The Annals of Mathematical Statistics
1971, Vol. 42, No. 6, 1992–2007

ASYMPTOTICALLY OPTIMAL TESTS FOR FINITE MARKOV CHAINS[1]

By Luis B. Boza

Bell Laboratories, Inc.

A discrete time, finite Markov chain with fixed initial state and stationary transition behavior is considered. Using Whittle's formula a large deviation result (similar to Hoeffding's result for one multinomial distribution) is obtained for the transition count matrix of a path of the chain of arbitrary length. This result is then used in the asymptotic comparison of a given sequence of tests about the transition probability matrix with a suitably constructed sequence of likelihood ratio tests. It is assumed that the sizes of these tests decrease to zero at a certain rate as the length of the observed path increases. The comparison is carried out at fixed alternatives in terms of the behavior of the ratio of type-II-error probabilities.

1. Introduction. One approach for the asymptotic theory of inference about the parameter of a Markov chain is described in Billingsley (1961a), (1961b). The transition behavior is assumed to be stationary. One path of the chain is observed and the asymptotic theory is developed as the length of the observed path increases. The parameter is supposed to be in some open subset of n-dimensional Euclidean space, and hence, under some conditions, the transition probability matrix itself can be regarded as the parameter in the finite case. For the testing problem about this matrix a suitable central limit theorem is obtained and then used to compute the limiting null distributions of several test statistics. These give asymptotic level results. A few considerations are made for power at "close alternatives."

Bahadur and Ragavachari (1970) consider the finite state space case and show that likelihood ratio tests are optimal in the exact slope sense, when testing hypotheses about the unknown transition probability matrix.

Johnson and Roussas (1969) consider a discrete time, real–valued, ergodic Markov process with a one-dimensional parameter space. The observation is again a path of the chain. In the testing problem the null hypothesis is assumed to be simple, and the alternatives one-sided. Wald's optimality criterion (Wald (1941)), essentially of "close-alternatives" type, is used. LeCam's contiguity techniques (LeCam (1960)) provide a way for relaxing Wald's regularity conditions and for obtaining asymptotically locally most powerful tests, and under a further assumption, asymptotically most powerful tests. Johnson and Roussas (1970) consider also the two-sided alternatives case and obtain similar results, still for a one-dimensional

Received September 15, 1970.

[1] Research partially supported by the National Science Foundation, Grants GP–8690 and GP–15283, at the University of California, Berkeley.

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

parameter. The extension of these results to the case of a multidimensional parameter will be considered by them in a forthcoming paper.

Another approach is that of Anderson and Goodman (1957), where the underlying process is a finite Markov chain. Instead of observing just one path of the chain and developing the asymptotic theory as the length of the path increases, they consider independent replicates of a stretch of the chain of fixed, finite length and then develop the asymptotic theory as the number of replicates increases. Relying deeply on multinomial arguments, they obtain limiting null distributions for certain test statistics, and hence, asymptotic level results. Some power considerations are made for close alternatives, but moreover, a stochastic comparison of tests is suggested. Although their setup will not be considered in detail in this paper, some remarks will be made in Section 8 about the applicability of the methods developed here to their case.

In 1965, W. Hoeffding's work (Hoeffding (1965)) on comparison of tests for multinomial distributions was published. In his approach, the asymptotic comparison of tests is made, at fixed alternatives and decreasing levels of the tests, in terms of the behavior of the ratio of type-II error probabilities. By the very nature of this method, large deviation theory rather than central limit theorems becomes the basic tool.

In this paper tests for the transition probability matrix of a discrete time, finite Markov chain are asymptotically compared. The observation consists of one path of the chain. As in Hoeffding's work, the asymptotic comparison is done in terms of the behavior of the ratio of the type-II-error probabilities at fixed alternatives, when the sizes of the tests go to zero at a certain rate as the length of the observed path increases.

The notation is introduced in Section 2. A basic large deviation result is obtained in Theorem 3.1 and then used for getting a preliminary result on comparison of tests (Theorem 4.4). The Kullback information function which appears in Theorem 3.1 is studied in Section 5. In Section 6 the concept of "equally informative sequences of sets" is introduced and then used in Section 7 to simplify the minimization problems needed for the applications of Theorem 4.4, thus leading to a new theorem on comparison of tests (Theorem 7.1). In Section 8 the applicability of this method to Anderson and Goodman's approach for the asymptotic theory of inference in finite chains is briefly discussed.

2. Notation. Let m be a fixed finite integer greater than one and $s \in \{1, 2, \cdots, m\}$. Let $\Omega_0 = \{s\}$, $\Omega_t = \{1, 2, \cdots, m\}$ for $t = 1, 2, \cdots$, and $\mathscr{B}_t$ the discrete σ-algebra on Ω_t for $t = 0, 1, 2, \cdots$. Let $\Omega = \prod_{t=0}^{\infty} \Omega_t$ and $\mathscr{B}$ the product σ-algebra on Ω. Let M_P be a probability measure on $(\Omega, \mathscr{B})$ under which the coordinate process X_0, X_1, X_2, $\cdots$ is a Markov chain with time homogeneous transition probability matrix $P = (p_{ij})$.

For $T = 0, 1, 2, \cdots$, let $\mathscr{A}_T$ be the sub-σ-algebra of $\mathscr{B}$ generated by $X_0, X_1, \cdots, X_T$, and $P_T = M_P \mid \mathscr{A}_T$. Occasionally in the course of this paper, the symbols $P_T{}^0$ and $\hat{P}_T$ will be used for $M_{P^0} \mid \mathscr{A}_T$, $M_{\hat{P}} \mid \mathscr{A}_T$, respectively.

For any path $\omega \in \Omega$, the corresponding *transition count matrix at time* T is denoted by $C(T)(\omega)$ and defined by

$$C(T)_{ij}(\omega) = \sum_{t=0}^{T-1} I_{[X_t(\omega)=i, X_{t+1}(\omega)=j]}$$

for any $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots m$. Let

$$C(T)_{i.}(\omega) = \sum_j C(T)_{ij}(\omega),$$

and

$$C(T)._i(\omega) = \sum_j C(T)_{ji}(\omega).$$

It is clear that for any $\omega \in \Omega$, the corresponding transition count matrix at time T fulfills the following conditions:

(1) $C(T)_{ij}(\omega)$ is a nonnegative integer for any i, j;

(2) $\sum_{ij} C(T)_{ij}(\omega) = T$;

(3) for any i, $C(T)_{i.}(\omega) - C(T)._i(\omega) = \delta_{is} - \delta_{if(\omega)}$ where $f(\omega) = X_T(\omega)$ and δ is the Kronecker delta function.

For $T = 1, 2, \cdots$ let $\mathscr{C}_T^* = \{C \mid C \text{ is an } m \times m \text{ matrix fulfilling (1), (2), (3)}\}$ and $\mathscr{F}_T^* = \{F \mid F = (F_{ij}) = (C_{ij}/T) \text{ for some } C \in \mathscr{C}_T^*\}$. Note that, for each T, there is a one to one correspondence between $\mathscr{C}_T^*$ and $\mathscr{F}_T^*$.

For a fixed $C \in \mathscr{C}_T^*$ let $A_T(C)$ be the set of all $\omega \in \Omega$ whose transition count matrix at time T is C. Clearly $A_T(C) \in \mathscr{A}_T$ for any such C. Abusing the notation for simplicity let, for any $C \in \mathscr{C}_T^*$ and the corresponding $F = (C_{ij}/T) \in \mathscr{F}_T^*$, $P_T(C) = P_T(F) = P_T(A_T(C))$.

Whittle (1955) and later on Billingsley (1961a) showed that for any $C \in \mathscr{C}_T^*$

$$P_T(C) = \Delta_{f(C,s),s}\left(\prod_i C_{i.}!/\prod_{ij} C_{ij}!\right)\prod_{ij} p_{ij}^{C_{ij}},$$

where

(a) $f(C, s)$ is the unique (time T) state defined by C and the initial state s through the restrictions (3);

(b) $\Delta_{f(C,s),s}$ is the cofactor of the $(f(C, s), s)$ entry in the matrix $\Gamma = (\Gamma_{ij})$ with $\Gamma_{ij} = \delta_{ij}$ if $C_{i.} = 0$ and $\Gamma_{ij} = \delta_{ij} - (C_{ij}/C_{i.})$ if $C_{i.} > 0$;

(c) by convention $p_{ij}^{C_{ij}} = 1$ if $p_{ij} = C_{ij} = 0$.

The above mentioned formula is usually referred to as "Whittle's formula". Note that $\sum_{C \in \mathscr{C}_T^*} P_T(C) = 1$ and also that the restriction $\Delta_{f(C,s),s} = 0$ removes from $\mathscr{C}_T^*$ a set of zero P_T-probability.

At time T, $C(T)$ (or equivalently the corresponding $F = (C(T)_{ij}/T)$) is a sufficient statistic for P, and the maximum likelihood estimator $\hat{P} = (\hat{p}_{ij})$ of P is given by

$$\begin{aligned} \hat{p}_{ij} &= C(T)_{ij}/C(T)_{i.} \qquad \text{if} \quad C(T)_{i.} > 0, \\ &= \alpha_{ij} \qquad\qquad\quad \text{if} \quad C(T)_{i.} = 0, \end{aligned}$$

for any i, j, where the α_{ij}'s are arbitrary within the restrictions $\alpha_{ij} \geqq 0$ for all i, j, and $\sum_j \alpha_{ij} = 1$ for all i.

DEFINITION.

(a) $\mathscr{C}_T = \{C \in C_T^* \mid \Delta_{f(C,s),s} > 0\}$.

(b) $\mathscr{F}_T = \{F \mid F = (C_{ij}/T) \text{ for some } C \in C_T\}$.
(Again note that for each T, there is a one to one correspondence between $\mathscr{C}_T$ and $\mathscr{F}_T$.)

(c) $\mathscr{F} = \{F \mid F \text{ is an } m \times m \text{ matrix } F_{ij} \geqq \text{all } i, j; \sum_{i,j} F_{ij} = 1\}$.

(d) For any set $A \subseteq \mathscr{F}$ let $A_T = A\mathscr{F}_T$, where $A\mathscr{F}_T$ denotes the intersection of the sets A and $\mathscr{F}_T$.

(e) $\mathscr{P} = \{P \mid P \text{ is } m \times m \text{ matrix}; p_{ij} \geqq 0 \text{ all } i, j; \sum_j p_{ij} = 1 \text{ all } i\}$.

(f) Φ is a function from $\mathscr{F} \times \mathscr{P}$ to the extended real line defined by: for any $F \in \mathscr{F}$, any $P \in \mathscr{P}$, $\Phi(F, P) = \sum_{i \ni F_{i\cdot} > 0} \sum_j F_{ij} \log (F_{ij}/F_{i\cdot}p_{ij})$, where by convention $F_{ij} \log (F_{ij}/F_{i\cdot}p_{ij})$ is taken to be zero whenever $F_{ij} = 0$.

3. Basic large deviation theorem. In this section a large deviation result is proved for the transition count matrix of the chain. It is similar to Theorem 2.1 of Hoeffding (1965) which, in turn, is a stronger version of a result due to Sanov (1957). This large deviation theorem will be used repeatedly in the rest of this paper.

THEOREM 3.1.

(a) *For any $F \in \mathscr{F}_T$ and any $P \in \mathscr{P}$*

$$P_T(F) = \exp (O(\log T) - T\Phi(F, P))$$

where $O(\log T)$ is uniform in $F \in \mathscr{F}_T$ and $P \in \mathscr{P}$.

(b) *For any set $A \subseteq \mathscr{F}$ and any $P \in \mathscr{P}$*

$$P_T(A) = P_T(A_T) = \exp (O(\log T) - T\Phi(A_T, P))$$

where $\Phi(A_T, P) = \inf_{F \in A_T} \Phi(F, P)$ (defined to be $+\infty$ if A_T is empty) and where $O(\log T)$ is uniform for $A \subseteq \mathscr{F}$ and $P \in \mathscr{P}$.

PROOF.

(i) Whittle's formula and the convention stated in the definition of Φ can be used to obtain, in a straightforward way, the following result: for any $F \in \mathscr{F}_T{}^*$,

$$P_T(F) = \hat{P}_T(F) \exp (-T\Phi(F, P)),$$

where $\hat{P}_T(F)$ is computed using Whittle's formula, but with $\hat{P}$ instead of P.

(ii) Clearly, for any $F \in \mathscr{F}_T{}^*$,

$$P_T(F) \leqq \exp (-T\Phi(F, P)).$$

(iii) To obtain a lower bound for $P_T(F)$ a two-step procedure is used. The first step is valid for any $F \in \mathscr{F}_T{}^*$, while the second is valid only on $\mathscr{F}_T$.

From Stirling's estimate for the factorials and the fact that $0 \leqq C_{ij} \leqq T$ for any i, j, it follows that for any $F \in \mathscr{F}_T{}^*$,

$$\hat{P}_T(F) \geqq \Delta_{f(C,s),s}(2\pi T)^{-m(m-1)/2},$$

where $C = (TF_{ij}) \in \mathscr{C}_T{}^*$.

A lower bound is now obtained for $\Delta_{f(C,s),s}$, where $C = (TF_{ij})$ and $F \in \mathscr{F}_T$. First assume that the set $I(C) = \{i \mid C_{i\cdot} = 0\}$ is empty. Then $\Delta_{f(C,s),s} = \Delta^*_{f(C,s),s} / \prod_{i \neq f} C_{i\cdot}$, where $\Delta^*_{f(C,s),s}$ is the cofactor of the $(f(C, s), s)$ entry in the matrix $\Gamma^* = (C_{i\cdot}\delta_{ij} - C_{ij})$. Since $F \in \mathscr{F}_T$, $\Delta^*_{f(C,s),s}$ is strictly positive. Moreover, since Γ^* has integer entries, $\Delta^*_{f(C,s),s} \geqq 1$. Finally, since $0 < C_{i\cdot} < T$ for any i, $\Delta_{f(C,s),s} \geqq T^{-(m-1)}$.

Consider now the case where $I(C)$ is nonempty. $F \in \mathscr{F}_T$ implies that $s \notin I(C)$, and hence that $1 \leqq \#I(C) \leqq (m-1)$, where $\#I(C)$ denotes the cardinality of the set $I(C)$. If $f(C, s) \notin I(C)$, the same argument of the preceding paragraph can be applied to a square submatrix of C, of dimension $(m - \#I(C))$, which contains all the positive entries of C. Hence, in this case,

$$\Delta_{f(C,s),s} \geqq T^{-(m-1-\#I(C))} \geqq T^{-(m-1)}.$$

On the other hand, if $f(C, s) \in I(C)$, it follows from the restrictions (3) and from $s \notin I(C)$, that $C_{\cdot f(C,s)} = 1$, and that there is a unique state $k \notin I(C)$ such that $C_{kf(C,s)} = C_{\cdot f(C,s)} = 1$. It is clear that, after deleting the $f(C, s)$th row and the sth column of Γ, the $f(C, s)$th column of Γ contains only one nonzero element, $\Gamma_{k,f(C,s)}$, not smaller–in absolute value–than T^{-1}. It now follows by the argument of the previous paragraph that

$$\Delta_{f(C,s),s} \geqq T^{-(m-\#I(C))} \geqq T^{-(m-1)}.$$

The result in (i), together with those already obtained in (iii), imply that

$$P_T(F) \geqq (2\pi)^{-m(m-1)/2} T^{-(m-1)(m+2)/2} \exp(-T\Phi(F, P)).$$

This result and the one in (ii) imply part (a) of Theorem 3.1.

(iv) Let A be a subset of $\mathscr{F}$ and $A_T = A\mathscr{F}_T$. Then,

$$P_T(A) = P_T(A_T) = \sum_{F \in A_T} P_T(F).$$

But, since A_T is a finite set, the infimum $\Phi(A_T, P)$ is attained for some $F^0 \in A_T$. Hence, by (iii),

$$P_T(A_T) \geqq P_T(F^0) \geqq (2\pi)^{-m(m-1)/2} T^{-(m-1)(m+2)/2} \exp(-T\Phi(A_T, P)).$$

For the upper bound note that

$$\sum_{F \in A_T} P_T(F) \leqq \exp(-T\Phi(A_T, P)) \sum_{F \in A_T} 1.$$

But the number of points in A_T is obviously smaller than the total number of ways

of placing exactly T ones into $m \times m$ places without any further restriction. It is a well–known combinatorial result Feller ((1962) page 36) that this number of ways is

$$\begin{bmatrix} T+m^2-1 \\ m^2-1 \end{bmatrix}.$$

Part (b) of this theorem follows.

4. Preliminary results on comparison of tests. Let Λ be a nonempty subset of $\mathscr{P}$. Consider the problem of testing $H_0: P \in \Lambda$ against $H_1: P \in \mathscr{P} - \Lambda$.

LEMMA 4.1. *For each T, the class of likelihood ratio* (L.R.) *tests of H_0 against H_1 is given by the class of rejection regions*

$$\{\{F \in \mathscr{F} \mid \Phi(F, \Lambda) \geqq b(T)\} \mid 0 \leqq b(T) \leqq \infty\},$$

where

$$\Phi(F, \Lambda) = \inf_{P \in \Lambda} \Phi(F, P).$$

PROOF.

$$\sup_{P \in \Lambda} P_T(F) = \sup_{P \in \Lambda} (\hat{P}_T(F) \exp(-T\Phi(F, P)))$$
$$= \hat{P}_T(F) \exp(-T\Phi(F, \Lambda)).$$

Similarly,

$$\sup_{P \in \mathscr{P}} P_T(F) = \hat{P}_T(F) \exp(-T\Phi(F, \mathscr{P})).$$

But, as will be seen in Lemma 5.1, $\Phi(F, \mathscr{P}) = 0$ for any $F \in \mathscr{F}$. The result follows.

Note that the actual rejection regions at time T are of the form $\{F \in \mathscr{F}_T \mid \Phi(F, \Lambda) \geqq b(T)\}$. The differences between these sets and the corresponding sets $\{F \in \mathscr{F} \mid \Phi(F, \Lambda) \geqq b(T)\}$ have P_T probability zero.

LEMMA 4.2. *The size of a* L.R. *test whose rejection region at time T is $R^{(T)} = \{F \in \mathscr{F} \mid \Phi(F, \Lambda) \geqq b(T)\}$ is* $\exp(O(\log T) - T\Phi(R_T^{(T)}, \Lambda))$, *where $R_T^{(T)} = R^{(T)}\mathscr{F}_T$, and*

$$\Phi(R_T^{(T)}, \Lambda) = \inf_{F \in R_T^{(T)}} \Phi(F, \Lambda)$$

(defined to be $+\infty$ if $R_T^{(T)}$ is empty).

PROOF. Follows from Theorem 3.1.

THEOREM 4.3. *Let $U^{(T)} \subseteq \mathscr{F} (T = 1, 2, \cdots)$ be a sequence of rejection regions corresponding to a certain sequence of tests of H_0 against H_1. Define a sequence of* L.R. *tests by their rejection regions*

$$R^{(T)} = \{F \in \mathscr{F} \mid \Phi(F, \Lambda) \geqq l_T^{(T)}\}, \qquad T = 1, 2, \cdots,$$

with

$$l_T^{(T)} = \Phi(U_T^{(T)}, \Lambda).$$

Then

(a) $U_T^{(T)} \subseteq R_T^{(T)}$, $T = 1, 2, \cdots$,

(b) *log size* $R^{(T)} = O(\log T) +$ *log size* $U^{(T)}$.

(c) *If* $Tl_T^{(T)}/\log T \to \infty$ *as* $T \to \infty$ *then, as* $T \to \infty$, *the sizes of the* $U^{(T)}$ *and the* $R^{(T)}$ *tests converge to zero faster than any power of* T.

(d) *For any set* $A \subseteq \mathscr{F}$ *let* $\sim A = \mathscr{F} - A$. *Let* $P \in \mathscr{P} - \Lambda$ *be such that* $P_T(\sim U^{(T)}) > 0$, *and that*

$$T\{\Phi((\sim R^{(T)})_T, P) - \Phi((\sim U^{(T)})_T, P)\}/\log T \to \infty \text{ as } T \to \infty.$$

Then $T^k P_T(\sim R^{(T)})/P_T(\sim U^{(T)}) \to 0$ *as* $T \to \infty$, *for any* k.

PROOF.

(a) $F \in U_T^{(T)} \Rightarrow \Phi(F, \Lambda) \geqq l_T^{(T)} \Rightarrow F \in R_T^{(T)}$.

(b) Since $U_T^{(T)}$ is a finite set, $\Phi(U_T^{(T)}, \Lambda)$ is attained in $U_T^{(T)}$, and hence, $\Phi(R_T^{(T)}, \Lambda) = \Phi(U_T^{(T)}, \Lambda) = l_T^{(T)}$.

(b) follows from Theorem 3.1(b).

(c) is a direct consequence of (b), and (d) follows from Theorem 4.1(b).

Notice that since $U_T^{(T)} \subseteq R_T^{(T)}$ for all T, the sizes of the $R^{(T)}$–tests are not smaller than those of the $U^{(T)}$–tests. The aim of the next theorem is to construct new L.R. tests, with sizes not larger than those of the corresponding $U^{(T)}$–tests, for which results similar to those in Theorem 4.3 can be shown.

THEOREM 4.4. *Let* $U^{(T)} \subseteq \mathscr{F}$, $T = 1, 2, \cdots$, *be a sequence of rejection regions for testing* H_0 *against* H_1. *Then*

(a) *there are numbers* λ_T, $0 \leqq \lambda_T = O(\log T/T)$ *for which the likelihood ratio tests with rejection regions* $S^{(T)} = \{F \in \mathscr{F} \mid \Phi(F, \Lambda) \geqq l_T^{(T)} + \lambda_T\}$ *have sizes not larger than those of the corresponding* $U^{(T)}$ *tests*;

(b) *let* $P \in \mathscr{P} - \Lambda$ *be such that* $P_T(\sim U^{(T)}) > 0$, $T\{\Phi((\sim R^{(T)})_T, P) - \Phi((\sim U^{(T)})_T, P)\}/\log T \to \infty$ *and*

$$\{\Phi((\sim R^{(T)})_T, P) - \Phi((\sim S^{(T)})_T, P)\}/\{\Phi((\sim R^{(T)})_T, P) - \Phi(\sim U^{(T)})_T, P)\} \to 0$$

as $T \to \infty$. *Then*

$$T^k P_T(\sim S^{(T)})/P_T(\sim U^{(T)}) \to 0$$

as $T \to \infty$ *for any* k.

PROOF. Similar to the proof of Theorem 3.1 of Hoeffding (1965).

5. Study of the function Φ. In this section we prove some facts about the function Φ which are relevant for the comparison of tests for simple hypotheses.

DEFINITION.

(a) A transition probability matrix $P \in \mathscr{P}$ is said to be regular iff there is a positive integer N such that $P^{(N)}$ is strictly positive entrywise, where $P^{(N)}$ is the matrix of N-step transition probabilities.

(b) $\mathscr{P}_0 = \{P \in \mathscr{P} \mid P \text{ is regular}\}$.

(c) For a fixed $P \in \mathscr{P}$ let $D(P) = \{(i,j) \mid p_{ij} = 0\}$, $\mathscr{F}(P) = \{F \in \mathscr{F} \mid F_{ij} = 0 \text{ if } (i,j) \in D(P)\}$, and $\mathscr{F}_0(P) = \{F \in \mathscr{F} \mid F_{ij} = 0 \text{ iff } (i,j) \in D(P)\}$.

Note that if P is strictly positive entrywise then $\mathscr{F}(P) = \mathscr{F}$. Otherwise $\mathscr{F}(P)$ is the intersection of those faces $\{F \in \mathscr{F} \mid F_{ij} = 0\}$ of the simplex $\mathscr{F}$ corresponding to pairs $(i,j) \in D(P)$. Moreover, for any $P \in \mathscr{P}$ and any T, $P_T(\mathscr{F}(P)) = 1$.

LEMMA 5.1.

(a) $0 \leqq \Phi(F,P) \leqq \infty$ *for any* $F \in \mathscr{F}$, *any* $P \in \mathscr{P}$.

(b) $\Phi(F,\dot{P}) = 0 \Leftrightarrow$ *for any* i *such that* $F_{i\cdot} > 0$,

$$(F_{ij}/F_{i\cdot}) = p_{ij} \quad \textit{all} \quad j.$$

(c) $\Phi(F,P) < \infty \Leftrightarrow F \in \mathscr{F}(P)$.

(d) *For any* $P \in \mathscr{P}$, *the function* $\Phi(\cdot, P)$ *is bounded and continuous on* $\mathscr{F}(P)$.

PROOF. Essentially similar to the proof of Lemma 4.1 of Hoeffding (1965).

LEMMA 5.2. *For any* $P \in \mathscr{P}$ *and* $A \subseteq \mathscr{F}$, $\Phi(A,P) < \infty$ iff $A\mathscr{F}(P)$ *is nonempty. In this case* $\Phi(A,P) = \Phi(A\mathscr{F}(P),P)$.

PROOF. The proof follows from Lemma 5.1.

LEMMA 5.3. *For any* $P \in \mathscr{P}$, *the function* $\Phi(\cdot, P)$ *is convex in* $\mathscr{F}$.

PROOF. Let G and H be arbitrary points in $\mathscr{F}$. For $\alpha \in (0,1)$ let $F(\alpha) = \alpha G + (1-\alpha)H$. Let $N_G = \{i \mid G_{i\cdot} > 0\}$, $N_H = \{i \mid H_{i\cdot} > 0\}$, $N = N_G \cup N_H$, $U = N_G N_H$, $V = (\sim N_G)N_H$, $W = N_G(\sim N_H)$, and $N(\alpha) = \{i \mid F(\alpha)_{i\cdot} > 0\}$.

Clearly, $N(\alpha) = N = U + V + W$ for any $\alpha \in (0,1)$. The lemma follows by studying separately the contributions to $\Phi(F(\alpha),P)$ made by entries in U, V and W, and then putting together these contributions.

DEFINITION.

(a) $\mathscr{L} = \{F \in \mathscr{F} \mid F_{i\cdot} - F_{\cdot i} = 0 \text{ all } i\}$.

(b) $\mathscr{K}_f^{(T)} = \{F \in \mathscr{F} \mid F_{s\cdot} - F_{\cdot s} = T^{-1}, F_{f\cdot} - F_{\cdot f} = -T^{-1}; F_{i\cdot} - F_{\cdot i} = 0 \text{ all } i \neq s,f\}$.

(c) $\mathscr{K}^{(T)} = \cup_{f \neq s} \mathscr{K}_f^{(T)}$.

REMARKS.

1. Any $\omega \in \Omega$ such that $X_T(\omega) = s$ produces, at time T, a transition count matrix $C(T)(\omega)$ for which the corresponding $F = (C(T)_{ij}(\omega)/T) \in \mathscr{L}$. Any $\omega \in \Omega$ such that $X_T(\omega) = f \neq s$ produces, at time T, a transition count matrix $C(T)(\omega)$ for which the corresponding $F = (C(T)_{ij}(\omega)/T) \in \mathscr{K}_f^{(T)}$. Hence it is only the Tth coordinate of the path that determines whether its normalized transition count matrix at time T is on $\mathscr{L}$ or on one of the sets $\mathscr{K}_f^{(T)}$. This remark, together with the facts that $\mathscr{L}$ and the $\mathscr{K}_f^{(T)}$'s are convex sets, and that $\mathscr{L}$ does not depend on T, suggests the very important role that the set $\mathscr{L}$ will play in the rest of this paper.

2. Clearly, for any T,

$$\mathscr{F}_T = \mathscr{L}\mathscr{F}_T \cup \mathscr{K}^{(T)}\mathscr{F}_T.$$

LEMMA 5.4. *Assume that $P \in \mathscr{P}_0$ and let A be a subset of $\mathscr{F}$ such that $A\mathscr{L}\mathscr{F}(P)$ is nonempty. Denote by A^c the closure of A in $\mathscr{F}$ and let $\mathscr{P}_0(A^c\mathscr{L}) = \{P \in \mathscr{P}_0 \mid$ there is an $F \in A^c\mathscr{L}$ such that $(F_{ij}/F_{i\cdot}) = p_{ij}$ for all j and all i such that $F_{i\cdot} > 0\}$. Then,*

(a) *there is an $F \in A^c\mathscr{L}\mathscr{F}(P)$ such that*

$$\Phi(F, P) = \Phi(A\mathscr{L}, P) = \Phi(A\mathscr{L}\mathscr{F}(P), P);$$

(b) *If $P \in \mathscr{P}_0(A^c\mathscr{L})$, then* (a) *is satisfied by any $F \in A^c\mathscr{L}\mathscr{F}(P)$ such that $(F_{ij}/F_{i\cdot}) = p_{ij}$ for all j and all i such that $F_{i\cdot} > 0$. In this case $\Phi(A\mathscr{L}, P) = 0$.*

(c) *If $P \in \mathscr{P}_0 - \mathscr{P}_0(A^c\mathscr{L})$ then $\Phi(A\mathscr{L}, P) > 0$ and any F satisfying* (a) *is in $\partial(A)\mathscr{L}\mathscr{F}(P)$, where $\partial(A)$ denotes the boundary of A in $\mathscr{F}$.*

PROOF.

(a) By Lemma 5.2, $\Phi(A\mathscr{L}, P) < \infty$ since $A\mathscr{L}\mathscr{F}(P)$ is nonempty by hypothesis. Part (a) follows by closedness of $A^c\mathscr{L}\mathscr{F}(P)$ and Lemma 5.1(d).

(b) Consider the system of equations

$$\psi_{i\cdot} = \sum_{k=1}^{m} \psi_{k\cdot} p_{ki} \qquad i = 1, 2, \cdots, m,$$

with the restrictions $\psi_{i\cdot} \geqq 0$ for all i, and $\sum_{i=1}^{m} \psi_{i\cdot} = 1$. Since $P \in \mathscr{P}_0$, there is a unique solution $(\psi_{1\cdot}^*, \cdots, \psi_{m\cdot}^*)$ to the system, and this solution is strictly positive (Kemeny and Snell (1963), pages 70, 71). Note that $(\psi_{1\cdot}^*, \cdots, \psi_{m\cdot}^*)$ are the stationary probabilities corresponding to P.

Hence $F^* = (\psi_{i\cdot}^* p_{ij})$ is the unique point in $\mathscr{L}$ such that $F_{ij}^*/F_{i\cdot}^* = p_{ij}$ for all i, j. The condition $P \in \mathscr{P}_0(A^c\mathscr{L})$ is equivalent to $F^* \in A^c\mathscr{L}$. Moreover it is clear that $F^* \in \mathscr{F}(P)$. Part (b) now follows from Lemma 5.1(b).

(c) In this case $\Phi(A\mathscr{L}, P) > 0$ by continuity of $\Phi(\cdot, P)$ in $\mathscr{F}(P)$. Suppose that there is $F_0 \in (\text{int } A)\mathscr{L}\mathscr{F}(P)$ such that $\Phi(F_0, P) = \Phi(A\mathscr{L}, P)$. Let F^* be the point defined in the proof of (b). Under the hypothesis $F^* \in \mathscr{L}\mathscr{F}(P) - A^c\mathscr{L}\mathscr{F}(P)$. If, for $\alpha \in (0, 1)$, we let $F(\alpha) = \alpha F_0 + (1-\alpha)F^*$, it follows from the convexity of $\Phi(\cdot, P)$ on $\mathscr{F}$ and from $\Phi(F^*, P) = 0$ that $\Phi(F(\alpha), P) \leqq \alpha\Phi(F_0, P) = \alpha\Phi(A\mathscr{L}, P)$. But, since $F_0 \in (\text{int } A)\mathscr{L}\mathscr{F}(P)$, it follows that there is an $\alpha_0 \in (0, 1)$ such that $F(\alpha_0) \in A\mathscr{L}\mathscr{F}(P)$. Hence

$$\Phi(F(\alpha_0), P) < \Phi(A\mathscr{L}, P),$$

and thus we have a contradiction.

Results similar to those of Lemma 5.4 can be obtained for each $\mathscr{K}_f^{(T)}$, instead of $\mathscr{L}$, for large T.

DEFINITION. Let $P \in \mathscr{P}$.

(a) A hyperplane in $\mathscr{F}(P)$ is a nonempty set

$$\{F \in \mathscr{F}(P) \mid \textstyle\sum_{(i,j) \notin D(P)} b_{ij} F_{ij} = b\}$$

with the b_{ij}'s not all equal.

(b) A hyperplane in $\mathscr{F}(P)\mathscr{L}$ is a nonempty set

$$\{F \in \mathscr{F}(P)\mathscr{L} \mid \textstyle\sum_{(i,j) \notin D(P)} b_{ij} F_{ij} = b\}$$

with the b_{ij}'s not all equal.

THEOREM 5.5. *Let* $P \in \mathscr{P}_0$, $0 < a < \max_{\mathscr{L}\mathscr{F}(P)} \Phi(F, P)$, *and* $R(a) = \{F \in \mathscr{F} \mid \Phi(F, P) \geqq a\}$. *Let* F^0 *be an arbitrary point in* $\partial(R(a))\mathscr{L}$.

(a) *If* $F^0 \in \partial(R(a))\mathscr{L}\mathscr{F}_0(P)$ *then, for any* $F \in \mathscr{F}(P)$,

$$\Phi(F,P) - a = \Phi(F, (F^0_{ij}/F^0_{i\cdot})) + \textstyle\sum_{i,j} (F_{ij} - F^0_{ij}) \log (F^0_{ij}/F^0_{i\cdot} p_{ij}).$$

The unique supporting hyperplane (in $\mathscr{L}\mathscr{F}(P)$*) to* $(\sim R(a))\mathscr{L}$ *at* F^0 *is*

$$H = \{F \in \mathscr{L}\mathscr{F}(P) \mid \textstyle\sum_{(i,j) \notin D(P)} (F_{ij} - F^0_{ij}) \log (F^0_{ij}/F^0_{i\cdot} p^0_{ij}) = 0\}.$$

(b) *If* $F^0 \in \partial(R(a))\mathscr{L} - \partial(R(a))\mathscr{L}\mathscr{F}(P)$ *then, for any* $(i, j) \notin D(P)$ *such that* $F^0_{ij} = 0$, $H_{ij} = \{F \in \mathscr{L}\mathscr{F}(P) \mid F_{ij} = 0\}$ *is a supporting hyperplane (in* $\mathscr{L}\mathscr{F}(P)$*) to* $(\sim R(a))\mathscr{L}$ *at* F^0. *These sets, and their intersections, are the only such supporting hyperplanes.*

PROOF. Since $\Phi(\cdot, P)$ is bounded on $\mathscr{F}(P)$, it follows from the hypothesis on a that $(\sim R(a))^c\mathscr{L} = \{F \in \mathscr{F} \mid \Phi(F, P) \leqq a\}\mathscr{L} \subset \mathscr{L}\mathscr{F}(P)$, thus any point in $\partial(R(a))\mathscr{L}$ is in $\mathscr{F}(P)$.

(a) The identity follows easily. From it, and from $\Phi(F, (F^0_{ij}/F^0_{i\cdot})) \geqq 0$, it follows that H is a supporting hyperplane (in $\mathscr{L}\mathscr{F}(P)$) to $(\sim R(a))\mathscr{L}$ at F^0. To show uniqueness let V^0 be a neighborhood of F^0 (in $\mathscr{L}\mathscr{F}(P)$) totally contained in $\mathscr{L}\mathscr{F}_0(P)$. For any $(i, j) \notin D(P)$, the derivatives of $\Phi(F, P)$ with respect to F_{ij} are continuous in V^0. Hence H is unique in V^0 and therefore in $\mathscr{L}\mathscr{F}(P)$.

(b) That the H_{ij}'s and their intersections are, in this case, supporting hyperplanes (in $\mathscr{L}\mathscr{F}(P)$) to $(\sim R(a))\mathscr{L}$ at F^0 is clear. To show that they are the only such hyperplanes it suffices to show that any straight line in $\mathscr{L}\mathscr{F}(P)$ passing through F^0 and an arbitrary point $X \in \mathscr{L}\mathscr{F}_0(P)$, intersects set $(\sim R(a))\mathscr{L}$.

Let $Z(\alpha) = (1-\alpha)F^0 + \alpha X$, $\alpha \in (0, 1)$. Let $N = \{i \mid F^0_{i\cdot} > 0\}$ and $M_i = \{j \mid F^0_{ij} > 0\}$. Then, as $\alpha \to 0$,

$$\frac{d}{d\alpha} \Phi(Z(\alpha), P) = \sum_{i \in N} \sum_{j \in (\sim M_i)} X_{ij} \log \alpha + O(1).$$

The range of the previous summation is nonempty since $X \in \mathscr{F}_0(P)$ and $F^0 \in \mathscr{F}(P) - F_0(P)$. So, as $\alpha \to 0$, $d/d\alpha\, \Phi(Z(\alpha), P) \to -\infty$.

On the other hand, by convexity of $\Phi(\cdot, P)$ and the fact that $\{i \mid X_{i\cdot} > 0, F^0_{i\cdot} > 0\}$ is not empty, it follows that $d^2/d\alpha^2\, \Phi(Z(\alpha), P) > 0$ for $\alpha \in (0, 1)$. Thus

$$a = \Phi(F^0, P) \geqq \Phi(Z(\alpha), P) - \alpha \frac{d}{d\alpha} \Phi(Z(\alpha), P),$$

and therefore, for small enough α, $\Phi(Z(\alpha), P) < a$ and $Z(\alpha) \in (\sim R(a))\mathscr{L}$.

THEOREM 5.6. *Let $P \in \mathscr{P}_0$ and let A be a subset of $\mathscr{F}$ such that $A^c\mathscr{L}\mathscr{F}_0(P)$ is nonempty. Assume further that $P \notin \mathscr{P}_0(A^c\mathscr{L})$. Then there is a unique point $Y \in \partial(A)\mathscr{L}$ such that $\Phi(Y, P) = \Phi(A\mathscr{L}, P)$.*

PROOF. Let $(\sim B) = \{F \in \mathscr{F} \mid \Phi(F, P) < \Phi(A\mathscr{L}, P)\}$. By the convexity of the function $\Phi(\cdot, P)$ on $\mathscr{F}$, and the convexity of $\mathscr{L}$, it follows that $(\sim B)\mathscr{L}$ is a convex set. Also, since $A^c\mathscr{L}\mathscr{F}(P)$ is nonempty, $\Phi(A\mathscr{L}, P) < \infty$, and hence $(\sim B)\mathscr{L} \subset \mathscr{L}\mathscr{F}(P)$. Furthermore, by definition, the convex sets $(\tilde{\sim} B)\mathscr{L}$ and $A\mathscr{L}$ have no points in common.

Let $(\psi^*_{1\cdot}, \cdots, \psi^*_{m\cdot})$ be the unique, strictly positive, stationary distribution of P, and $F^* = (\psi^*_{i\cdot} p_{ij})$. Clearly $F^* \in (\sim B)\mathscr{L}\mathscr{F}(P)$.

The point Y is obviously in $\mathscr{F}(P)$. If $Y \in \mathscr{F}(P) - \mathscr{F}_0(P)$ then the only supporting hyperplanes (in $\mathscr{L}\mathscr{F}(P)$) to $(\sim B)\mathscr{L}$ at Y would be those described in Theorem 5.5(b). No one of these planes separates $(\sim B)\mathscr{L}$ and $A\mathscr{L}$ because both these sets contain points in $\mathscr{F}_0(P)$. Since a separating hyperplane in $\mathscr{L}\mathscr{F}(P)$ must exist, it follows that $Y \in \mathscr{F}_0(P)$, and therefore, by Theorem 5.5(a), this hyperplane is unique and has an explicit form. The theorem follows by noting that Y is the unique point in $(\sim B)^c\mathscr{L}$ that lies in this hyperplane.

REMARKS.

1. Theorem 5.6 is very useful in the computation of type-II error probabilities of tests for $H_0: P = P_0$ against $H_1: P \neq P_0$, which have convex acceptance regions.

2. The previous results are all that is needed for the comparison of tests for simple hypotheses. If composite hypotheses were to be considered, double infima of the type $\Phi(A, \Lambda)$, $(A \subset \mathscr{F}, \Lambda \subset \mathscr{P})$ would be needed. Results similar to those in Hoeffding (1965) can also be proved in this case.

6. Almost equally informative sequences of sets. Even in the case of simple hypotheses (which the rest of this paper will consider) the computation of the infima that appear in the theorems of Section 4 may prove to be extremely difficult. It is thus desirable to look for conditions under which these discrete minimization problems could be replaced by simpler ones. Heuristically, the problem can be motivated as follows: suppose we wish to minimize a certain smooth function over the discrete set $A\mathscr{F}_T$, where $A(\subseteq \mathscr{F})$ is smooth enough. For large T, since the set $\mathscr{L}$ is "close" to all the sets $\mathscr{K}_f^{(T)}$, the infimum of the function over

$$A\mathscr{F}_T = A\mathscr{L}\mathscr{F}_T \cup (\cup_{f \neq s} A\mathscr{K}_f^{(T)}\mathscr{F}_T)$$

should not be "too different" from the infimum over $A\mathscr{L}\mathscr{F}_T$.

Furthermore, by definition of the sets $\mathscr{F}_T$, the infimum over $A\mathscr{L}\mathscr{F}_T$ should not be "too different", for large T, from the infimum over $A\mathscr{L}$.

DEFINITION. Let $A^{(T)}$, $B^{(T)}$, $T = 1, 2, \cdots$ be two sequences of subjects of $\mathscr{F}$ and let $P \in \mathscr{P}$. These sequences are said to be "almost equally informative with respect to P (a.e.i.–P)" if there is a T_0 such that for all $T \geqq T_0$,

$$\left|\Phi(A^{(T)}, P) - \Phi(B^{(T)}, P)\right| \leqq O(\log T/T).$$

LEMMA 6.1. *Let $A^{(T)}$, $T = 1, 2, \cdots$ be a sequence of subsets of $\mathscr{F}$ and let $P \in \mathscr{P}$. If (1) $A^{(T)}\mathscr{L}$ and $A^{(T)}(\mathscr{L} \cup \mathscr{K}^{(T)})$ are* a.e.i.–P, *(2) $A^{(T)}\mathscr{L}$ and $A_T{}^{(T)}\mathscr{L}$ are* a.e.i.–P, *then $A_T{}^{(T)}$ and $A^{(T)}\mathscr{L}$ are* a.e.i.–P.

PROOF. $A_T{}^{(T)}\mathscr{L} \subseteq A_T{}^{(T)} \subseteq A^{(T)}(\mathscr{L} \cup \mathscr{K}^{(T)})$ and hence

$$\Phi(A^{(T)}(\mathscr{L} \cup \mathscr{K}^{(T)}), P) \leqq \Phi(A_T{}^{(T)}, P) \leqq \Phi(A_T{}^{(T)}\mathscr{L}, P).$$

So by hypothesis, for large T,

$$\Phi(A^{(T)}\mathscr{L}, P) - O(\log T/T) \leqq \Phi(A_T{}^{(T)}, P) \leqq \Phi(A^{(T)}\mathscr{L}, P) + O(\log T/T).$$

DEFINITION. Let $F^{(T)}$, $T = 1, 2, \cdots$ be a sequence of points in $\mathscr{F}$. It will be said to be "bounded away from zero as $T \to \infty$" if there are numbers $\gamma > 0$ (independent of T) and T_0 such that $F_{ij}^{(T)} \geqq \gamma$ for any i, j, and any $T \geqq T_0$.

THEOREM 6.2. *Let $A^{(T)}$, $B^{(T)}$, $T = 1, 2, \cdots$ be two sequences of subsets of $\mathscr{F}$ and let $P \in \mathscr{P}_0$. Assume that for any T, $B^{(T)} \subseteq A^{(T)}$. If there is a T_0 such that for any $T \geqq T_0$*

(a) *there are points $F^{(T)} \in A^{(T)}$ bounded away from zero as $T \to \infty$ such that $\Phi(F^{(T)}, P) = \Phi(A^{(T)}, P)$;*

(b) *there are points $G^{(T)} \in B^{(T)}$ such that $\left|G_{ij}^{(T)} - F_{ij}^{(T)}\right| < aT^{-1}$ where $a \geqq 0$ independent of T; then the sequences $A^{(T)}$, $B^{(T)}$, $T = 1, 2, \cdots$ are* a.e.i.–P.

PROOF. It suffices to show that under the hypothesis $\Phi(G^{(T)}, P) - \Phi(F^{(T)}, P) \leqq O(\log T/T)$, since by hypothesis $\Phi(G^{(T)}, P) - \Phi(F^{(T)}, P) \geqq 0$. In most of what follows the superindex T will be deleted for simplicity.

Let $N_{FG} = \{i \mid F_{i\cdot} > 0, G_{i\cdot} > 0\}$, $N_F = \{i \mid F_{i\cdot} > 0, G_{i\cdot} = 0\}$, $N_G = \{i \mid F_{i\cdot} = 0, \ G_{i\cdot} > 0\}$, $M_{FG}(i) = \{j \mid F_{ij} > 0, \ G_{ij} > 0\}$, $M_F(i) = \{j \mid F_{ij} > 0, \ G_{ij} = 0\}$, $M_G(i) = \{j \mid F_{ij} = 0, G_{ij} > 0\}$. From the hypothesis that F is bounded away from 0, it follows that the sets N_G, $M_G(i)$ are empty for sufficiently large T. From the hypotheses that F is bounded away from 0 and that $\left|G_{ij} - F_{ij}\right| < aT^{-1}$ it follows that, for sufficiently large T, the sets N_F and $M_F(i)$ are empty. So for large T

$$\Phi(G, P) - \Phi(F, P) = \textstyle\sum_{i \in N_{FG}} \sum_{j \in M_{FG}(i)} d_{ij} = \sum_i \sum_j d_{ij},$$

where $d_{ij} = G_{ij} \log (G_{ij}/G_{i\cdot}p_{ij}) - F_{ij} \log (F_{ij}/F_{i\cdot}p_{ij})$. The rest of the proof follows in a manner similar to that of Lemma A.1 of Hoeffding (1965).

7. Further results on comparison of tests. In this section we study theorems like those of Section 4, in the cases where the approximation procedures described in Section 6 can be applied.

THEOREM 7.1. *Let* $U^{(T)}$, $T = 1, 2, \cdots$ *be a sequence of subsets of* $\mathscr{F}$. *Assume that the sequences* ${U_T}^{(T)}$ *and* $U^{(T)}\mathscr{L}$, $T = 1, 2, \cdots$ *are* a.e.i. *with respect to* $P \in \mathscr{P}$. *Then*

$$P_T(U^{(T)}) = P_T({U_T}^{(T)}) = \exp\,(O(\log T) - T\Phi(U^{(T)}\mathscr{L}, P)).$$

PROOF. It follows from the definition of a.e.i.–P sequences and Theorem 3.1(b).

THEOREM 7.2. *Let* $U^{(T)}$, $T = 1, 2, \cdots$ *be a sequence of subsets of* $\mathscr{F}$ *to be used as rejection regions of some sequence of tests for* $H_0: P = P_0$ *against* $H_1: P \neq P_0$.

Let $c^{(T)} = \Phi(U^{(T)}\mathscr{L}, P_0)$, ${c_T}^{(T)} = \Phi({U_T}^{(T)}, P_0)$ *and* $R(x) = \{F \in \mathscr{F} \mid \Phi(F, P_0) \geqq x\}$.

Assume that ${U_T}^{(T)}$, $U^{(T)}\mathscr{L}$, $T = 1, 2, \cdots$ *are* a.e.i.–P_0 *and that* $R(c^{(T)})\mathscr{L}$ *and* $R(c^{(T)})_T$, $T = 1, 2, \cdots$ *are* a.e.i.–P_0. *Then*:

(a) *There are numbers* λ_T, $0 \leqq \lambda_T = O(\log T/T)$ *such that*

$$P_T{}^0(R(c^{(T)} + \lambda_T)) \leqq P_T{}^0(U^{(T)}).$$

(b) *Let* $P \in \mathscr{P}$, $P \neq P_0$. *Assume that the sequences* $(\sim U^T)_T$ *and* $(\sim U^T)\mathscr{L}$, $T = 1, 2, \cdots$ *are* a.e.i.–P *and that* $(\sim R(c^{(T)} + \lambda_T))(\mathscr{L} \cup \mathscr{K}^{(T)})$ *and* $(\sim R(c^{(T)} + \lambda_T))\mathscr{L}$, $T = 1, 2, \cdots$ *are* a.e.i.–P. *Assume further that* $P_T(\sim R(c^{(T)} + \lambda_T)) > 0$ *and* $P_T(\sim U^{(T)}) > 0$. *Then*

$$P_T(\sim R(c^{(T)} + \lambda_T)) \leqq \{\exp\,[O(\log T) - Td_T(P) + Te_T(P)]\}P_T(\sim U^{(T)})$$

for sufficiently large T, *where*

$$e_T(P) = \Phi((\sim R(c^{(T)}))\mathscr{L}, P) - \Phi((\sim R(c^{(T)} + \lambda_T))\mathscr{L}, P) \geqq 0,$$

$$d_T(P) = \Phi((\sim R(c^{(T)}))\mathscr{L}, P) - \Phi((\sim U^{(T)})\mathscr{L}, P) \geqq 0.$$

PROOF. $c^{(T)} - O(\log T/T) \leqq {c_T}^{(T)} \leqq c^{(T)} + O(\log T/T)$ for large enough T, since ${U_T}^{(T)}$ and $U^{(T)}\mathscr{L}$ are a.e.i.–P_0.

Case 1. $c^{(T)} \leqq {c_T}^{(T)}$.

In this case $0 \leqq {c_T}^{(T)} - c^{(T)} \leqq O(\log T/T)$. So

$$F \in U^{(T)}\mathscr{L} \Rightarrow \Phi(F, P_0) \geqq c^{(T)} \Rightarrow F \in R(c^{(T)}),$$

$$F \in {U_T}^{(T)} \Rightarrow \Phi(F, P_0) \geqq {c_T}^{(T)} \geqq c^{(T)} \Rightarrow F \in R(c^{(T)}).$$

Hence $U^{(T)}\mathscr{L} \subseteq R(c^{(T)})$; ${U_T}^{(T)} \subseteq R(c^{(T)})$; ${U_T}^{(T)} \subseteq R(c^{(T)})_T$, which implies that $P_T{}^0(U^{(T)}) \leqq P_T{}^0(R(c^{(T)}))$. So the size of $R(c^{(T)})$ is not smaller than that of $U^{(T)}$. However, log size $U^{(T)} = O(\log T) - T{c_T}^{(T)} = O(\log T) - Tc^{(T)}$, log size $R(c^{(T)}) = O(\log T) - T\Phi(R(c^{(T)})_T, P_0) = O(\log T) - Tc^{(T)}$ since $R(c^{(T)})_T$ and $R(c^{(T)})\mathscr{L}$ are a.e.i.–P_0. (a) follows in this case. Also $R(c^{(T)} + \lambda_T) \subseteq R(c^{(T)})$ since $\lambda_T \geqq 0$. Under the hypothesis that $P_T(\sim R(c^{(T)} + \lambda_T)) > 0$ and $P_T(\sim U^{(T)}) > 0$ it follows that

$$P_T(\sim R(c^{(T)} + \lambda_T))$$
$$= \{\exp\,[O(\log T) - T(\Phi((\sim R(c^{(T)} + \lambda_T))_T, P) - \Phi((\sim U^{(T)})_T, P))]\}P_T(\sim U^{(T)}).$$

But by hypothesis

$$\Phi((\sim U^{(T)})_T, P) \leqq \Phi((\sim U^{(T)})\mathscr{L}, P) + O(\log T/T)$$

and

$$\begin{aligned}\Phi((\sim R(c^{(T)}+\lambda_T))_T, P) &\geqq \Phi((\sim R(c^{(T)}+\lambda_T))(\mathscr{L}U\mathscr{K}^{(T)}), P)\\ &\geqq \Phi((\sim R(c^{(T)}+\lambda_T))\mathscr{L}, P) - O(\log T/T).\end{aligned}$$

Thus

$$\begin{aligned}&P_T(\sim R(c^{(T)}+\lambda_T))\\ &\quad\leqq \{\exp[O(\log T) - T(\Phi((\sim R(c^{(T)}+\lambda_T)\mathscr{L}, P) - \Phi((\sim U^{(T)})\mathscr{L}, P))]\}P_T(\sim U^{(T)})\end{aligned}$$

and (b) follows by adding and subtracting $T\Phi((\sim R(c^{(T)})\mathscr{L}, P)$ in the exponent. Moreover, since $(\sim R(c^{(T)}))\mathscr{L} \subseteq (\sim R(c^{(T)}+\lambda_T))\mathscr{L}$, $e_T(P) \geqq 0$. And since $(\sim R(c^{(T)}))\mathscr{L} \subseteq (\sim U^{(T)})\mathscr{L}$, $d_T(P) \geqq 0$ also.

Case 2. $c^{(T)} \geqq c_T{}^{(T)}$.

In this case, $0 \leqq c^{(T)} - c_T{}^{(T)} \leqq O(\log T/T)$. So $F \in U^{(T)}\mathscr{L} \Rightarrow \Phi(F, P_0) \geqq c^{(T)} \Rightarrow F \in R(c^{(T)})\mathscr{L}$ which implies that $U^{(T)}\mathscr{L} \subseteq R(c^{(T)})\mathscr{L} \subseteq R(c^{(T)})$. (Note that $U_T{}^{(T)} \subseteq R(c^{(T)})$ is not necessarily true.)

So the size of the $U^{(T)}$ test need not be smaller than the size of $R(c^{(T)})$ test in this case.

However, log size $U^{(T)} = O(\log T) - Tc^{(T)}$, log size $R(c^{(T)}) = O(\log T) - Tc^{(T)}$ because $U_T{}^{(T)}$, $U^{(T)}\mathscr{L}$ and $R(c^{(T)})_T$, $R(c^{(T)})\mathscr{L}$ are a.e.i.–P_0.

So in any case it is true that there are λ_T's $0 \leqq \lambda_T = O(\log T/T)$ such that size of $R(c^{(T)}+\lambda_T)$ is not larger than the size of U^T.

The rest of the proof follows exactly in the same way as in Case 1.

COROLLARY 7.4. *Under the assumptions of Theorem* 7.3, *if* $Td_T(P)/\log T \to \infty$ *as* $T \to \infty$ *and* $e_T(P)/d_T(P) \to 0$ *as* $T \to \infty$, *then*

$$P_T(\sim R(c^{(T)}+\lambda_T))/P_T(\sim U^{(T)}) \to 0 \text{ as } T \to \infty$$

faster than any power of T.

Theorem 7.3 can be used to obtain results similar to those in Hoeffding (1965). In the case of testing $H_0: P = P_0 (P_0 \in \mathscr{P}_0)$ against $H_1: P \neq P_0$ it is clear that any reasonable test will eventually discriminate between the null hypothesis and any alternative P for which $D(P) \neq D(P_0)$. For alternatives P such that $D(P) = D(P_0)$, the theory developed in this paper can be used to show results similar to those in Hoeffding (1965). In particular it can be shown that given a sequence of chi-square tests for the above described problem, whose sizes go to zero at a not too fast rate, there is a sequence of L.R. tests with sizes not larger than those of the chi-square tests, and with type-II error probabilities much smaller, for large T, than those of the chi-square tests, *at every alternative* P *of interest* (i.e. such that $D(P) = D(P_0)$) *except for those lying on a well-defined curve.* This, and other related results, will be presented in a forthcoming paper.

8. Remarks about replicates. A different approach for the asymptotic theory of inference in finite state space, discrete time Markov chains was developed by Anderson and Goodman (1957).

In this approach a fixed, finite T is considered. Independent replicates $X_0^{(n)}$, $X_1^{(n)}$, $\cdots$, $X_T^{(n)}$; $n = 1, 2, \cdots, N$, are observed. The transition count vector is $(C(1), \cdots, C(T))$ where $C(t) = (C_{ij}(t))$ and $C_{ij}(t) = \sum_{n=1}^{N} I_{(X_{t-1}^{(n)} = i, X_t^{(n)} = j)}$, $t = 1, 2, \cdots, T$.

If the chain has a nonstationary transition behavior characterized by the vector of stochastic matrices $P(1), \cdots, P(T)$ then $(C(1), \cdots, C(T))$ is a sufficient statistic for $P(1), \cdots, P(\mathrm{T})$ at stage N.

If the chain has stationary transition behavior characterized by P then the pooled transition count matrix C defined by $C = (C_{ij})$, $C_{ij} = \sum_{t=1}^{T} C_{ij}(t)$ is a sufficient statistic for P at stage N.

In the nonstationary case, it is clear that, at stage N, $C_{ij}(t) \geqq 0$, $\sum_{ij} C_{ij}(t) = N$, $\sum_j C_{ij}(t) = \sum_j C_{ji}(t-1)$ for $t = 1, 2, \cdots, T$.

Under the assumption that $C_{i\cdot}(1)$, $i = 1, \cdots, m$ are known, and that $C_{i\cdot}(1)/N \to \eta_i > 0$ for $i = 1, \cdots, m$ as $N \to \infty$, Anderson and Goodman derive asymptotic distributions for $(C(1), \cdots, C(T))$ in the nonstationary case and for C in the stationary case.

Not surprisingly, because of the implicit multinomial structure, a large deviation result (similar to that of Theorem 3.1 of this paper and to that of Theorem 2.1 of Hoeffding (1965) for $P_N((C(1), C(2), \cdots, C(T)) \in A)$ can be derived in the nonstationary case. Also a general theory like that in Sections 5, 6, 7 can be developed from the large deviation result. The notation, though, becomes extremely cumbersome. It can be conjectured, however, that results like those mentioned at the end of Section 7 of this paper should also hold in this case.

In the stationary case, the large deviation result for $P_N(C \in A)$ does not hold uniformly in A. An example of this fact can be easily constructed. The class of sets on which the large deviation result holds uniformly must thus be restricted. This smaller class of sets should include the acceptance and rejection regions of the tests to be compared so that the theory developed in this paper carries over to this case.

Acknowledgment. The results in this paper are part of the author's Ph.D. dissertation, submitted to the University of California, Berkeley. The author wishes to thank his advisor Professor Peter J. Bickel for his guidance throughout this research, and Professor Lucien LeCam for many useful conversations. He also wishes to thank the referee for his useful comments.

REFERENCES

ANDERSON, T. W. and GOODMAN, L. A. (1957). Statistical methods about Markov chains. *Ann. Math. Statist.* **28** 89–110.

BAHADUR, P. R. and RAGAVACHARI, M. (1970). Some asymptotic properties of likelihood ratios on general sample spaces. To appear in *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*

BILLINGSLEY, P. (1961a). Statistical methods in Markov chains. *Ann. Math. Statist.* **32** 12–39.

BILLINGSLEY, P. (1961b). *Statistical Inference for Markov Processes*. Univ. of Chicago Press.

FELLER, W. (1962). An Introduction to Probability Theory and its Applications **1** (2nd ed.) Wiley, New York.

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–408.

JOHNSON, R. A. and ROUSSAS, G. G. (1969). Asymptotically most powerful tests in Markov processes. *Ann. Math. Statist.* **40** 1207–1215.

JOHNSON, R. A., ROUSSAS, G. G. (1970). Asymptotically optimal tests in Markov processes. *Ann. Math. Statist.* **41** 918–938.

KEMENY, J. G. and Snell, J. L. (1963). *Finite Markov Chains* (1st ed.). Van Nostrand, Princeton.

LECAM, L. (1960). Locally asymptotically normal families of distributions. *University Calif. Publ. Statist.* **3** 37–98.

SANOV, I. N. (1957). On the probability of large deviations of random variables. *Mat. Sb.* **42** (84) 11–44. (English translation in *Selected Transl. Math. Statist. Prob.* **1** (1961) 213–244.)

WALD, A. (1941). Asymptotically most powerful tests of statistical hypotheses. *Ann. Math. Statist.* **12** 1–19.

WHITTLE, P. (1955). Some distribution and moment formulae for the Markov chain. *J. Roy. Statist. Soc. B* **17** 235–242.