ASYMPTOTICALLY OPTIMAL TESTS FOR MULTINOMIAL DISTRIBUTIONS[1]

By Wassily Hoeffding

University of North Carolina

Summary. Tests of simple and composite hypotheses for multinomial distributions are considered. It is assumed that the size α_N of a test tends to 0 as the sample size N increases. The main concern of this paper is to substantiate the following proposition: If a given test of size α_N is "sufficiently different" from a likelihood ratio test then there is a likelihood ratio test of size $\leqq \alpha_N$ which is considerably more powerful than the given test at "most" points in the set of alternatives when N is large enough, provided that $\alpha_N \to 0$ at a suitable rate. In particular, it is shown that chi-square tests of simple and of some composite hypotheses are inferior, in the sense described, to the corresponding likelihood ratio tests. Certain Bayes tests are shown to share the above-mentioned property of a likelihood ratio test.

1. Introduction. This paper is concerned with asymptotic properties of tests of simple and composite hypotheses concerning the parameter vector $p = (p_1, \cdots, p_k)$ of a multinomial distribution as the sample size N tends to infinity. In traditional asymptotic test theory the size α of the test is held fixed and its power is investigated at alternatives $p = p^{(N)}$ which approach the hypothesis set as $N \to \infty$, in such a way as to keep the error probability away from 0. These restrictions make it possible to apply the central limit theorem and its extensions. However, it seems reasonable to let the size α_N of a test tend to 0 as the number N of observations increases. It is also of interest to consider alternatives not very close to the hypothesis, at which, typically, the error probabilities will tend to zero. To attack these problems, the theory of probabilities of large deviations is needed. For the case of sums of independent real-valued random variables this theory is by now well developed. It has been used by Chernoff [2] to compare the performance of tests based on sums of independent, identically distributed random variables when the error probabilities tend to zero. Sanov [7] made an interesting contribution to a general theory of probabilities of large deviations. He studied the asymptotic behavior of the probability that the empirical distribution function is contained in a given set A of distribution functions when the true distribution function is not in A. For the special case of a multinomial distribution a slight elaboration of one of Sanov's results implies the following.

Let the random vector $Z^{(N)}$ take the values $z^{(N)} = (n_1/N, \cdots, n_k/N)$, where $n_1, \cdots, n_k$ are nonnegative integers whose sum is N, and let the probability of $Z^{(N)} = z^{(N)}$ be $N! \prod_{i=1}^{k} (p_i^{n_i}/n_i!)$, where $p = (p_1, \cdots, p_k) \varepsilon \Omega$, the set of points

Received 21 September 1964.

[1] This research was supported by the Mathematics Division of the Air Force Office of Scientific Research.

p with $p_i \geqq 0, p_1 + \cdots + p_k = 1$. Let A be any subset of Ω, and let $A^{(N)}$ denote the set of points $z^{(N)}$ which are in A. Then for the probability $P_N(A \mid p)$ of $Z^{(N)} \subset A$ we have (see Theorem 2.1)

$$(1.1) \qquad P_N(A \mid p) = \exp \{-NI(A^{(N)}, p) + O(\log N)\},$$

uniformly for $A \subset \Omega$ and $p \varepsilon \Omega$, where

$$(1.2) \qquad I(x, p) = \sum_{i=1}^{k} x_i \log (x_i/p_i),$$

$$(1.3) \qquad I(A, p) = \inf \{I(x, p) \mid x \varepsilon A\}.$$

This elementary and crude estimate of the probability $P_N(A \mid p)$ makes it possible to study, to a first approximation, the asymptotic behavior of the error probabilities of an arbitrary (non-randomized) test of a hypothesis concerning p when these probabilities tend to 0 at a sufficiently rapid rate.

In Section 3 the special role of the likelihood ratio test is brought out. Let H be the hypothesis that $p \varepsilon \Lambda (\Lambda \subset \Omega)$. The likelihood ratio test, based on an observation $z^{(N)}$ of $Z^{(N)}$, for testing H against the alternatives $p \varepsilon \Omega - \Lambda$ rejects H when

$$I(z^{(N)}, \Lambda) > \text{const},$$

where $I(x, \Lambda) = \inf \{I(x, p) \mid p \varepsilon \Lambda\}$. For the size of an arbitrary test which rejects H when $z^{(N)} \varepsilon A$ we have from (1.1)

$$(1.4) \qquad \sup_{p \varepsilon \Lambda} P_N(A \mid p) = \exp \{-NI(A^{(N)}, \Lambda) + O(\log N)\},$$

uniformly for $A \subset \Omega$ and $\Lambda \subset \Omega$, where $I(A, \Lambda) = \inf \{I(x, p) \mid x \varepsilon A, p \varepsilon \Lambda\}$. This easily implies the following: The union of the critical regions $A^{(N)}$ of all tests of size $\leqq \alpha_N$ for testing H is contained in the critical region $B^{(N)}$ of a likelihood ratio test for testing H against $p \varepsilon \Omega - \Lambda$ whose size α_N' satisfies

$$\log \alpha_N' = \log \alpha_N + O(\log N).$$

Thus if α_N tends to 0 faster than any power of N, the size of the $B^{(N)}$ test is, to a first approximation, α_N. Of course, $\alpha_N' \geqq \alpha_N$. It is trivial that the $B^{(N)}$ test is uniformly at least as powerful as any test of size $\leqq \alpha_N$.

We can also define a likelihood ratio test of size $\leqq \alpha_N$ whose critical region does not differ much from $B^{(N)}$ in the sense that both critical regions are of the form

$$NI(z^{(N)}, \Lambda) \geqq -\log \alpha_N + O(\log N).$$

The main concern of this paper is to substantiate the following proposition: *If a given test of size α_N is "sufficiently different" from a likelihood ratio test, then there is a likelihood ratio test of size $\leqq \alpha_N$ which is considerably more powerful than the given test at "most" points p in the set of alternatives when N is large enough, provided that $\alpha_N \to 0$ at a suitable rate.* The meaning of the words in quotation marks will have to be made precise. By "considerably more powerful" we mean that the ratio of the error probabilities at p of the two tests tends to 0 more rapidly than any power of N.

A general characterization of the set Γ_N of alternatives p at which a given test is considerably less powerful than a comparable likelihood ratio test is contained in Theorem 3.1. Sections 4 and 5 are preparatory to what follows and deal with properties of the function $I(x, p)$ and its infima. In Section 6 we restrict ourselves to tests whose critical regions are regular in a sense which implies that the expression (1.4) for the size of a test remains true with $I(A^{(N)}, \Lambda)$ replaced by $I(A, \Lambda)$, the infimum of $I(x, \Lambda)$ with respect to all $x \varepsilon A$ (not only with respect to the lattice points $z^{(N)}$ contained in A), and an analogous replacement may be made in the expressions for the error probabilities $P_N(A' \mid p)$, where $A' = \Omega - A$ and $p \varepsilon \Lambda' = \Omega - \Lambda$. (Sufficient conditions for regularity are given in the Appendix.) Consider a test which rejects H when $z^{(N)} \varepsilon A$ (where $A = A_N$ may depend on N). Let $B = \{x \mid I(x, \Lambda) \geqq I(A, \Lambda)\}$, so that $B^{(N)}$ is the critical region of a likelihood ratio test. Note that $A \subset B$. It is shown that the set Γ_N essentially depends on the set of common boundary points of the sets A and B. In particular, if the A test differs sufficiently from a likelihood ratio test in the sense that the sets A and B have only finitely many boundary points in common then, under certain additional conditions, a likelihood ratio test whose size does not exceed the size of the A test is considerably more powerful than the latter at all alternatives except those points p which lie on certain curves in the $(k - 1)$-dimensional simplex Ω and those at which both tests have zero error probabilities.

Approximations for the error probabilities of a likelihood ratio test of a simple hypothesis are given in Section 7.

In Section 8 the result just described is shown to be true for a chi-square test of a simple hypothesis whose size tends to 0 at a suitable rate (Theorem 8.4). This is of special interest in view of the fact that if the size of the chi-square test tends to a positive limit, its critical region and power differ little from those of a likelihood ratio test. In Section 9 chi-square tests of composite hypotheses are briefly discussed. An example shows that at least in some cases the situation is similar to that in the case of a simple hypothesis. It is noted that one common version of the chi-square test may have the property that its size cannot be smaller than some power of N, which makes the theory of this paper inapplicable. Certain competitors of the chi-square test are considered in Section 10.

It is pointed out that certain Bayes tests have the same asymptotic power properties as the corresponding likelihood ratio test (Section 11).

The likelihood ratio test was introduced by J. Neyman and E. S. Pearson in 1928 [6]. It is known that the likelihood ratio test has certain asymptotically optimal properties when the error probabilities are bounded away from 0 (Wald [8]). The present results are of a different nature and appear to be of a novel type.

An extension of the results of this paper to certain classes of distributions other than the multinomial class should be possible. (The extension to the case of several independent multinomial random vectors is quite straightforward.)

It should be emphasized that throughout this paper the number k is regarded as fixed and is not allowed to increase with N. In particular, the results of Section 8, which suggest that the likelihood ratio test of a simple hypothesis is asymptotically either equivalent or superior in a global sense to the chi-square test, are subject to the limitation that k is fixed or does not increase rapidly with N. Otherwise the relation between the two tests may be reversed. This is shown by the following unpublished result of Charles Stein, who kindly permitted to include it here.

For testing the hypothesis $p_1 = \cdots = p_k = 1/k$ consider the class C of symmetric tests whose critical regions are of the form $\sum a_q M_q \geqq c$, where the a_q and c are constants (which may depend on N) and M_q is the number of n_j which are equal to q. Both the chi-square test and the likelihood ratio test belong to C. If the significance level is moderate, k is large and N/k is moderate, then the chi-square test is nearly most powerful in C against alternatives for which all of the $|p_j - 1/k|$ are small compared with $1/k$. In particular, it is appreciably more powerful than the likelihood ratio test.

I wish to express my gratitude to Professor R. A. Wijsman whose comments on the original manuscript led to substantial improvements of this paper. In particular, a result of his (Lemma 4.4 below), which is of independent interest, helped to fill a gap in the author's original proof of Lemma 5.1.

2. Probabilities of large deviations in multinomial distributions. Let $Z^{(N)} = (Z_1^{(N)}, \cdots, Z_k^{(N)})$ be a random vector whose values are

$$z^{(N)} = (z_1^{(N)}, \cdots, z_k^{(N)}) = (n_1/N, \cdots, n_k/N), \tag{2.1}$$

where $n_1, \cdots, n_k$ are any nonnegative integers such that $n_1 + \cdots + n_k = N$, and whose distribution is given by

$$\Pr\{Z^{(N)} = z^{(N)}\} = P_N(z^{(N)} \mid p) = [N!/(n_1! \cdots n_k!)]p_1^{n_1} \cdots p_k^{n_k}. \tag{2.2}$$

Here $p = (p_1, \cdots, p_k)$ is any point in the simplex

$$\Omega = \{(x_1, \cdots, x_k) \mid x_1 \geqq 0, \cdots, x_k \geqq 0, \quad x_1 + \cdots + x_k = 1\}. \tag{2.3}$$

By convention, $p_i^{n_i} = 1$ if $p_i = n_i = 0$.

We can write

$$P_N(z^{(N)} \mid p) = P_N(z^{(N)} \mid z^{(N)}) \exp\{-NI(z^{(N)}, p)\}, \tag{2.4}$$

where, for any two points x and p in Ω,

$$I(x, p) = \textstyle\sum_{i=1}^k x_i \log(x_i/p_i). \tag{2.5}$$

Here it is understood that $x_i \log(x_i/p_i) = 0$ if $x_i = 0$.

We note that $I(x, p) > 0$ unless $x = p$ (since $\log u > 1 - u^{-1}$ for $u > 0$, $u \neq 1$). Also, $I(x, p) < \infty$ unless $p_i = 0$ and $x_i > 0$ for some i.

For any subset A of Ω let

$$P_N(A \mid p) = \Pr\{Z^{(N)} \varepsilon A\} = \textstyle\sum_{z^{(N)} \varepsilon A} P(z^{(N)} \mid p). \tag{2.6}$$

The set of lattice points $z^{(N)}$ contained in A will be denoted by $A^{(N)}$. We define

$$I(A, p) = \inf\{I(x, p) \mid x \varepsilon A\}, \qquad I(A, p) = +\infty \text{ if } A \text{ is empty}. \tag{2.7}$$

The following theorem is a slight elaboration of a result due to Sanov [7].

THEOREM 2.1. *For any set $A \subset \Omega$ and any point $p \in \Omega$ we have*

$$(2.8) \qquad \begin{aligned} C_0 N^{-(k-1)/2} \exp\{-NI(A^{(N)}, p)\} &\leqq P_N(A \mid p) \\ &\leqq \binom{N+k-1}{k-1} \exp\{-NI(A^{(N)}, p)\}, \end{aligned}$$

where C_0 is a positive absolute constant. Hence

$$(2.9) \qquad P_N(A \mid p) = \exp\{-NI(A^{(N)}, p) + O(\log N)\},$$

uniformly for $A \subset \Omega$ and $p \in \Omega$. Also,

$$(2.10) \qquad P_N(A \mid p) \leqq \exp\{-NI(A, p) + O(\log N)\},$$

uniformly for $A \subset \Omega$ and $p \in \Omega$.

PROOF. Clearly (2.8) implies (2.9) and since $A^{(N)} \subset A$, (2.8) implies (2.10). It is sufficient to prove (2.8).

If $A^{(N)}$ is empty, $P_N(A \mid p) = 0$ and (2.8) is trivially true. Assume that $A^{(N)}$ is not empty.

The number of points $z^{(N)}$ in Ω is easily found to be $\binom{N+k-1}{k-1}$. By (2.4), $z^{(N)} \in A$ implies $P_N(z^{(N)} \mid p) \leqq \exp\{-NI(A^{(N)}, p)\}$. Hence the second inequality (2.8) follows from (2.6).

By Stirling's formula, for $m \geqq 1$,

$$m! = m^m (2\pi m)^{\frac{1}{2}} \exp[-m + (\theta/12m)], \qquad 0 < \theta < 1.$$

Hence it easily follows that if $n_i \geqq 1$ for all i,

$$(2.11) \qquad P_N(z^{(N)} \mid z^{(N)}) = (N!/N^N) \textstyle\prod_{i=1}^{k} n_i^{n_i}/n_i! \geqq C_0 N^{-(k-1)/2},$$

where C_0 is a positive absolute constant. (We can take $C_0 = \frac{1}{2}$.) If $n_i = 0$ for some i, (2.11) is *a fortiori* true. The first inequality (2.8) follows from (2.4), (2.6) and (2.11).

Theorem 2.1 is nontrivial only if the set A contains no points $z^{(N)}$ which are too close to p. In this sense the theorem is concerned with probabilities of large deviations of $Z^{(N)}$ from its mean p.

It should be noted that (2.9) gives an asymptotic expression for the logarithm of the probability on the left but not for the probability itself. This crude result is sufficient to study asymptotically the main features of any test whose size tends to 0 fast enough as N increases.

The precise order of magnitude of $P_N(A \mid p)$ for certain sets A will be considered in another paper. For the case $A = \{x \mid \sum a_i x_i \leqq c\}$, so that $P_N(A \mid p)$ is a value of the distribution function of a sum of N independent random variables, see Bahadur and Rao [1] and the references there given. For $A = \{x \mid F(x) \geqq 0\}$, where $F(x)$ satisfies certain regularity conditions, Sanov ([7], Theorem 4) gave without proof a result which, however, is inaccurate in the stated generality. (Compare the author's abstract [3].)

3. The role of the likelihood ratio test. Consider the problem of testing, on the basis of an observation $z^{(N)}$ of the random vector $Z^{(N)}$, the hypothesis H that the parameter vector p is contained in a subset Λ of Ω. The likelihood ratio test for testing H against the alternative $p \varepsilon \Lambda' = \Omega - \Lambda$ is based on the statistic

$$\frac{\sup \{P_N(z^{(N)} \mid p) \mid p \varepsilon \Lambda\}}{\sup \{P_N(z^{(N)} \mid p) \mid p \varepsilon \Omega\}} = \exp \{-NI(z^{(N)}, \Lambda)\}, \tag{3.1}$$

where

$$I(x, \Lambda) = \inf \{I(x, p) \mid p \varepsilon \Lambda\}. \tag{3.2}$$

The equality in (3.1) follows from (2.4). Thus the likelihood ratio test rejects H if $I(z^{(N)}, \Lambda)$ exceeds a constant.

Now consider an arbitrary test which rejects H if $z^{(N)} \varepsilon A_N$, where A_N is any subset of Ω. For the size of the test (the supremum of its error probability for $p \varepsilon \Lambda$) we have from Theorem 2.1

$$\sup \{P_N(A_N \mid p) \mid p \varepsilon \Lambda\} = \exp \{-NI(A_N^{(N)}, \Lambda) + O(\log N)\}, \tag{3.3}$$

uniformly for $A_N \subset \Omega$ and $\Lambda \subset \Omega$, where

$$I(A, \Lambda) = \inf \{I(x, p) \mid x \varepsilon A, p \varepsilon \Lambda\}. \tag{3.4}$$

Clearly $I(A, \Lambda) = \inf \{I(x, \Lambda) \mid x \varepsilon A\} = \inf \{I(A, p) \mid p \varepsilon \Lambda)\}$.

The test of the preceding paragraph will be referred to as test A_N. The set $A_N^{(N)}$ of all points $z^{(N)}$ contained in A_N will be called its critical region. (We could have assumed that A_N contains no other points than the lattice points $z^{(N)}$, but it is often convenient to define the critical region in terms of a more inclusive set.)

We may compare the test A_N with the likelihood ratio test which rejects H if $z^{(N)} \varepsilon B_N$, where

$$B_N = \{x \mid I(x, \Lambda) \geqq c_N\}, \qquad c_N = I(A_N^{(N)}, \Lambda). \tag{3.5}$$

Its critical region $B_N^{(N)}$ contains the critical region $A_N^{(N)}$. In fact, $B_N^{(N)}$ is the union of the critical regions of all tests A_N^* for which $I(A_N^{*(N)}, \Lambda) \geqq c_N$. Moreover, the size of the test B_N is $\exp \{-Nc_N + O(\log N)\}$, since $I(B_N^{(N)}, \Lambda) = c_N$. Thus if $Nc_N/\log N$ tends to infinity with N, which means that the size α_N of the test A_N tends to 0 faster than any power of N, then the size α_N' of the test B_N is approximately equal to α_N in the sense that $\log \alpha_N' = \log \alpha_N + O(\log N)$.

In a similar way we can obtain the following conclusion: The union of the critical regions of all tests of size $\leqq \alpha_N$ for testing the hypothesis $p \varepsilon \Lambda$ is contained in the critical region of a likelihood ratio test for testing $p \varepsilon \Lambda$ against $p \not\varepsilon \Lambda$ whose size α_N' satisfies $\log \alpha_N' = \log \alpha_N + O(\log N)$. The simple proof is left to the reader.

Since $A_N^{(N)} \subset B_N^{(N)}$, the probability that test B_N rejects H is never smaller than the probability that test A_N rejects H. Thus B_N is uniformly at least as powerful as A_N, but the size of B_N is in general somewhat larger than the size of A_N. It may be more appropriate to compare a given test with a likelihood ratio test whose size does not exceed the size of the former. Now it easily follows from (3.3) that we can choose numbers

$$0 \leqq \delta_N = O(N^{-1} \log N) \tag{3.6}$$

such that the size of the likelihood ratio test

$$B_N^* = \{x \mid I(x, \Lambda) \geqq c_N + \delta_N\} \tag{3.7}$$

is not larger than the size of test A_N. If $Nc_N/\log N$ tends to infinity fast enough, we may expect that the power of the test B_N^* will not be much smaller than the power of B_N.

For any $p \varepsilon \Lambda'(\Lambda' = \Omega - \Lambda)$ the probabilities that the tests A_N and B_N falsely accept the hypothesis are given by

$$P_N(A_N' \mid p) = \exp\{-NI(A_N'^{(N)}, p) + O(\log N)\}, \tag{3.8}$$

$$P_N(B_N' \mid p) = \exp\{-NI(B_N'^{(N)}, p) + O(\log N)\}. \tag{3.9}$$

Always $P_N(B_N' \mid p) \leqq P_N(A_N' \mid p)$ and $I(B_N'^{(N)}, p) \geqq I(A_N'^{(N)}, p)$. At those points $p \varepsilon \Lambda'$ for which $P_N(A_N' \mid p) \neq 0$ and

$$\lim_{N\to\infty} N\{I(B_N'^{(N)}, p) - I(A_N'^{(N)}, p)\}/\log N = +\infty, \tag{3.10}$$

the test B_N is considerably more powerful than A_N in the sense that the ratio of the error probabilities at p, $P_N(B_N' \mid p)/P_N(A_N' \mid p)$, tends to 0 more rapidly than any power of N.

For the test B_N^* whose size does not exceed the size of A_N we have a similar conclusion. Note that B_N^* is not necessarily more powerful than A_N, and the difference in (3.10) with B_N replaced by B_N^* may be negative. However, if $P_N(A_N', p) \neq 0$, if (3.10) is satisfied, and

$$\lim_{N\to\infty} \frac{I(B_N'^{(N)}, p) - I(B_N^{*\prime(N)}, p)}{I(B_N'^{(N)}, p) - I(A_N'^{(N)}, p)} = 0 \tag{3.11}$$

then the ratio $P_N(B_N^{*\prime} \mid p)/P_N(A_N' \mid p)$ tends to 0 more rapidly than any power of N.

The main conclusions of the preceding discussion are summarized in the following theorem.

THEOREM 3.1. *Let Λ and A_N be non-empty subsets of Ω. Then*

(a) *the size of the test A_N for testing the hypothesis $p \varepsilon \Lambda$ is given by* (3.3) *and its error probability at $p \varepsilon \Lambda'$ by* (3.8).

(b) *There exist positive numbers $\delta_N = O(N^{-1} \log N)$ such that the size of the likelihood ratio test $B_N^* = \{x \mid I(x, \Lambda) \geqq I(A_N^{(N)}, \Lambda) + \delta_N\}$ does not exceed the size of the test A_N.*

(c) *For each $p \varepsilon \Lambda'$ such that $P_N(A_N' \mid p) \neq 0$ and Conditions* (3.10) *and* (3.11) *are satisfied, the ratio $P_N(B_N^{*\prime} \mid p)/P_N(A_N' \mid p)$ of the error probabilities at p tends to* 0 *faster than any power of N.*

In Section 6 we shall continue in more detail the study of the set of alternatives at which a given test is less powerful than a comparable likelihood ratio test, assuming that the sets A_N are regular in a certain sense. The following two sections are preparatory to what follows.

4. The function $I(x, p)$ and its infima. In this section properties of the function

$$I(x, p) = \sum_{i=1}^{k} x_i \log (x_i/p_i)$$

and its infima for $x \varepsilon A$ or $p \varepsilon \Lambda$ are studied.

The function $I(x, p)$ is denfied for x and p in the simplex Ω given by (2.3). Throughout, Ω is considered as the space of the points x and p, with the Euclidean metric. Thus the complement A' of a subset A of Ω is $\Omega - A$. The closure of A is denoted by $\bar{A}$. The boundary of A is $\bar{A} \cap \bar{A'}$.

We define the subsets Ω_0 and $\Omega(p)$ (for $p \varepsilon \Omega$) of Ω by

$$\Omega_0 = \{x \mid x_i > 0, \quad i = 1, \cdots, k\}, \tag{4.1}$$

$$\Omega(p) = \{x \mid x_i = 0 \quad \text{if} \quad p_i = 0\}. \tag{4.2}$$

Thus if $p \varepsilon \Omega_0$, $\Omega(p) = \Omega$. If $p \varepsilon \Omega_0'$, $\Omega(p)$ is the intersection of those faces $\{x \mid x_i = 0\}$ of the simplex Ω for which $p_i = 0$.

LEMMA 4.1.

(a) $0 \leqq I(x, p) \leqq \infty$. *$I(x, p) = 0$ if and only if $x = p$. $I(x, p) < \infty$ if and only if $x \varepsilon \Omega(p)$.*

(b) *For each $p \varepsilon \Omega_0$, $I(\cdot, p)$ is continuous and bounded in Ω. For each $p \varepsilon \Omega_0'$, $I(\cdot, p)$ is continuous and bounded in $\Omega(p)$.*

(c) *For each $x \varepsilon \Omega$, $I(x, \cdot)$ is continuous in Ω. That is, $p^j \to p$ implies $I(x, p^j) \to I(x, p)$, even when $I(x, p) = \infty$.*

(d) *For each $p \varepsilon \Omega$, $I(\cdot, p)$ is convex in Ω. For each $x \varepsilon \Omega$, $I(x, \cdot)$ is convex in Ω.*

PROOF.

(a) See Section 2 after (2.5).

(b) If $p \varepsilon \Omega_0$, $I(\cdot, p)$ is bounded since $I(x, p) \leqq \sum x_i \log (1/p_i) \leqq \max_i \log (1/p_i)$. The proof of continuity is obvious. For $p \varepsilon \Omega_0'$ the proof is similar.

(c) If $x \varepsilon \Omega(p)$ then $I(x, p) < \infty$ and the continuity at p is obvious. If $x \varepsilon \Omega'(p)$ then $I(x, p) = \infty$. If $p' \to p$ then $p_i' \to 0$ for some i with $x_i > 0$. Hence $I(x, p') \to \infty = I(x, p)$.

(d) The convexity of $I(\cdot, p)$ and $I(x, \cdot)$ follows from the convexity of $u \log u$ and $-\log u$ for $u > 0$.

The next lemma is concerned with $I(A, p)$, the infimum of $I(x, p)$ for $x \varepsilon A$. The relevance of $I(A^{(N)}, p)$ for the approximation of $P_N(A \mid p)$ is clear from Theorem 2.1. If the set A is sufficiently regular, the approximation (2.9) is true with $I(A^{(N)}, p)$ replaced by $I(A, p)$ (see Section 6 and the Appendix).

LEMMA 4.2. *Let A be a non-empty subset of Ω.*

(a) *Let $p \varepsilon \Omega_0$. Then there is at least one point y such that*

$$y \varepsilon \bar{A}, \qquad I(y, p) = I(A, p). \tag{4.3}$$

If $p \varepsilon \bar{A}$ then $I(A, p) = 0$ and (4.3) is satisfied only with $y = p$. If $p \not\varepsilon \bar{A}$ then $I(A, p) > 0$ and any y which satisfies (4.3) is in the boundary of A.

(b) *Let* $p \not\in \Omega_0$. *Then* $I(A, p) < \infty$ *if and only if the intersection* $A \cap \Omega(p)$ *is not empty. If this is the case, then* $I(A, p) = I(A \cap \Omega(p), p)$ *and the statements of Part* (a) *are true with* A *replaced by* $A \cap \Omega(p)$.

PROOF. The lemma follows easily from Lemma 4.1. We prove only the last assertion of Part (a). Let $p \in \Omega_0$, $p \not\in \bar{A}$. Then (since $I(\cdot, p)$ is continuous) $I(A, p) > 0$. Suppose that $I(y, p) = I(A, p)$ for some y in the interior of A. Then the point $z = (1 - t)y + tp$ is in A for some positive $t < 1$. Since $I(\cdot, p)$ is convex,

$$I(z, p) \leqq (1 - t)I(y, p) + tI(p, p) = (1 - t)I(A, p) < I(A, p).$$

This contradicts the definition of $I(A, p)$. Hence any y which satisfies (4.3) is in the boundary of A.

A maximum likelihood estimate of p under the assumption $p \in \Lambda$ is a point $\hat{p} = \hat{p}(z^{(N)})$ which maximizes $P_N(z^{(N)} \mid p)$ for $p \in \Lambda$ (or $\bar{\Lambda}$). From (2.4) we see that $\hat{p}$ minimizes $I(z^{(N)}, p)$, so that $I(z^{(N)}, \hat{p}) = I(z^{(N)}, \Lambda)$. By extension, we may define $\hat{p}(x)$ for any $x \in \Omega$ as a point in $\bar{\Lambda}$ for which $I(x, \hat{p}(x)) = I(x, \Lambda)$. The next lemma asserts the existence of at least one $\hat{p}(x)$ for each x.

LEMMA 4.3. *Let* Λ *be a non-empty subset of* Ω.

(a) *For each* $x \in \Omega$ *there is at least one point* $\hat{p}(x)$ *such that*

$$\hat{p}(x) \in \bar{\Lambda}, \qquad I(x, \hat{p}(x)) = I(x, \Lambda). \tag{4.4}$$

(b) *If* $x \in \bar{\Lambda}$ *then* $I(x, \Lambda) = 0$ *and* (4.4) *is satisfied only for* $\hat{p}(x) = x$.

(c) *If* $x \not\in \bar{\Lambda}$ *then* $I(x, \Lambda) > 0$ *and any* $\hat{p}(x)$ *which satisfies* (4.4) *is in the boundary of* Λ.

(d) $I(x, \Lambda)$ *is bounded in* Ω *if and only if* $\Lambda \cap \Omega_0$ *is not empty.*

The lemma follows easily from Lemma 4.1.

The function $I(\cdot, \Lambda)$ may be bounded and not continuous. For example, let $k = 3$ and let Λ consist of the two points $p^1 = (\frac{1}{2}, \frac{1}{2}, 0)$ and $p^2 = (\frac{1}{5}, \frac{1}{5}, \frac{3}{5})$. If $x_3 = 0$, $I(x, p^1) \leqq \log 2 < \log(\frac{5}{2}) \leqq I(x, p^2)$. If $x_3 \neq 0$, $I(x, p^1) = \infty$. Hence $I(x, \Lambda) = I(x, p^1)$ or $I(x, p^2)$ according as $x_3 = 0$ or $x_3 \neq 0$, and $I(\cdot, \Lambda)$ is discontinuous at the points x with $x_3 = 0$.

The following lemma is due to R. A. Wijsman, who kindly permitted to include it here with his proof.

LEMMA 4.4 (Wijsman).

(a) *The function* $I(\cdot, \Lambda)$ *is lower semicontinuous.*

(b) *If* $\Lambda \subset \Omega_0$, *or if* Λ *is the closure of a subset of* Ω_0, *then* $I(\cdot, \Lambda)$ *is continuous.*

(c) $I(\cdot, \Lambda)$ *is continuous in* Ω_0.

PROOF.

(a) It follows from Lemma 4.3 that $I(\cdot, \Lambda) = I(\cdot, \bar{\Lambda})$, so that we may assume Λ closed and therefore compact. Put $J(x, p) = \sum x_i(-\log p_i)$, $J(x, \Lambda) = \inf\{J(x, p) \mid p \in \Lambda\}$. Then $I(x, \Lambda) = \sum x_i \log x_i + J(x, \Lambda)$, where the first sum on the right is continuous. Hence it is sufficient to show that $J(\cdot, \Lambda)$ is lower semicontinuous, that is, $x^n \to y$ implies $J(y, \Lambda) \leqq \liminf J(x^n, \Lambda)$.

Let $x^n \to y$ and $\liminf J(x^n, \Lambda) = c$. If $c = \infty$ the claim is trivially true, so assume $c < \infty$. By taking a suitable subsequence we may assume $J(x^n, \Lambda) \to c$.

Using compactness of Λ, for each n there exists $p^n \varepsilon \Lambda$ such that $J(x^n, \Lambda) = J(x^n, p^n)$, and by taking a subsequence if necessary we may assume $p^n \to p^0$, say, where $p^0 \varepsilon \Lambda$.

All terms in the sum $J(x^n, p^n) = \sum x_i^n(-\log p_i^n)$ are $\geqq 0$. Since the sum converges to c, each sequence $\{x_i^n(-\log p_i^n)\}$ is bounded from some n on. Suppose now $y_i > 0$ if $i \varepsilon M, y_i = 0$ if $i \varepsilon M'$. If for some $i \varepsilon M$ we would have $p_i^0 = 0$, then since $x_i^n \to y_i > 0$ and $p_i^n \to p_i^0 = 0$, we would have $x_i^n(-\log p_i^n) \to \infty$ which was excluded. Therefore if $i \varepsilon M$ then $p_i^0 > 0$ and $x_i^n(-\log p_i^n) \to y_i(-\log p_i^0)$. For $i \varepsilon M'$ we have $y_i \log p_i^0 = 0$. Hence

$$\begin{aligned} J(y, \Lambda) \leqq J(y, p^0) &= \textstyle\sum_{i \varepsilon M} y_i(-\log p_i^0) \\ &= \lim \textstyle\sum_{i \varepsilon M} x_i^n(-\log p_i^n) \leqq \lim \sum_{i=1}^k x_i^n(-\log p_i^n) = c. \end{aligned}$$

(b) Suppose that $\Lambda \subset \Omega_0$. For every $p \varepsilon \Omega_0$, $I(\cdot, p)$ is continuous, so that $I(\cdot, \Lambda)$ is the infimum of a family of continuous functions, therefore upper semi-continuous and, by Part (a), continuous. Since $I(\cdot, \Lambda) = I(\cdot, \bar{\Lambda})$, the same result holds if Λ is the closure of a subset of Ω_0.

(c) If $\Lambda \cap \Omega_0$ is empty then $I(\cdot, \Lambda)$ is identically ∞ on Ω_0. Otherwise $I(x, \Lambda) = I(x, \Lambda \cap \Omega_0)$ for $x \varepsilon \Omega_0$, and $I(\cdot, \Lambda \cap \Omega_0)$ is continuous by Part (b).

LEMMA 4.5 *Let A and Λ be non-empty subsets of Ω. Suppose that $I(x, \Lambda)$ is continuous in Ω and $I(A, \Lambda) > 0$. Then there is at least one point y such that*

$$y \varepsilon \bar{A}, \qquad I(y, \Lambda) = I(A, \Lambda), \tag{4.5}$$

and any point y which satisfies (4.5) *is in the boundary of A.*

PROOF. The existence of a point y which satisfies (4.5) follows from the assumed continuity of $I(x, \Lambda)$. Suppose that $I(y, \Lambda) = I(A, \Lambda)$ for some y in the interior of A. By Lemma 4.3 there is a $\hat{p} \varepsilon \bar{\Lambda}$ such that $I(y, \hat{p}) = I(y, \Lambda)$. The point $z = (1 - t)y + t\hat{p}$ is in A for some positive $t < 1$. Since $I(\cdot, \hat{p})$ is convex,

$$I(z, \hat{p}) \leqq (1 - t)I(y, \hat{p}) = (1 - t)I(A, \Lambda) < I(A, \Lambda), \tag{4.6}$$

due to $I(A, \Lambda) > 0$. But since $\hat{p} \varepsilon \bar{\Lambda}$, since $I(z, \cdot)$ is continuous, and $z \varepsilon A$, we have $I(z, \hat{p}) \geqq I(z, \Lambda) \geqq I(A, \Lambda)$, which contradicts (4.6). This implies the lemma.

We conclude this section with some remarks on the determination of the infimum $I(A, p)$ and on the set of points in $\bar{A}$ at which the infimum is attained.

We restrict ourselves to the case $p \varepsilon \Omega_0$. (Lemma 4.2 implies that the general case can be reduced to this case.) The set A is contained in the set $\{x \mid I(x, p) \geqq I(A, p)\}$, whose complement C is convex. The following lemma gives information about the boundary of C. A hyperplane (briefly: plane) in Ω is a non-empty set $\{x \mid \sum a_i x_i = c\}$, where $a_1, \cdots, a_k$ are not all equal. The dimension of a hyperplane is at most $k - 2$; in degenerate cases, such as $\{x \mid x_1 = 1\}$, the dimension may be less than $k - 2$.

LEMMA 4.6. *Let* $p \varepsilon \Omega_0$,

$$C = \{x \mid I(x, p) < c\}, \qquad 0 < c < \max_x I(x, p), \tag{4.7}$$

and let y be a boundary point of C.

(a) *If* $y \varepsilon \Omega_0$ *then*

$$I(x, p) - c = \textstyle\sum_{i=1}^k (\log (y_i/p_i))(x_i - y_i) + I(x, y) \tag{4.8}$$

and the unique tangent plane of C *at* y *is* $T = \{x \mid \sum (\log (y_i/p_i))(x_i - y_i) = 0\}$.

(b) *If* $y \not\varepsilon \Omega_0$ *then, for each* j *with* $y_j = 0$, $T_j = \{x \mid x_j = 0\}$ *is a tangent plane of* C *at* y, *and* C *has no tangent planes at* y *other than these* T_j *and their intersections.*

(c) *All boundary points of* C *are in* Ω_0 *if and only if*

$$c < -\log (1 - p_{\min}), \qquad p_{\min} = \min_i p_i . \tag{4.9}$$

PROOF.

(a) The identity (4.8) follows immediately from the fact that $I(y, p) = c$ and y and p are in Ω_0. Hence T is a tangent plane of C at y. It is unique since the derivatives $\partial I(x, p)/\partial x_i = \log (x_i/p_i) + 1$ are continuous in Ω_0.

(b) Clearly, if $y_j = 0$, then $T_j = \{x \mid x_j = 0\}$ is a tangent plane of C at y. It is sufficient to prove that no hyperplane containing points in Ω_0 is a tangent plane at y. This will follow if we show that every straight line containing y and some point in Ω_0 intersects the open convex set C.

Let x^0 be a point in Ω_0,

$$z(t) = (1 - t)y + tx^0, \qquad F(t) = I(z(t), p).$$

We must show that $F(t) < c$ for some $t \varepsilon (0, 1)$.

We have for $t \varepsilon (0, 1)$

$$\begin{aligned} F'(t) &= \partial I(z(t), p)/\partial t = \textstyle\sum_{i=1}^k (x_i^0 - y_i) \log (z_i(t)/p_i) \\ &= \textstyle\sum_{y_i=0} x_i^0 \log (tx_i^0/p_i) + O(1) \\ &= \textstyle\sum_{y_i=0} x_i^0 \log t + O(1) \qquad \text{as } t \to 0. \end{aligned}$$

Since $F''(t) > 0$, we have by Taylor's formula for $t \varepsilon (0, 1)$

$$c = I(y, p) = F(0) \geqq F(t) - tF'(t).$$

It follows that $F(t) < c$ for t positive and sufficiently small, as was to be proved.

(c) Let x be a point with $x_j = 0$. Let $\bar{p}$ be defined by $\bar{p}_j = 0$, $\bar{p}_i = p_i/(1 - p_j)$, $i \neq j$. Then

$$I(x, p) = \textstyle\sum_i x_i \log (x_i/\bar{p}_i) - \log (1 - p_j) \geqq -\log (1 - p_j),$$

with equality for $x = \bar{p}$. Hence $\{x \mid I(x, p) \leqq c\} \subset \Omega_0$ if, and only if, $c < -\log (1 - p_{\min})$. Part (c) follows.

LEMMA 4.7. *Let*

$$A = \{x \mid f(x) > 0\}, \tag{4.10}$$

where the function $f(x)$ *is continuous in* Ω *and* $\max f(x) > 0$. *Let* p *be a point in* Ω_0 *such that* $f(p) < 0$. *Suppose further that the derivatives* $f_i'(x) = \partial f(x)/\partial x_i$, $i = 1, \cdots, k$, *exist and are continuous at all* x *in* Ω_0 *for which* $f(x) = 0$.

Let y be any point in $\bar{A}$ such that $I(y, p) = I(A, p)$. Then if $y \varepsilon \Omega_0$, it is necessary that $f(y) = 0$ and

$$\log (y_i/p_i) = af_i'(y) + b \qquad i = 1, \cdots, k, \tag{4.11}$$

where $a > 0$ and b are constants.

PROOF. By Lemma 4.2, any point y in $\bar{A}$ for which $I(y, p) = I(A, p)$ is in the boundary of A. Since $f(x)$ is continuous, this means that $f(y) = 0$. The method of Lagrange multipliers yields the necessary condition (4.11) with some constants a and b. That a must be positive follows from

$$f(x) = f(x) - f(y) = \sum f_i'(y)(x_i - y_i) + o(|x - y|)$$

and (4.8) since $f(x) > 0$ implies $I(x, p) > I(y, p)$. (Note that in (4.8), $I(x, y) = o(|x - y|)$.)

LEMMA 4.8. *If A is convex and $A \cap \Omega_0$ is not empty, and if $p \varepsilon \Omega_0$, $p \not\varepsilon \bar{A}$, then there is exactly one point $y \varepsilon \bar{A}$ such that $I(y, p) = I(A, p)$.*

PROOF. The point y is a common boundary point of the disjoint convex sets A and $B' = \{x \mid I(x, p) < I(A, p)\}$. Since A and B' contain points in Ω_0, it follows from Lemma 4.6(b) that y is in Ω_0. Lemma 4.6(a) implies that the separating hyperplane of the sets A and B' is unique, and y is the unique point in $\overline{B'}$ which is in that hyperplane.

5. The infimum of $I(x, p)$ subject to the condition $I(x, \Lambda) < c$. The infimum $I(B', p)$, where $B' = \{x \mid I(x, \Lambda) < c\}$, is needed for the approximation of the power of a likelihood ratio test for testing the hypothesis $p \varepsilon \Lambda$. For the case of a simple hypothesis, where Λ consists of a single point p^0, the problem is solved explicitly (Theorem 5.1) and an asymptotic expression for the infimum is obtained (Theorem 5.2). The case of an arbitrary Λ is then briefly discussed.

THEOREM 5.1. *Let p^0 and p be points in Ω, c a finite positive number,*

$$B' = \{x \mid I(x, p^0) < c\}. \tag{5.1}$$

(I) *We have $I(B', p) < \infty$ if and only if*

$$\max_i p_i^0 p_i \neq 0 \quad \text{and} \quad -\log \textstyle\sum_{p_i \neq 0} p_i^0 < c. \tag{5.2}$$

(II) *Suppose that Condition (5.2) is satisfied. Then there is a unique point y such that*

$$I(y, p^0) \leqq c, \qquad I(B', p) = I(y, p). \tag{5.3}$$

If $I(\bar{p}, p^0) \leqq c$ then $y = \bar{p}$, where

$$\bar{p}_i = p_i / \textstyle\sum_{p_j^0 \neq 0} p_j \quad \text{if } p_i^0 \neq 0; \qquad \bar{p}_i = 0 \quad \text{if} \quad p_i^0 = 0. \tag{5.4}$$

If $c < I(\bar{p}, p^0)$ then

$$y_i = (p_i^0)^{1-s} p_i^s / M(s), \qquad i = 1, \cdots, k, \tag{5.5}$$

and

$$I(B', p) = c - M'(s)/M(s), \tag{5.6}$$

where, for $0 < t < 1$,

$$M(t) = \sum_{i=1}^{k} (p_i^0)^{1-t} p_i^t, \qquad M'(t) = dM(t)/dt, \tag{5.7}$$

and the number $s(0 < s < 1)$ *is uniquely determined by*

$$[sM'(s)/M(s)] - \log M(s) = c. \tag{5.8}$$

PROOF. First assume that p^0 and p are in Ω_0. Then $\bar{p} = p$ and the functions $I(\cdot, p^0)$ and $I(\cdot, p)$ are continuous and bounded in Ω. By Lemma 4.2 there is at least one point y such that (5.3) is satisfied, and if $I(p, p^0) \leqq c$, then necessarily $y = p$.

Now suppose that $0 < c < I(p, p^0)$. Then y is a common boundary point of the disjoint convex sets B' and $C = \{x \mid I(x, p) < I(B', p)\}$. Since B' contains points in Ω_0, Lemma 4.6(b) implies that y must be in Ω_0. By Lemma 4.7 with $f(x) = -I(x, p^0) + c$ we must have

$$\log (y_i/p_i) = -a \log (y_i/p_i^0) + b, \qquad i = 1, \cdots, k'$$

where $a > 0$. This is equivalent to (5.5) with $s = 1/(1 + a) > 0$. The point y must satisfy the conditions $\sum y_i = 1$ and $I(y, p^0) = c$. This implies that $M(s)$ is given by (5.7) and s must satisfy (5.8). Thus s is a positive root of the equation $F(t) = c$, where $F(t) = tL'(t) - L(t)$, $L(t) = \log M(t)$. Now $F'(t) = tL''(t) > 0$ for $t > 0$. Also, $F(1) = L'(1) - L(1) = M'(1) = I(p, p^0) > c$. Hence s is uniquely determined by (5.8), and $0 < s < 1$.

One easily calculates that $I(B', p) = I(y, p)$ is equal to the right-hand side of (5.6). This completes the proof for the case where p^0 and p are in Ω_0.

Now consider the general case. Define $\bar{p}^0$ by

$$\bar{p}_i^0 = p_i^0 / \sum_{p_j \neq 0} p_j^0 \quad \text{if } p_i \neq 0; \qquad \bar{p}_i^0 = 0 \quad \text{if } p_i = 0.$$

In order that $I(x, p)$ be finite for some x such that $I(x, p^0) < c$, it is necessary that $x \,\varepsilon\, \Omega(p^0) \cap \Omega(p)$. If this is the case, then

$$I(x, p^0) = I(x, \bar{p}^0) + I(\bar{p}^0, p^0) \geqq I(\bar{p}^0, p^0) = -\log \sum_{p_i \neq 0} p_i^0.$$

These facts imply Part (I) of the theorem.

If Condition (5.2) is satisfied, it follows from the preceding paragraph and the identity

$$I(x, p) = I(x, \bar{p}) - \log \sum_{p_i^0 \neq 0} p_i \quad \text{for} \quad x \,\varepsilon\, \Omega(p^0) \cap \Omega(p)$$

that $I(B', p)$ is the infimum of $I(x, \bar{p}) - \log \sum_{p_i^0 \neq 0} p_i$ subject to the conditions $x \,\varepsilon\, \Omega(p^0) \cap \Omega(p)$ and $I(x, \bar{p}^0) < c + \log \sum_{p_i \neq 0} p_i^0 = \bar{c}$, say. The solution of this problem follows immediately from the first part of the proof, with Ω, p^0, p, c replaced by $\Omega(p^0) \cap \Omega(p)$, $\bar{p}^0$, $\bar{p}$, $\bar{c}$. It can be verified that the result is equivalent to that stated in the theorem.

We now derive an asymptotic expression for the infimum $I(B', p)$ of Theorem 5.1 as $c \to 0$. We confine ourselves to the case $p \,\varepsilon\, \Omega_0$. In this case, by Theorem 5.1, $I(B', p)$ is finite for small values of c only if $p^0 \,\varepsilon\, \Omega_0$. To emphasize the dependence on c we write $B'(c)$ for B'.

THEOREM 5.2. *Let* $p^0 \varepsilon \Omega_0$, $p \varepsilon \Omega_0$, $B'(c) = \{x \mid I(x, p^0) < c\}$. *Then as* $c \to 0$,

$$(5.9) \quad I(B'(c), p) = I(p^0, p) - (2m_2)^{\frac{1}{2}}c^{\frac{1}{2}} + [1 + (m_3/3m_2)]\, c + O(c^{\frac{3}{2}}),$$

where

$$(5.10) \quad m_j = \sum_{i=1}^{k} p_i^0 [\log (p_i^0/p_i) - I(p^0, p)]^j.$$

PROOF. By Theorem 5.1,

$$(5.11) \quad I(B'(c), p) = c - L'(s_c),$$

where $L(t) = \log M(t)$, $M(t) = \sum (p_i^0)^{1-t} p_i^{\,t}$, and $s_c > 0$ is determined by

$$(5.12) \quad F(s_c) = c, \qquad F(t) = tL'(t) - L(t).$$

All derivatives of $L(t)$ and $F(t)$ exist for all real t, and we have

$$F'(t) = tL''(t), \qquad F''(t) = L''(t) + tL'''(t),$$
$$F'''(t) = 2L'''(t) + tL^{(4)}(t).$$

Since $F(0) = 0$ and $F(t)$ is strictly increasing for $t > 0$, we have $s_c \to 0$ as $c \to 0$.

As $t \to 0$,

$$(5.13) \quad F(t) = \tfrac{1}{2} L''(0)t^2 + \tfrac{1}{3} L'''(0)t^3 + O(t^4).$$

It is easy to calculate that

$$(5.14) \quad L(0) = 0, \quad L'(0) = -I(p^0, p), \quad L''(0) = m_2, \quad L'''(0) = -m_3,$$

where m_j is defined by (5.10). Hence as $c \to 0$,

$$(5.15) \quad c = \tfrac{1}{2} m_2 s_c^{\,2} - \tfrac{1}{3} m_3 s_c^{\,3} + O(s_c^{\,4}).$$

This implies

$$(5.16) \quad s_c = (2/m_2)^{\frac{1}{2}} c^{\frac{1}{2}} + \tfrac{2}{3}(m_3/m_2^{\,2})\, c + O(c^{\frac{3}{2}}).$$

Now

$$L'(s_c) = L'(0) + L''(0)s_c + \tfrac{1}{2}L'''(0)s_c^{\,2} + O(s_c^{\,3})$$
$$= -I(p^0, p) + m_2 s_c - \tfrac{1}{2} m_3 s_c^{\,2} + O(s_c^{\,3}).$$

With (5.16) this yields

$$(5.17) \quad L'(s_c) = -I(p^0, p) + (2m_2)^{\frac{1}{2}} c^{\frac{1}{2}} - \tfrac{1}{3}(m_3/m_2)c + O(c^{\frac{3}{2}}).$$

The expansion (5.9) follows from (5.11) and (5.17).

Now let Λ be a non-empty subset of Ω and

$$(5.18) \quad B' = \{x \mid I(x, \Lambda) < c\}.$$

We have

$$(5.19) \quad B' = \bigcup_{p^0 \varepsilon \Lambda} B'(p^0), \qquad B'(p^0) = \{x \mid I(x, p^0) < c\}.$$

Hence

$$I(B', p) = \inf \{I(B'(p^0), p) \mid p^0 \varepsilon \Lambda\}. \tag{5.20}$$

For each p^0, $I(B'(p^0), p)$ can be obtained from Theorem 5.1. Thus the problem of evaluating $I(B', p)$ is reduced to that of minimizing $I(B'(p^0), p)$ for $p^0 \varepsilon \Lambda$.

Alternatively, if the function $I(x, \Lambda)$ is sufficiently regular, $I(B', p)$ can be evaluated by applying Lemma 4.7 with $f(x) = c - I(x, \Lambda)$.

We conclude this section with a lemma which will be used in Section 6. Suppose that $p \varepsilon \Omega_0$ and

$$0 < c < I(p, \Lambda). \tag{5.21}$$

By Lemma 4.2 there is at least one point y such that

$$y \varepsilon \bar{B'}, \qquad I(y, p) = I(B', p), \tag{5.22}$$

and y must be in the boundary of B'. (Note that in general the set B' is not convex and there may be more than one minimizing point y.)

By Lemma 4.3, for each y there is at least one point $\hat{p}(y)$ such that

$$\hat{p}(y) \varepsilon \bar{\Lambda}, \qquad I(y, \hat{p}(y)) = I(y, \Lambda). \tag{5.23}$$

Let

$$B_y' = \{x \mid I(x, \hat{p}(y)) < I(y, \hat{p}(y))\}. \tag{5.24}$$

(Note that if $I(\cdot, \Lambda)$ is continuous then $I(y, \Lambda) = c$.)

LEMMA 5.1. *Let B' be defined by* (5.18). *Suppose that $p \varepsilon \Omega_0$ and Condition* (5.21) *is fulfilled. Let y and $\hat{p}(y)$ be points which satisfy* (5.22) *and* (5.23). *Then the set B_y' defined by* (5.24) *is a subset of B' and*

$$I(B', p) = I(B_y', p). \tag{5.25}$$

PROOF. Since $\hat{p}(y) \varepsilon \bar{\Lambda}$ and, by Lemma 4.3, $I(x, \Lambda) = I(x, \bar{\Lambda})$, we have $I(x, \Lambda) \leqq I(x, \hat{p}(y))$ for all x. Since $y \varepsilon \overline{B'}$ and $I(\cdot, \Lambda)$ is lower semicontinuous by Lemma 4.4, $I(y, \hat{p}(y)) = I(y, \Lambda) \leqq c$. It follows that $B_y' \subset B'$.

This implies $I(B', p) \leqq I(B_y', p)$. On the other hand, since $I(y, \hat{p}(y)) \leqq c < \infty$ and $I(\cdot, \hat{p}(y))$ is continuous in $\Omega(\hat{p}(y))$, we have $y \varepsilon \overline{B_y'}$. Hence $I(B', p) = I(y, p) \geqq I(\overline{B_y'}, p) = I(B_y', p)$, and (5.25) follows.

6. The set of alternatives at which a likelihood ratio test is better than a given test. In this section we shall consider tests which satisfy certain regularity conditions and shall investigate the set of alternatives at which a likelihood ratio test of approximately the same size has a smaller error probability than the given test when N is sufficiently large.

DEFINITION 6.1. A sequence $\{A_N\}$ of subsets of Ω is said to be regular relative to a point p in Ω if

$$I(A_N^{(N)}, p) = I(A_N, p) + O(N^{-1} \log N). \tag{6.1}$$

A sequence $\{A_N\}$ is said to be regular relative to a subset Λ of Ω if

$$I(A_N^{(N)}, \Lambda) = I(A_N, \Lambda) + O(N^{-1} \log N). \tag{6.2}$$

A subset A of Ω is said to be regular relative to p (or Λ) if (6.1) (or (6.2)) holds with $A_N = A$.

Sufficient conditions for a sequence of sets to be regular relative to p are derived in the Appendix.

From Theorem 2.1 and Definition 6.1 we immediately obtain

THEOREM 6.1. *If the sequence $\{A_N\}$ is regular relative to p then*

$$P_N(A_N \mid p) = \exp\{-NI(A_N, p) + O(\log N)\}. \tag{6.3}$$

If $\{A_N\}$ is regular relative to Λ then

$$\sup_{p \varepsilon \Lambda} P_N(A_N \mid p) = \exp\{-NI(A_N, \Lambda) + O(\log N)\}. \tag{6.4}$$

We now state another version of Theorem 3.1 which compares a test A_N for testing the hypothesis $p \varepsilon \Lambda$ with a likelihood ratio test. Let

$$B(c) = \{x \mid I(x, \Lambda) \geqq c\}. \tag{6.5}$$

THEOREM 6.2. *Let $\{A_N\}$ be a sequence of sets regular relative to Λ and let*

$$c_N = I(A_N, \Lambda). \tag{6.6}$$

There exist positive numbers $\delta_N = O(N^{-1} \log N)$ such that

$$\sup_{p \varepsilon \Lambda} P_N(B(c_N + \delta_N) \mid p) \leqq \sup_{p \varepsilon \Lambda} P_N(A_N \mid p), \tag{6.7}$$

and for any $p \varepsilon \Omega$ such that the sequence $\{A_N'\}$ is regular relative to p,

$$\begin{aligned} &P_N(B'(c_N + \delta_N) \mid p) \\ &\qquad \leqq \exp\{-Nd_N(p) + Ne_N(p) + O(\log N)\} P_N(A_N' \mid p), \end{aligned} \tag{6.8}$$

provided that the two probabilities in (6.8) are different from 0, where

$$d_N(p) = I(B'(c_N), p) - I(A_N', p) \geqq 0, \tag{6.9}$$

$$e_N(p) = I(B'(c_N), p) - I(B'(c_N + \delta_N), p) \geqq 0. \tag{6.10}$$

The proof is clear is we note that relations (6.3) and (6.4) with $=$ replaced by $\leqq$ are true for arbitrary sets A_N. Hence to obtain inequalities (6.7) and (6.8) it is not necessary to assume that the sets $B(c_N + \delta_N)$ and their complements are regular.

The assumption that the probabilities in (6.8) are positive implies that $I(A_N', p)$ and $I(B'(c_N + \delta_N), p)$ are finite, so that the differences $d_N(p)$ and $e_N(p)$ are defined.

By (6.8), if (i) $N\, d_N(p)/\log N \to \infty$ and (ii) $e_N(p)/d_N(p) \to 0$, then the ratio $P_N(B'(c_N + \delta_N) \mid p)/P_N(A_N' \mid p)$ tends to 0 faster than any power of N. If $A_N = A$ is independent of N, so are $c_N = c$ and $d_N(p) = d(p)$, and Conditions (i) and (ii) reduce to $d(p) > 0$ and $e_N(p) \to 0$. The latter is true if $I(B'(c), p)$

is a continuous function of c, and then we need only determine the set of points p for which $d(p)$ is positive. In the general case the set where $d_N(p) > 0$ is also of primary importance, as will be seen in the sequel. The following theorem gives a characterization of this set. To simplify the notation we omit the subscripts N.

THEOREM 6.3. *Let A and Λ be non-empty subsets of Ω such that* $0 < I(A, \Lambda) < \infty$. *Let*

$$B = \{x \mid I(x, \Lambda) \geqq I(A, \Lambda)\}, \tag{6.11}$$

and for any p such that $I(A', p) < \infty$, let

$$d(p) = I(B', p) - I(A', p). \tag{6.12}$$

(I) *Always $d(p) \geqq 0$; $d(p) = 0$ if and only if*

$$\{x \mid I(x, p) < I(B', p)\} \subset A. \tag{6.13}$$

(II) *If $d(p) = 0$ and $0 < I(B', p) < \infty$ then $I(B', p) = I(y, p)$ for some common boundary point y of A* and *B.*

(III) *Suppose that $d(p) = 0$ and $0 < I(B', p) < \infty$. Let y be a common boundary point of A and B such that $I(y, p) = I(B', p)$, and let $\hat{p}(y)$ be a point in $\bar{\Lambda}$ such that $I(y, \hat{p}(y)) = I(y, \Lambda)$. If p and y are in Ω_0 then $\hat{p}(y) \varepsilon \Omega_0$ and p is on the curve*

$$p = p(t), \qquad -\infty < t < 0, \tag{6.14}$$

where

$$p_i(t) = \hat{p}_i(y)^t y_i^{1-t} / \textstyle\sum_{j=1}^k \hat{p}_j(y)^t y_j^{1-t}, \qquad i = 1, \cdots, k. \tag{6.15}$$

PROOF.

(I) Since $A \subset B$, $d(p) \geqq 0$. If $d(p) = 0$ then $x \varepsilon A'$ implies $I(x, p) \geqq I(B', p)$, which is equivalent to (6.13). If (6.13) is satisfied then $x \varepsilon A'$ implies $I(x, p) \geqq I(B', p)$, hence $I(A', p) \geqq I(B', p)$ and therefore $d(p) = 0$.

(II) Suppose that $d(p) = 0$ and $0 < I(B', p) < \infty$. First assume $p \varepsilon \Omega_0$. By Lemma 4.2, $I(B', p) = I(y, p)$, where y is in the boundary of B'. Since $B' \subset A'$, y is in $\overline{A'}$, and $I(y, p) = I(A', p)$. Again by Lemma 4.2, y is in the boundary of A'. Thus y is a common boundary point of A and B.

If $p \not\varepsilon \Omega_0$, the proof is analogous, with reference to Lemma 4.2(b).

(III) Under the assumptions of Part (III) the conditions of Lemma 5.1 with $c = I(A, \Lambda)$ are satisfied. Hence the set

$$B_y' = \{x \mid I(x, \hat{p}(y)) < I(y, \hat{p}(y))\}$$

is a subset of B' and $I(y, p) = I(B', p) = I(B_y', p)$. (This is true without the assumption $y \varepsilon \Omega_0$.) Since $y \varepsilon \overline{B'}$, we have $I(y, \hat{p}(y)) = I(y, \Lambda) \leqq I(A, \Lambda) < \infty$. Hence $y \varepsilon \Omega(\hat{p}(y))$. In particular, if $y \varepsilon \Omega_0$ then $\hat{p}(y) \varepsilon \Omega_0$.

It follows from Theorem 5.1 with $p^0 = \hat{p}(y)$ (or, more directly, by an argument used in the proof of that theorem) that

$$\log (y_i/p_i) = -a \log (y_i/\hat{p}_i(y)) + b, \qquad i = 1, \cdots, k,$$

where $a > 0$. This is equivalent to (6.14) and (6.15) with $t = -a < 0$. The proof is complete.

REMARKS ON THEOREM 6.3. We have excluded the case $I(A', p) = \infty$, which implies $I(B', p) = \infty$ and $P_N(A', p) = P_N(B', p) = 0$.

If $I(B', p) = 0$, that is, $p \,\varepsilon\, \bar{B'}$, then clearly $d(p) = 0$. (In this case the set on the left of (6.13) is empty.) At such alternatives p the error probabilities $P_N(A' \mid p)$ and $P_N(B' \mid p)$ can not be very small.

The alternatives p of interest to us are those for which $I(B', p) > 0$. The conditions $I(A', p) < \infty$ and $d(p) = 0$ imply $I(B', p) < \infty$. If $I(\cdot, \Lambda)$ is continuous then $I(B', p) > 0$ if and only if $I(p, \Lambda) > I(A, \Lambda)$. (Note that if $A = A_N$ depends on N in such a way that $I(A_N, \Lambda) \to 0$ then $I(p, \Lambda) > I(A_N, \Lambda)$ for each $p \not\varepsilon\, \bar{\Lambda}$ for N large enough.)

Theorem 6.3 shows that the set of points p for which $I(B', p) > 0$ and $d(p) = 0$ essentially depends on the set of common boundary points of A and B. Suppose, in particular, that the test with critical region $A^{(N)}$ differs sufficiently from a likelihood ratio test in the sense that the sets A and B have only finitely many common boundary points y. Under some additional conditions Theorem 6.3 implies that the set of points p with $d(p) = 0$ is small in a specified sense; this is made precise in the corollary stated below.

To simplify the statement of the theorem we have assumed in Part (III) that y as well as p are in Ω_0. For the general case Theorem 5.1 implies a similar result except that, for given points y and $\hat{p}(y)$, the set where $d(p) = 0$ may be of more than one dimension. (Compare Example 9.2 in Section 9.)

Under the assumptions of Part (III) the condition that p is on the curve (6.14) is necessary but not in general sufficient for $d(p) = 0$. It is sufficient if, for instance, the complement A' of A is convex.

We state the following implication of Theorem 6.3.

COROLLARY 6.3.1. *Let* $0 < I(A, \Lambda) < \infty$ *and let* B *and* $d(p)$ *be defined by* (6.11) *and* (6.12). *Suppose that the number of common boundary points* y *of* A *and* B *is finite; that all these points* y *are in* Ω_0; *and that for each* y *there are only finitely many points* $\hat{p}(y) \,\varepsilon\, \bar{\Lambda}$ *such that* $I(y, \hat{p}(y)) = I(y, \Lambda)$. *Then if* $I(B', p) > 0$ *and* $p \,\varepsilon\, \Omega_0$, *we have* $d(p) > 0$ *except perhaps when* p *is on one of the finitely many curves* (*one for each pair* $(y, \hat{p}(y))$) *defined by* (6.14) *and* (6.15).

In the special case of a simple hypothesis, where Λ consists of a single point p^0, we have $B = \{x \mid I(x, p^0) \geqq I(A, p^0)\}$. Here $\hat{p}(y) = p^0$ for all y. If $p^0 \,\varepsilon\, \Omega_0$ then, by Lemma 4.6, the condition $I(A, p^0) < -\log(1 - p^0_{\min})$ is sufficient for all common boundary points of A and B to be in Ω_0.

We conclude this section with a lemma concerning the behavior of $e_N(p)$ as defined in (6.10) for the case where Λ consists of a single point p^0.

LEMMA 6.1. *Let* $p^0 \,\varepsilon\, \Omega_0$, $p \,\varepsilon\, \Omega_0$, $B'(c) = \{x \mid I(x, p^0) < c\}$. *Then as* $\delta \to 0+$,

$$I(B'(c), p) - I(B'(c + \delta), p) = O(\delta c^{-\frac{1}{2}}) \tag{6.16}$$

uniformly for $0 < c < I(p, p^0) - \gamma$, *where* γ *is any fixed positive number.*

PROOF. Let $J(c) = I(B'(c), p)$. By Theorem 5.1, $J(c) = c - L'(s_c)$ for

$0 < c < I(p, p^0)$, where $0 < s_c < 1, F(s_c) = c$, $F(t) = tL'(t) - L(t)$, $L(t) = \log M(t)$, and $M(t)$ is defined in (5.7). For the derivative $s_c' = ds_c/dc$ we have $F'(s_c)s_c' = 1$. Since $F'(t) = tL''(t)$, we obtain $L''(s_c)s_c s_c' = 1$. Hence

$$J'(c) = 1 - L''(s_c)s_c' = 1 - s_c^{-1},$$
$$J''(c) = s_c^{-2}s_c' = s_c^{-3}L''(s_c)^{-1} > 0.$$

Therefore for $\delta > 0$,

$$0 > J(c + \delta) - J(c) \geqq \delta J'(c) = -\delta(1 - s_c)/s_c .$$

For c bounded away from 0 and $I(p, p^0)$, s_c is bounded away from 0 and 1. As $c \to 0$, $s_c \sim (2/m_2)^{\frac{1}{2}}c^{\frac{1}{2}}$ by (5.16), where $m_2 > 0$. This implies the lemma.

7. The likelihood ratio test of a simple hypothesis. The likelihood ratio test for testing the simple hypothesis $p = p^0$ rejects the hypothesis if $z^{(N)} \varepsilon B(c_N)$, where

$$B(c) = \{x \mid I(x, p^0) \geqq c\}. \tag{7.1}$$

The following theorem gives approximations for the error probabilities of this test.

THEOREM 7.1. *For any $p^0 \varepsilon \Omega$ and any number sequence $\{c_N\}$ we have*

$$P_N(B(c_N) \mid p^0) = \exp\{-Nc_N + O(\log N)\}. \tag{7.2}$$

If $p^0 \varepsilon \Omega_0$, $p \varepsilon \Omega_0$, and

$$N^2 c_N \to \infty \quad as \quad N \to \infty, \tag{7.3}$$

then

$$P_N(B'(c_N) \mid p) = \exp\{-NI(B'(c_N), p) + O(\log N)\}, \tag{7.4}$$

where $I(B'(c_N), p) = I(B', p)$ is evaluated in Theorem 5.1 with $c = c_N$.

PROOF. Since the sets $B'(c_N)$ are convex, the sequence $\{B(c_N)\}$ is regular relative to p^0 by Theorem A.1 of the Appendix. We have $I(B(c_N), p^0) = c_N$. Hence (7.2) follows from Theorem 6.1.

Now suppose that $p^0 \varepsilon \Omega_0$ and $p \varepsilon \Omega_0$. Let $c_N > 0$ and assume with no loss of generality that $c_N < \infty$. We apply Theorem A.2 of the Appendix with $f(x) = -I(x, p^0)$ and c_N replaced by $-c_N$. It follows from Theorem 5.1 with $c = c_N$ that for each N there is a unique point $y^{(N)}$ such that $I(y^{(N)}, p^0) \leqq c_N$ and $I(y^{(N)}, p) = I(B'(c_N), p)$. Moreover, $y_i^{(N)} \geqq \min(p_i^0, p_i) > 0$ for all i and all N. Hence the conditions of Theorem A.2 up to (A.9) are satisfied. Condition (A.10) is satisfied if $Ns_N \to \infty$ as $N \to \infty$, where $s_N = s > 0$ is defined in Theorem 5.1. If c_N is bounded away from 0, so is s_N. If $c_N \to 0$ then, by (5.16), s_N is asymptotically proportional to $c_N^{\frac{1}{2}}$. Thus Condition (7.3) is sufficient for the sequence $\{B'(c_N)\}$ to be regular relative to p. Equation (7.4) now follows from Theorem 6.1.

REMARK. The case where Condition (7.3) is not satisfied is of no statistical interest since if $N^2 c_N$ is bounded, the size of the test tends to 1. If $c_N < a^2/N^2$

and a is sufficiently small, the set $B'(c_N)^{(N)}$ is empty for infinitely many N and hence the sequence $\{B'(c_N)\}$ is not regular relative to any p in Ω_0.

8. Chi-square and likelihood ratio tests of a simple hypothesis. Let p^0 be a point in Ω_0. Let

$$Q^2(x, p^0) = \sum_{i=1}^k (x_i - p_i^0)^2/p_i^0. \tag{8.1}$$

The chi-square test for testing the simple hypothesis $H: p = p^0$ rejects H if $Q^2(z^{(N)}, p^0) \geqq \epsilon_N^2$, where ϵ_N is a positive number. We shall compare this test with the likelihood ratio test which rejects H if $I(z^{(N)}, p^0) \geqq c_N$, where c_N is so chosen that the two tests have approximately the same size.

It is well known that if $p = p^0$ then the random variables $NQ^2(Z^{(N)}, p^0)$ and $2NI(Z^{(N)}, p^0)$ have the same limiting χ^2 distribution with $k - 1$ degrees of freedom. Hence if $\epsilon_N^2 = 2c_N = 2c/N$, where c is a positive constant, the sizes of the two tests converge to the same positive limit. In fact, in this case the critical regions of the two tests differ very little from each other when N is large. Indeed, we have

$$I(x, p^0) = \tfrac{1}{2}Q^2(x, p^0) + O(|x - p^0|^3), \tag{8.2}$$

where $|x - p^0|$ denotes the Euclidean distance between x and p^0. This implies that the set $\{x \mid Q^2(x, p^0) < 2c/N\}$ both contains and is contained in a set of the form $\{x \mid I(x, p^0) < c/N + O(N^{-\frac{3}{2}})\}$. Hence it can be shown that at any point $p \neq p^0$ which is in Ω_0 the ratio of the error probabilities of the two tests is bounded away from 0 and ∞ as $N \rightarrow \infty$. (If $p \not\in \Omega_0$, the error probabilities at p of both tests are zero for N sufficiently large.)

In this section it will be shown that if ϵ_N tends to 0 not too rapidly, then at "most" points p the error probability of the likelihood ratio test is much smaller than that of the chi-square test when N is large enough.

We first observe that

$$\begin{aligned} Q^2(x, p^0) &= \sum_{i=1}^k x_i(x_i - p_i^0)/p_i^0 \leqq \sum_{i=1}^k x_i(1 - p_i^0)/p_i^0 \\ &\leqq \max\,(1 - p_i^0)/p_i^0 = (1 - p_{\min}^0)/p_{\min}^0, \end{aligned} \tag{8.3}$$

where $p_{\min}^0 = \min p_i^0$. The upper bound is attained if and only if $x_j = 1$ and $x_i = 0, i \neq j$, for some j such that $p_j^0 = p_{\min}^0$.

Hence when we consider the test defined by $Q^2(x, p^0) \geqq \epsilon^2$, we may assume that $\epsilon^2 \leqq (1 - p_{\min}^0)/p_{\min}^0$. The case $\epsilon^2 = (1 - p_{\min}^0)/p_{\min}^0$ is trivial, and we shall restrict ourselves to the case of strict inequality. Note that $p_{\min}^0 \leqq 1/k$, and $p_{\min}^0 < \frac{1}{2}$ unless $k = 2$ and $p^0 = (\frac{1}{2}, \frac{1}{2})$.

Let

$$A(\epsilon) = \{x \mid Q^2(x, p^0) \geqq \epsilon^2\}. \tag{8.4}$$

THEOREM 8.1. *Suppose that* $p^0 \in \Omega_0$ *and*

$$0 < \epsilon < ((1 - p_{\min}^0)/p_{\min}^0)^{\frac{1}{2}}. \tag{8.5}$$

If r is the number of components p_j^0 *equal to* $p_{\min}^0$, *there are exactly r points y such that*

$$(8.6) \qquad y \varepsilon A(\epsilon), \qquad I(A(\epsilon), p^0) = I(y, p^0).$$

Explicitly, for each j such that $p_j^0 = p_{\min}^0$ *the corresponding point y is given by*

$$(8.7) \qquad y_i = b p_{\min}^0 \quad \text{if } i = j, \qquad y_i = a p_i^0 \quad \text{if } i \neq j,$$

$$(8.8) \qquad a = 1 - (p_{\min}^0/(1 - p_{\min}^0))^{\frac{1}{2}}\epsilon, \qquad b = 1 + ((1 - p_{\min}^0)/p_{\min}^0)^{\frac{1}{2}}\epsilon,$$

and we have

$$(8.9) \qquad I(A(\epsilon), p^0) = p_{\min}^0 b \log b + (1 - p_{\min}^0) a \log a.$$

Furthermore,

$$(8.10) \quad 2p_{\min}^0(1 - p_{\min}^0)\epsilon^2 \leqq \frac{p_{\min}^0(1 - p_{\min}^0)}{1 - 2p_{\min}^0} \log \frac{1 - p_{\min}^0}{p_{\min}^0} \epsilon^2 \leqq I(A(\epsilon), p^0) \leqq \epsilon^2,$$

where the second expression is to be replaced by $\frac{1}{2}\epsilon^2$ *if* $p_{\min}^0 = \frac{1}{2}$. *As* $\epsilon \to 0$,

$$(8.11) \qquad I(A(\epsilon), p^0) = \frac{1}{2}\epsilon^2 + \frac{1}{6}\frac{2p_{\min}^0 - 1}{(p_{\min}^0(1 - p_{\min}^0))^{\frac{1}{2}}}\epsilon^3 + O(\epsilon^4).$$

PROOF. Let y denote any point which satisfies (8.6). By Lemma 4.2 we must have $Q^2(y, p^0) = \epsilon^2$. It can be shown that necessarily $y \varepsilon \Omega_0$. (For ϵ small enough, $Q^2(y, p^0) = \epsilon^2$ implies $y \epsilon \Omega_0$. In general this result can be proved with the help of Lemma 4.6(b). The details are left to the reader.) By Lemma 4.7 we must have $\log\,(y_i/p_i^0) = sy_i/p_i^0 + t$, $i = 1, \cdots, k$, where $s > 0$. Hence y_i/p_i^0 can take at most two different values, say

$$(8.12) \qquad y_i = ap_i^0 \quad \text{if } i \varepsilon M, \qquad y_i = bp_i^0 \quad \text{if } i \not\varepsilon M,$$

where M is a non-empty proper subset of $\{1, \cdots, k\}$. The conditions $\sum y_i = 1$ and $Q^2(y, p^0) = \epsilon^2$ are equivalent to

$$(8.13) \qquad ah + b(1 - h) = 1, \qquad (a - 1)^2 h + (b - 1)^2(1 - h) = \epsilon^2,$$

$$(8.14) \qquad h = \textstyle\sum_{i \varepsilon M} p_i^0.$$

We may assume $a < b$. Then

$$(8.15) \qquad a = 1 - ((1 - h)/h)^{\frac{1}{2}}\epsilon, \qquad b = 1 + (h/(1 - h))^{\frac{1}{2}}\epsilon.$$

To satisfy $y_i > 0$ we must have $a > 0$, that is, $\epsilon^2 < h/(1 - h)$. If ϵ^2 is close to its upper bound $(1 - p_{\min}^0)/p_{\min}^0$, this condition is satisfied only when h takes its largest possible value, $1 - p_{\min}^0$. It will be shown that, for any ϵ, y satisfies (8.6) if and only if $h = 1 - p_{\min}^0$.

For y defined by (8.12) we have $I(y, p^0) = ha \log a + (1 - h)b \log b = f(h)$, say, where $a = a(h)$ and $b = b(h)$ are given by (8.15). By a straightforward calculation we obtain for the derivative of $f(h)$

$$f'(h) = b\{1 - (a/b) + \tfrac{1}{2}[1 + (a/b)] \log (a/b)\}.$$

The expression on the right is negative. Hence as h ranges over the values (8.14), $f(h)$ attains its minimum at $h = 1 - p^0_{\min}$. This implies that Condition (8.6) is satisfied if and only if y is one of the points defined by (8.7) and (8.8), and that $I(A(\epsilon), p^0)$ is given by (8.9).

The inequality $I(A(\epsilon), p^0) \leqq \epsilon^2$ in (8.10) follows from the general inequality

$$(8.16)\qquad I(x, p) = \sum x_i \log (x_i/p_i) \leqq \sum x_i((x_i/p_i) - 1) = Q^2(x, p).$$

The first two inequalities in (8.10) are contained in Theorem 1 of Hoeffding [4]. (Note that the closer lower bound in (8.10) is attained for $\epsilon = (1 - 2p^0_{\min})/(p^0_{\min}(1 - p^0_{\min}))^{\frac{1}{2}}$.)

The expansion (8.11) is easily verified. The proof is complete.

The next theorem gives the infimum of $I(x, p)$ subject to the condition $x \,\varepsilon\, A'(\epsilon)$, that is, $Q^2(x, p^0) < \epsilon^2$.

THEOREM 8.2. *Let $p^0 \,\varepsilon\, \Omega_0$, $p \,\varepsilon\, \Omega_0$, $\epsilon > 0$, $\epsilon^2 < Q^2(p, p^0)$. Then there is a unique point z such that*

$$(8.17)\qquad z \,\varepsilon\, \overline{A'(\epsilon)}, \qquad I(A'(\epsilon), p) = I(z, p).$$

The point z is determined by the conditions

$$(8.18)\qquad (z_i - p_i^0)/p_i^0 = -s_\epsilon \log (z_i/p_i) + t_\epsilon, \quad i = 1, \cdots, k;\ s_\epsilon > 0,$$

$$(8.19)\qquad \textstyle\sum_{i=1}^k z_i = 1, \qquad Q^2(z, p^0) = \epsilon^2.$$

As $\epsilon \to 0$,

$$(8.20)\qquad z_i = p_i^0 - m_2(p)^{-\frac{1}{2}} p_i^0(\log (p_i^0/p_i) - I(p^0, p))\epsilon + O(\epsilon^3),$$

$$(8.21)\qquad I(A'(\epsilon), p) = I(p^0, p) - m_2(p)^{\frac{1}{2}}\epsilon + \tfrac{1}{2}\epsilon^2 + O(\epsilon^3),$$

where

$$(8.22)\qquad m_j(p) = \textstyle\sum_{i=1}^k p_i^0(\log (p_i^0/p_i) - I(p^0, p))^j.$$

PROOF. By Lemma 4.8 there is exactly one point z which satisfies (8.17). By Lemma 4.7, z must satisfy (8.18) with $s_\epsilon > 0$. The constants s_ϵ and t_ϵ are determined by (8.19).

Now let $\epsilon \to 0$. The condition $Q^2(z, p^0) = \epsilon^2$ implies $z_i - p_i^0 = O(\epsilon)$, $i = 1, \cdots, k$. Hence

$$\log (z_i/p_i) = \log (p_i^0/p_i) + (z_i - p_i^0)/p_i^0 + O(\epsilon^2).$$

With (8.18) this gives

$$(8.23)\qquad (1 + s_\epsilon)(z_i - p_i^0)/p_i^0 = t_\epsilon - s_\epsilon(\log (p_i^0/p_i) + O(\epsilon^2)).$$

Multiplication of both sides with p_i^0 and summation yields

$$(8.24)\qquad t_\epsilon = s_\epsilon(I(p^0, p) + O(\epsilon^2)).$$

From (8.23) and (8.24) we obtain

$$(8.25)\qquad (z_i - p_i^0)/p_i^0 = -[s_\epsilon/(1 + s_\epsilon)](\log (p_i^0/p_i) - I(p^0, p) + O(\epsilon^2)).$$

If we square both sides of (8.25), multiply with p_i^0 and sum with respect to i, we find that $\epsilon^2 = [s_\epsilon/(1 + s_\epsilon)]^2(m_2(p) + O(\epsilon^2))$. Hence

$$s_\epsilon/(1 + s_\epsilon) = m_2(p)^{-\frac{1}{2}}\epsilon + O(\epsilon^3). \tag{8.26}$$

From (8.25) and (8.26) we obtain relation (8.20).

Now

$$I(z, p) = I(p^0, p) + \sum (z_i - p_i^0) \log (p_i^0/p_i) + \tfrac{1}{2}\epsilon^2 + O(\epsilon^3). \tag{8.27}$$

With (8.17) and (8.20) this implies (8.21). This completes the proof.

Let $A(\epsilon)$ be defined by (8.4) and let

$$B(\epsilon) = \{x \mid I(x, p^0) \geqq I(A(\epsilon), p^0)\}, \tag{8.28}$$

$$d(p, \epsilon) = I(B'(\epsilon), p) - I(A'(\epsilon), p). \tag{8.29}$$

THEOREM 8.3. *Let* $p^0 \varepsilon \Omega_0$, $0 < \epsilon < ((1 - p_{\min}^0)/p_{\min}^0)^{\frac{1}{2}}$.

(I) *If* $p \varepsilon \Omega_0$ *and* $Q^2(p, p^0) > \epsilon^2$ *then* $d(p, \epsilon) > 0$ *unless for some* j *with* $p_j^0 = p_{\min}^0$

$$p_j = 1 - a + ap_j^0; \qquad p_i = ap_i^0, \qquad i \neq j, \tag{8.30}$$

$$0 < a < 1 - (p_{\min}^0/(1 - p_{\min}^0))^{\frac{1}{2}}\,\epsilon. \tag{8.31}$$

(II) *As* $\epsilon \to 0$,

$$d(p, \epsilon) = \tfrac{1}{6}\, m_2(p)^{\frac{1}{2}}\Delta(p)\epsilon^2 + O(\epsilon^3), \tag{8.32}$$

where $m_j(p)$ *is defined in* (8.22) *and*

$$\Delta(p) = \frac{m_3(p)}{m_2(p)^{\frac{3}{2}}} + \frac{1 - 2p_{\min}^0}{(p_{\min}^0(1 - p_{\min}^0))^{\frac{1}{2}}}. \tag{8.33}$$

(III) *We have* $\Delta(p) \geqq 0$ *for all* $p \varepsilon \Omega_0$, $p \neq p^0$; *and* $\Delta(p) > 0$ *unless* p *satisfies* (8.30) *with* $a \neq 1$, $0 < a < (1 - p_{\min}^0)^{-1}$ *for some* j *such that* $p_j^0 = p_{\min}^0$.

PROOF. Part (I) follows from Theorems 6.3, 8.1 and 8.2. (The parameter a in (8.30) is a function of the parameter t in Theorem 6.3.)

(II) By Theorem 8.1, as $\epsilon \to 0$,

$$I(A(\epsilon), p^0) = \frac{1}{2}\epsilon^2 + \frac{1}{6}\frac{2p_{\min}^0 - 1}{(p_{\min}^0(1 - p_{\min}^0))^{\frac{1}{2}}}\epsilon^3 + O(\epsilon^4). \tag{8.34}$$

By Theorem 5.2,

$$I(B'(\epsilon), p) = I(p^0, p) - (2m_2(p))^{\frac{1}{2}}c^{\frac{1}{2}} + \left(1 + \frac{m_3(p)}{3m_2(p)}\right)c + O(c^{\frac{3}{2}}), \tag{8.35}$$

where $c = I(A(\epsilon), p^0)$. From (8.34) and (8.35) we obtain after simplification

$$\begin{aligned} I(B'(\epsilon), p) &= I(p^0, p) - m_2(p)^{\frac{1}{2}}\epsilon \\ &\quad + \left\{\frac{1}{2} + \frac{1}{6}\frac{m_3(p)}{m_2(p)} - \frac{1}{6}\, m_2(p)^{\frac{1}{2}}\frac{2p_{\min}^0 - 1}{(p_{\min}^0(1 - p_{\min}^0))^{\frac{1}{2}}}\right\}\epsilon^2 + O(\epsilon^3). \end{aligned} \tag{8.36}$$

By Theorem 8.2,

$$I(A'(\epsilon), p) = I(p^0, p) - m^2(p)^{\frac{1}{2}} \epsilon + \tfrac{1}{2}\epsilon^2 + O(\epsilon^3). \tag{8.37}$$

The expression (8.32) for $d(p, \epsilon)$ follows from (8.36) and (8.37).

(III) Let $u_i = (\log (p_i^0/p_i) - I(p^0, p))/m_2(p)^{\frac{1}{2}}$,

$$u = (u_1, \cdots, u_k), \qquad \mu_j(u) = \sum_{i=1}^k p_i^0 u_i^j. \tag{8.38}$$

Then $\mu_1(u) = 0$, $\mu_2(u) = 1$, $\mu_3(u) = m_3(p)/m_2(p)^{\frac{3}{2}}$. Part (III) of Theorem 8.3 is an immediate consequence of the following lemma. (Note that $\Delta(p) \geqq 0$ is implied by $d(p, \epsilon) \geqq 0$. The lemma gives the conditions for equality.)

LEMMA 8.1. *Let $\mu_j(u)$ be defined by (8.38), where $p^0 \varepsilon \Omega_0$ and $u_1, \cdots, u_k$ are any real numbers such that $\mu_1(u) = 0$ and $\mu_2(u) = 1$. Then*

$$\mu_3(u) \geqq (2p_{\min}^0 - 1)/(p_{\min}^0(1 - p_{\min}^0))^{\frac{1}{2}}. \tag{8.39}$$

The sign of equality holds if and only if for some j such that $p_j^0 = p_{\min}^0$

$$u_j = -((1 - p_{\min}^0)/p_{\min}^0)^{\frac{1}{2}}; \qquad u_i = (p_{\min}^0/(1 - p_{\min}^0))^{\frac{1}{2}}, \qquad i \neq j. \tag{8.40}$$

PROOF. Since $p^0 \varepsilon \Omega_0$, the set of points u defined by $\mu_1(u) = 0$, $\mu_2(u) = 1$ is bounded and closed. Hence $\mu_3(u)$ has a finite minimum in this set. An application of the method of Lagrange multipliers shows that for u to be a minimizing point it is necessary that u_i take only two values, say $u_i = a$ if $i \varepsilon M$, $u_i = b$ if $i \not\varepsilon M$, $a > b$. The conditions $\mu_1(u) = 0$, $\mu_2(u) = 1$ imply $\mu_3(u) = (1 - 2h)/(h(1 - h))^{\frac{1}{2}}$, where $h = \sum_{i \varepsilon M} p_i^0$. The minimum with respect to h of this ratio is attained at $h = 1 - p_{\min}^0$, and the lemma follows.

The following lemma establishes the regularity of the sequences of sets $\{A(\epsilon_N)\}$ and $\{A'(\epsilon_N)\}$ under general conditions.

LEMMA 8.2. *Let $A(\epsilon) = \{x \mid Q^2(x, p^0) \geqq \epsilon^2\}$. For any p^0 and any ϵ_N the sequence $\{A(\epsilon_N)\}$ is regular relative to p^0. If $p_0 \varepsilon \Omega_0$, $p \varepsilon \Omega_0$, and*

$$N^2 \epsilon_N^2 \to \infty \quad as\ N \to \infty, \tag{8.41}$$

the sequence $\{A'(\epsilon_N)\}$ is regular relative to p.

The proof closely parallels the proof of Theorem 7.1 and uses Theorem 8.2 and Equation (8.20). A remark analogous to that after the proof of Theorem 7.1 applies to the present case.

We now can state the following result about the relative performance of a chi-square test and a likelihood ratio test of a simple hypothesis.

THEOREM 8.4. *Let $p^0 \varepsilon \Omega_0$ and $0 < \epsilon_N < ((1 - p_{\min}^0)/p_{\min}^0)^{\frac{1}{2}}$.*

(I) *For the error probabilities of the chi-square test which rejects the hypothesis $p = p^0$ if $z^{(N)} \varepsilon A(\epsilon_N) = \{x \mid Q^2(x, p^0) \geqq \epsilon_N^2\}$ we have*

$$P_N(A(\epsilon_N) \mid p^0) = \exp\{-NI(A(\epsilon_N), p^0) + O(\log N)\}, \tag{8.42}$$

where $I(A(\epsilon), p^0)$ is given explicitly in Theorem 8.1; and if

$$p \varepsilon \Omega_0, \qquad N^2 \epsilon_N^2 \to \infty \quad as\ N \to \infty, \tag{8.43}$$

then

$$(8.44)\qquad P_N(A'(\varepsilon_N) \mid p) = \exp\{-NI(A'(\epsilon_N), p) + O(\log N)\},$$

where $I(A'(\epsilon_N), p)$ *is given in Theorem* 8.2.

(II) *There exist positive constants* $\delta_N = O(N^{-1}\log N)$ *such that for the likelihood ratio test which rejects* $p = p^0$ *if*

$$z^{(N)} \varepsilon B_N = \{x \mid I(x, p^0) \geqq I(A(\epsilon_N), p^0) + \delta_N\}$$

we have

$$(8.45)\qquad P_N(B_N \mid p^0) \leqq P_N(A(\epsilon_N) \mid p^0);$$

and if Conditions (8.43) *are satisfied and* $\epsilon_N^2 < Q^2(p, p^0) - \beta$ *for some* $\beta > 0$, *then*

$$(8.46)\qquad \begin{aligned} P_N(B_N' \mid p) = \exp\{&-N\,d(p, \epsilon_N) \\ &+ O(\log N/I^{\frac{1}{2}}(A(\epsilon_N), p^0))\}P_N(A'(\epsilon_N) \mid p), \end{aligned}$$

where $d(p, \epsilon)$ *is defined in* (8.29) *and has the properties stated in Theorem* 8.3.

In particular, if

$$(8.47)\qquad \epsilon_N \to 0, \qquad N\epsilon_N^3/\log N \to \infty \qquad as\ N \to \infty,$$

then at each point $p \varepsilon \Omega_0$, $p \neq p^0$ *which does not lie on one of the line segments*

$$(8.48)\qquad p_j = 1 - a + ap_j^0; \qquad p_i = ap_i^0,\ i \neq j; \qquad 0 < a < 1;\ p_j^0 = p_{\min}^0,$$

the likelihood ratio test B_N *is more powerful than the chi-square test* $A(\epsilon_N)$ *when* N *is sufficiently large.*

Proof. Part (I) follows from Theorem 6.1 and Lemma 8.2.

Part (II) follows from Theorem 6.2, Lemma 6.1, and Theorems 7.1, 8.1 and 8.3. The assumption $\epsilon_N^2 < Q^2(p, p^0) - \beta, \beta > 0$, implies $I(A(\epsilon_N), p^0) < I(p, p^0) - \gamma$ for some $\gamma > 0$, as required in Lemma 6.1. The equality in (8.46) follows from Theorem 7.1.

Remark. The line segments (8.48) connect the point p^0 with some of the vertices of the simplex Ω. For any finite N the likelihood ratio test is more powerful than the chi-square test except in a certain neighborhood of these line segments, which depends on ϵ_N. It would be interesting to determine the extent of this neighborhood for moderate values of N and selected values of ϵ_N (that is, of the size of the test).

9. Chi-square and likelihood ratio tests of a composite hypothesis. There is reason to believe that in the case of a composite hypothesis the relation between a chi-square test and a likelihood ratio test in general is analogous to that in the case of a simple hypothesis (see Section 8), with a notable exception mentioned below. For a chi-square test of a composite hypothesis the determination of the common boundary points of the sets A and B (in the notation of Theorem 6.3) is somewhat cumbersome. We therefore present no general results. We first shall show by an example that for one common version of the chi-square test it may happen that the size of the test is never smaller than some power of N; if this is the case, our theory is not applicable. We then give a simple example where the situation is analogous to the case of a simple hypothesis.

There are several versions of the chi-square test for testing a composite hypothesis, $p \varepsilon \Lambda$. One is the minimum chi-square test which is based on the statistic $Q^2(z^{(N)}, \Lambda)$, where

$$(9.1) \qquad Q^2(x, \Lambda) = \inf \{Q^2(x, p) \mid p \varepsilon \Lambda\}.$$

Here $Q^2(x, p)$ is defined by (8.1), with the convention that $(x_i - p_i)^2/p_i = 0$ if $x_i = p_i = 0$. The calculation of $Q^2(x, \Lambda)$ is cumbersome for some of the common hypotheses. When a maximum likelihood estimator $\hat{p}(x)$ of p under the assumption $p \varepsilon \Lambda$ (as defined in Lemma 4.3) is available one often resorts to the test based on

$$(9.2) \qquad \hat{Q}^2(x) = Q^2(x, \hat{p}(x)).$$

If the size of the test is held fixed as N increases and the set Λ is sufficiently regular, the tests based on $Q^2(x, \Lambda)$ or $\hat{Q}^2(x)$ differ little from a likelihood ratio test based on $I(x, \Lambda)$, just as in the case of a simple hypothesis. However, if we require that the size of the test tend to 0 more rapidly than a certain power of N, it turns out that this requirement can not in general be satisfied with a $\hat{Q}^2$ test.

Let $\bar{\epsilon}_N^2$ denote the maximum of $\hat{Q}^2(z^{(N)})$ for $z^{(N)} \varepsilon \Omega^{(N)}$. The $\hat{Q}^2$ test of smallest positive size for testing the hypothesis $H: p \varepsilon \Lambda$ rejects H if and only if $\hat{Q}^2(z^{(N)}) = \bar{\epsilon}_N^2$. Suppose that this critical region contains a point $z^{(N)}$ which is close to Λ in the sense that $I(z^{(N)}, \Lambda) = I(z^{(N)}, \hat{p}(z^{(N)}))$ is of order $N^{-1} \log N$. Then, by (3.3), the smallest positive size of a $\hat{Q}^2$ test is not smaller than some power of N. The following example serves to illustrate this phenomenon.

EXAMPLE 9.1: *Hypothesis of independence in a contingency table.* Let the $k = rs$ components of $x \varepsilon \Omega$ be denoted by x_{ij}, $i = 1, \cdots, r; j = 1, \cdots, s$, where $r \geqq 2$, $s \geqq 2$. Define $x_i^{(1)} = \sum_j x_{ij}$, $x_j^{(2)} = \sum_i x_{ij}$. Let

$$(9.3) \qquad \Lambda = \{p \mid p_{ij} = p_i^{(1)} p_j^{(2)}, i = 1, \cdots, r; j = 1, \cdots, s\}.$$

Then $\hat{p}_{ij}(x) = x_i^{(1)} x_j^{(2)}$ and

$$\hat{Q}^2(x) = \textstyle\sum_{i=1}^r \sum_{j=1}^s [x_{ij}^2/(x_i^{(1)} x_j^{(2)})] - 1,$$

where, by definition, the terms with $x_i^{(1)} x_j^{(2)} = 0$ are zero. For simplicity let $r = s$. Then, due to $x_{ij}^2 \leqq x_{ij} x_j^{(2)}$, $\hat{Q}^2(x) \leqq r - 1$, with equality holding if and only if each row and each column of the matrix x_{ij} contains exactly one non-zero element. Let $z^{(N)}$ denote the point defined by $z_{11}^{(N)} = 1 - (r - 1)/N$; $z_{ii}^{(N)} = 1/N$, $i = 2, \cdots, r$; $z_{ij}^{(N)} = 0$, $i \neq j$. Then $\hat{Q}^2(z^{(N)}) = r - 1$ and

$$P_N(z^{(N)} \mid \hat{p}(z^{(N)})) = [N!/(N - r + 1)!]\{1 - [(r - 1)/N]\}^{2(N-r+1)} N^{-2(r-1)}$$
$$\sim e^{-2r+2} N^{-r+1}.$$

Thus the smallest positive size of a $\hat{Q}^2$ test is proportional to N^{-r+1}.

In contrast, the size of the likelihood ratio test which rejects H if $I(z^{(N)}, \Lambda) \geqq c$ does not exceed $\exp\{-Nc + O(\log N)\}$.

In the following example the situation is similar to the case of a simple hypothesis.

EXAMPLE 9.2. Let $x = (x_1, \cdots, x_k)$, $k \geqq 3$,

$$\Lambda = \{p \mid p_1 = p_2\}.$$

We have $\hat{p}_i(x) = (x_1 + x_2)/2, i = 1, 2; \hat{p}_i(x) = x_i, i > 2$. Hence

$$\hat{Q}^2(x) = 4(x_1 + x_2)\{[x_1/(x_1 + x_2)] - \tfrac{1}{2}\}^2,$$

$\max \hat{Q}^2(x) = 1$. Let $A = \{x \mid \hat{Q}^2(x) \geqq \epsilon^2\}, 0 < \epsilon < 1$. It can be shown that the sets A and $B = \{x \mid I(x, \Lambda) \geqq I(A, \Lambda)\}$ have exactly two common boundary points, $y^1 = ((1 - \epsilon)/2, (1 + \epsilon)/2, 0, \cdots, 0)$, $y^2 = ((1 + \epsilon)/2, (1 - \epsilon)/2, 0, \cdots, 0)$. Since these points are not in Ω_0, Part (III) of Theorem 6.3 is not directly applicable. It is not difficult to show that if $I(p, \Lambda) > I(A, \Lambda)$ then $d(p) > 0$ unless $p_1 = 0$ or $p_2 = 0$. Thus if $k \geqq 4$, the set of points p such that $I(B', p) > 0$ and $d(p) = 0$ is of more than one dimension. Since, however, the present hypothesis set Λ is such that we have effectively $k = 3$ components, the result may be said to be analogous to that in the case of a simple hypothesis.

10. Some competitors of the chi-square test. There are a number of test statistics for testing a simple hypothesis which have the same asymptotic distribution as the chi-square statistic when the hypothesis is true. As an example consider the test which rejects the hypothesis $p = p^0$ if $D_1^2(z^{(N)}, p^0)$ exceeds a constant, where

$$D_1^2(x, p) = \textstyle\sum_{i=1}^k (x_i^{\frac{1}{2}} - p_i^{\frac{1}{2}})^2 \tag{10.1}$$

(see Matusita [5]).

For $p^0 \in \Omega_0$ we have

$$D_1^2(x, p^0) = \tfrac{1}{4}Q^2(x, p^0) + O(|x - p^0|^3).$$

Thus if the size of the test is bounded away from zero, the test behaves asymptotically as the chi-square test and differs little from the likelihood ratio test.

Let $A = \{x \mid D_1^2(x, p^0) \geqq \epsilon^2\}, 0 < \epsilon^2 < 2$. It is easily seen that there are only finitely many points y in A for which $I(y, p^0) = I(A, p^0)$. Just as in the case of the chi-square test they are such that the ratio y_i/p_i^0 takes only two different values. If the size of the test tends to 0 at an appropriate rate, the test compares with the likelihood ratio test in a similar way as the chi-square test.

For testing a composite hypothesis we may use the test based on $D_1^2(x, \hat{p}(x))$. It can be shown that in the case of Example 9.1 the size of this test may decrease at an exponential rate, in contrast to the analogous chi-square test.

Another interesting class of tests is defined in terms of the distances

$$D(x, p) = \max_{M \in \mathfrak{M}} \textstyle\sum_{i \in M}(x_i - p_i), \tag{10.2}$$

where $\mathfrak{M}$ is a family of subsets of $\{1, \cdots, k\}$. If $\mathfrak{M}$ contains all these subsets then $D(x, p) = \frac{1}{2}\sum |x_i - p_i|$. If $\mathfrak{M}$ consists of the sets $\{1, \cdots, i\}$ and $\{i, \cdots, k\}$ for $i = 1, \cdots, k$, then $D(z^{(N)}, p^0)$ may be identified with the Kolmogorov statistic (discrete case).

Let $p^0 \varepsilon \Omega_0$, $0 < \epsilon < \max_x D(x, p^0)$. The set $A = \{x \mid D(x, p^0) \geqq \epsilon\}$ is the union of the half-spaces $A_M = \{x \mid \sum_{i \varepsilon M}(x_i - p_i^0) \geqq \epsilon\}$, $M \varepsilon \mathfrak{M}$. Hence we obtain

$$I(A, p^0) = \min_{M \varepsilon \mathfrak{M}} I(A_M, p^0) = \min_{M \varepsilon \mathfrak{M}} J(h_M),$$

where $h_M = \sum_{i \varepsilon M} p_i^0$ and

$$J(h) = (h + \epsilon) \log [(h + \epsilon)/h] + (1 - h - \epsilon) \log (1 - h - \epsilon)/(1 - h).$$

Again the minimizing points y are such that the ratio y_i/p_i^0 takes only two values. The function $J(h)$ has a unique minimum at a point h_0 which is close to $\frac{1}{2}$ if ϵ is small. This implies that if $D(x, p^0) = \frac{1}{2}\sum |x_i - p_i^0|$ and $p_i^0 = 1/k$, $i = 1, \cdots, k$, then, for ϵ small, there are close to $\binom{k}{k/2}$ minimizing points. For the chi-square test this number is only k.

11. Bayes tests and likelihood ratio tests. In this section it will be shown that certain Bayes tests differ little from the corresponding likelihood ratio test if N is large, not only when the size α_N of the test is bounded away from 0 (in which case a chi-square test has a similar property) but also when α_N tends to zero.

Let G be a distribution function on the simplex Ω and let

$$P_N(z^{(N)} \mid G) = \int_\Omega P_N(z^{(N)} \mid p) \, dG(p). \tag{11.1}$$

The Bayes test for testing the hypothesis H: $p = p^0$ against the alternative that p is distributed according to G rejects H if the ratio $P_N(z^{(N)} \mid G)/P_N(z^{(N)} \mid p^0)$ exceeds a constant. This ratio is $\leqq \exp \{NI(z^{(N)}, p^0)\}$.

Let U denote the uniform distribution on Ω, so that the vector $(p_1, \cdots, p_{k-1})$ has a constant probability density. We have

$$P_N(z^{(N)} \mid U) = \binom{N+k-1}{k-1}^{-1} \tag{11.2}$$

for all $z^{(N)}$. Hence

$$\begin{aligned} NI(z^{(N)}, p^0) - \log [P_N(z^{(N)} \mid U)/P_N(z^{(N)} \mid p^0)] \\ = \log \binom{N+k-1}{k-1} + \log P_N(z^{(N)} \mid z^{(N)}). \end{aligned} \tag{11.3}$$

Here the left side is the difference between the test statistics for the likelihood ratio test and the Bayes test. An application of Stirling's formula to the last term in (11.3) (see (2.11)) shows that if the components of $z^{(N)}$ are bounded away from 0, the right side of (11.3) is of the form $c_N + O(1)$, where c_N does not depend on $z^{(N)}$. This implies that the critical regions of the two tests (when they are of approximately the same size) and their error probabilities at the points in Ω_0 differ little from each other.

The uniform distribution U has been chosen for simplicity. We obtain a similar result if U is replaced by a distribution G such that, for example, the probability density of $(p_1, \cdots, p_{k-1})$ is positive and bounded.

Now consider a composite hypothesis, H: $p \in \Lambda$. Let G_0 be a distribution on Ω such that the set Λ has probability one. We may expect that for suitable choices of G_0 and G the Bayes test based on the ratio $P_N(z^{(N)} \mid G)/P_N(z^{(N)} \mid G_0)$ will differ little from the likelihood ratio test based on $I(z^{(N)}, \Lambda)$. This is here illustrated by two examples.

EXAMPLE 11.1. *Binomial hypothesis.* Let $k = m + 1$ and denote the points of Ω by $x = (x_0, x_1, \cdots, x_m)$. Let

$$\Lambda = \{p(\theta) \mid 0 \leqq \theta \leqq 1\}, \tag{11.4}$$
$$p_i(\theta) = \binom{m}{i}\theta^i(1 - \theta)^{m-i}, \qquad i = 0, 1, \cdots, m.$$

Then $I(x, \Lambda) = I(x, \hat{p}(x))$, where $\hat{p}(x) = p(\hat{\theta}(x))$, $\hat{\theta}(x) = \sum ix_i/m$.

Let U_0 denote the distribution on Λ induced by the uniform distribution of θ on $(0, 1)$. Then

$$P_N(z^{(N)} \mid U_0) = (N!/\prod n_i!)[(mN + 1)\binom{mN}{s}]^{-1} \prod \binom{m}{i}^{n_i},$$

where $s = \sum in_i = mN\hat{\theta}(z^{(N)})$. Let, as before, U be the uniform distribution on Ω. After simplification we obtain

$$\begin{aligned} NI(z^{(N)}, \Lambda) &- \log [P_N(z^{(N)} \mid U)/P_N(z^{(N)} \mid U_0)] \\ &= \log \binom{N+m}{m} - \log (mN + 1) + \log P_N(z^{(N)} \mid z^{(N)}) - \log P_N^*, \end{aligned} \tag{11.5}$$

where

$$P_N^* = \binom{mN}{s}(s/mN)^s[1 - (s/mN)]^{mN-s}. \tag{11.6}$$

Relation (11.5) is analogous to (11.3) and implies a similar conclusion.

EXAMPLE 11.2. *Hypothesis of independence in a contingency table.* Let Λ be defined as in Example 9.1. Let U_0 be the distribution on Λ such that the random vectors $(p_1^{(1)}, \cdots, p_r^{(1)})$ and $(p_1^{(2)}, \cdots, p_s^{(2)})$ are independent and each is uniformly distributed on the respective probability simplex. Let $z_{ij}^{(N)} = n_{ij}/N$, $n_i^{(1)} = \sum_j n_{ij}$, $n_j^{(2)} = \sum_i n_{ij}$. We obtain

$$\begin{aligned} NI(z^{(N)}, \Lambda) &- \log [P_N(z^{(N)} \mid U)/P_N(z^{(N)} \mid U_0)] \\ &= \log [\binom{N+rs-1}{rs-1}/\binom{N+r-1}{r-1}\binom{N+s-1}{s-1}] + \log [P_N(z^{(N)} \mid z^{(N)})/P_N^{(1)}P_N^{(2)}], \end{aligned} \tag{11.7}$$

where

$$P_N^{(1)} = (N!/N^N) \prod_{i=1}^r (n_i^{(1)})^{n_i^{(1)}}/n_i^{(1)}!$$

and $P_N^{(2)}$ is defined in an analogous way in terms of the $n_j^{(2)}$.

The result is quite similar to that of Example 11.1.

The hypothesis sets Λ of Examples 11.1 and 11.2 are special cases of a class of subsets of Ω for which relations analogous to (11.5) and (11.7) hold true.

Appendix. Regular sequences of sets. In this appendix sufficient conditions are derived for a sequence of subsets of Ω to be regular relative to a point in Ω. We recall that, by Definition 6.1, the sequence $\{A_N\}$ is regular relative to p if

$$I(A_N^{(N)}, p) = I(A_N, p) + O(N^{-1} \log N). \tag{A.1}$$

(Sanov [7] considered the weaker regularity condition where the remainder term in (A.1) is replaced by $o(1)$.)

Since $A_N^{(N)} \subset A_N$, we have $I(A_N^{(N)}, p) \geqq I(A_N, p)$. Hence for those N for which $I(A_N, p) = \infty$ Condition A.1 is satisfied. Thus $\{A_N\}$ is regular relative to p if Condition (A.1), with $=$ replaced by $\leqq$, is fulfilled for those sets A_N for which $I(A_N, p) < \infty$.

LEMMA A.1. *The sequence $\{A_N\}$ is regular relative to p if there exist constants N_0 and c such that for each $N > N_0$ with $I(A_N, p) < \infty$ there is a point $y \in \Omega$ for which*

$$I(y, p) \leqq I(A_N, p) \tag{A.2}$$

and a point $z \in A_N^{(N)}$ for which

$$|z_i - y_i| < cN^{-1} \quad \text{if} \quad p_i > 0, \qquad z_i = 0 \quad \text{if} \quad p_i = 0. \tag{A.3}$$

PROOF. We may restrict ourselves to values N for which $I(A_N, p) < \infty$. It is sufficient to show that

$$I(z, p) - I(y, p) \leqq O(N^{-1} \log N). \tag{A.4}$$

The assumptions imply that y and z are in $\Omega(p)$ for $N > N_0$. Hence for $N > N_0$

$$I(z, p) - I(y, p) = \textstyle\sum_{p_i \neq 0} d_i, \qquad d_i = z_i \log(z_i/p_i) - y_i \log(y_i/p_i).$$

If $z_i = 0$ then $y_i < cN^{-1}$ and $d_i = -y_i \log(y_i/p_i) = O(N^{-1} \log N)$.

If $z_i \neq 0$ then $z_i \geqq N^{-1}$ and

$$\begin{aligned} d_i &= (z_i - y_i) \log(z_i/p_i) + y_i \log(z_i/y_i) \\ &\leqq (z_i - y_i) \log(z_i/p_i) + y_i((z_i/y_i) - 1) \\ &\leqq |z_i - y_i|\, |\log N^{-1}| + O(|z_i - y_i|) = O(N^{-1} \log N). \end{aligned}$$

Hence $d_i \leqq O(N^{-1} \log N)$ for all i with $p_i \neq 0$, and (A.4) follows.

For any real numbers $a_1, \cdots, a_k, c$ the subset of Ω defined by $\sum a_i x_i > c$ or $\sum a_i x_i \geqq c$ will be called a *half-space*. (It is convenient here not to exclude the case where all a_i are equal. Thus the proof of the next lemma for the case $p \not\in \Omega_0$ is strictly analogous to the proof for $p \in \Omega_0$.)

LEMMA A.2. *If p is any point in Ω and A is any half-space such that $I(A, p) < \infty$ then there is a point $y \in \Omega$ for which $I(y, p) = I(A, p)$, and for each $N \geqq k(k - 1)$ there is a point $z \in A^{(N)}$ such that $|z_i - y_i| \leqq (k - 1)N^{-1}$ if $p_i > 0$ and $z_i = 0$ if $p_i = 0$.*

PROOF. First assume that $p \in \Omega_0$ and

$$A = \{x \mid \textstyle\sum a_i x_i > c\}.$$

Since $I(A, p) < \infty$, A is not empty, so that $\max a_i > c$. By Lemma 4.2 there is a point y such that $I(y, p) = I(A, p)$ and $\sum a_i y_i \geqq c$. It is easy to show that $y \varepsilon \Omega_0$.

We have $y_i \geqq k^{-1}$ for some i. For definiteness assume that $y_k \geqq k^{-1}$. Define $z = (z_1, \cdots, z_k)$ as follows. For $i = 1, \cdots, k-1$ let Nz_i be an integer such that

$$\text{(A.5)} \qquad z_i \geqq 0, \quad z_i \neq y_i, \quad |z_i - y_i| \leqq N^{-1}, \quad (a_i - a_k)(z_i - y_i) \geqq 0.$$

These conditions can be satisfied since $y \varepsilon \Omega_0$. Let $z_k = 1 - z_1 - \cdots - z_{k-1}$. Then

$$z_k \geqq 1 - \sum_{i=1}^{k-1} (y_i + N^{-1}) = y_k - (k-1)N^{-1} \geqq k^{-1} - (k-1)N^{-1}.$$

Hence if $N \geqq k(k-1)$ then $z_k \geqq 0$ and $z \varepsilon \Omega^{(N)}$. Moreover, $|z_i - y_i| \leqq (k-1)N^{-1}$ for all i. Now

$$\text{(A.6)} \qquad \sum_{i=1}^{k} a_i z_i - c \geqq \sum_{i=1}^{k} a_i z_i - \sum_{i=1}^{k} a_i y_i = \sum_{i=1}^{k-1} (a_i - a_k)(z_i - y_i).$$

If the a_i are not all equal, the last sum is strictly positive by (A.5). Otherwise the inequality in (A.6) is strict. Hence $z \varepsilon A^{(N)}$ for $N \geqq k(k-1)$. The lemma is proved for the present case.

If $p \varepsilon \Omega_0$ and $A = \{x \mid \sum a_i x_i \geqq c\}$, the conclusion of the lemma follows from the first part of the proof provided that the set $\{x \mid \sum a_i x_i > c\}$ is not empty. If it is empty then, since A must be non-empty, we have $\max a_i = c$ and A is the set of all points x such that $x_i = 0$ if $a_i < c$. We have $I(A, p) = I(y, p)$ where $y_i = p_i / \sum_{a_j = c} p_j$ if $a_i = c$, $y_i = 0$ otherwise. It is trivial to show that the conclusion of the lemma is true in this case.

If $p \not\varepsilon \Omega_0$, the assumption $I(A, p) < \infty$ implies $I(A, p) = I(A \cap \Omega(p), p)$ and the proof is similar to that for the case $p \varepsilon \Omega_0$.

Lemmas A.1 and A.2 imply that any sequence of half-spaces is regular relative to any point in Ω. More generally we have

THEOREM A.1. *Any sequence of subsets of Ω whose complements are convex is regular relative to any point in Ω.*

PROOF. Let A be a set whose complement is convex and p a point such that $I(A, p) < \infty$. We restrict ourselves to the case $p \varepsilon \Omega_0$ since the case $p \not\varepsilon \Omega_0$ is treated in an analogous way, as in the proof of Lemma A.2.

Let y be a point in $\bar{A}$ such that $I(y, p) = I(A, p)$. If $p \not\varepsilon \bar{A}$ then y is in the boundary of the convex set A'; hence there exists an open half-space H defined by a supporting hyperplane of A' through y such that $A' \subset H'$, so $H \subset A$. If $p \varepsilon \bar{A}$ then $y = p$, and again there is a half-space $H \subset A$ such that y is in the boundary of H.

If $y \varepsilon \Omega_0$ then H is not empty, so that $y \varepsilon \bar{H}$ and therefore $I(A, p) = I(H, p)$. By Lemma A.2 and its proof, for each $N \geqq k(k-1)$ there is a point z in $H^{(N)}$, hence in $A^{(N)}$, with the property stated in that lemma.

Now suppose that $y \not\varepsilon \Omega_0$. We first show that the set $A \cap \Omega(y)$ is not empty.

Every neighborhood of y contains a point $x \varepsilon A$. Let $h = \sum_{y_i = 0} x_i$. As

$x \to y$, $h \to 0$. We may assume $h > 0$ since otherwise $x \varepsilon A \cap \Omega(y)$. The point $\bar{x}$ defined by $\bar{x}_i = x_i/(1 - h)$ if $y_i > 0$, $\bar{x}_i = 0$ if $y_i = 0$, is in $\Omega(y)$. It is sufficient to show that $\bar{x} \varepsilon A$ for $|x - y|$ small enough. This, in turn, will follow if we show that for $|x - y|$ small enough there is a number $t < 0$ such that the point $z = (1 - t)x + t\bar{x}$ is in Ω and satisfies $I(z, p) < I(y, p)$. Indeed, the latter implies $z \varepsilon A'$, and if $\bar{x} \varepsilon A'$, it would follow from the convexity of A' that $x \varepsilon A'$, contrary to our assumption.

Fix $g \varepsilon (0, 1)$ and define t by $g = h(1 - t)$. We have $t < 0$ for $h < g$, and $z_i = x_i(1 - g)/(1 - h)$ if $y_i > 0$, $z_i = x_i g/h$ if $y_i = 0$. Hence as $x \to y$,

$$I(z, p) \sim (1 - g) \textstyle\sum_{y_i > 0} y_i \log [y_i(1 - g)/p_i] + (g/h) \sum_{y_i = 0} x_i \log (x_i g)/(hp_i)$$
$$\leqq (1 - g)I(y, p) + \log (g/p_{\min}).$$

With $g = p_{\min}$ this implies that $I(z, p) < I(y, p)$ for $|x - y|$ small enough, as was to be proved.

Since A $\cap$ $\Omega(y)$ is not empty, $I(A, p) = I(A \cap \Omega(y), p)$. The set $A' \cap \Omega(y)$ is a convex subset of $\Omega(y)$. For $x \varepsilon \Omega(y)$ we have $I(x, p) = I(x, \bar{p}) - \log \sum_{y_j > 0} p_j$, where $\bar{p}_i = p_i / \sum_{y_j > 0} p_j$ if $y_i > 0$, $\bar{p}_i = 0$ otherwise. The argument used in the case $y \varepsilon \Omega_0$, with Ω replaced by the subspace $\Omega(y)$, leads to the conclusion reached for that case.

Thus the conclusion of Lemma A.2 is true for any subset A of Ω whose complement is convex. With Lemma A.1 this implies the theorem.

Define the subset Ω_ϵ of Ω by

$$\Omega_\epsilon = \{x \mid x_i > \epsilon, i = 1, \cdots, k\}. \tag{A.7}$$

THEOREM A.2. *Let*

$$A_N = \{x \mid f(x) > c_N\}, \tag{A.8}$$

where the c_N are real numbers and $f(x)$ is a function defined on Ω whose derivatives $f_i'(x) = \partial f(x)/\partial x_i$ and $f_{ij}''(x) = \partial^2 f(x)/\partial x_i \partial x_j$ exist and are continuous in Ω_0. Let $p \varepsilon \Omega_0$. Suppose that there exist positive numbers N_0 and ϵ such that for each $N > N_0$ there is a point $y^{(N)}$ with the properties

$$y^{(N)} \varepsilon \Omega_\epsilon, \qquad f(y^{(N)}) \geqq c_N, \qquad I(y^{(N)}, p) = I(A_N, p) \tag{A.9}$$

and that

$$\lim_{N\to\infty} N \max_{ij} |f_i'(y^{(N)}) - f_j'(y^{(N)})| = +\infty. \tag{A.10}$$

Then the sequence $\{A_N\}$ is regular relative to p.

PROOF. The assumptions imply that for $N > N_0$

$$\begin{aligned} f(x) - c_N &\geqq f(x) - f(y^{(N)}) \\ &= \textstyle\sum_{i=1}^{k} f_i'(y^{(N)})(x_i - y_i^{(N)}) + O(|x - y^{(N)}|^2) \\ &= \textstyle\sum_{i=1}^{k-1} a_i^{(N)}(x_i - y_i^{(N)}) + O(|x - y^{(N)}|^2) \end{aligned}$$

uniformly for $x \varepsilon \Omega_{\epsilon/2}$, where $a_i^{(N)} = f_i'(y^{(N)}) - f_k'(y^{(N)})$.

For $i = 1, \cdots, k - 1$ let $m_i^{(N)}$ denote the largest integer $\leqq N y_i^{(N)}$ and let $m_k^{(N)} = N - m_1^{(N)} - \cdots - m_{k-1}^{(N)}$. Define the point $z^{(N)}$ by

$$N z_i^{(N)} = m_i^{(N)} + 2 \quad \text{if} \quad a_i^{(N)} \geqq 0, \qquad N z_i^{(N)} = m_i^{(N)} - 1 \quad \text{if} \quad a_i^{(N)} < 0,$$

for $i \leqq k - 1$ and $z_k^{(N)} = 1 - z_1^{(N)} - \cdots - z_{k-1}^{(N)}$. Then

$$\text{(A.11)} \qquad |z_i^{(N)} - y_i^{(N)}| < 2k/N, \qquad i = 1, \cdots, k.$$

Since $y^{(N)} \varepsilon \Omega_\epsilon$, $z^{(N)}$ is in $\Omega_{\epsilon/2}$ for $N > N_1 \geqq N_0$. Moreover,

$$a_i^{(N)}(z_i^{(N)} - y_i^{(N)}) \geqq N^{-1} |a_i^{(N)}|, \qquad i = 1, \cdots, k - 1.$$

Hence for $N > N_1$

$$\begin{aligned} f(z^{(N)}) - c_N &\geqq N^{-1} \textstyle\sum_{i=1}^{k-1} |a_i^{(N)}| + O(N^{-2}) \\ &\geqq \tfrac{1}{2} N^{-1} \max_{ij} |f_i'(y^{(N)}) - f_j'(y^{(N)})| + O(N^{-2}). \end{aligned}$$

Condition (A.10) implies that for N large enough we have $f(z^{(N)}) > c_N$, that is, $z^{(N)} \varepsilon A_N$.

Thus the conditions of Lemma A.1 are satisfied. The proof is complete.

REFERENCES

[1] Bahadur, R. R. and Rao, R. Ranga (1960). On deviations of the sample mean. *Ann. Math. Statist.* **31** 1015–1027.

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

[3] Hoeffding, Wassily (1963). Large deviations in multinomial distributions. (Abstract.) *Ann. Math. Statist.* **34** 1620.

[4] Hoeffding, Wassily (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

[5] Matusita, Kameo (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation. *Ann. Math. Statist.* **26** 631–640.

[6] Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20-A** 175–240, 264–299.

[7] Sanov, I. N. (1957). On the probability of large deviations of random variables (Russian) *Mat. Sbornik N. S.* **42** (**84**), 11–44. English translation: *Select. Transl. Math. Statist. and Probability* **1** (1961) 213–244.

[8] Wald, Abraham (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

DISCUSSION OF HOEFFDING'S PAPER

Jerzy Neyman[1]: Professor Hoeffding is to be heartily congratulated on his very interesting paper. His results as explicitly formulated are important enough. It *is* important to know that out of the several tests of the same hypothesis, the tests whose certain asymptotic properties are identical and which, therefore, were considered equivalent, one particular test has an asymptotic property, not pre-

[1] Prepared with the partial support of the National Science Foundation Grant GP-2593.