STATISTICAL METHODS IN MARKOV CHAINS*

P. Billingsley**
Consultant, Mathematics Division
The RAND Corporation

P-2092

September 6, 1960

**Professor, Department of Statistics
University of Chicago

*Special invited paper read before
the Institute of Mathematical Statistics
at Stanford University, August 23, 1960.

SUMMARY

This paper is an expository survey of the mathematical aspects of statistical inference as it applies to finite Markov chains, the problem being to draw inferences about the transition probabilities from one long, unbroken observation $\{x_1, x_2, \ldots, x_n\}$ on the chain. The topics covered include Whittle's formula, chi-square and maximum-likelihood methods, estimation of parameters, and multiple Markov chains. At the end of the paper it is briefly indicated how these methods can be applied to a process with an arbitrary state space or a continuous time parameter.

Section 2 contains a simple proof of Whittle's formula; Section 3 provides an elementary and self-contained development of the limit theory required for the application of chi-square methods to finite chains. In the remainder of the paper, the results are accompanied by references to the literature, rather than by complete proofs.

As is usual in a review paper, the emphasis reflects the author's interests. Other general accounts of statistical inference on Markov processes will be found in Grenander [53], Bartlett [9] and [10], Fortet [35], and in my monograph [18].

I would like to thank Paul Meier for a number of very helpful discussions on the topics treated in this paper, particularly those of Section 3.

1. INTRODUCTION

Let $\{x_1, x_2, \ldots\}$ be a stochastic process or sequence of random variables taking values in some finite set. The variable x_n is to be thought of as the state at time n of some system the evolution of which is governed by a set of probability laws. The finite set of values which the random variables assume, called the state space of the process, may be taken for notational convenience to consist of the first s positive integers.

The process $\{x_n\}$ is a Markov chain of order t if the conditional probability

$$P\{x_n = a_n \mid\mid x_m = a_m,\ m < n\}$$

is independent of the values of a_m for $m < n - t$. (A t-th order Markov process should be carefully distinguished from a t-dependent process, the defining property of the latter being that $(x_1, x_2, \ldots, x_m)$ and $(x_n, x_{n+1}, \ldots, x_{n+r})$ are independent if $n - m > t$. The terminology in the statistical literature is sometimes confusing.) A Markov chain of order 1 is also called a simple Markov chain. Throughout what follows it will be assumed that the Markov chain has stationary transition probabilities, that is,

$$(1.1) \qquad P\{x_n = a_{t+1} \mid\mid x_{n-t} = a_1, \ldots, x_{n-1} = a_t\} = p_{a_1, \ldots, a_t : a_{t+1}}$$

is independent of n. If t = 1, these quantities form an sxs stochastic matrix (p_{ij}), the transition matrix of the process.

If the transition probabilities are unknown, or else are specified functions of an unknown parameter, there arises the problem of making inferences about them from empirical data. It is therefore supposed that n + 1 successive states have been observed in an unbroken sequence; thus one has at hand a realization (or sample) $\{x_1, x_2, \ldots, x_{n+1}\}$ of the first n + 1 random variables. (The use of n + 1 instead of n simplifies later formulas.) The succeeding sections will deal with the large-sample theory of drawing inferences in this situation. The theory is based on chi-square methods, or the Neyman-Pearson criterion; any objections which can be made of these methods in the independent case apply a fortoriori in the present case (see Cochran [23]).

Since any probabilistic question about t-th order Markov chains is reducible by a standard device to a corresponding question about simple Markov chains, and since the same is essentially true of statistical questions (see Section 6), only simple chains will be considered in the next four sections. The following definitions and facts concerning such chains will be needed; see Feller [33] for a systematic account. The chain is said to be irreducible if for any pair i and j of states, $p_{ij}^{(n)} > 0$ for some n, where

$$p_{ij}^{(n)} = P\left\{x_{m+n} = j \,||\, x_m = i\right\}$$

are the n-th order transition probabilities. If the chain is irreducible then there is a unique set of (positive) stationary

probabilities, given by the solution of the system

$$\begin{cases} \sum_i p_i p_{ij} = 1 \\ \\ \sum_i p_i = 1. \end{cases}$$

If $P\left\{x_n = i\right\} = p_i$ holds for n = 1, then it holds for all n, so that the chain is stationary. The chain is said to be ergodic if it is irreducible and if its period (the greatest common divisor of the set of integers n such that $p_{11}^{(n)} > 0$) is 1. In the ergodic case there exist positive constants γ and ρ, $\rho < 1$, such that

$$|p_{ij}^{(n)} - p_j| < \gamma\rho^n \tag{1.2}$$

holds for all i,j and n. An elementary proof of this last fact will be found on p. 173 of Doob [32]. In most of what follows it will be assumed that the chain is stationary and ergodic.

2. WHITTLE'S FORMULA

Let $\left\{x_1, x_2, \ldots, x_{n+1}\right\}$ be a sample from a first order Markov process with transition probabilities p_{ij} and initial probabilities p_i. If $\left\{a_1, a_2, \ldots, a_{n+1}\right\}$ is a sequence of n + 1 states, then the probability that $x_1, x_2, \ldots, x_{n+1}$ is this sequence is just $p_{a_1} p_{a_1 a_2} \cdots p_{a_n a_{n+1}}$. For i,j=1,...s, let f_{ij} be the number of m, with $1 \leq m \leq n$, for which $a_m = i$ and $a_{m+1} = j$. The sxs matrix $F = \left\{f_{ij}\right\}$ will be called the transition count

of the sequence. Since

$$(2.1) \qquad p_{a_1} p_{a_1 a_2} \cdots p_{a_n a_{n+1}} = p_{a_1} \Pi_{ij} p_{ij}^{f_{ij}},$$

the transition count together with the initial state forms a sufficient statistic. The distribution of this statistic, which will now be derived, plays in the analysis of samples from Markov chains a role analogous to that played by the multinomial distribution in the analysis of independent samples.

Since the probability of obtaining any particular sequence which begins with a_1 and has transition count F is given by (2.1), it is necessary only to count the number of such sequences, in order to find the distribution of the sufficient statistic. If $f_{i.} = \sum_j f_{ij}$ and $f_{.j} = \sum_i f_{ij}$, then $\{f_{i.}\}$ and $\{f_{.j}\}$ are the frequency counts of $\{a_1,\ldots,a_n\}$ and $\{a_2,\ldots,a_{n+1}\}$ respectively, from which it follows that

$$f_{i.} - f_{.i} = \delta_{ia_1} - \delta_{ia_{n+1}}$$

$$\sum_{ij} f_{ij} = \sum_i f_{i.} = \sum_j f_{.j} = n.$$

It is clear from the first of these relations that F and the initial state completely determine the terminal state; similarly, F and the terminal state determine the initial state. (However, F alone does not determine both the initial and final states: $\{1,2,1\}$ and $\{2,1,2\}$ have identical transition counts, for example.) The following answer to the combinatorial problem posed above is due to Whittle.

THEOREM 2.1: Let F be an sxs matrix of nonnegative integers such that $\sum_{ij} f_{ij} = n$ and such that $f_{i.} - f_{.i} = \delta_{iu} - \delta_{iv}$, $i=1,\ldots,s$, for some pair u,v. If $N_{uv}^{(n)}(F)$ is the number of sequences $(a_1,a_2,\ldots,a_{n+1})$ having transition count F and satisfying $a_1 = u$ and $a_{n+1} = v$, then

$$N_{uv}^{(n)}(F) = \frac{\Pi_i f_{i.}!}{\Pi_{ij} f_{ij}!} F_{vu}^{*} \tag{2.2}$$

where F_{vu}^{*} is the (v,u)-th cofactor of the matrix $F^{*} = \left\{f_{ij}^{*}\right\}$ with components

$$f_{ij}^{*} = \begin{cases} \delta_{ij} - f_{ij}/f_{i.} & \text{if } f_{i.} > 0 \\ \delta_{ij} & \text{if } f_{i.} = 0 \end{cases} \tag{2.3}$$

The proof goes by induction. The result being easy to establish if n = 1 (in which case both sides of (2.2) are 1), assume it holds if n is replaced by n − 1. If F(u,w) is F with its (u,w)-th entry diminished by 1, then clearly

$$N_{uv}^{(n)}(F) = \sum_w N_{wv}^{(n-1)}(F(u,w)),$$

where the summation extends over those w for which $f_{uw} > 0$. Hence it suffices to show that the right-hand side of (2.2) satisfies this same relation, or that

$$F_{vu}^{*} = \sum_w f_{uw} f_{u.}^{-1} F_{vw}^{*}(u,w). \tag{2.4}$$

Since $F^*(u,w)$ and F^* agree outside the w-th column, $F^*_{vw}(u,w) = F^*_{vw}$. From this fact together with the definition (2.3), it follows that (2.4) is equivalent to $\sum_w f^*_{uw} F^*_{vw} = 0$, where the summation now extends over all w. Since $\sum_w f^*_{uw} F^*_{vw} = \delta_{uv} \det F^*$, (2.4) follows immediately for the case in which $u \neq v$ and it is necessary only to show that $\det F^* = 0$ if $u = v$. Suppose for notational convenience that $f_{i.} = f_{.i}$ is positive for $i \leq r$ and zero for $i > r$. Then F has the form

$$F = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix},$$

where A is an rxr matrix. By the definition (2.3),

$$F^* = \begin{bmatrix} A^* & 0 \\ 0 & I \end{bmatrix},$$

where the rows of A^* sum to 0. Thus $\det F^* = \det A^* = 0$. (If $u \neq v$, it may happen that F^* is nonsingular.)

Whittle's original proof of this theorem [78] involved integration methods. Subsequent proofs were given by Dawson and Good [30] and by Goodman [49], who derived the result from known theorems (due to van Aarden-Ehrenfest and de Bruijn [1] and to Smith and Tutte [76]) on the number of unicursal paths in an oriented linear graph. The proof given above is a corrected version of the one on p. 195 of my paper [17]. It is possible to reverse the steps of the proofs in [30] and [49] and deduce the graph-theoretic result from (2.2). (It should be pointed out that Dawson

and Good considered not the transition count F, but the circularized transition count, which is obtained from F by adding an extra tally in the (v,u)-th cell if $a_{n+1} = v$ and $a_1 = u$.)

From (2.1) and (2.2) it now follows that the probability that $\{x_1, x_2, \ldots, x_{n+1}\}$ has F as its transition count and that $x_1 = u$ (and hence, $x_{n+1} = v$) is just

$$p_u F^*_{vu} \frac{\Pi_i f_{i\cdot}!}{\Pi_{ij} f_{ij}!} \Pi_{ij} p_{ij}^{f_{ij}} \tag{2.5}$$

which is Whittle's formula. Note that for the validity of (2.5) it is not necessary to assume that the initial probabilities are stationary, or even that the transition matrix (p_{ij}) has any particular ergodic structure. Whittle's formula can be made the starting point of a number of investigations; I will indicate two of them.

Suppose that the process $\{x_n\}$ is actually independent with $P\{x_n = i\} = p_i$. Then (2.5) reduces to

$$p_u F^*_{vu} \frac{\Pi_i f_{i\cdot}!}{\Pi_{ij} f_{ij}!} \Pi_j p_j^{f_{\cdot j}} .$$

Now the probability that $\{x_2, \ldots, x_{n+1}\}$ has $\{f_{\cdot j}\}$ as its frequency count and that $x_1 = u$, $x_{n+1} = v$, is

$$p_u (n-1)! \frac{f_{\cdot v}}{\Pi_j f_{\cdot j}!} \Pi_j p_j^{f_{\cdot j}},$$

by the multinomial formula. Therefore the conditional probability of the transition count F, given the frequency count $\{f_{\cdot j}\}$ and

the fact that $x_1 = u$, $x_{n+1} = v$, is

$$(2.6) \qquad \frac{nF^*_{vu}}{f_{\cdot v}} \quad \frac{\Pi_i f_{i\cdot}! \Pi_j f_{\cdot j}!}{n! \Pi_{ij} f_{ij}!} ,$$

a formula due to Dawson and Good [30] and to Goodman [49]. Note that (2.6) is independent of the p_i. Now the second factor in (2.6) is just the conditional probability of obtaining cell frequencies f_{ij} in an ordinary contingency table, given that the marginal frequencies are $f_{i\cdot}$ and $f_{\cdot j}$. Further, it follows from the weak law of large numbers for independent trials that the first factor in (2.6) goes in probability to a constant (namely, p_v^{-1} times the (v,u)-th cofactor of the matrix $(\delta_{ij} - p_j)$). Since (2.6), as well as (2.6) with the first factor removed, yields 1 when summed over F, it is intuitively clear that this constant must be 1. Let S be any statistic which would test the hypothesis of independence in the contingency table $\{f_{ij}\}$ if it really were a contingency table instead of a transition count. If the first factor in (2.6) goes to 1 in probability then it is also intuitively clear that the asymptotic distribution of S is the same in the present case, that is, if $\{f_{ij}\}$ is the transition count of an independent sequence, as it would be in the standard contingency case. These facts are proved rigorously in Dawson and Good [30] and in Goodman [49]. For example, the chi-square statistic

$$(2.7) \qquad \sum_{ij} \frac{\left(f_{ij} - f_{i\cdot} f_{\cdot j}/n\right)^2}{f_{i\cdot} f_{\cdot j}/n}$$

has asymptotically the chi-square distribution with $(s-1)^2$ degrees of freedom. Thus (2.7) can be used to test the hypothesis that $\{x_n\}$ is independent (and stationary) within* the hypothesis that $\{x_n\}$ is a first-order Markov process. This fact has been proved also by Hoel [55] and Good [44] and will be a corollary of the more general results of Section 4 below.

A second application of Whittle's formula is to run theory. Suppose once more that $\{x_n\}$ is a Markov process but that $s = 2$. In this case the transition count

$$\begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}$$

is determined by f_{12}, f_1 and f_2 (dropping for the moment the distinction between $f_{i.}$ and $f_{.i}$). But f_{12} is essentially the number r of runs of 1's in the sample. Thus (r, f_1, f_2) is essentially a sufficient statistic and its distribution is a special case of Whittle's formula. This fact has been used by Goodman [48] to derive the distributions of a number of runs tests. Most of these runs tests turn out to be tests of the Markov property. See [48] and its forerunners: David [29]; Barton and David [11], [12] and [13]; and Moore [69] and [70].

---

*If H is a hypothesis contained in the larger hypothesis H', I will, following Good [44], speak of testing H <u>within</u> H', rather than of testing H against alternatives in H'-H.

Whittle's formula can also be used to derive the moments and cumulants of various distributions; see Whittle [78], Patankar [73], Good [45], Gabriel [38], and Krishna Iyer [60].

3. CHI-SQUARE METHODS

A more systematic way of attacking the problem of statistical analysis of Markov chains is to carry over to the Markov case the chi-square methods applicable in the multinomial case, the methods treated for example in Chapter 30 of Cramer [26]. To simplify the discussion it will be assumed at first that the chain is stationary and ergodic; later it will be indicated how these requirements can be relaxed.

Ignoring the factor $p_u F^*_{vu}$ in Whittle's formula (2.5), one can say roughly that the probability of the transition count F is

$$\Pi_i \left[\frac{f_i!}{\Pi_j f_{ij}!} \Pi_j p_{ij}^{f_{ij}} \right] . \tag{3.1}$$

(In this section $f_{i.}$ will be denoted by f_i; this quantity is still to be distinguished from $f_{.i}$.) Now (3.1) is formally the same as the probability of obtaining the s frequency counts $(f_{i1},\ldots,f_{is})$ in s independent samples of sizes f_i respectively from multinomial populations with cell probabilities $(p_{i1},\ldots,p_{is})$. Let

$$\xi_{ij} = \left(f_{ij} - f_i p_{ij}\right)/\sqrt{f_i} . \tag{3.2}$$

If this multinomial situation really did obtain, then the s random vectors $\xi_i = (\xi_{i1},\ldots,\xi_{is})$ would be independent of each other, the covariance structure of ξ_i would be $E\{\xi_{ij}\xi_{i\ell}\} = \delta_{j\ell}p_{ij} - p_{ij}p_{i\ell}$, and, if f_i were large, ξ_i would be approximately normally distributed. Now in the Markov case, the f_i will be large with high probability, provided n is large. Hence it is reasonable to conjecture the following result.

THEOREM 3.1: In the stationary, ergodic case, the distribution of the s^2-dimensional random vector $\xi = (\xi_{ij})$ converges as $n \to \infty$ to the normal distribution* with covariance matrix $(\lambda_{ij,k\ell})$, where

$$\lambda_{ij,k\ell} = \delta_{ik}(\delta_{j\ell}p_{ij} - p_{ij}p_{i\ell}) \cdot \tag{3.3}$$

Assuming for the moment the truth of this theorem, it follows from the ordinary chi-square theory that each of the statistics

$$\sum_j \frac{(f_{ij} - f_i p_{ij})^2}{f_i p_{ij}}, \qquad i=1,\ldots,s, \tag{3.4}$$

has asymptotically the chi-square distribution. The summation in (3.4) must be restricted to those indices j for which $p_{ij} > 0$; if the number of these is d_i, then the number of degrees of freedom in the limiting distribution is $d_i - 1$.

*All normal distributions considered here are centered at the origin.

(The degenerate case $d_i = 1$ is possible.) Moreover, the s statistics are asymptotically independent, so that their sum

$$(3.5) \qquad \sum_{ij} \frac{(f_{ij} - f_i p_{ij})^2}{f_i p_{ij}}$$

has asymptotically a chi-square distribution with d-s degrees of freedom, where $d = \sum_i d_i$ is the number of positive entries in the transition matrix (p_{ij}). The statistic (3.5), first considered by Bartlett [7], provides a measure of the goodness of fit of the sample with the assumed transition probabilities p_{ij}.

A number of different proofs of Theorem 3.1 are possible (see Bartlett [7] and Whittle [78]); for example, it can be proved via the central limit theorem for Markov chains. The following proof, which was suggested to me by Paul Meier, simply makes precise the heuristic arguments which preceded the statement of the theorem. It is very simple, direct, and, from the statistical point of view, natural. It has the further advantage that it can be made the basis of a new proof of the central limit theorem for Markov chains. The following preliminary result is needed.

LEMMA 3.2: Assume that the chain is stationary and ergodic and let $\zeta = (\zeta_1, \ldots, \zeta_s)$ be the random vector with components

$$(3.6) \qquad \zeta_i = (f_i - n\, p_i)/\sqrt{n}\,.$$

Then

$$(3.7) \qquad \begin{cases} E\{\zeta_i\} = 0 \\ E\{\zeta_i \zeta_j\} = \alpha_{ij} + O(1/n), \end{cases}$$

where

$$(3.8)\quad \alpha_{ij} = \delta_{ij}p_i - p_ip_j + p_i \sum_{m=1}^{\infty}(p_{ij}^{(m)} - p_j) + p_j \sum_{m=1}^{\infty}(p_{ji}^{(m)} - p_i).$$

Moreover, the weak law of large numbers holds:

$$(3.9)\qquad \operatorname*{p\,lim}_{n\to\infty} f_i/n = p_i.$$

To prove (3.7), define the random variable $c_m(i)$ to be 1 or 0 according as x_m equals i or not. Then $f_i = \sum_{m=1}^{n} c_m(i)$. From the stationarity of the chain it follows that $E\{c_m(i)\} = p_i$, so that $E\{\zeta_i\} = 0$. Now

$$E\{\zeta_i\zeta_j\} = n^{-1}\sum_{\ell=1}^{n}\sum_{m=1}^{n} E\{(c_\ell(i) - p_i)(c_m(j) - p_j)\}\ .$$

Again using the stationarity, one sees that

$$E\{(c_\ell(i)-p_i)(c_m(j)-p_j)\} = P\{x_\ell=i,\ x_m=j\} - p_ip_j = \begin{cases} p_ip_{ij}^{(m-\ell)} - p_ip_j & \text{if } m>\ell \\ p_i\delta_{ij} - p_ip_j & \text{if } m=\ell \\ p_jp_{ji}^{(\ell-m)} - p_jp_i & \text{if } m<\ell. \end{cases}$$

Therefore,

$$(3.10)\quad E\{\zeta_i\zeta_j\} = (p_i\delta_{ij}-p_ip_j) + n^{-1}\sum_{m=1}^{n-1}(n-m)(p_ip_{ij}^{(m)}-p_ip_j) + n^{-1}\sum_{m=1}^{n-1}(n-m)(p_jp_{ji}^{(m)}-p_jp_i).$$

The first sum on the right-hand side of this equation differs from the corresponding sum in the definition of α_{ij} by the amount

$$(3.11) \qquad n^{-1} p_i \sum_{m=n}^{\infty} (p_{ij}^{(m)} - p_j) + n^{-1} p_i \sum_{m=1}^{n-1} m(p_{ij}^{(m)} - p_j).$$

From (1.2) it follows that the series $\sum_{m=1}^{\infty} (p_{ij}^{(m)} - p_j)$ and $\sum_{m=1}^{\infty} m(p_{ij}^{(m)} - p_j)$ converge absolutely. Therefore, the difference (3.11) is of the order $O(1/n)$. The second sum in (3.10) is treated similarly and (3.7) is thus established. (This sort of computation is standard; see p. 225 of Doob [32].) And now (3.9) follows by Chebyshev's inequality.

The weak law of large numbers (3.9), the only part of Lemma 3.2 needed for the proof of Theorem 3.1, follows also from recurrent event theory; see p. 297 of Feller [33]. However, the computation (3.8) is needed for the central limit theorem (Theorem 3.3 below).

Theorem 3.1 will now be proved. The process $\{x_n\}$ can be viewed as having been generated in the following fashion. Consider an <u>independent</u> collection of random variables x_1 and w_{in} $(i=1,2,\ldots,s;\ n=1,2,\ldots)$ such that $P\{x_1 = i\} = p_i$ and $P\{w_{in} = j\} = p_{ij}$. Imagine the variables w_{in} set out in the following array:

$$\begin{cases} w_{11}, w_{12}, \cdots, w_{1n}, \cdots \\ w_{21}, w_{22}, \cdots, w_{2n}, \cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ w_{s1}, w_{s2}, \cdots, w_{sn}, \cdots \end{cases}.$$

First, x_1 is sampled. If $x_1 = i$, then the first variable in the i-th row of the array is sampled, the result being x_2 by definition. If $x_2 = j$, then the first variable in the j-th row is sampled, unless $j = i$, in which case the second variable is sampled. In any case, the result of the sampling is by definition x_3. The next variable sampled is the first one in row x_3 which has not yet been sampled. The process continues in the obvious way. More formally, x_2 is defined to be $w_{x_1 1}$, and, if $x_1, x_2, \ldots, x_n$ have been defined, then x_{n+1} is taken to be $w_{x_n, m}$, where $m - 1$ is the number of ℓ, $1 \leq \ell < n$, such that $x_\ell = x_n$. It is intuitively clear that

$$(3.12) \qquad P\{x_k = a_k,\ 1 \leq k \leq n+1\} = p_{a_1} p_{a_1 a_2} \cdots p_{a_n a_{n+1}}.$$

For a rigorous proof, note that by definition

$$\{x_k = a_k,\ 1 \leq k \leq n+1\} = \{x_1 = a_1,\ w_{a_{k-1} m_k} = a_k,\ 2 \leq k \leq n+1\},$$

where $m_k - 1$ is the number of elements among $\{a_1, \ldots, a_{k-1}\}$ which are equal to a_k. Since the variables involved are all distinct and independent,

$$P\{x_k = a_k,\ 1 \leq k \leq n+1\} = P\{x_1 = a_1\} P\{w_{a_1 m_2} = a_2\} \cdots P\{w_{a_n m_{n+1}} = a_{n+1}\},$$

and (3.12) follows.

Since the process produced according to the above prescription has, by (3.12), the proper joint distributions, it can be used to compute the distributions of the f_{ij}. Clearly $(f_{i1}, \ldots, f_{is})$ is

the frequency count of $\{w_{i1},\ldots,w_{if_i}\}$. Since, by the weak law of large numbers (3.9), f_i is near np_i with high probability, it is natural to compare $(f_{i1},\ldots,f_{is})$ with the frequency count $(g_{i1},\ldots,g_{is})$ of $\{w_{i1},\ldots,w_{i[np_i]}\}$. From the independence of the array $\{w_{in}\}$ and the central limit theorem for multinomial trials, it follows that the s^2 random variables

$$(g_{ij} - [np_i]p_{ij})/\sqrt{np_i}$$

are asymptotically jointly normally distributed with covariance matrix given by (3.3). Now it will follow by Section 20.6 of Cramér [26] that the s^2-dimensional random vector η, with components

$$\eta_{ij} = (f_{ij} - f_i p_{ij})/\sqrt{np_i}\ , \tag{3.13}$$

will have this same limiting distribution, if it is shown that for each fixed i and j, the difference

$$\frac{g_{ij} - [np_i]p_{ij}}{\sqrt{n}} - \frac{f_{ij} - f_i p_{ij}}{\sqrt{n}} \tag{3.14}$$

goes to 0 in probability. Since the ratio of ξ_{ij} (defined by (3.2)) and η_{ij} goes to 1 in probability by (3.9), it will then follow (by Section 20.6 of [26] again) that ξ has this limiting distribution as well, which will complete the proof of Theorem 3.1.

To show that (3.14) goes to 0 in probability it will be convenient to change the notation; let e_m be defined by

$$e_m = \begin{cases} 1 - p_{ij} & \text{if } w_{im} = j \\ -p_{ij} & \text{if } w_{im} \neq j \end{cases}$$

and put $S_m = e_1 + \ldots + e_m$. Then the e_m are independent and identically distributed with mean 0 and variance $\sigma^2 = p_{ij}(1-p_{ij})$, and the difference (3.14) becomes

$$(3.15) \qquad (S_{[np_i]} - S_{f_i})/\sqrt{n} \ .$$

Given $\epsilon > 0$, choose n_o so that if $n \geq n_o$, then

$$P\left\{\left|f_i - [np_i]\right| > n\epsilon^3\right\} < \epsilon,$$

which is possible by (3.9). If $n \geq n_o$, then

$$P\left\{\left|S_{[np_i]} - S_{f_i}\right|/\sqrt{n} > \epsilon\right\}$$

$$\leq P\left\{\left|f_i - [np_i]\right| > n\epsilon^3\right\} + P\left\{\max_{|m-[np_i]| \leq n\epsilon^3} \left|S_{[np_i]} - S_m\right| > \epsilon\sqrt{n}\right\}$$

$$\leq \epsilon + 2\, P\left\{\max_{1 \leq m \leq n\epsilon^3} |S_m| > \epsilon\sqrt{n}/2\right\}$$

$$\leq \epsilon + 2\,(4/\epsilon^2 n)(n\epsilon^3\sigma^2) = (1 + 8\,\sigma^2)\epsilon,$$

where the last inequality follows from that of Kolmogorov (see p. 220 of Feller [33]). Since ϵ was arbitrary, (3.15) goes to 0 in probability. (This sort of argument is used in sequential problems; see Anscombe [5].) This completes the proof of Theorem 3.1.

It is possible to show that the covariance matrix of η, defined by (3.13), is exactly that of its limiting distribution. In fact, if

$$d_m(i,j) = \begin{cases} 1 - p_{ij} & \text{if } x_m = i \text{ and } x_{m+1} = j \\ -p_{ij} & \text{if } x_m = i \text{ and } x_{m+1} \neq j \\ 0 & \text{if } x_m \neq i \end{cases}$$

then $f_{ij} - f_i p_{ij} = \sum_{m=1}^{n} d_m(i,j)$. A straightforward computation shows that if $m \neq r$, then $d_m(i,j)$ and $d_r(k,\ell)$ are uncorrelated. From this fact together with stationarity it follows that

$$E\left\{(f_{ij} - f_i p_{ij})(f_{k\ell} - f_k p_{k\ell})\right\} = n \; E\left\{d_1(i,j) d_1(k,\ell)\right\} .$$

The proof is completed by showing that

$$E\left\{d_1(i,j)\; d_1(k,\ell)\right\} = p_i \lambda_{ij,k\ell} ,$$

which is again just a matter of computation.

Although Theorem 3.1 is all that is needed for the statistical analysis of Markov chains, it is interesting to see how it leads to a simple proof of the asymptotic normality of the random vector ζ defined by (3.6).

THEOREM 3.3: Under the assumptions of Lemma 3.2, the distribution of the random vector ζ converges to the normal distribution with covariance matrix (α_{ij}).

Now it has been shown that the distribution of the random vector η, defined by (3.13), approaches the normal distribution

with covariance matrix $\Lambda = (\lambda_{ij,k\ell})$. Moreover, the covariance matrix of η is exactly Λ for all n. Since f_j and $f_{\cdot j}$ differ at most by 1, if

$$\phi_j = (f_j - \sum_i f_i p_{ij})/\sqrt{n}$$

then

$$\phi_j + O(1/\sqrt{n}) = (f_{\cdot j} - \sum_i f_i p_{ij})/\sqrt{n} = \sum_i p_i^{1/2} \eta_{ij}.$$

Therefore the distribution of $\phi = (\phi_j)$ approaches a normal distribution with some covariance matrix M, and the covariance matrix of ϕ itself has the form $M + O(1/n)$. But the relation

$$\phi_j = \sum_i (\delta_{ij} - p_{ij})\zeta_i \tag{3.16}$$

is easy to verify. Thus ϕ, known to be asymptotically normal, is a linear transformation of ζ. If this transformation were invertible, the asymptotic normality of ζ would follow immediately. Actually, although the transformation (3.16) is singular, ζ can be recovered in a linear fashion from ϕ because (3.16) is one-to-one on that (s-1)-dimensional subspace of R_s in which ζ and ϕ must lie, namely, the subspace $H = \{z \epsilon R_s : \sum_i z_i = 0\}$. Suppose in fact that z is a (nonrandom) element of H such that

$$\sum_i (\delta_{ij} - p_{ij}) z_i = 0, \qquad i=1,\ldots,s. \tag{3.17}$$

Since the transition matrix (p_{ij}) is ergodic, the solutions of the system $z_j = \sum_i z_i p_{ij}$ form a one-dimensional subspace of R_s, that spanned by $(p_1,\ldots,p_s)$. Therefore (3.17) implies that

$z_i = \alpha p_i$, where α is a scalar. If $\sum_i z_i = 0$, then α must be 0. Therefore the transformation (3.16) is nonsingular when restricted to H, so that ζ is a linear function of ϕ. This implies, in the first place, that the distribution of ζ approaches a normal distribution with some covariance matrix N, and, in the second place, that the covariance matrix of ζ has the form $N + O(1/n)$. But by Lemma 3.1, the covariance matrix of ζ is $(\alpha_{ij}) + O(1/n)$. Therefore $N = (\alpha_{ij})$ and the proof is complete.

The central limit theorem for Markov chains is usually stated in a different form. Let $\psi(1),\ldots,\psi(s)$ be s numbers such that $E\{\psi(x_n)\} = \sum_i p_i \psi(i) = 0$. Then the distribution of $n^{-1/2}S_n = n^{-1/2} \sum_{k=1}^{n} \psi(x_k)$ approaches a normal distribution with mean 0 and variance

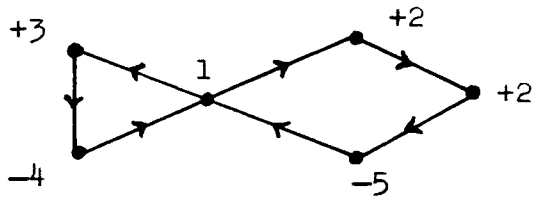
$$\sigma^2 = E\{\psi(x_1)^2\} + 2\sum_{k=1}^{\infty} E\{\psi(x_1)\ (x_{k+1})\}. \tag{3.18}$$

(In this form the theorem can be proved under much more general conditions; see p. 228 of Doob [32].) This theorem is a consequence of Theorem 3.3, since $n^{-1/2}S_n = \sum_i \zeta_i \psi(i)$ and since (3.18) is just another way of writing

$$\sigma^2 = \sum_{ij} \alpha_{ij} \psi(i) \psi(j).$$

Note that if the vector $(\psi(i),\ldots,\psi(s))$ is annihilated by the matrix (α_{ij}) then σ^2 will be zero, so that $n^{-1/2}S_n$ will go to zero in probability. This, the so-called degenerate case, can arise in circumstances of which the following example is the prototype. Consider a six-state chain represented by the

following diagram.



Here the points represent the states, the arrows represent the possible transitions and the numbers are the values of the $\psi(i)$. The chain is ergodic, but clearly $|S_n| \leq 5$ for all n, since the sum of the $\psi(i)$ around any circuit is 0.

Now that Theorem 3.3 has been proved, the results stated in the paragraph following it are established. The goodness of fit statistic (3.5), which has now been proved to have asymptotically a chi-square distribution with d-s degrees of freedom, can be shown to be equivalent to the appropriate Neyman-Pearson criterion as Bartlett [7] pointed out. In fact, using the methods of Wilks [79] (see [18] for the details) it can be shown that

$$(3.19) \qquad \sum_{ij} \frac{(f_{ij} - f_i p_{ij})^2}{f_i p_{ij}} \sim 2 \sum_{ij} f_{ij} \lg \frac{f_{ij}}{f_i p_{ij}} .$$

(Here and in what follows, the notation $\xi \sim \eta$ is used to indicate that the difference $\xi - \eta$ goes to 0 in probability.) Now the log-likelihood of the sample $\{x_1, \ldots, x_{n+1}\}$ is essentially

$$\Sigma_{ij} f_{ij} \lg p_{ij} .$$

Here the term $\lg p_{x_1}$ has been suppressed, since it is small

compared with this sum. If this expression is maximized subject to the constraints $\sum_j p_{ij} = 1$ by the method of Lagrange multipliers, it is found that the maximum occurs at $\hat{p}_{ij} = f_{ij}/f_i$ and that the maximum value is

$$\sum_{ij} f_{ij} \lg (f_{ij}/f_i).$$

Thus the right-hand member of (3.16) is just the Neyman-Pearson criterion, that is, twice the difference of the maximum of the log-likelihood and its actual value.

Throughout this section it has been assumed that the chain is stationary and ergodic. If the assumption of stationarity is removed, and any initial distribution allowed, then the results still hold, since the initial effects wear off as n become large. The only difference now is that the expected values of the various random variables (3.6), etc., are asymptotically 0, rather than exactly 0.

Suppose there is just one ergodic class, say $\{1,2,\ldots,r\}$, but that there exist transient states $\{r+1,\ldots,s\}$. The transition matrix P then has the form

$$P = \begin{bmatrix} A & 0 \\ B & C \end{bmatrix}.$$

The process very quickly leaves the transient set (the probability of being in a transient state at time n goes to 0 exponentially fast) and once the ergodic class is entered, it is never left. Thus the large sample theory above makes it

possible to do inference on the elements of the rxr stochastic matrix A. Large sample theory is not applicable to the elements of B and C, however, since the process stays among the transient states such a short time. A systematic analysis of this situation would be interesting.

The assumption that the chain is aperiodic can certainly be removed; all that happens is that the formula (3.8) becomes more complicated. The easiest way to see that the assumption of aperiodicity is inessential is to consider, if the chain has period λ, a new chain $\{(x_n, x_{n+1}, \ldots, x_{n+\lambda-1});\ n=1,2,\ldots\}$. This new chain is aperiodic, and a knowledge of its evolution the equivalent to a knowledge of the evolution of the original chain $\{x_n\}$.

The only assumption which cannot be relaxed is, of course, that of irreducibility. However, if the chain has more than one ergodic class, it is still possible to derive the limiting distributions of the various statistics considered here, conditional on a knowledge of which ergodic class the initial state lies in. This is all that is necessary for purposes of inference.

This section has dealt with the problem of testing, within the hypothesis that $\{x_n\}$ is a Markov chain, the hypothesis that it has specified transition probabilities. It is possible also to test one simple hypothesis against another. If (p_{ij}) and (q_{ij}) are two ergodic stochastic matrices with stationary distributions (p_i) and (q_i), then the logarithm of the likelihood

ratio appropriate to the test is

$$\lg \left(q_{x_1}/p_{x_1}\right) + \sum_{k=1}^{n} \lg \left(q_{x_k x_{k+1}}/p_{x_k x_{k+1}}\right)$$

$$= \lg \left(q_{x_1}/p_{x_1}\right) + \sum_{ij} f_{ij} \lg \left(q_{ij}/p_{ij}\right).$$

The limiting distribution of this statistic (properly normed) is normal, but it is hard to get simple expressions for the mean and variance; see Goodman [48].

Further papers related to the topics treated in this section are Romanovskii [74]; Bartlett [8]; Smirnov [75]; Cox [24] and [25]; Mihoc [67]; Firescu [34]; Broadbent [21]; and Cane [22]. A few results on power will be found in my monograph [18].

4. ESTIMATION OF PARAMETERS

In the preceding section it was shown that

$$(4.1) \qquad \sum_{ij} \frac{(f_{ij} - f_i p_{ij})^2}{f_i p_{ij}}$$

is asymptotically chi-square in distribution. If all the p_{ij} are positive, as will be assumed throughout this section to simplify matters, then the number of degrees of freedom is $s(s-1)$. This chi-square statistic is useful for testing whether the transition probabilities of the process have specified values p_{ij}. There arises naturally the problem of testing whether these transition probabilities have a specified

form $p_{ij}(\theta)$, where θ is an unknown parameter which must be estimated from the sample. Now if the process is really governed by the transition matrix $(p_{ij}(\theta))$, the log-likelihood of the observation $\{x_1,\ldots,x_{n+1}\}$ is (essentially)

$$\sum_{ij} f_{ij} \lg p_{ij}(\theta). \tag{4.2}$$

If the parameter is a vector $\theta = (\theta_1,\ldots,\theta_r)$ with r real components, then the maximum likelihood equations are

$$\sum_{ij} \frac{f_{ij}}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_u} = 0, \qquad u=1,\ldots,r. \tag{4.3}$$

If this system of equations has a solution $\hat{\theta}$, then the insertion of $p_{ij}(\hat{\theta})$ into (4.1) yields a statistic appropriate to the testing problem in question, namely,

$$\sum_{ij} \frac{(f_{ij} - f_i p_{ij}(\hat{\theta}))^2}{f_i p_{ij}(\hat{\theta})} . \tag{4.4}$$

One expects this statistic to be approximately chi-square with $s(s-1)-r$ degrees of freedom; the following theorem shows that this is true under appropriate regularity conditions.

THEOREM 4.1: Suppose that for each θ in an open subset Θ of r-dimensional Euclidean space, $(p_{ij}(\theta))$ is an sxs stochastic matrix with positive entries. Suppose that each $p_{ij}(\theta)$ has continuous partial derivatives of first and second order in Θ and that the s^2xr matrix D with entries

$$d_{ij,u} = \partial p_{ij}(\theta)/\partial \theta_u \tag{4.5}$$

has rank r *throughout* Θ. *Suppose further that* $\{x_n\}$ *is a Markov chain with transition probabilities* $p_{ij}(\theta)$ *for some* $\theta \in \Theta$. *Then there exists a random vector* $\hat{\theta}$ *in* $\bar{\Phi}$ *such that* $\hat{\theta}$ *is, with probability going to* 1, *a solution of the system* (4.3) *and such that* $\hat{\theta}$ *converges in probability to the true value of* θ. *Finally, the statistic* (4.4) *has asymptotically the chi-square distribution with* s(s-1)-r *degrees of freedom.*

It should be pointed out that in this theorem certain possible pathologies are ignored. There is only one consistent solution to (4.3), but there may be others which are not consistent; the theorem provides no means of selecting that solution which is near the true value of θ. Further, while it is true that if n is large, then $\hat{\theta}$ is, with high probability, a local maximum of (4.2), there is no assurance that it is an absolute maximum. These difficulties usually do not arise in actual applications; see Kraft and LeCam [59].

The assumption that the matrix D has rank r is made to ensure that there is no redundancy among the parameters $\theta_1 \ldots \theta_r$. Since $\sum_j p_{ij}(\theta) = 1$ for all θ, $\sum_j \partial p_{ij}(\theta)/\partial\theta_u = 0$ for all i and u. Thus there are s independent constraints on the rows of D, which implies that $r \leq s^2 - s$.

Theorem 4.1 can be proved by the methods of Section 30.3 of Cramér [26]. In fact, by virtue of Theorem 3.1, the random variables f_{ij} may as well (from the asymptotic point of view) have arisen from s independent samples of sizes f_i from multinomial populations $(p_{i1},\ldots,p_{is})$. Thus Theorem 4.1

reduces to the results of [26]. (Cramèr actually carries through the proof only for the case of one multinomial sample, but he indicates (and uses) the more general result.)
A somewhat simpler proof of Theorem 4.1, under the additional assumption that the $p_{ij}(\theta)$ have continuous third order partial derivatives, will be found in my monograph [18]. This proof makes use of the methods of Section 7 below.

Just as in the case in which there are no parameters to estimate, the chi-square statistic derived above can be transformed into a Neyman-Pearson criterion. As was seen in Section 3, the maximum of $\sum_{ij} f_{ij} \lg p_{ij}$, as (p_{ij}) ranges over all stochastic matrices, is $\sum_{ij} f_{ij} \lg (f_{ij}/f_i)$. And the maximum of $\sum_{ij} f_{ij} \lg p_{ij}(\theta)$, as θ ranges over Θ, is $\sum_{ij} f_{ij} \lg p_{ij}(\hat{\theta})$, (ignoring the difficulties mentioned above). Therefore, $2 \sum_{ij} f_{ij} \lg (f_{ij}/f_i\, p_{ij}(\hat{\theta}))$ is the Neyman-Pearson statistic for testing, within the hypothesis that $\{x_n\}$ is a Markov process, the smaller hypothesis that the transition probabilities are $p_{ij}(\theta)$ for some value of θ. It can be shown (see [18]) that

$$\sum_{ij} \frac{(f_{ij} - f_i p_{ij}(\hat{\theta}))^2}{f_i p_{ij}(\hat{\theta})} \sim 2 \sum_{ij} f_{ij} \lg(f_{ij}/f_i p_{ij}(\hat{\theta}))$$

if the smaller (null) hypothesis is true.

As an example, suppose one wants to test whether $p_{ij} = p_j$ is independent of i; that is, whether the Markov chain is really an independent sequence. Let $r = s-1$, let Θ consist of the set of vectors $\theta = (\theta_1, \ldots, \theta_{s-1})$ with positive components

the sum of which is less than 1, put $p_{ij}(\theta) = \theta_j$ for $j < s$ and put $p_{is}(\theta) = 1 - \sum_{i=1}^{s-1} \theta_i$. Then the conditions of the theorem can be verified and the equations (4.3) can be solved explicitly. (It is of course actually easier to maximize $\sum_{ij} f_{ij} \lg p_j$ by Lagrange multipliers.) The solution is $\theta_j = f_{\cdot j}/n$, as could have been anticipated. In this case the chi-square and Neyman-Pearson statistics become

$$\sum_{ij} \frac{(f_{ij} - f_i f_{\cdot j}/n)^2}{f_i f_{\cdot j}/n} \sim \sum_{ij} f_{ij} \lg \frac{f_{ij}}{f_i f_{\cdot j}/n} .$$

Each one has in the limit a chi-square distribution with $s(s-1) - s(s-1) - (s-1) = (s-1)^2$ degrees of freedom. This chi-square statistic was derived from Whittle's formula in Section 1.

Tests of various other hypotheses can be derived in a routine manner from Theorem 4.1. For instance, one can test the hypothesis that the process has given stationary probabilities; that is, that the transition probabilities p_{ij} satisfy $\sum_i p_i p_{ij} = p_j$, where the p_i are prescribed numbers. A number of such examples will be found in [18]. Other papers relevant in this connection are Bartlett [7], Patankar [72], and Gani [39] and [40].

The theory of this and the preceding sections can be extended to cover the case of two samples. Let $\{f_{ij}\}$ and $\{g_{ij}\}$ be the transition counts of two samples, independent of each other, from Markov chains with transition matrices (p_{ij}) and (q_{ij}). The estimates of p_{ij} and of q_{ij} are f_{ij}/f_i and g_{ij}/g_i, respectively, while if it is hypothesized that

$p_{ij} = q_{ij}$, then the common estimate is $(f_{ij} + g_{ij})/(f_i + g_i)$. It is easily shown that the chi-square statistic for testing the hypothesis that $p_{ij} = q_{ij}$ (homogeneity) is

$$\sum_{ij} \frac{\left[f_{ij} - f_i \frac{f_{ij} + g_{ij}}{f_i + g_i}\right]^2}{f_i \frac{f_{ij} + g_{ij}}{f_i + g_i}} + \sum_{ij} \frac{\left[g_{ij} + g_i \frac{f_{ij} + g_{ij}}{f_i + g_i}\right]^2}{g_i \frac{f_{ij} + g_{ij}}{f_i + g_i}}$$

$$= \sum_{ij} \frac{f_i g_i}{f_{ij} + g_{ij}} \left(\frac{f_{ij}}{f_i} - \frac{g_{ij}}{g_i}\right)^2 .$$

The asymptotic distribution has $s(s-1)$ degrees of freedom. This sort of problem has been treated by Darwin [28] and by me [16] and [18].

Results of this sort apply equally well, of course, if the number of samples is three or more. It must be assumed, however, that the number of samples is fixed, while the sample sizes go to infinity. A different theory is needed in the opposite case, that in which the samples are of fixed length (say ℓ), while the number n of them goes to infinity. In principle, the standard multinomial theory applies in this case. Suppose in fact that for $k=1,\ldots,n$, $\{x_{k1},\ldots,x_{k\ell}\}$ is a sample from a Markov chain with transition probabilities (p_{ij}). (It is possible in this case to let the transition probabilities vary from trial to trial.) The n samples together can be regarded as one independent sample of size n from a multinomial population with s^ℓ categories, the category

$(a_1,\ldots,a_\ell)$ having probability $p_{a_1}p_{a_1a_2}\cdots p_{a_{\ell-1}a_\ell}$. Various special problems arise, however. If one has only partial information, for example the frequency count of $\{x_{1i},\ldots,x_{ni}\}$ for each $i=1,\ldots,\ell$, then special methods are required. Papers on the analysis of many short samples are Miller [68], Goodman [47], Kao [56], Anderson [3], Anderson and Goodman [4], and Madansky [65].

5. PSI-SQUARE STATISTICS

The chi-square statistic (3.4) treated in Section 3 has a direct appeal as a goodness of fit criterion, quite aside from its connection with the Neyman-Pearson criterion. A statistic which at first sight perhaps seems even more natural from this point of view is

$$(5.1) \qquad \sum_{ij} \frac{(f_{ij} - np_ip_{ij})^2}{np_ip_{ij}} .$$

Aside from the fact that this statistic has no simple interpretation in terms of likelihood theory, it is not very useful because its limiting distribution is not free of the parameters (p_{ij}). If $p_{ij} = p_j$, that is, if the process is independent, then (5.1) reduces to

$$(5.2) \qquad S = \sum_{ij} \frac{(f_{ij} - np_ip_j)^2}{np_ip_j} ,$$

a so-called psi-square statistic. Although (5.2) also lacks a likelihood interpretation, at least its limiting distribution

is free of the parameters (p_i). This psi-square statistic was first used by Kendall and Smith [57], [58] as a test for serial correlation in their random number tables, but it was incorrectly assumed by them to have asymptotically a chi-square distribution. It is the purpose of this section to show that the asymptotic distribution function of (5.2) is

$$(5.3) \qquad K_{s-1}(x/2) * K_{(s-1)^2}(x),$$

where $K_d(x)$ is the chi-square distribution function for d degrees of freedom.

Let H_1 denote the hypothesis that $\{x_n\}$ is an independent process with specified probabilities $p_i = P\{x_n=i\}$; let H_2 be the hypothesis that $\{x_n\}$ is an independent, stationary process with the probabilities $P\{x_n=i\}$ unspecified; finally, let H_3 be the hypothesis that $\{x_n\}$ is a Markov process. By the results of Section 3 the statistic for testing H_1 within H_3 is

$$(5.4) \qquad \sum_{ij} f_{ij} \lg \frac{f_{ij}}{f_i p_j} \sim S_{13} = \sum_{ij} \frac{(f_{ij} - f_i p_j)^2}{f_i p_j}.$$

By Section 4, the statistic for testing H_2 within H_3 is

$$(5.5) \qquad \sum_{ij} f_{ij} \lg \frac{f_{ij}}{f_i f_j/n} \sim S_{23} = \sum_{ij} \frac{(f_{ij} - f_i f_j/n)^2}{f_i f_j/n}.$$

(Here the distinction between f_i and $f_{.i}$ has been dropped.) It is known from the ordinary multinomial theory that the statistic for testing H_1 within H_2 is

$$(5.6) \qquad \sum_i f_i \lg \frac{f_i}{np_i} \quad S_{12} = \sum_i \frac{(f_i - np_i)^2}{np_i} .$$

Since the left-hand members of (5.5) and (5.6) sum to the left-hand member of (5.4), it follows that

$$(5.7) \qquad S_{13} \sim S_{12} + S_{23}.$$

In fact, if the denominators in S_{12} and S_{13} are replaced by f_i and $f_i f_j/n$ respectively, which is legitimate (see Section 20.6 of Cramér [26]), then (5.7) becomes an equality. Since the three hypotheses stand in the relation $H_1 \subset H_2 \subset H_3$, the statistics S_{12} and S_{23} are asymptotically independent. (This phenomenon is familiar in analysis of variance; see [18] for a proof.) That the limiting distributions of S_{12}, S_{23} and S_{13}, which are respectively chi-square with $s-1, (s-1)^2$ and $s(s-1)$ degrees of freedom, convolve properly is a reflection of this fact together with (5.7).

Now S, defined by (5.2), is related to S_{12} and S_{13} by

$$(5.8) \qquad S \sim S_{12} + S_{13}.$$

This relation is proved by noting that if the denominator in S_{13} is changed to np_ip_j (use Section 20.6 of [26] again) then the two members of the relation become algebraically identical. From (5.7) and (5.8) it follows that

$$S \sim 2S_{12} + S_{23}.$$

Since S_{12} and S_{23} are asymptotically independent and chi-square with s-1 and $(s-1)^2$ degrees of freedom, it follows by an obvious generalization of the result of Section 24.5 of [26], that the limiting distribution of S is given by (5.3). This theorem was first proved for the case in which $p_i \equiv 1/s$ and s is a prime number by Good [43], and in the general case (by methods very different from the ones above) by me [15]. Various extensions are to be found in Stepanov [77]; Good [46]; Basharin [14]; Goodman [50], [51] and [52]; and in my papers [15], [16] and [18].

If L_{ij} is the Neyman-Pearson statistic for testing the hypothesis H_i (above) within H_j, then it is obvious that

$$S \sim 2L_{12} + L_{23}.$$

It is thus hard to see what interpretation is to be put on S.

6. MULTIPLE MARKOV CHAINS

Let $\{x_n\}$ be a t-th order Markov chain (as defined in Section 1) with transition probabilities

$$p_{a_1 \ldots a_t : a_{t+1}} = P\{x_n = a_{t+1} \,\|\, x_{n-t} = a_1, \ldots, x_{n-1} = a_t\},$$

assumed for simplicity to be positive. If $t > 1$, $\{x_n\}$ is called a multiple Markov chain. Problems involving multiple Markov chains are easily reduced to problems about simple ones by the following device; see p. 89 and p. 185 of Doob [32]. Consider the process $\{y_m; m=1,2,\ldots\}$, where $y_m = (x_m, x_{m+1}, \ldots, x_{m+t-1})$. Then $\{y_m\}$ is a first-order Markov

chain the state space of which consists of the s^t different t-tuples, the transition probabilities being

$$(6.1)\qquad p_{(a_1\ldots a_t)(b_1\ldots b_t)} = \begin{cases} p_{a_1\ldots a_t:b_t} & \text{if } b_i=a_{i+1},\ i=1,\ldots,t-1 \\ 0 & \text{otherwise.} \end{cases}$$

A knowledge of the first n+t steps of the original process $\{x_m\}$ is obviously equivalent to a knowledge of the first n+1 steps of the new process $\{y_m\}$. For example, let $f_{a_1\ldots a_\nu}$ be the number of m, with $1 \leq m \leq n$, such that $(x_m,\ldots,x_{m+\nu-1}) = (a_1,\ldots,a_\nu)$. Then the rôles played by the f_i and the f_{ij} in the paragraph following Theorem 3.1 are assumed here by the $f_{a_1\ldots a_t}$ and the $f_{a_1\ldots a_{t+1}}$. Clearly the s there is to be replaced by s^t here. Finally, the number of positive entries in the s^t x s^t matrix defined by (6.1) is s^{t+1}, a number which plays the role of the d of Section 3. It follows that the statistic

$$(6.2)\qquad \sum_{a_1\ldots a_{t+1}} \frac{\left(f_{a_1\ldots a_{t+1}} - f_{a_1\ldots a_t}\, p_{a_1\ldots a_t:a_{t+1}}\right)^2}{f_{a_1\ldots a_t}\, p_{a_1\ldots a_t:a_{t+1}}}$$

is asymptotically chi-square with $s^{t+1} - s^t$ degrees of freedom. As in Section 3, it can be shown that this statistic is asymptotically equivalent to the appropriate Neyman-Pearson criterion.

The results of Section 4 can be carried over so as to take into account the possibility of estimating parameters upon which the $p_{a_1\ldots a_t:a_{t+1}}$ may depend. For example, if $r < t$,

then the parameters may be so defined as to correspond to the hypothesis that $\{x_m\}$ is a Markov chain of order r. In this case the $p_{a_1 \cdots a_t : a_{t+1}}$ in (6.2) are to be replaced by

$$\hat{p}_{a_1 \cdots a_t : a_{t+1}} = f_{a_{t-r+1} \cdots a_{t+1}} / f_{a_{t-r+1} \cdots a_t}.$$

If this is done, the resulting statistic, appropriate for testing the null nypothesis that $\{x_m\}$ is an r-th order Markov chain within the hypothesis that it is of t-th order, is asymptotically chi-square with $(s^{t+1} - s^t) - (s^{r+1} - s^r)$ degrees of freedom, provided the null hypothesis is true. Papers on this subject are Bartlett [7]; Good [44]; Dawson and Good [30]; Goodman [49]; and my papers [16] and [18].

Generalized versions of the psi-square statistic (5.2) can be treated by applying the method of Section 5 to the process $\{y_m\}$ defined above. It turns out, for example, that if $\{x_m\}$ is an independent process with $P\{x_m=i\} = p_i$, then the asymptotic distribution function of

$$\sum_{a_1 \cdots a_t} \frac{\left(f_{a_1 \cdots a_t} - np_{a_1} \cdots p_{a_t}\right)^2}{np_{a_1} \cdots p_{a_t}}$$

is given by

$$\overset{t-1}{\underset{k=1}{*}} K_s t-k-1 (s-1)^2(x/k) * K_{s-1}(x/t),$$

where the first * stands for iterated convolution. If t=2, this result reduces to that of Section 5. If $\{x_m\}$ is a

first-order Markov chain then the distribution function of

$$(6.3) \qquad \sum_{a_1 \cdots a_t} \frac{\left(f_{a_1 \cdots a_t} - f_{a_1} p_{a_1 a_2} \cdots p_{a_{t-1} a_t}\right)^2}{f_{a_1} p_{a_1 a_2} \cdots p_{a_{t-1} a_t}}$$

approaches

$$\overset{t-2}{\underset{k=1}{*}} K_s t-k-1_{(s-1)^2} (x/k) * K_{s(s-1)} (x/t-1).$$

If t=2, only the final factor remains and this result becomes that of Section 3. If, however, the f_{v_1} in (6.3) is replaced by np_{v_1}, the statistic is no longer asymptotically distribution-free. In this connection, see the references given at the end of the preceding section.

7. EXTENSION TO GENERAL STATE SPACES

The problem of analyzing a simple from a first-order Markov chain was approached in Section 2 through Whittle's formula and in Sections 3 and 4 by extending the multinomial chi-square methods. There is a third possibility. Suppose the transition probabilities are functions of θ, as in Section 4, so that the log-likelihood function is

$$(7.1) \qquad L(\theta) = \sum_{ij} f_{ij} \lg p_{ij}(\theta).$$

If the regularity conditions of Theorem 4.1 are satisfied, there exists a consistent solution $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)$ of the maximum-likelihood equations

$$(7.2) \qquad \sum_{ij} f_{ij} \frac{\partial}{\partial \theta_u} \lg p_{ij}(\theta) = 0, \qquad u=1,\ldots,r.$$

It can be shown that if θ is the true value of the parameter then the random vector $\sqrt{n}\,(\hat{\theta} - \theta)$ is asymptotically normal. In fact, if $z = (z_1,\ldots,z_r)$ is the "score", that is, if

$$(7.3) \qquad z_u = \sum_{ij} f_{ij} \frac{\partial}{\partial \theta_u} \lg p_{ij}(\theta), \qquad u=1,\ldots,r,$$

then it can be shown that $z/\sqrt{n}$ converges in distribution to that normal distribution with covariance matrix $\sigma = (\sigma_{uv})$, where

$$\sigma_{uv} = \sum_{ij} p_i(\theta) p_{ij}(\theta) \left[\frac{\partial}{\partial \theta_u} \lg p_{ij}(\theta)\right] \left[\frac{\partial}{\partial \theta_v} \lg p_{ij}(\theta)\right]$$

$$= E\left\{\left[\frac{\partial}{\partial \theta_u} \lg p_{x_1 x_2}(\theta)\right] \left[\frac{\partial}{\partial \theta_v} \lg p_{x_1 x_2}(\theta)\right]\right\}.$$

Moreover,

$$(7.4) \qquad z/\sqrt{n} \sim \sigma\sqrt{n}\,(\hat{\theta} - \theta),$$

and, since σ is nonsingular, as follows from the assumption that the matrix D defined by (4.5) has rank r, the vector $\sqrt{n}\,(\hat{\theta} - \theta)$ is itself normal in the limit, with covariance matrix σ^{-1}. Finally, it can be shown that

$$(7.5) \qquad 2[L(\hat{\theta}) - L(\theta)] \sim n \sum_{uv} \sigma_{uv} (\hat{\theta}_u - \theta_u)(\hat{\theta}_v - \theta_v),$$

from which it follows that the Neyman–Pearson statistic on the left has asymptotically a chi-square distribution with r degrees

of freedom. If the $p_{ij}(\theta)$ are chosen in such a way that $(p_{ij}(\theta))$ ranges over all stochastic matrices as θ ranges over Θ, then this statistic reduces to

$$(7.6) \qquad 2 \sum_{ij} f_{ij} \lg(f_{ij}/f_i p_{ij}(\theta)).$$

Since (7.6) can be converted into the chi-square form, one has a new derivation of the result of Section 3. This method can be used to obtain all the statistics of the preceding sections.

This approach has the advantage that it admits of an extension to the case in which the state space of the process $\{x_n\}$ is no longer finite. This extension, carried through in detail in my monograph [18], will be briefly sketched here. Suppose that $\{x_n\}$ is a Markov process taking values in some general space X. The structure of the process is then specified by transition measures

$$p(\xi, A) = P\{x_{n+1} \in A \,||\, x_n = \xi\},$$

where for each $\xi \in X$, $p(\xi, \cdot)$ is a probability measure on an appropriate Borel field of subsets of X. Now suppose that these transition measures have densities with respect to another measure λ, and that these densities depend on an unknown parameter $\theta = (\theta_1, \ldots, \theta_r)$:

$$p(\xi, A) = \int_A f(\xi, \eta; \theta)\, \lambda(d\eta).$$

If X is finite and if λ is taken to be counting measure, each point of X having λ-measure 1, then the densities $f(\xi,\eta;\ \theta)$ reduce to the transition probabilities $p_{ij}(\theta)$ of the preceding sections. The cases of greatest interest other than the finite one are those in which X is countable, λ being counting measure again, and in which X is Euclidean, λ being Lebergue measure. It is important, however, to admit more general spaces, as will be seen in Section 8.

In this general situation, the log-likelihood (7.1) is to be replaced by

$$L(\theta) = \sum_{k=1}^{n} \lg\ f(x_k,x_{k+1};\ \theta),$$

the maximum-likelihood system (7.2) becomes

$$\sum_{k=1}^{n} \frac{\partial}{\partial\theta_u} \lg\ f(x_k,x_{k+1};\ \theta) = 0, \qquad u=1,\ldots,r,$$

while the "score" (7.3) becomes

$$z_u = \sum_{k=1}^{n} \frac{\partial}{\partial\theta_u} \lg\ f(x_k,x_{k+1};\ \theta). \tag{7.7}$$

It can be shown under suitable regularity conditions that there is a consistent solution $\hat{\theta}$ of the maximum-likelihood system, that $z/\sqrt{n}$ is asymptotically normal, and that (7.4) holds, where σ, the covariance matrix of the limiting distribution of $z/\sqrt{n}$, is given by

$$\sigma_{uv} = E\left\{\left[\frac{\partial}{\partial\theta_u} \lg\ f(x_1,x_2;\ \theta)\right]\left[\frac{\partial}{\partial\theta_v} \lg\ f(x_1,x_2;\ \theta\right]\right\}.$$

If it is assumed that σ is nonsingular, then (7.5) holds as well.

What are the regularity conditions which lead to these results? In the first place, it must be assumed that the densities $f(\xi,\eta;\theta)$, as functions of θ, satisfy smoothness conditions like those of Section 33.3 of Cramér [26]. In the second place, it is necessary to impose some set of conditions on the process $\{x_n\}$ which will ensure that the random variables $z_u/\sqrt{n}$ defined by (7.7) are asymptotically normal. Now while the summands in (7.7) are functions of the successive states of a Markov process, and while there exist central limit theorems for sums of such functions, there is no single theorem of this sort which covers all cases of interest. Fortunately, however, the summands in (7.7) are not just any functions of the states of the process; it can be shown that their partial sums form (for each u) a martingale. Lévy ([63] and pp. 237 ff. of [64]) has proved interesting central limit theorems for martingales; a suitable modification of his results yields the asymptotic normality of (7.7) for the case in which the summands have moments of some order greater than 2. See [18] for the details.

The sets of conditions sketched in the preceding paragraph cover many Markov processes (with stationary transition measures) which are of interest, in addition to those with finite state spaces. Suppose, for example, that $\{x_n\}$ is an autoregressive process.

$$x_n = \sum_{k=0} \alpha^k y_{n-k}, \tag{7.8}$$

where y_k is an independent sequence of identically, normally distributed random variables, the mean and the variance of the y_k, as well as α, where $|\alpha| < 1$, being unknown parameters. This process satisfies the conditions outlined above, so that the theory of this section contains the essentials of the Mann–Wald theory [66]. (The Mann–Wald theory is the intersection of time–series analysis and likelihood theory for Markov processes, in the following sense. In time series analysis, that is, in correlation and spectral theory, only wide–sense properties of the process are made use of. This reduces to likelihood theory only if the second–order moments completely determine the structure of the process; that is, if the process is Gaussian. But the most general stationary, Gaussian Markov process is given by (7.8).)

In the theory outlined above, the state space X is arbitrary, but the parameter θ is assumed to have only finitely many components. If the state space is finite then finitely many parameters suffice to describe any hypothesis on the process. In the general case, however, infinitely many parameters may be necessary; the complete structure of a Markov process on the space of integers is specified by the infinite matrix (p_{ij}),

for example. This difficulty cannot be gotten around by lumping the states into finitely many classes, since this in general destroys the Markov property. (For the problem of inference on grouped chains, see Blackwell and Koopmans [20] and Gilbert [41].) While the infinite matrix (p_{ij}) has been treated by Derman [31] (his proofs can be simplified by using the methods of Section 3), no general attack on the problem of infinitely many parameters is known to me.

For a very general approach to likelihood theory, see LeCam [62].

8. PROCESSES CONTINUOUS IN TIME

Suppose $\{x_t;\ t \geq 0\}$ is a time-continuous process, the random variables x_t taking their values in a finite set $X = \{1,2,\ldots s\}$. If

$$P\{x_{\tau+t} = j \,||\, x_u,\ u \leq \tau\} = P\{x_{\tau+t} = j \,||\, x_\tau\}, \qquad t > 0,$$

then $\{x_t\}$ is a Markov process and its probability structure is specified by the transition probabilities

$$p_{ij}(t) = P\{x_{\tau+t} = j \,||\, x_\tau = i\}, \qquad t > 0,$$

which are assumed to be independent of τ. Models in many fields of application have this structure. If the $p_{ij}(t)$ depend on an unknown parameter θ, there arises the problem of drawing statistical inferences about θ from a sample $\{x_\tau;\ 0 \leq \tau \leq t\}$ from the process.

If

$$\lim_{t \to 0} p_{ij}(t) = \delta_{ij},$$

then it can be shown that the limits

$$q_i = \lim_{t \to 0} (1 - p_{ii}(t))/t$$

$$q_{ij} = \lim_{t \to 0} p_{ij}(t)/t \qquad (i \neq j)$$

exist; see Doob [32]. The quantities q_i and q_{ij} have the following important probabilistic significance. Under a suitable regularity condition on $\{x_t\}$, namely that it is separable [32], the process starts out in some state $x_o = i$, chosen according to an initial distribution p_i; it stays in the initial state i for a length of time ρ_1, where ρ_1 is a random variable which is exponentially distributed with parameter q_i $\left(P\{\rho_1 \geq \alpha\} = e^{-q_i \alpha}\right)$; at time ρ_1 the process jumps instantaneously to a different state j, chosen according to the distribution q_{ij}/q_i $(j \neq i)$, where it stays a random length of time ρ_2 which is exponentially distributed with parameter q_j; at time $\rho_1 + \rho_2$ the process jumps to a new state k chosen according to the distribution q_{jk}/q_j $(k \neq j)$; and so on. Let $z_1, z_2, \ldots$ be the succession of distinct states the process passes through and let $\rho_1, \rho_2, \ldots$ be the lengths of time the process stays in these states. If $\nu(t)$ is the number of jumps which have occurred up to time t, that is, if

$$\nu(t) = \max\{n: \rho_1 + \ldots + \rho_n < t\}, \tag{8.1}$$

then clearly $x_t = z_{\nu(t)}$. The important point is that the process of pairs $\{(z_n, \rho_n);\ m=1,2,\ldots\}$, which may be called the imbedded process, is a time-discrete Markov process with state space $X \times (0,\infty)$ and transition measures

$$(8.2)\quad P\{z_{n+1} = j,\ \rho_{n+1} \geq \alpha \,||\, z_n = i,\ \rho_n = \beta\} = (q_{ij}/q_i)\ e^{-q_i \alpha};$$

see Doob [32]. Particular processes are usually described by specifying the q_i and the q_{ij}, rather than the $p_{ij}(t)$.

Thus the evolution of the time-continuous Markov process $\{x_t\}$ is determined by that of the time-discrete imbedded process $\{(z_n, \rho_n)\}$. If the quantities q_i and q_{ij} depend on an unknown parameter θ and if one has at hand a sample $\{(z_1, \rho_1), \ldots, (z_n, \rho_n)\}$ from the imbedded process, then it is possible to draw inferences about θ by applying the methods of the preceding section to the transition measures (8.2), which also depend on θ. However, if it is supposed that one has a sample $\{x_\tau;\ 0 \leq \tau \leq t\}$ from the original process, rather than one from the imbedded process, the situation is slightly different. In this case the sample $\{x_\tau;\ 0 \leq \tau \leq t\}$ is essentially equivalent to a sample $\{(z_1, \rho_1), \ldots, (z_{\nu(t)}, \rho_{\nu(t)})\}$ from the imbedded process, where $\nu(t)$ is the random variable defined by (8.1). (These two samples give the same information if one neglects the knowledge of what state the process is in during the time interval from $\rho_1 + \ldots + \rho_{\nu(t)}$ to t; the error committed is negligible if t is large.) Therefore a sequential version of the theory of Section 7 will enable one to perform

statistical inference on time-continuous processes with finite state space. Such a theory is developed in my monograph [18].

Even if the state space X of the process $\{x_t\}$ is finite, as has been assumed above, the state space $X \times (0,\infty)$ of the imbedded process is, while not pathological in any sense, neither discrete nor Euclidean. In order to reduce the problems of this section to those of the preceding one, it is therefore essential there not to make restrictive assumptions about the state space. In view of the generality of Section 7, one can treat, by the method of this section, time continuous processes with infinite state spaces X; see [18]. It must be assumed, however, that $\{x_t\}$ is a process of the completely discontinuous type, that is, that the sample paths are step functions; this excludes diffusion processes.

L. LeCam has pointed out (orally) that diffusion processes involve, from the point of view of statistics, an excessive amount of idealization. Suppose that x_t is a Brownian motion with $E\{x_t\} = 0$ and $E\{x_t^2\} = \theta t$. Then, no matter how small t is, the measures on the space of paths $\{x_\tau;\ 0 \leq \tau \leq t\}$ corresponding to different values of θ are mutually singular. It is therefore, in principle, possible to determine θ exactly from an observation of arbitrarily short duration, which is nonsense from the practical point of view. It should be pointed out that processes of the completely discontinuous type, while they certainly involve idealization, at least do not have this unfortunate singularity property.

Previous work on time-continuous chains has been done by Lange [61]; Fortet [36] and [37]; Hayward [54]; Benês [19]; and by Albert [2]. Papers on the estimation of the parameters of a birth-and-death process are: Anscombe [6], Moran [71] and Darwin [27]. Birth-and-death processes differ from the ones treated in the present paper in that they are either transient or absorbing. A systematic investigation of inference in such cases would be valuable.

REFERENCES

A notation (MR.u.v) refers to page v of volume u of Mathematical Reviews, where a review of the paper in question is to be found.

1. T. van Aardenne-Ehrenfest and N. G. de Bruijn, "Circuits and trees in oriented linear graphs," Simon Stevin, Vol. 28 (1951), pp. 203-217 (MR.13.857).

2. A. Albert, "Estimating the infinitesimal generator of a finite-state, continuous time Markov process," presented to the Institute of Mathematical Statistics, August 24, 1960.

3. T. W. Anderson, "Probability models for analyzing time changes in attitudes," in Mathematical Thinking in the Social Sciences, Glencoe, 1954 (MR.16.496).

4. T. W. Anderson and L. A. Goodman, "Statistical inference about Markov chains," Ann. Math. Stat., Vol. 28 (1957), pp. 89-109 (MR.18.944).

5. F. J. Anscombe, "Large sample theory of sequential estimation," Proc. Camb. Phil. Soc., Vol. 48 (1952), pp. 600-607 (MR.14.487).

6. ———, "Sequential estimation," J. Roy. Stat. Soc., Ser. B, Vol. 15 (1953), pp. 1-21 (MR.15.142).

7. M. S. Bartlett, "The frequency goodness of fit test for probability chains," Proc. Camb. Phil. Soc., Vol. 47 (1951), pp. 86-95 (MR.12.512).

8. ———, "A sampling test of the X^2 theory for probability chains," Biometrika, Vol. 39 (1952), pp. 118-121, (MR.13.962).

9. ———, An Introduction to Stochastic Processes, Cambridge, 1956 (MR.16.939).

10. ———, "The statistical analysis of stochastic processes," Colloque sur l'Analyse Statistique, Bruxelles, 1954, pp. 113-132 (MR.17.506).

11. D. E. Barton and F. N. David, "Multiple runs," Biometrika, Vol. 44 (1957), pp. 168-177, and "Corrigenda," ibid., p. 534 (MR.19.70).

12. ———, "Runs in a ring," Biometrika, Vol. 45 (1958), pp. 572-578.

13. ———, "Non-randomness in a sequence of two alternatives II. Runs test," Biometrika, Vol. 45 (1958), pp. 253-256.

14. G. P. Basharin, "The use of the chi-square criterion as a test for the independence of events," Dokl. Akad. Nank SSSR (N.S.), Vol. 117 (1957), pp. 167-170 (Russian), (MR.20.64).

15. P. Billingsley, "Asymptotic distributions of two goodness of fit criteria," Ann. Math. Stat., Vol. 27 (1956), pp. 1123-1129 (MR.18.607).

16. ———, "On testing Markov chains," presented to the Institute of Mathematical Statistics, September 10, 1957, (unpublished manuscript).

17. ———, "Hausdorff dimension in probability theory," Ill. J. of Math., Vol. 4 (1960), pp. 187-209.

18. ———, Statistical Inference for Markov Processes, Institute of Mathematical Statistics - University of Chicago Monographs in Statistics, to appear.

19. V. E. Beneŝ, "A sufficient set of statistics for a simple telephone exchange model," Bell System Tech. J., Vol. 36 (1957), pp. 939-964.

20. D. Blackwell and L. Koopmans, "On the identifiability problem for functions of finite Markov chains," Ann. Math. Stat., Vol. 28 (1957), pp. 1011-1015, (MR.20.916).

21. S. R. Broadbent, "The inspection of a Markov process," J. Roy. Stat. Soc., Ser. B, Vol. 20 (1958), pp. 111-119, (MR.20.1022).

22. V. R. Cane, "Behavior sequences as semi-Markov chains," J. Roy. Stat. Soc., Ser. B, Vol. 21 (1959), pp. 36-49.

23. W. G. Cochran, "The X^2 test of goodness of fit," Ann. Math. Stat., Vol. 23 (1952), pp. 315-345, (MR.19.1094).

24. D. R. Cox, "Some statistical methods connected with series of events," J. Roy. Stat. Soc., Ser. B, Vol. 21 (1959), pp. 129-157 (MR.19.1094).

25. ———, "The regression analysis of binary sequences," J. Roy. Stat. Soc., Ser. B, Vol. 20 (1958), pp. 215-231, (MR.20.918).

26. H. Cramér, Mathematical Methods of Statistics, Princeton, 1946.

27. J. H. Darwin, "The behaviour of an estimator for a simple birth and death process," Biometrika, Vol. 43 (1956), pp. 23–31 (MR.17.1102).

28. ———, "Note on the comparison of several realizations of a Markov chain," Biometrika, Vol. 46 (1959), pp. 412–419.

29. F. N. David, "A power function for tests of randomness in a sequence of alternatives," Biometrika, Vol. 34 (1947), pp. 335–339 (MR.9.600).

30. R. Dawson and I. J. Good, "Exact Markov probabilities from oriented linear graphs," Ann. Math. Stat., Vol. 28 (1957), pp. 946–956 (MR.20.58).

31. C. Derman, "Some asymptotic distribution theory for Markov chains with a denumerable number of states," Biometrika, Vol. 43 (1956), pp. 285–294 (MR.18.519).

32. J. L. Doob, Stochastic Processes, New York, 1953.

33. W. Feller, An Introduction to Probability Theory and Its Applications, 2nd. ed., New York, 1957.

34. D. Firescu, "Sur les fonctions d'estimation des probabilités de passage d'une chaîne de Markov," An. Univ. "C. I. Parhon" Bucuresti. Ser. Şti. Nat., Vol. 7 (1958), No. 18, pp. 9–18 (Romanian, Russian and French summaries), (MR.20.1209).

35. Robert Fortet, "Recent advances in probability," in Some Aspects of Analysis and Probability, Surveys in Applied Mathematics IV, New York, 1958, pp. 171–243, (MR.20.1017).

36. ———, "Tests et estimations pour des processus de Markov," in Colloque de Recherche Operationnelle de Bruxelles, 1958.

37. ———, "Observations discretes periodiques," Trabajos de Estadistica, Vol. 10 (1959), pp. 209–232.

38. K. R. Gabriel, "The distribution of the number of successes in a sequence of dependent trials," Biometrika, Vol. 46 (1959), pp. 454–460.

39. J. Gani, "Some theorems and sufficiency conditions for the maximum likelihood estimator of an unknown parameter in a simple Markov chain," Biometrika, Vol. 42 (1955), pp. 342–359, "Corrigendum," ibid., Vol. 43 (1956), pp. 497–498, (MR.17.640, MR.18.342).

40. J. Gani, "Sufficiency conditions in regular Markov chains and certain random walks," Biometrika, Vol. 43 (1956), pp. 276–284, (MR.18.342).

41. E. J. Gilbert, "On the identifiability problem for functions of finite Markov chains," Ann. Math. Stat., Vol. 30 (1959), pp. 688–697.

42. R. Gold, "Inference about Markov chains with nonstationary transition probabilities," Doctoral Thesis, Columbia University, 1960 (Abstract: Ann. Math. Stat., Vol. 31 (1960), p. 533.)

43. I. J. Good, "The serial test for random sampling numbers and other tests for randomness," Proc. Camb. Phil. Soc., Vol. 49_2 (1953), pp. 276–284 (MR.15.727).

44. ———, "The likelihood ratio test for Markov chains," Biometrika, Vol. 42 (1955), pp. 531–533, and "Corrigenda," ibid., Vol. 44 (1957), p. 301, (MR.17.381).

45. ———, "On the serial test for random sequences," Ann. Math. Stat., Vol. 28 (1957), pp. 262–264 (MR.19.73).

46. ———, Review of [15], Math. Rev., Vol. 18 (1957), p. 607.

47. L. A. Goodman, "A further note on 'Finite Markov processes in psychology'," Psychometrika, Vol. 18 (1953), pp. 245–248 (MR.15.333).

48. ———, "Simplified runs tests and likelihood ratio tests for Markov chains," Biometrika, Vol. 45 (1958), pp. 181–197 (MR.19.1090).

49. ———, "Exact probabilities and asymptotic relationships for some statistics from m–th order Markov chains," Ann. Math. Stat., Vol. 29 (1958), pp. 476–490 (MR.20.225).

50. ———, "Asymptotic distributions of 'psi–squared' goodness of fit criteria for m–th order Markov chains," Ann. Math. Stat., Vol. 29 (1958), pp. 1123–1133, (MR.20.1022).

51. ———, "A note on Stepanov's tests for Markov chains," Prob. th. and its appl., Vol. IV (1959), pp. 89–92 (SIAM translation), (MR.21.322).

52. ———, "On some statistical tests for m–th order Markov chains," Ann. Math. Stat., Vol. 30 (1959), pp. 154–164, (MR.21.78).

53. U. Grenander, "Stochastic processes and statistical inference," Arkiv för Math., Vol. 1 (1950), pp. 195–277, (MR.12.511).

54. W. S. Hayward, "The reliability of telephone traffic switch counts," Bell Telephone System Techn. Public., Monograph 1975.

55. P. G. Hoel, "A test for Markov Chains," Biometrika, Vol. 41 (1954), pp. 430–433 (MR.16.498).

56. R. C. W. Kao, "Note on Miller's 'Finite Markov processes in psychology'," Psychometrika, Vol. 18 (1953), pp. 24–243 (MR.15.333).

57. M. G. Kendall and B. Babington Smith, "Randomness and random sampling numbers," J. Roy. Stat. Soc., Vol. 101 (1938), pp. 147–166.

58. ———, "Second paper on random sampling numbers," Suppl. J. Roy. Stat. Soc., Vol. 6 (1939), pp. 51–61.

59. C. Kraft and L. LeCam, "A remark on the roots of the maximum likelihood equation," Ann. Math. Stat., Vol. 27 (1956), pp. 1174–1177 (MR.18.772).

60. P. V. Krishna Iyer and N. S. Shakuntala, "Cumulants of some distributions arising from a two-state Markoff chain," Proc. Camb. Phil. Soc., Vol. 55 (1959), pp. 273–276 (MR.21.725).

61. O. Lange, "Statistical investigation of parameters in Markov processes," Colloq. Math., Vol. 3 (1955), pp. 147–160 (MR.16.1039).

62. L. LeCam, "Locally asymptotically normal families of distributions," Univ. California Publ. Statist., to appear.

63. P. Lévy, "Propriétés asymplotiques des sommes de variables aléatoires enchainés," Bull. Sci. Math., Vol. 59 (1935), pp. 84–96, 109–128.

64. ———, Theorie de l'Addition des Variables Aléatoires, Paris, 1937.

65. A. Madansky, "Least squares estimation in finite Markov processes," Psychometrika, Vol. 24 (1959), pp. 137–144.

66. H. B. Mann and A. Wald, "On the treatment of linear stochastic difference equations," Econometrica, Vol. 11 (1943), pp. 173–220 (MR.5.129).

67. G. Mihoc, "Fonctions d'estimation efficients pour les suites de variables dépendantes," Bull. Math. Soc. Sci. Math. Phys. R. P. Roumaine (N.S.), Vol. 1(49), (1957), pp. 449–456 (MR.21.726).

68. George A. Miller, "Finite Markov processes in psychology," Psychometrika, Vol. 17 (1952), pp. 149–167 (MR.14.188).

69. P. G. Moore, "A test for randomness in a sequence of two alternatives involving a 2 x 2 table," Biometrika, Vol. 36 (1949), pp. 305–316 (MR.11.447).

70. ———, "A sequential test for randomness," Biometrika, Vol. 40 (1953), pp. 111–115 (MR.14.1104).

71. P. A. P. Moran, "The estimation of the parameters of a birth and death process," J. Roy. Stat. Soc. Ser. B, Vol. 15 (1953), pp. 241–245 (MR.15.545).

72. V. N. Patankar, "The goodness of fit of frequency distributions obtained from stochastic processes," Biometrika, Vol. 41 (1954), pp. 450–462 (MR.16.731).

73. ———, "A note on recurrent events," Proc. Camb. Phil. Soc., Vol. 51 (1955), pp. 96–102 (MR.16.494).

74. V. I. Romanovskii, Discrete Markov Chains, Gosudarstvennoe Izdatel'stvo Tehniko–Teoretičeskoĭ Literatury, Moscow–Leningrad, 1949, 436 pp. (in Russian), (MR.11.445).

75. N. V. Smirnov, "The statistical estimation of transition probabilities in Markov chains," Vestnik Leningradskoyo Universiteta, Vol. 1 (1955), pp. 47–48 (MR.17.757).

76. C. A. B. Smith and W. T. Tutte, "On unicursal paths in a network of degree 4," Amer. Math. Monthly, Vol. 48 (1941), pp. 233–237.

77. V. E. Stepanov, "Certain statistical criteria for Markov chains," Teor. Veroyatnost. i Primenen, Vol. 2 (1957), pp. 143–144.

78. P. Whittle, "Some distribution and moment formulae for the Markov chain," J. Roy. Stat. Soc. Ser. B, Vol. 17 (1955), pp. 235–242 (MR.17.982).

79. S. S. Wilks, "The likelihood test of independence in contingency tables," Ann. Math. Stat., Vol. 6 (1955), pp. 190–196.

The following additional references have been supplied by A. T. Bharucha-Reid.

80. B. Adhikari, "Tests d'hypothèses pour processus stochastiques," Doctoral Thesis, Paris.

81. N. T. J. Bailey, The Mathematical Theory of Epidemics, New York, 1957.

82. A. T. Bharucha-Reid, "Note on estimation of the number of states in a discrete Markov chain," Experientia, Vol. 12 (1956), p. 176.

83. ———, "On the stochastic theory of epidemics," Proc. Third Berkeley Symposium on Math. Statistics and Probability, Vol. 4 (1956), pp. 111-119 (MR.18.951).

84. ———, "Sequential decision problems for a class of stochastic processes. Testing hypotheses," (Abstract) Ann. Math. Stat., Vol. 27 (1956), pp. 217-218.

85. ———, An Introduction to the Stochastic Theory of Epidemics and Some Related Statistical Problems, Randolph AFB: School of Aviation Medicine, 1957.

86. ———, "Comparison of populations whose growth can be described by a branching stochastic process— with special reference to a problem in epidemiology," Sankhya, Vol. 19 (1958), pp. 1-14.

87. A. B. Clarke, "Maximum likelihood estimates in a simple queue," Ann. Math. Stat., Vol. 28 (1957), pp. 1036-1040.

88. R. B. Dawson, "Exact probabilities in a test for Markov dependency," (Abstract) Ann. Math. Stat., Vol. 27 (1956), p. 219.

89. A. Dvoretsky, J. Kiefer and J. Wolfowitz, "Sequential decision problems for processes with continuous parameter. Testing hypotheses," Ann. Math. Stat., Vol. 24 (1953), pp. 254-264 (MR.14.997).

90. ———, "Sequential decision problems for processes with continuous time parameter. Problems of estimation," Ann. Math. Stat., Vol. 24 (1953), pp. 403-415 (MR.15.242).

91. D. Firescu, "Fonctions d'estimation pour les probabilités fondamentales d'une châine de Markov multiple, homogène, d'ordre fini," Bull. Mathematique de la Soc. Sci. Math. Phys. de la R. P. R., Vol. 2(50), no. 4 (1958), pp. 401-410.

92. D. Firescu, "Fonctions d'estimation pour les probabilités de passage d'une châine de Markov simple, homogène, d'ordre fini," Anal. Univ. "C. I. Parhon" Bucuresti, Vol. 20 (1958).

93. ———, "Fonctions d'estimation pour les probabilités de passage inverses d'une châine de Markov simple, homogène, d'ordre fini," Anal. Univ. "C. I. Parhon" Bucuresti, Vol. 21 (1959).

94. R. Fortet, "Hypothesis testing on random elements in functional spaces," Proc. Fourth Berkeley Symposium on Math. Statistics and Probability, 1960, to appear.

95. R. Fortet and E. Mourier, "Les fonctions aléatoires comme élements aléatiores dans un espace de Banach," J. Math. Pure Appl., Vol. 38 (1959), pp. 347–364.

96. T. E. Harris, "Branching processes," Ann. Math. Stat., Vol. 19 (1948), pp. 474–494 (MR.10.311).

97. E. R. Immel, "Problems of estimation and hypothesis testing in connection with birth–and–death stochastic processes," Doctoral Thesis, University of California, Los Angeles, 1951 (Abstract: Ann. Math. Stat., Vol. 22 (1951), p. 485).

98. D. D. Joshi, "Les processus stochastiques en demographie," Publ. Inst. Statist. Univ. Paris, Vol. 3 (1954), pp. 153–177 (MR.16.731).

99. A. Kazami, "Asymptotic properties of the estimates of an unknown parameter in a stationary Markov process," Ann. Inst. Stat. Math., Tokyo, Vol. 4 (1952), pp. 1–6 (MR.14.569).

100. D. G. Kendall, "Stochastic processes and population growth," J. Roy. Stat. Soc., Ser. B, Vol. 11 (1949), pp. 230–264, (MR.11.672).

101. ———, "Les processus stochastiques de croissance en biologie," Ann. Inst. Henri Poincare, Vol. 13 (1952), pp. 43–108 (MR.15.243).

102. J. Kiefer and J. Wolfowitz, "Sequential tests of hypotheses about the mean occurrence time of a continuous parameter Poisson process," Naval Research Logistics Quarterly, Vol. 3 (1956), pp. 205–219 (MR.18.833).

103. L. H. Koopmans, "Asymptotic rate of discrimination for Markov processes," (Abstract) Ann. Math. Stat., Vol. 30 (1959), p. 622.

104. S. Luvsanceren, "Maximum likelihood estimators and confidence regions for unknown parameters of a stationary process of Markov type," Doklady Akad. Nauk SSSR (N.S.), Vol. 98 (1954), pp. 723–726 (MR.16.385).

105. P. A. P. Moran, "Estimation methods for evolutive processes," J. Roy. Stat. Soc., Ser. B, Vol. 13 (1951), pp. 141–146, (MR.13.667).

106. M. Ogawara, "On the normal stationary Markov process of higher order," Bull. Math. Statist., Vol. 2 (1946), pp. 101–119.

107. ———, "A note on the test of serial correlation coefficients," Ann. Math. Stat., Vol. 22 (1951), pp. 115–118 (MR.12.726).

108. O. Onicescu, "La repartition limite des sommes de variables aléatoires d'un processus Markov fini, homogène et continue," An. Univ. "C. I. Parhon" Bucuresti Ser. Str. Nat., Vol. 5, no. 12 (1956).

109. B. N. Singh, "Use of complex Markov's chain in testing randomness," Jour. Indian Soc. Agric. Statistics, Vol. 4 (1952), pp. 145–148 (MR.14.777).

110. A. Wald, "Asymptotic properties of the maximum likelihood estimate of an unknown parameter of a discrete stochastic process," Ann. Math. Stat., Vol. 19 (1948), pp. 40–46 (MR.9.454).