AN OPTIMAL PROPERTY OF THE LIKELIHOOD RATIO STATISTIC

R. R. BAHADUR
UNIVERSITY OF CHICAGO

1. Introduction

Let $s = (x_1, x_2, \cdots, \text{ad inf})$ be a sequence of independent and identically distributed observations on a variable x with distribution depending on a parameter θ taking values in a set Θ. Let Θ_0 be a subset of Θ and consider the null hypothesis that θ is in Θ_0. For each n, let $T_n = T_n(x_1, \cdots, x_n)$ be a real-valued statistic such that, in testing the hypothesis, large values of T_n are significant. For any given s, let $L_n(s)$ be the level attained by T_n in the given case; that is, $L_n(s)$ is the maximum probability (consistent with θ in Θ_0) of obtaining a value of T_n as large or larger than $T_n(s)$. Then, in typical cases, L_n is asymptotically distributed uniformly over (0, 1) in the null case, and L_n tends to zero in probability, or perhaps even with probability one, in the nonnull case. The rate at which L_n tends to zero when a given nonnull θ obtains is a measure of the asymptotic efficiency of T_n against that θ. It is shown in this paper (under very mild restrictions on the family of possible distributions of x) that L_n cannot tend to zero at a rate faster than $[\rho(\theta)]^n$ when a nonnull θ obtains; here ρ is a parametric function defined in terms of the Kullback-Leibler information numbers such that, in typical cases, $0 < \rho < 1$ (theorem 1). It is also shown (under much more restrictive conditions on the distributions of x) that if $\hat{T}_n$ is (any strictly decreasing function of) the likelihood ratio statistic of Neyman and Pearson [1], and $\hat{L}_n$ is the level attained by $\hat{T}_n$, then $\hat{L}_n$ tends to zero at the rate $[\rho(\theta)]^n$ in the nonnull case (theorem 2). In short, the likelihood ratio statistic is an optimal sequence in terms of exact stochastic comparison as described and exemplified in [2], [3], and [4].

Theorems 1 and 2 are stated more precisely in section 2. Section 3 contains a discussion of these theorems. Proofs are given in sections 4 and 5.

2. Theorems

Let X be a space of points x, $\mathcal{B}$ a σ-field of sets of X, and for each point θ in a set Θ, let P_θ be a probability measure on $\mathcal{B}$. Let Θ_0 be a given subset of Θ.

This research was supported in part by Research Grant No. NSF-GP3707 from the Division of Mathematical, Physical and Engineering Sciences of the National Science Foundation and in part by the Statistics Branch of the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ASSUMPTION 1. *There exists a σ-finite measure λ on $\mathcal{B}$ such that each P_θ admits a probability density with respect to λ, say $dP_\theta = f(x, \theta)\, d\lambda$, $0 \leq f < \infty$.*

For any θ in Θ and θ_0 in Θ_0 let

$$K(\theta, \theta_0) = -\int_X \log\,[f(x, \theta_0)/f(x, \theta)]\, dP_\theta. \tag{1}$$

K is one of the information numbers introduced by Kullback and Leibler [5], [6]. It is easily seen that K is well-defined by (1); $0 \leq K \leq \infty$; $K = 0$ if and only if $P_\theta = P_{\theta_0}$ on $\mathcal{B}$; and $K < \infty$ implies that P_θ is absolutely continuous with respect to P_{θ_0}. Even if P_θ and P_{θ_0} are mutually absolutely continuous, K can be infinite.

ASSUMPTION 2. *For each θ in $\Theta - \Theta_0$ and θ_0 in Θ_0 such that $K(\theta, \theta_0) < \infty$, there exists a $t = t(\theta, \theta_0) > 0$ such that $\int_X [f(x, \theta)/f(x, \theta_0)]^t \cdot dP_\theta < \infty$.*

If $K(\theta, \theta_0) < \infty$, then $0 < f(x, \theta)/f(x, \theta_0) < \infty$ with probability one when θ obtains, so that the integral in the statement of assumption 2 is well defined for every t.

Let

$$J(\theta) = \inf\,\{K(\theta, \theta_0) : \theta_0 \in \Theta_0\}, \qquad \rho(\theta) = \exp\,[-J(\theta)]. \tag{2}$$

As stated in the introduction, in typical cases $0 < \rho < 1$ for θ in $\Theta - \Theta_0$, but we shall include the cases $J = 0$ and $J = \infty$ in the discussion because theorem 1 [theorem 2] is not entirely vacuous in case $J = 0$ [$J = \infty$].

Now let $s = (x_1, x_2, \cdots, \text{ad inf})$ be a sequence of independent and identically distributed observations on x. The probability distribution of s in its sample space when θ obtains is denoted by $P_\theta^{(\infty)}$, but we shall usually abbreviate $P_\theta^{(\infty}$ to P_θ.

For each $n = 1, 2, \cdots$, let $T_n(s)$ be an extended real-valued measurable function of s such that T_n depends on s only through $(x_1, \cdots, x_n)$. For each θ let $F_n(t, \theta)$ denote the left-continuous probability distribution function of T_n when θ obtains; that is,

$$F_n(t, \theta) = P_\theta(T_n(s) < t), \tag{3}$$

and let

$$G_n(t) = \inf\,\{F_n(t, \theta) : \theta \in \Theta_0\}, \qquad (-\infty \leq t \leq \infty). \tag{4}$$

Define

$$L_n(s) = 1 - G_n(T_n(s)). \tag{5}$$

For any ϵ with $0 < \epsilon < 1$ let $N(\epsilon, s)$ = the least positive integer m such that $L_n \leq \epsilon$ for all $n \geq m$, and let $N(\epsilon, s) = +\infty$ if no such m exists. As just defined, N is then the sample size required in order that the sequence $\{T_n\}$ of test statistics becomes (and remains) significant at the level ϵ.

The following theorem 1 is a generalization and extension of theorem 4.1 of [4] in the following respects: the null hypothesis is not necessarily simple, and no restrictions other than measurability are imposed on the sequence $\{T_n\}$.

THEOREM 1. *For each* θ *in* $\Theta - \Theta_0$

$$\liminf_{n\to\infty} \frac{1}{n} \log L_n(s) \geq -J(\theta), \tag{6}$$

and

$$\liminf_{\epsilon\to 0} \frac{N(\epsilon, s)}{\log\left(\frac{1}{\epsilon}\right)} \geq \frac{1}{J(\theta)} \tag{7}$$

with probability one when θ *obtains.*

It follows from (6) that, for each nonnull θ,

$$\liminf_{n\to\infty} \frac{1}{n} \log E_\theta(L_n) \geq -J(\theta) \tag{8}$$

and

$$\lim_{n\to\infty} P_\theta(L_n > r^n) = 1 \qquad \text{if} \quad 0 < r < \rho(\theta). \tag{9}$$

The conclusions (8) and (9) are more useful than (6) or (7) in case L_n does not necessarily tend to 0 with probability one in the nonnull case.

For each n, let λ_n be the likelihood ratio statistic; that is,

$$\lambda_n(s) = \frac{\sup\left\{\prod_{i=1}^n f(x_i, \theta_0) : \theta_0 \in \Theta_0\right\}}{\sup\left\{\prod_{i=1}^n f(x_i, \theta) : \theta \in \Theta\right\}}. \tag{10}$$

In case the numerator and denominator in (10) are both 0, or both ∞, let $\lambda_n = 1$. Then λ_n is well-defined, with $0 \leq \lambda_n \leq 1$. It is assumed that λ_n is measurable for each n.

Since small values of λ_n are significant, we consider instead an equivalent statistic, $\hat{T}_n$ say, such that $\hat{T}_n$ is a strictly decreasing function of λ_n for each n. The particular choice of $\hat{T}_n$ is immaterial since only the exact levels attained are being considered, and we choose

$$\hat{T}_n(s) = -n^{-1} \log \lambda_n(s) \tag{11}$$

mainly because this choice facilitates some of the writing. Let $\hat{F}_n$, $\hat{G}_n$, and $\hat{L}_n$ be defined by (3), (4), and (5) by taking T_n to be $\hat{T}_n$, and let $\hat{N}$ be determined as above by the sequence $\{\hat{L}_n\}$.

Suppose now that, in addition to assumptions 1 and 2, assumptions 3–6 of section 5 are also satisfied.

THEOREM 2. *For each* θ *in* $\Theta - \Theta_0$

$$\lim_{n\to\infty} \frac{1}{n} \log \hat{L}_n(s) = -J(\theta), \tag{12}$$

and

$$\lim_{\epsilon\to 0} \frac{\hat{N}(\epsilon, s)}{\log\left(\frac{1}{\epsilon}\right)} = \frac{1}{J(\theta)} \tag{13}$$

with probability one when θ *obtains.*

It follows from (7) and (13) that for any given sequence $\{T_n\}$, the resulting sample size N required to attain the level ϵ satisfies

$$\liminf_{\epsilon \to 0} \frac{N(\epsilon, s)}{\hat{N}(\epsilon, s)} \geq 1 \tag{14}$$

with probability one whenever a nonnull θ with $0 < J(\theta) < \infty$ obtains.

It follows from (12) that for each nonnull θ,

$$\lim_{n \to \infty} \frac{1}{n} \log E_\theta(\hat{L}_n) = -J(\theta) \tag{15}$$

and

$$\lim_{n \to \infty} P_\theta(r_1^n < \hat{L}_n < r_2^n) = 1 \qquad \text{if} \quad r_1 < \rho(\theta) < r_2. \tag{16}$$

The likelihood ratio statistic is sometimes defined to be the right-hand side of (10) but with Θ replaced with $\Theta - \Theta_0$ in the denominator. This modified definition of λ_n is usually, but not always, equivalent to the definition (10). It can be seen from section 5 that under the same assumptions 1–6, theorem 2 holds also for the modified $\hat{T}_n$.

3. Remarks

(a) Let us say that a sequence $\{T_n\}$ is optimal when a given $\theta \in \Theta - \Theta_0$ obtains if, with L_n the level attained by T_n, $n^{-1} \log L_n \to -J(\theta)$ with probability one. According to theorems 1 and 2, this definition of optimality is plausible and $\{\hat{T}_n\}$ is an optimal sequence for every nonnull θ. Optimality in the present sense is, however, a rather weak property and is enjoyed, presumably, by a fairly wide class of statistics. An example of an optimal sequence other than $\{\hat{T}_n\}$ has already been mentioned at the end of section 2, and other examples are described in the following remarks (b) and (c). Further comparison of two optimal sequences requires, in general, an analysis very much deeper than is available at present. A similar difficulty arises in a theory of estimation closely related to the stochastic comparison of tests (cf. [4], section 6).

(b) The optimal exponential rate of convergence of levels, namely ρ^n, depends on the null set Θ_0 and on the particular alternative θ in $\Theta - \Theta_0$ under consideration, but not on the entire set of alternatives $\Theta - \Theta_0$. It follows, in particular, that if Δ is a subset of $\Theta - \Theta_0$, and if $T_n^* = -n^{-1} \log \lambda_n^*$, where λ_n^* is the likelihood ratio statistic for testing Θ_0 against Δ, then $\{T_n^*\}$ and $\{\hat{T}_n\}$ are both optimal sequences whenever a θ in Δ obtains. To consider the matter from another viewpoint, suppose that the initial nonnull set $\Theta - \Theta_0$ is enlarged to a set Σ by admitting certain additional nonnull distributions, and suppose that assumptions 1–6 are satisfied in the enlarged framework. Let $T_n^0 = -n^{-1} \log \lambda_n^0$, where λ_n^0 is the likelihood ratio statistic for testing Θ_0 against Σ. Then $\{T_n^0\}$ is an optimal sequence everywhere on Σ and hence also on $\Theta - \Theta_0$. Presumably, however, closer analysis will show that when a θ in $\Theta - \Theta_0$ obtains, T_n^0 is distinctly inferior to

$\hat{T}_n$ in the sense that $L_n^0 > \hat{L}_n$ with probability one for all sufficiently large n. This last is the case, for example, if X is the real line, P_θ denotes the normal distribution with mean θ and variance 1, $\Theta = [0, \infty)$, $\Theta_0 = \{0\}$, and $\Sigma = (-\infty, \infty) - \{0\}$; in this example, $\hat{L}_n = \frac{1}{2}L_n^0$ for all sufficiently large n when a positive θ obtains.

(c) Suppose that the maximum likelihoods in (10) are replaced by average likelihoods over Θ_0 and Θ with respect to appropriate averaging distributions. Then, under certain conditions, the resulting statistic remains optimal against each Θ in $\Theta - \Theta_0$. This important remark was suggested by Dr. P. J. Bickel at the reading of this paper at the Symposium. Dr. Bickel and the author hope to present an adequate treatment of the remark elsewhere, but it may be worthwhile to state here the following. Suppose that assumptions 1–6 and some additional assumptions are satisfied. Then $\Theta - \Theta_0$ and Θ_0 are metric spaces. Let ξ be a fixed prior probability distribution such that each neighborhood of each point in either space has positive probability. For each n let $\pi_n(s)$ be the posterior probability of Θ_0 given $(x_1, \cdots, x_n)$, and let $\overline{T}_n(s) = n^{-1} \log [(1 - \pi_n)/\pi_n]$. Then the relevant asymptotic properties (cf. (19) and (20) below) of $\overline{T}_n$ are exactly the same as those of $\hat{T}_n$.

(d) For given n and s let $L_n(s)$ defined by (3), (4), and (5) be written temporarily as $L_n(s, T_n)$ to indicate its dependence on T_n. Let $M_n(s) = \inf \{L_n(s, T_n)\}$, the infimum being taken over the class of all measurable statistics T_n which depend on s only through $(x_1, \cdots, x_n)$. Although M_n is not the level attained by any statistic (that is, there exists no T_n such that $M_n(s) = L_n(s, T_n)$ for all s), it is of some theoretical interest to study the behavior of M_n. We consider two special cases.

Suppose first that for each x in X the set $\{x\}$ is $\mathcal{B}$-measurable and $P_\theta(\{x\}) = 0$ for all θ in Θ_0. In this case $M_n(s) = 0$ for all s and all n.

Suppose next that X is a finite set and that $\mathcal{B}$ is the class of all subsets of X. In this case $M_n(s) = \sup \{\prod_{i=1}^n f(x_i, \theta) : \theta \in \Theta_0\}$ where $f(x, \theta) = P_\theta(\{x\})$. It follows hence by lemma 4 of section 5 that

$$\lim_{n \to \infty} \frac{1}{n} \log M_n = -J(\theta) - H(\theta) \tag{17}$$

with probability one when θ obtains, where

$$H(\theta) = -\sum_{x \in X} f(x, \theta) \log f(x, \theta) \tag{18}$$

is the Shannon information number. It follows that with $N'(\epsilon, s)$ the sample size required to make $M_n \leq \epsilon$, we have $N' \leq \hat{N}$ and $\lim_{\epsilon \to 0} \{N'/\hat{N}\} = J(\theta)/[J(\theta) + H(\theta)]$ with probability one in the nonnull case. If X contains k points, $H(\theta) \leq \log k$, so that $J/[J + H] \leq J/[J + \log k]$ for all θ.

To consider a simple example, suppose that X consists of the two points 0 and 1, $\Theta = (0, 1)$, $P_\theta(\{1\}) = 1 - P_\theta(\{0\}) = \theta$, and $\Theta_0 = \{\frac{1}{2}\}$. In this example, J and H are functions of $|\theta - \frac{1}{2}|$, and the values of $J/[J + H]$ for $|\theta - \frac{1}{2}| = .0(.1).5$ are .00, .03, .12, .27, .53, and 1.00.

(e) Assumptions 1 and 2 of theorem 1 are very weak and even these can be dispensed with to a certain extent (cf. the last paragraph of section 4). Unfortunately, some of the additional assumptions 3–6 required by the present proof of theorem 2 are quite restrictive, and what is perhaps worse, it is often difficult to determine whether they hold in a given case. The troublesome assumptions include versions of the compactifiability and integrability conditions introduced by Wald [7] in his proof of the consistency of maximum likelihood estimates. As is pointed out in [8], it is often difficult and sometimes impossible to verify such conditions, even in certain apparently simple cases where the estimates themselves are visibly consistent, and the likelihood function behaves as it should. It may be added here that at least some of the conditions embodied in assumptions 3–6 are indispensable to a general proof of theorem 2; this may be seen from [9].

In many examples it is a relatively simple matter to show directly that (12) and (13) are satisfied, as follows. First it is shown that

$$\hat{T}_n \to J(\theta) \tag{19}$$

with probability one when θ obtains. Next it is shown that the distribution function $\hat{G}_n$ satisfies the following condition: for each positive t in some neighborhood of J,

$$\frac{1}{n} \log [1 - \hat{G}_n(t)] \to -t \quad \text{as} \quad n \to \infty. \tag{20}$$

It is then immediate from (19) and (20) that (12) holds, and (12) implies (13). Of course, (20) is not quite necessary for (12); in fact, there are simple examples where even assumptions 1–6 hold but (20) as stated does not.

The proof of (19) is troublesome in the general case (cf. section 5) but quite trivial in many examples. Proofs of (20), or of versions thereof, are always nontrivial since (20) is an assertion about very small tail probabilities of the exact null distribution of $\hat{T}_n$.

The present regularity assumptions give little or no trouble in certain fairly general circumstances. Assumptions 1–6 are satisfied in case X is a finite set (that is, the multinomial case) no matter what Θ and Θ_0 may be, provided that $\theta_1 \neq \theta_2$ implies $P_{\theta_1} \neq P_{\theta_2}$ for θ_1 and θ_2 in Θ. (This last proviso is harmless in the present context.) Only assumption 2 requires verification in case Θ is a finite set, no matter what X may be. Assumptions 2–6 are usually satisfied but require verification in case Θ is an interval on the real line, assumption 1 holds, and $f(x, \theta)$ is continuous in θ over Θ for each x.

It is worthwhile to note that the regularity conditions under discussion do not include conditions required by the asymptotic null distribution theory of maximum likelihood and likelihood ratios; consequently, the present conditions are satisfied in many so-called irregular cases. For example, if X is the real line, $\Theta = (-\infty, +\infty)$, $\Theta_0 = \{0\}$, and P_θ represents the uniform distribution over $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, then assumptions 1–6 hold with $J(\theta) = \infty$ for each nonnull θ. In this example there exists a random variable $m = m(s)$ with $1 \leq m \leq \infty$ such that $P_\theta(m < \infty) = 1$ for each nonnull θ, and such that $\hat{L}_n = 1$ for $n < m$ and $\hat{L}_n = 0$ for $n \geq m$ for every s; hence, $\hat{N} = m$ for every ϵ and s.

(f) As pointed out in [2], [3], and [4], stochastic comparison has several connections with power function considerations. In particular, theorems 1 and 2 can be shown to yield the following conclusions concerning the asymptotic properties of critical regions. Consider a particular nonnull θ. For each n, let W_n be a critical region in the sample space of $(x_1, \cdots, x_n)$ such that $P_\theta(W_n) \to p$ as $n \to \infty$, where $0 < p < 1$. Let $\alpha_n = \sup \{P_{\theta_0}(W_n): \theta_0 \in \Theta_0\}$ be the size of W_n. Then $\lim \inf_{n\to\infty} n^{-1} \log \alpha_n \geq -J(\theta)$. Next, let $\hat{W}_n$ be a critical region of the form $\{s: \hat{T}_n \geq \hat{k}_n\}$, with the constants $\hat{k}_n$ chosen so that $P_\theta(\hat{W}_n) \to \hat{p}$ where $0 < \hat{p} < 1$; then $n^{-1} \log \hat{\alpha}_n \to -J(\theta)$. In other words, if the power of the critical region against a given alternative is held fixed, the rate of convergence to zero of the resulting size is optimal for regions based on $\hat{T}_n$. Related but much deeper optimality conclusions concerning critical regions based on $\hat{T}_n$ have been obtained previously by Hoeffding [10] in the case when X is a finite set.

4. Proof of theorem 1

The following lemma 1 is required in the proofs of theorems 1 and 2. Let z be an extended real-valued random variable such that $P(-\infty \leq z < \infty) = 1$, and let $\varphi(t) = E(e^{tz})$ be the moment generating function (m.g.f.) of z, $0 \leq \varphi \leq \infty$.

LEMMA 1. *Let n be a positive integer, and let $z_1, \cdots, z_n$ be mutually independent replicates of z. Then $P(z_1 +, \cdots, + z_n \geq 0) \leq [\varphi(t)]^n$ for $t > 0$.*

PROOF. The lemma (and much more) is well known (cf. [11], [12], [13]), but for the sake of completeness we include here the proof given in [11]. Let $Z_n = \sum_{i=1}^n z_i$. Then $P(Z_n \geq 0) = P(\exp(tZ_n) \geq 1) \leq E(\exp(tZ_n)) = [\varphi(t)]^n$.

Now choose and fix a θ in $\Theta - \Theta_0$, a θ_0 in Θ_0, and an $\epsilon > 0$. Let $r_1 = \exp[-K(\theta, \theta_0) - \epsilon]$, $0 \leq r_1 < 1$. Let W_n denote an event which depends on s only through $x_1, \cdots, x_n$. The following lemma is closely related to a theorem of C. Stein (cf. [6], pp. 76–77).

LEMMA 2. *There exists $r_2 = r_2(\theta, \theta_0, \epsilon)$, $0 < r_2 < 1$, such that for each n and W_n,*

$$P_{\theta_0}(W_n) \geq r_1^n[P_\theta(W_n) - r_2^n]. \tag{21}$$

PROOF. Consider a fixed n. If $K = \infty$, then $r_1 = 0$ and (21) holds trivially with $r_2 = \frac{1}{2}$ (say). Suppose then that $K < \infty$. Let

$$A_n = \left\{s: \prod_{i=1}^n f(x_i, \theta_0) \geq r_1^n \prod_{i=1}^n f(x_i, \theta)\right\}. \tag{22}$$

Then

$$\begin{aligned} P_{\theta_0}(W_n) &\geq P_{\theta_0}(A_n \cap W_n) \\ &\geq r_1^n P_\theta(A_n \cap W_n) \qquad \text{by (22)} \\ &\geq r_1^n[P_\theta(W_n) - [1 - P_\theta(A_n)]]. \end{aligned} \tag{23}$$

Now consider the random variable $y = \log [f(x, \theta)/f(x, \theta_0)]$ when θ obtains; y is well-defined and $P_\theta(-\infty < y < \infty) = 1$. The m.g.f. of y is ≤ 1 at $t = -1$, and is finite for a positive t by assumption 2. Thus the m.g.f. of y is finite in a neighborhood of $t = 0$. Let $z = y - K - \epsilon$. Then the m.g.f. of z, $\varphi(t)$ say, is finite in a neighborhood of $t = 0$ and $\varphi'(0) = E_\theta(z) = -\epsilon < 0$ by (1). Since $\varphi(0) = 1$, there exists a $t_2 > 0$ such that with $r_2 = \varphi(t_2)$ we have $0 < r_2 < 1$. It follows from (22) that, in an obvious notation, $1 - P_\theta(A_n) = P_\theta\left(\sum_{i=1}^n z_i > 0\right)$. Hence

$$1 - P_\theta(A_n) \leq r_2^n \tag{24}$$

by lemma 1. It is plain from (23) and (24) that (21) holds.

Let there be given a sequence of measurable statistics T_n as in section 2. By putting $W_n = \{s: T_n \geq t\}$ in (21) it follows from (3) that

$$1 - F_n(t, \theta_0) \geq r_1^n[1 - F_n(t, \theta) - r_2^n] \tag{25}$$

for all t and all n.

LEMMA 3. *With probability one when θ obtains,*

$$1 - F_n(T_n(s), \theta) \geq n^{-2} \tag{26}$$

for all sufficiently large n.

PROOF. It is easily verified that if T is an extended real-valued random variable, and $F(t) = P(T < t)$, then $P(1 - F(T) < r) \leq r$ for all r in $[0, 1]$. It follows hence that $\sum_{n=1}^\infty P_\theta(1 - F_n(T_n, \theta) < n^{-2}) \leq \sum_{n=1}^\infty n^{-2} < \infty$.

PROOF OF THEOREM 1. Choose and fix a θ in $\Theta - \Theta_0$. Let $B = B(\theta)$ be the set of all s such that (26) holds for all sufficiently large n. Then B is a measurable set with $P_\theta^{(\infty)}(B) = 1$. We shall show that (6) and (7) hold for each s in B. Choose and fix an s in B, and let $m = m(s)$ be an integer such that (26) holds for all $n \geq m$.

Let θ_0 be a point in Θ_0 and let ϵ be a positive constant. For each n let $t = T_n(s)$ in (25). It then follows from (25) and (26) that $1 - F_n(T_n(s), \theta_0) \geq r_1^n[n^{-2} - r_2^n]$ for $n \geq m$. Since $L_n(s)$ defined by (3), (4), and (5) cannot be less than $1 - F_n(T_n(s), \theta_0)$, it follows that $L_n(s) \geq r_1^n[n^{-2} - r_2^n]$ for $n \geq m$. Hence

$$\liminf_{n\to\infty} \frac{1}{n} \log L_n(s) \geq -K(\theta, \theta_0) - \epsilon \tag{27}$$

by the definition of r_1. Since θ_0 and ϵ in (27) are arbitrary, (6) holds.

Since (7) holds trivially if $J = \infty$, suppose that $0 \leq J < \infty$. It then follows from (6) that $L_n > 0$ for all sufficiently large n. If $\limsup_{n\to\infty} L_n(s) > 0$, then $N = \infty$ for all sufficiently small ϵ and (7) again holds trivially. Suppose then that $\lim_{n\to\infty} L_n = 0$. In this case $1 \leq N < \infty$ for all ϵ; $N \to \infty$ through a subsequence of the integers as $\epsilon \to 0$; and $L_N \leq \epsilon$ for all ϵ. It follows hence that

$$\overline{\lim_{\epsilon\to 0}} \{N^{-1} \log (1/\epsilon)\} \leq \overline{\lim_{\epsilon\to 0}} \{-N^{-1} \log L_N\} \leq \overline{\lim_{n\to\infty}} \{-n^{-1} \log L_n\} \leq J(\theta) \tag{28}$$

by (6), and this establishes (7). This completes the proof of theorem 1.

It is plain from the preceding proof that assumptions 1 and 2 can be weakened considerably. Indeed, there is a version of theorem 1 which holds without any regularity assumptions whatsoever. To describe this version, for any θ and θ_0 in Θ let $K(\theta, \theta_0) = \int_X [\log (dP_\theta/dP_{\theta_0})] \, dP_\theta$ if P_θ is absolutely continuous with respect to P_{θ_0}, and let $K = \infty$ otherwise. Let J be defined by (2). It then follows by a slight modification of the preceding proof (using the law of large numbers instead of lemma 1, and $E_\theta[1 - F_n(T_n, \theta)] \geq \frac{1}{2}$ instead of lemma 3) that (8) holds for each nonnull θ. It follows from (8) that (6) and (7) are satisfied with both inferior limits replaced by superior limits. It would be interesting to know whether theorem 1 as stated holds (with the present definition of J) without any assumptions whatsoever.

5. Proof of theorem 2

We shall first state the additional assumptions required of the given framework X, $\mathcal{B}$, $\{P_\theta : \theta \in \Theta\}$, $\Theta_0 \subset \Theta$, and $f(x, \theta) = dP_\theta/d\lambda$. In order to avoid needless loss of generality, most of these assumptions are stated below in more or less the forms required by the proof itself. Certain stronger but more readily verifiable conditions are also given.

Let $\bar{\Theta}$ be a metric space of points θ, and let δ denote the given metric on $\bar{\Theta}$. We shall say that $\bar{\Theta}$ is a suitable compactification of Θ if the following conditions (i)–(iv) are satisfied: (i) $\bar{\Theta}$ is compact; (ii) $\Theta \subset \bar{\Theta}$, and Θ is everywhere dense in $\bar{\Theta}$; (iii) for each $\theta \in \bar{\Theta}$ there exists $d_1 = d_1(\theta) > 0$ such that, for each d in $(0, d_1)$,

$$g(x, \theta, d) = \sup \{f(x, \theta_1) : \theta_1 \in \Theta, \delta(\theta, \theta_1) < d\} \tag{29}$$

is $\mathcal{B}$-measurable, $0 \leq g \leq \infty$; and (iv) for each $\theta \in \bar{\Theta}$,

$$\int_X g(x, \theta, 0) \, d\lambda \leq 1, \tag{30}$$

where $g(x, \theta, 0) = \lim_{d \to 0} g(x, \theta, d)$. In typical cases $g(x, \theta, 0) = f(x, \theta)$ for $\theta \in \Theta$, so that g is an extension of the given function f on $X \times \Theta$ to a function on $X \times \bar{\Theta}$.

A slightly different formulation of the notion of suitable compactification, and many nontrivial examples, are given in [8].

Assumption 3. *There exists a suitable compactification of* Θ, *say* $\bar{\Theta}$. *With* $\bar{\Theta}_0$ *the closure of* Θ_0 *in* $\bar{\Theta}$, $\bar{\Theta}_0$ *is a suitable compactification of* Θ_0.

The second part of this assumption is to the effect that, for each $\theta_0 \in \bar{\Theta}_0$,

$$g_0(x, \theta_0, d) = \sup \{f(x, \theta_1) : \theta_1 \in \Theta_0, \delta(\theta_0, \theta_1) < d\} \tag{31}$$

is $\mathcal{B}$-measurable for all sufficiently small $d > 0$. With $g_0(x, \theta_0, 0) = \lim_{d \to 0} g_0(x, \theta_0, d)$, it is plain from (29) and (31) that $g_0(x, \theta_0, 0) \leq g(x, \theta_0, 0)$ for all x and $\theta_0 \in \bar{\Theta}_0$; consequently, in view of (30), the required condition

$$\int_X g_0(x, \theta_0, 0) \, d\lambda \leq 1 \tag{32}$$

is automatically satisfied.

For any $\theta \in \Theta$ and $\theta_0 \in \bar{\Theta}_0$, let $\bar{K}(\theta, \theta_0)$ be defined by (1) with $f(x, \theta_0)$ replaced by $g_0(x, \theta_0, 0)$. It follows from (32) that $\bar{K}$ is well-defined and $0 \leq \bar{K} \leq \infty$. Since g_0 may be thought of as an extension of the function $f(x, \theta_0)$ on $X \times \Theta_0$ to $X \times \bar{\Theta}_0$, $\bar{K}$ is to be thought of as an extension of K on $\Theta \times \Theta_0$ to $\Theta \times \bar{\Theta}_0$. An alternative method of extending K is to use g instead of g_0, but the present approach is preferable in that the following assumptions 4 and 5 are weaker than the corresponding assumptions in terms of g.

ASSUMPTION 4. *For each θ in $\Theta - \Theta_0$,*

$$J(\theta) = \inf \{\bar{K}(\theta, \theta_0) : \theta_0 \in \bar{\Theta}_0\}. \tag{33}$$

It is plain from (2) that (33) holds if $\bar{K}$ is indeed an extension of K and if, for the given θ, $\bar{K}(\theta, \theta_0)$ is either continuous in θ_0 over $\bar{\Theta}_0$, or $= \infty$ for θ_0 in $\bar{\Theta}_0 - \Theta_0$.

ASSUMPTION 5. *For given θ in $\Theta - \Theta_0$ and θ_0 in $\bar{\Theta}_0$, there exists $d = d(\theta, \theta_0) > 0$ such that*

$$\int_X \log^+ [g_0(x, \theta_0, d)/f(x, \theta)] \, dP_\theta < \infty. \tag{34}$$

Assumptions 4 and 5 are automatically satisfied if Θ_0 is a finite set, and in particular, if the null hypothesis is simple.

It is convenient to restate assumption 5 here as follows. For given $\theta \in \Theta - \Theta_0$ and $\theta_0 \in \bar{\Theta}_0$, let d be restricted to sufficiently small values so that $g_0(x, \theta_0, d)$ is measurable. Consider

$$y_0 = y_0(x, \theta_0, d : \theta) = \log [g_0(x, \theta_0, d)/f(x, \theta)] \tag{35}$$

when θ obtains. Then y_0 is well-defined and $-\infty \leq y_0 < \infty$ with probability one. The condition (34) is that $E_\theta(y_0)$ exists and $-\infty \leq E_\theta(y_0) < \infty$. Since $g_0(x, \theta_0, d)$ decreases to $g_0(x, \theta_0, 0)$ as d decreases to zero, and since $-\bar{K}$ is by definition the expected value of $y_0(x, \theta_0, 0 : \theta)$ when θ obtains, (34) implies (and is implied by)

$$\lim_{d \to 0} E_\theta(y_0(x, \theta_0, d : \theta)) = -\bar{K}(\theta, \theta_0), \tag{36}$$

even if $\bar{K} = \infty$. (Cf. [8], section 2.)

ASSUMPTION 6. *Given τ, $0 < \tau < 1$, $\epsilon > 0$, and θ in $\bar{\Theta}$, there exists $d = d(\tau, \epsilon, \theta) > 0$ such that*

$$\int_X [g(x, \theta, d)/f(x, \theta_0)]^\tau \, dP_{\theta_0} < 1 + \epsilon \tag{37}$$

for all θ_0 in Θ_0.

In order to discuss this assumption, consider a particular $\theta_0 \in \Theta_0$ and sufficiently small $d > 0$. Consider

$$y = y(x, \theta, d : \theta_0) = \log [g(x, \theta, d)/f(x, \theta_0)] \tag{38}$$

when θ_0 obtains. Then y is well-defined and $-\infty \leq y < \infty$ with probability one. Let the integral in (37) be denoted by $\psi(\tau | \theta, d, \theta_0)$; ψ is the m.g.f. of y. It follows from the convexity of m.g.f.'s that for $0 < \tau < 1$, $\psi(\tau)$ cannot exceed

$\max\{P_{\theta_0}(y > -\infty),\ \int_X g(x, \theta, d)\, d\lambda\}$. Hence $\psi(\tau) \leq \max\{1, \int_X g(x, \theta, d)\, d\lambda\}$; this bound does not depend on θ_0 (or on τ). We observe next that

$$\int_X g(x, \theta, d)\, d\lambda \to \int_X g(x, \theta, 0)\, d\lambda \leq 1 \quad \text{as} \quad d \to 0,$$

provided that

$$\int_X g(x, \theta, d_1)\, d\lambda < \infty \qquad \text{for some} \quad d_1 = d_1(\theta) > 0. \tag{39}$$

It follows that (39) is a sufficient condition for the validity of assumption 6 at the given $\theta \in \bar{\Theta}$, no matter what the null set Θ_0 may be. Condition (39) is satisfied if, for example, X is a countable set, $\mathcal{B}$ is the class of all subsets of X, and there exists $h(x)$ such that $P_\theta(\{x\}) \leq h(x)$ for all θ and all x, and $\sum_x h(x) < \infty$. It is plain that condition (39) is satisfied whenever Θ is a finite set.

Condition (39) is, however, much stronger than is generally necessary. To obtain weaker or different sufficient conditions, suppose that $\psi(\tau|\theta, d, \theta_0) < \infty$ for some $d > 0$. It then follows that

$$\lim_{d \to 0} \psi(\tau|\theta, d, \theta_0) = \psi(\tau|\theta, 0, \theta_0). \tag{40}$$

The right-hand side in (40) is the m.g.f. of $y(x, \theta, 0{:}\theta_0)$. Since this last m.g.f. does not exceed one for $0 < \tau < 1$, assumption 6 will hold at the given θ if (40) holds uniformly for θ_0 in Θ_0. Uniformity is guaranteed by Dini's theorem if for each d in some interval $[0, d_1)$ with $d_1 > 0$, $\psi(\tau|\theta, d, \theta_0)$ is continuous in θ_0 over Θ_0 and has a continuous extension to $\bar{\Theta}_0$, and (40) holds for the extended functions for each θ_0 in $\bar{\Theta}_0$. This last condition is satisfied, in particular, if there exists a $\mathcal{B}$-measurable $h(x)$ such that $f(x, \theta_0) \leq h(x)$ for all x and all $\theta_0 \in \Theta_0$ and such that $\int_X [g(x, \theta, d_1)]^\tau [h(x)]^{1-\tau}\, d\lambda < \infty$, and if $g_0(x, \theta_0, 0)$ is an extension of $f(x, \theta_0)$ and is continuous over $\bar{\Theta}_0$ for each x.

We proceed to establish theorem 2. Assumptions 3, 4, and 5 are used to obtain lemma 4 below, and assumptions 3 and 6 to obtain lemma 5. Theorem 2 is a straightforward consequence of theorem 1 and lemmas 4 and 5.

For any set $\Gamma \subset \bar{\Theta}$ such that $\Gamma \cap \Theta$ is nonempty and any $\theta \in \Theta$, let

$$\begin{aligned} R_n(\Gamma, \theta) &= R_n(s{:}\Gamma, \theta) \\ &= n^{-1} \log \frac{\sup\left\{\prod_{i=1}^n f(x_i, \theta_1) : \theta_1 \in \Gamma \cap \Theta\right\}}{\left\{\prod_{i=1}^n f(x_i, \theta)\right\}}. \end{aligned} \tag{41}$$

R_n is well-defined (with $-\infty \leq R_n \leq \infty$) with probability one when θ obtains. It is not required, however, that R_n be a measurable function of s.

LEMMA 4. *For each* $\theta \in \Theta - \Theta_0$,

$$R_n(\Theta_0, \theta) \to -J(\theta) \tag{42}$$

with probability one when θ *obtains.*

PROOF. Choose and fix $\theta \in \Theta - \Theta_0$ and suppose θ obtains. Let $a > 0, b > 0$ be constants, and let $H = \max\{-J(\theta) + a, -b\}$. Let θ_0 be a point in $\bar{\Theta}_0$. According to (36), there exists $d = d(\theta_0) > 0$ such that, with $y_0(x)$ defined by (35), $E_\theta(y_0) < \max\{-\bar{K}(\theta, \theta_0) + a, -b\}$. Hence $E_\theta(y_0) \leq H$, by assumption 4. Let Γ be the open sphere in $\bar{\Theta}_0$ with center θ_0 and radius d, and let $\Gamma^0 = \Gamma \cap \Theta_0$. It is then plain from (31), (35), and (41) that $R_n(s:\Gamma^0, \theta) \leq n^{-1}\sum_{i=1}^n y_0(x_i)$ for every s and n. Hence $\limsup_{n\to\infty} R_n(\Gamma^0, \theta) \leq H$ with probability one.

Since $\bar{\Theta}_0$ is compact, we can find a finite number of spheres $\Gamma_1, \cdots, \Gamma_k$ such that $\bigcup_j \Gamma_j = \bar{\Theta}_0$, and such that the conclusion of the preceding paragraph holds for each $\Gamma_j^0 = \Gamma_j \cap \Theta_0$. Since $R_n(\Theta_0, \theta) = \max\{R_n(\Gamma_j^0, \theta): j = 1, \cdots, k\}$, it follows that $\limsup_{n\to\infty} R_n(\Theta_0, \theta) \leq H$ with probability one. Since a and b are arbitrary, we conclude that $\limsup_{n\to\infty} R_n(\Theta_0, \theta) \leq -J$ with probability one.

With θ_0 a point in Θ_0, $R_n(\Theta_0, \theta) \geq R_n(\{\theta_0\}, \theta)$ by (41); hence,

$$\liminf_{n\to\infty} R_n(\Theta_0, \theta) \geq -K(\theta, \theta_0)$$

with probability one, by (1) and (41). Since θ_0 is arbitrary, we see from (2) that $\liminf_{n\to\infty} R_n(\Theta_0, \theta) \geq -J(\theta)$ with probability one.

LEMMA 5. *Given $\epsilon > 0$ and τ, $0 < \tau < 1$, there exists a positive integer $k = k(\epsilon, \tau)$ such that*

$$(43) \qquad 1 - \hat{G}_n(t) \leq k \cdot (1 + \epsilon)^n \cdot e^{-n\tau t}$$

for all n and t.

PROOF. Let θ be a point in $\bar{\Theta}$, and $d = d(\theta) > 0$ be such that, with g defined by (29), (37) holds for all θ_0 in Θ_0. Let Γ denote the open sphere in $\bar{\Theta}$ with center θ and radius d.

Consider a particular $\theta_0 \in \Theta_0$ and suppose that θ_0 obtains. Let $y(x)$ be given by (38), and let ψ be the m.g.f. of y. According to (37), $\psi(\tau) < 1 + \epsilon$. It is plain from (29), (38), and (41) that $R_n(\Gamma, \theta_0) \leq n^{-1}\sum_{i=1}^n y(x_i) = S_n$, say. An application of lemma 1 (with $z = y - t$, $t = \tau$, and $\varphi(\tau) = \psi(\tau)\exp(-t\tau)$) shows that $P_{\theta_0}(S_n \geq t) \leq (1 + \epsilon)^n \exp(-n\tau t) = b_n(t)$, say, for all n and t.

Since $\bar{\Theta}$ is compact, we can find a finite number of open spheres $\Gamma_1, \cdots, \Gamma_k$ such that $\bigcup_j \Gamma_j = \bar{\Theta}$, and such that, for each $\theta_0 \in \Theta_0$ and j, there exists a random variable $S_{nj} = S_{nj}(\theta_0)$ with $R_n(\Gamma_j, \theta_0) \leq S_{nj}$ and $P_{\theta_0}(S_{nj} \geq t) \leq b_n(t)$. Now, it is clear from (10), (11), and (41) that, when a given θ_0 obtains,

$$(44) \qquad \begin{aligned} \hat{T}_n \leq R_n(\Theta, \theta_0) &= \max\{R_m(\Gamma_j, \theta_0): j = 1, \cdots, k\} \\ &\leq \max\{S_{nj}(\theta_0): j = 1, \cdots, k\}. \end{aligned}$$

Since $\hat{T}_n$ is measurable by assumption, it follows from (44) that

$$(45) \qquad P_{\theta_0}(\hat{T}_n \geq t) \leq \sum_j P_{\theta_0}(S_{nj} \geq t) \leq \sum_j b_n(t) = k \cdot b_n(t).$$

Thus

$$(46) \qquad 1 - \hat{F}_n(t, \theta_0) \leq k \cdot (1 + \epsilon)^n \cdot \exp(-n\tau t).$$

Since (46) holds for every finite t, it follows by letting $t \to \infty$ that $P_{\theta_0}(\hat{T}_n = \infty) = 0$, that is, (46) holds for $t = \infty$ also. Since θ_0 in (46) is arbitrary, we see from (3) and (4) that (43) holds.

PROOF OF THEOREM 2. Suppose that a given θ in $\Theta - \Theta_0$ obtains. Since $\hat{T}_n \geq -R_n(\Theta_0, \theta)$ by (10), (11), and (41), it follows from (42) that

$$\liminf_{n\to\infty} \hat{T}_n \geq J(\theta) \tag{47}$$

with probability one. It will be shown later that in fact (19) holds.

Choose ϵ and τ as in lemma 5. Since $\hat{L}_n \equiv 1 - \hat{G}_n(\hat{T}_n)$, we see from (43) that

$$n^{-1} \log \hat{L}_n \leq -\tau \hat{T}_n + n^{-1} \log k + \log (1 + \epsilon) \tag{48}$$

for every s and n. It follows from (47) and (48) that $\limsup_{n\to\infty} \{n^{-1} \log \hat{L}_n\} \leq -\tau J(\theta) + \log (1 + \epsilon)$ with probability one. Since ϵ and τ are arbitrary, $\limsup \{n^{-1} \log \hat{L}_n\} \leq -J(\theta)$ with probability one. Theorem 1 applied to $\hat{T}_n$ now shows that (12) holds with probability one.

If $J = 0$ for the given θ, theorem 1 applied to $\hat{T}_n$ shows that (13) holds with probability one. Suppose then that $0 < J \leq \infty$, and choose and fix an s such that (12) is satisfied. Suppose first that $\hat{L}_n = 0$ for all sufficiently large n. Then $\hat{N}$ is a bounded function of ϵ, and $J = \infty$ by (12), so (13) holds. Suppose now that $\hat{L}_n > 0$ for infinitely many n. It is plain from (12) and $J > 0$ that $\hat{L}_n \to 0$. Consequently, $1 \leq \hat{N}(\epsilon, s) < \infty$ for every ϵ; $\hat{N}$ increases to ∞ through a subsequence of the integers as ϵ decreases to zero; and

$$\hat{L}_{\hat{N}-1} > \epsilon \geq \hat{L}_{\hat{N}} \tag{49}$$

for all ϵ such that $\hat{N} \geq 2$. It follows easily from (49) by using (12) that $\lim_{\epsilon\to 0} \{\hat{N}^{-1} \log (1/\epsilon)\} = J$. This completes the proof of theorem 2.

It may be worthwhile to note that the present assumptions imply that (19) holds with probability one. Choose ϵ and τ as in lemma 5. It follows from (48) by theorem 1 applied to $\hat{T}_n$ that

$$\liminf_{n\to\infty} (-\tau \hat{T}_n) + \log (1 + \epsilon) \geq -J(\theta) \tag{50}$$

with probability one. Since ϵ and τ are arbitrary, (50) implies that $\limsup \hat{T}_n \leq J(\theta)$ with probability one, and (19) now follows from (47).

In view of (42), (19) is equivalent to

$$R_n(\Theta, \theta) \to 0 \tag{51}$$

in the case when $J(\theta) < \infty$. Condition (51) is of the same formal structure as (42), since J vanishes when Θ_0 is replaced by Θ on the right-hand side of (2). It follows that a direct proof of (51) (and thereby of (19)) can be given along the lines of the proof of lemma 4. This direct proof requires, however, that the integrability condition of assumption 5 hold for each θ_0 in $\bar{\Theta}$ and with g_0 replaced by g.

REFERENCES

[1] J. NEYMAN and E. S. PEARSON, "On the use and interpretation of certain test criteria for purposes of statistical inference," *Biometrika*, Vol. 20-A (1928), pp. 175–240 and 264–299.

[2] R. R. BAHADUR, "Simultaneous comparison of the optimum and sign tests of a normal mean," *Contributions to Probability and Statistics*, Stanford, Stanford University Press, 1960, pp. 79–88.

[3] ———, "Stochastic comparison of tests," *Ann. Math. Statist.*, Vol. 31 (1960), pp. 276–295.

[4] ———, "Asymptotic efficiency of tests and estimates," *Sankhyā*, Vol. 22 (1960), pp. 229–252.

[5] S. KULLBACK and R. A. LEIBLER, "On information and sufficiency," *Ann. Math. Statist.*, Vol. 22 (1951), pp. 79–86.

[6] S. KULLBACK, *Information Theory and Statistics*, New York, Wiley, 1959.

[7] A. WALD, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, Vol. 20 (1949), pp. 595–601.

[8] J. KIEFER and J. WOLFOWITZ, "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *Ann. Math. Statist.*, Vol. 27 (1956), pp. 887–906.

[9] R. R. BAHADUR, "Examples of inconsistency of maximum likelihood estimates," *Sankhyā*, Vol. 20 (1958), pp. 207–210.

[10] W. HOEFFDING, "Asymptotically optimal tests for multinominal distributions," *Ann. Math. Statist.*, Vol. 36 (1965), pp. 369–408.

[11] H. CHERNOFF, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, Vol. 23 (1952), pp. 493–507.

[12] R. R. BAHADUR and R. RANGA RAO, "On deviations of the sample mean," *Ann. Math. Statist.*, Vol. 31 (1960), pp. 1015–1027.

[13] W. HOEFFDING, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 13–30.