2, and f. By a trivial extension of equation (4) in [10], if d is any test in D_s such that

$$\sup_{F \in \mathcal{Q}^1} Pr\{a_2|F, d\} \leqq \alpha_1, \qquad \sup_{F \in \mathcal{Q}^2} Pr\{a_1|F, d\} \leqq \alpha_2 \tag{23}$$

where $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha_1 + \alpha_2 < 1$, then

$$E(n|F, d) \geqq \frac{-\log[\alpha_1^c(1-\alpha_2)^{1-c} + (1-\alpha_1)^c \alpha_2^{1-c}]}{c\int f \log \frac{f}{f_1}\, d\nu + (1-c)\int f \log \frac{f}{f_2}\, d\nu} \tag{24}$$

for $0 < c < 1$, where the integrals are taken over the entire space. If, in particular, $F \in \mathcal{Q}^1$, the right side of (24) is maximized with $F_1 = F$ and $c \to 0$, and we obtain

$$E(n|F, d) \geqq \frac{\alpha_1 \log \frac{\alpha_1}{1-\alpha_2} + (1-\alpha_1)\log \frac{1-\alpha_1}{\alpha_2}}{\int f \log \frac{f}{f_2}\, d\nu} \quad \text{if } F \in \mathcal{Q}^1. \tag{25}$$

We note that the numerators and denominators in (24) and (25) are positive; the denominators may be infinite.

Hence if for any positive number M and any two positive numbers α_1 and α_2 with $\alpha_1 + \alpha_2 < 1$ the distributions $F_1 \in \mathcal{Q}^1$ and $F_2 \in \mathcal{Q}^2$ and the number c can be so chosen that the right side of (24) exceeds M, the sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$ are sequentially indistinguishable at F. If $F \in \mathcal{Q}^1$, the two sets are sequentially indistinguishable at F if for any $\epsilon > 0$ we can find an $F_2 \in \mathcal{Q}^2$ such that

$$\int f \log \frac{f}{f_2}\, d\nu < \epsilon. \tag{26}$$

By example 5.1 two sets of normal distributions with fixed means and unrestricted variances are finitely indistinguishable. On the other hand, by a well-known result of Stein [14], these sets are sequentially absolutely distinguishable in the class of all normal distributions. However, if the requirement $E(n|F, d) < \infty$ is replaced by the stronger condition that $E(n|F, d) = E(n|\mu, \sigma; d)$ be bounded in σ for μ fixed, inequality (24) easily implies that condition (22) cannot be satisfied.

As an application of condition (26) we shall show that the sets $\mathcal{Q}^1$ and $\mathcal{Q}^2$ of example 5.3, with $\gamma_1 = 0 < \gamma_2$, are sequentially indistinguishable in $\mathcal{Q}^2$. Let F be any distribution in $\mathcal{Q}^2$, so that $\mu(F)/\sigma(F) = \gamma_2$. Let $F_1 = (1-t)F + tG$, where $0 < t < 1$ and G is the distribution which assigns probability one to the point $a = -\mu(F)(1-t)/t$. Then

$F_1 \in \mathcal{Q}^1$. Both F_1 and F are absolutely continuous relative to $\nu = F_1$, with respective densities $f_1(x) = 1$ and

$$f(x) = \begin{cases} \dfrac{1}{1-t} & \text{if} \quad x \neq a, \\ \dfrac{b}{b+t-bt} & \text{if} \quad x = a, \end{cases} \tag{27}$$

where $b = b(t)$ is the F-probability of the point a. Hence

$$\int f \log \frac{f}{f_1}\, d\nu = -(1-b)\log(1-t) + b \log \frac{b}{b+t-bt}, \tag{28}$$

where the last term is to be omitted if $b = 0$. The right side of (28) tends to 0 as $t \to 0$. Thus condition (26) (with $\mathcal{Q}^1$ and $\mathcal{Q}^2$ interchanged) can be satisfied for any $\epsilon > 0$.

The proof shows that this result still holds if $\mathcal{Q}^1$ and $\mathcal{Q}^2$ consist only of the mixtures $H = (1-t)F + tG$ of a normal distribution F and an arbitrary distribution G, where $0 \leqq t < \epsilon$ and ϵ is positive and as small as we please. The distributions H are, in a sense, very close to normal distributions.

REFERENCES

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

[2] A. BERGER and A. WALD, "On distinct hypotheses," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 104–109.

[3] G. E. P. BOX, "Non-normality and tests on variances," *Biometrika*, Vol. 40 (1953), pp. 318–335.

[4] G. E. P. BOX and S. L. ANDERSEN, "The robust tests for variances and effect of non-normality and variance heterogeneity on standard tests," *Institute of Statistics Mimeo Series, No. 101*, Consolidated University of North Carolina, 1954.

[5] G. B. DANTZIG, "On the non-existence of tests of 'Student's' hypothesis having power functions independent of σ," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 186–192.

[6] F. N. DAVID and N. L. JOHNSON, "Extension of a method of investigating the properties of analysis of variance tests to the case of random and mixed models," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 594–601.

[7] B. V. GNEDENKO and A. N. KOLMOGOROV, *Limit Distributions for Sums of Independent Random Variables*, Cambridge, Mass., Addison-Wesley, 1954.

[8] J. L. HODGES, JR. and E. L. LEHMANN, "Some problems in minimax point estimation," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 182–197.

[9] W. HOEFFDING, "The large-sample power of tests based on permutations of observations," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 169–192.

[10] ———, "A lower bound for the average sample number of a sequential test," *Annals of Math. Stat.*, Vol. 24 (1953), pp. 127–130.

[11] E. L. LEHMANN, "Consistency and unbiasedness of certain nonparametric tests," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 165–179.

[12] E. L. LEHMANN and C. STEIN, "On the theory of some non-parametric hypotheses," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 28–45.

[13] H. ROBBINS, "Mixture of distributions," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 360–369.

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

[15] C. KRAFT, "Some conditions for consistency and uniform consistency of statistical procedures," *Univ. Calif. Publ. Stat.*, Vol. 2 (1956), pp. 125–142.