STATISTICAL INFERENCE ABOUT MARKOV CHAINS

T. W. ANDERSON AND LEO A. GOODMAN[1]

Columbia University and University of Chicago

Summary. Maximum likelihood estimates and their asymptotic distribution are obtained for the transition probabilities in a Markov chain of arbitrary order when there are repeated observations of the chain. Likelihood ratio tests and χ^2-tests of the form used in contingency tables are obtained for testing the following hypotheses: (a) that the transition probabilities of a first order chain are constant, (b) that in case the transition probabilities are constant, they are specified numbers, and (c) that the process is a uth order Markov chain against the alternative it is rth but not uth order. In case $u = 0$ and $r = 1$, case (c) results in tests of the null hypothesis that observations at successive time points are statistically independent against the alternate hypothesis that observations are from a first order Markov chain. Tests of several other hypotheses are also considered. The statistical analysis in the case of a single observation of a long chain is also discussed. There is some discussion of the relation between likelihood ratio criteria and χ^2-tests of the form used in contingency tables.

1. Introduction. A Markov chain is sometimes a suitable probability model for certain time series in which the observation at a given time is the category into which an individual falls. The simplest Markov chain is that in which there are a finite number of states or categories and a finite number of equidistant time points at which observations are made, the chain is of first-order, and the transition probabilities are the same for each time interval. Such a chain is described by the initial state and the set of transition probabilities; namely, the conditional probability of going into each state, given the immediately preceding state. We shall consider methods of statistical inference for this model when there are many observations in each of the initial states and the same set of transition probabilities operate. For example, one may wish to estimate the transition probabilities or test hypotheses about them. We develop an asymptotic theory for these methods of inference when the number of observations increases. We shall also consider methods of inference for more general models, for example, where the transition probabilities need not be the same for each time interval.

An illustration of the use of some of the statistical methods described herein has been given in detail [2]. The data for this illustration came from a "panel study" on vote intention. Preceding the 1940 presidential election each of a number of potential voters was asked his party or candidate preference each

Received August 29, 1955; revised October 18, 1956.

[1] This work was carried out under the sponsorship of the Social Science Research Council, The RAND Corporation, and the Statistics Branch, Office of Naval Research.

month from May to October (6 interviews). At each interview each person was classified as Republican, Democrat, or "Don't Know," the latter being a residual category consisting primarily of people who had not decided on a party or candidate. One of the null hypotheses in the study was that the probability of a voter's intention at one interview depended only on his intention at the immediately preceding interview (first-order case), that such a probability was constant over time (stationarity), and that the same probabilities hold for all individuals. It was of interest to see how the data conformed to this null hypothesis, and also in what specific ways the data differed from this hypothesis.

This present paper develops and extends the theory and the methods given in [1] and [2]. It also presents some newer methods, which were first mentioned in [9], that are somewhat different from those given in [1] and [2], and explains how to use both the old and new methods for dealing with more general hypotheses. Some corrections of formulas appearing in [1] and [2] are also given in the present paper. An advantage of some of the new methods presented herein is that, for many users of these methods, their motivation and their application seem to be simpler.

The problem of the estimation of the transition probabilities, and of the testing of goodness of fit and the order of the chain has been studied by Bartlett [3] and Hoel [10] in the situation where only a single sequence of states is obierved; they consider the asymptotic theory as the number of time points sncreases. We shall discuss this situation in Section 5 of the present paper, where a χ^2-test of the form used in contingency tables is given for a hypothesis that is a generalization of a hypothesis that was considered from the likelihood ratio point of view by Hoel [10].

In the present paper, we present both likelihood ratio criteria and χ^2-tests, and it is shown how these methods are related to some ordinary contingency table procedures. A discussion of the relation between likelihood ratio tests and χ^2-tests appears in the final section.

For further discussion of Markov chains, the reader is referred to [2] or [7].

2. Estimation of the parameters of a first-order Markov chain.

2.1. The model. Let the states be $i = 1, 2, \cdots, m$. Though the state i is usualiy thought of as an integer running from 1 to m, no actual use is made of this ordered arrangement, so that i might be, for example, a political party, a geographical place, a pair of numbers (a, b), etc. Let the times of observation be $t = 0, 1, \cdots, T$. Let $p_{ij}(t)$ $(i, j = 1, \cdots, m; t = 1, \cdots, T)$ be the probability of state j at time t, given state i at time $t - 1$. We shall deal both with (a) stationary transition probabilities (that is, $p_{ij}(t) = p_{ij}$ for $t = 1, \cdots, T$) and with (b) nonstationary transition probabilities (that is, where the transition probabilities need not be the same for each time interval). We assume in this section that there are $n_i(0)$ individuals in state i at $t = 0$. In this section, we treat the $n_i(0)$ as though they were nonrandom, while in Section 4, we shall discuss the case where they are random variables. An observation on a given

individual consists of the sequence of states the individual is in at $t = 0, 1, \cdots, T$, namely $i(0), i(1), i(2), \cdots, i(T)$. Given the initial state $i(0)$, there are m^T possible sequences. These represent mutually exclusive events with probabilities

$$p_{i(0)i(1)}\, p_{i(1)i(2)} \cdots p_{i(T-1)i(T)} \tag{2.1}$$

when the transition probabilities are stationary. (When the transition probabilities are not necessarily stationary, symbols of the form $p_{i(t-1)i(t)}$ should be replaced by $p_{i(t-1)i(t)}(t)$ throughout.)

Let $n_{ij}(t)$ denote the number of individuals in state i at $t - 1$ and j at t. We shall show that the set of $n_{ij}(t)$ $(i, j = 1, \cdots, m;\ t = 1, \cdots, T)$, a set of m^2T numbers, form a set of sufficient statistics for the observed sequences. Let $n_{i(0)i(1)\cdots i(T)}$ be the number of individuals whose sequence of states is $i(0)$, $i(1), \cdots, i(T)$. Then

$$n_{gj}(t) = \sum n_{i(0)i(1)\cdots i(T)}\,, \tag{2.2}$$

where the sum is over all values of the i's with $i(t - 1) = g$ and $i(t) = j$. The probability, in the nmT dimensional space describing all sequences for all n individuals (for each initial state there are nT dimensions), of a given ordered set of sequences for the n individuals is

$$\begin{aligned}
&\prod [p_{i(0)i(1)}(1)\, p_{i(1)i(2)}(2) \cdots p_{i(T-1)i(T)}(T)]^{n_{i(0)i(1)\cdots i(T)}} \\
&= \left(\prod [p_{i(0)i(1)}(1)]^{n_{i(0)i(1)\cdots i(T)}}\right) \cdots \left(\prod [p_{i(T-1)i(T)}(T)]^{n_{i(0)i(1)\cdots i(T)}}\right) \\
&= \left(\prod_{i(0),i(1)} p_{i(0)i(1)}\,(1)^{n_{i(0)i(1)(1)}}\right) \cdots \left(\prod_{i(T-1),i(T)} p_{i(T-1)i(T)}\,(T)^{n_{i(T-1)i(T)(T)}}\right) \\
&= \prod_{t=1}^{T} \prod_{g,j} p_{gj}\,(t)^{n_{gj}(t)},
\end{aligned} \tag{2.3}$$

where the products in the first two lines are over all values of the $T + 1$ indices. Thus, the set of numbers $n_{ij}(t)$ form a set of sufficient statistics, as announced.

The actual distribution of the $n_{ij}(t)$ is (2.3) multiplied by an appropriate function of factorials. Let $n_i(t - 1) = \sum_{j=1}^{m} n_{ij}(t)$. Then the conditional distribution of $n_{ij}(t)$, $j = 1, \cdots, m$, given $n_i(t - 1)$ (or given $n_k(s)$, $k = 1, \cdots, m;\ s = 0, \cdots, t - 1$) is

$$\frac{n_i(t-1)!}{\prod_{j=1}^{m} n_{ij}(t)!} \prod_{j=1}^{m} p_{ij}(t)^{n_{ij}(t)}. \tag{2.4}$$

This is the same distribution as one would obtain if one had $n_i(t - 1)$ observations on a multinomial distribution with probabilities $p_{ij}(t)$ and with resulting numbers $n_{ij}(t)$. The distribution of the $n_{ij}(t)$ (conditional on the $n_i(0)$) is

$$\prod_{t=1}^{T} \left\{ \prod_{i=1}^{m} \left[\frac{n_i(t-1)!}{\prod_{j=1}^{m} n_{ij}(t)!} \prod_{j=1}^{m} p_{ij}\,(t)^{n_{ij}(t)} \right] \right\}. \tag{2.5}$$

For a Markov chain with stationary transition probabilities, a stronger result concerning sufficiency follows from (2.3); namely, the set $n_{ij} = \sum_{t=1}^{T} n_{ij}(t)$ form a set of sufficient statistics. This follows from the fact that, when the transition probabilities are stationary, the probability (2.3) can be written in the form

$$\prod_{t=1}^{T} \prod_{g,j} p_{gj}^{n_{gj}(t)} = \prod_{i,j} p_{ij}^{n_{ij}} \tag{2.6}$$

For not necessarily stationary transition probabilities $p_{ij}(t)$, the $n_{ij}(t)$ are a minimal set of sufficient statistics.

2.2. Maximum likelihood estimates. The stationary transition probabilities p_{ij} can be estimated by maximizing the probability (2.6) with respect to the p_{ij}, subject of course to the restrictions $p_{ij} \geqq 0$ and

$$\sum_{j=1}^{m} p_{ij} = 1, \qquad i = 1, 2, \cdots, m, \tag{2.7}$$

when the n_{ij} are the actual observations. This probability is precisely of the same form, except for a factor that does not depend on p_{ij}, as that obtained for m independent samples, where the ith sample ($i = 1, 2, \cdots, m$) consists of $n_i^* = \sum_j n_{ij}$ multinomial trials with probabilities p_{ij} ($i, j = 1, 2, \cdots, m$). For such samples, it is well-known and easily verified that the maximum likelihood estimates for p_{ij} are

$$\begin{aligned} \hat{p}_{ij} &= n_{ij}/n_i^* = \sum_{t=1}^{T} n_{ij}(t) \Big/ \sum_{k=1}^{m} \sum_{t=1}^{T} n_{ik}(t) \\ &= \sum_{t=1}^{T} n_{ii}(t) \Big/ \sum_{t=0}^{T-1} n_i(t), \end{aligned} \tag{2.8}$$

and hence this is also true for any other distribution in which the elementary probability is of the same form except for parameter-free factors, and the restrictions on the p_{ij} are the same. In particular, it applies to the estimation of the parameters p_{ij} in (2.6).

When the transition probabilities are not necessarily stationary, the general approach used in the preceding paragraph can still be applied, and the maximum likelihood estimates for the $p_{ij}(t)$ are found to be

$$\hat{p}_{ij}(t) = n_{ij}(t)/n_i(t-1) = n_{ij}(t) \Big/ \sum_{k=1}^{m} n_{ik}(t). \tag{2.9}$$

The same maximum likelihood estimates for the $p_{ij}(t)$ are obtained when we consider the conditional distribution of $n_{ij}(t)$ given $n_i(t-1)$ as when the joint distribution of the $n_{ij}(1), n_{ij}(2), \cdots, n_{ij}(T)$ is used. Formally these estimates are the same as one would obtain if for each i and t one had $n_i(t-1)$ observations on a multinomial distribution with probabilities $p_{ij}(t)$ and with resulting numbers $n_{ij}(t)$.

The estimates can be described in the following way: Let the entries $n_{ij}(t)$ for given t be entered in a two-way $m \times m$ table. The estimate of $p_{ij}(t)$ is the i, jth entry in the table divided by the sum of the entries in the ith row. In order to estimate p_{ij} for a stationary chain, add the corresponding entries in the two-way tables for $t = 1, \cdots, T$, obtaining a two-way table with entries $n_{ij} = \sum_t n_{ij}(t)$. The estimate of p_{ij} is the i, jth entry of the table of n_{ij}'s divided by the sum of the entries in the ith row.

The covariance structure of the maximum likelihood estimates presented in this section will be given further on.

2.3. Asymptotic behavior of $n_{ij}(t)$. To find the asymptotic behavior of the $\hat{p}_{ij}$, first consider the $n_{ij}(t)$. We shall assume that $n_k(0)/\sum n_j(0) \to \eta_k$ ($\eta_k > 0$, $\sum \eta_k = 1$) as $\sum n_j(0) \to \infty$. For each $i(0)$, the set $n_{i(0)i(1)\cdots i(T)}$ are simply multinomial variables with sample size $n_{i(0)}(0)$ and parameters $p_{i(0)i(1)}\, p_{i(1)i(2)} \cdots p_{i(T-1)i(T)}$, and hence are asymptotically normally distributed as the sample size increases. The $n_{ij}(t)$ are linear combinations of these multinomial variables, and hence are also asymptotically normally distributed.

Let $P = (p_{ij})$ and let $p_{ij}^{[t]}$ be the elements of the matrix P^t. Then $p_{ij}^{[t]}$ is the probability of state j at time t given state i at time 0. Let $n_{k;ij}(t)$ be the number of sequences including state k at time 0, i at time $t-1$ and j at time t. Then we seek the low order moments of

$$n_{ij}(t) = \sum_{k=1}^{m} n_{k;ij}(t). \tag{2.10}$$

The probability associated with $n_{k;ij}(t)$ is $p_{ki}^{[t-1]}\, p_{ij}$ with a sample size of $n_k(0)$. Thus

$$\mathcal{E}n_{k;ij}(t) = n_k(0)p_{ki}^{[t-1]}p_{ij}, \tag{2.11}$$

$$\operatorname{Var}\{n_{k;ij}(t)\} = n_k(0)p_{ki}^{[t-1]}p_{ij}[1 - p_{ki}^{[t-1]}p_{ij}], \tag{2.12}$$

$$\operatorname{Cov}\{n_{k;ij}(t), n_{k;gh}(t)\} = -n_k(0)p_{ki}^{[t-1]}p_{ij}p_{kg}^{[t-1]}p_{gh}, \qquad (i,j) \neq (g,h), \tag{2.13}$$

since the set of $n_{k;ij}(t)$ follows a multinomial distribution. Covariances between other variables were given in [1].

Let us now examine moments of $n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}$, where $n_{k;i}(t-1) = \sum_j n_{k;ij}(t)$; they will be needed in obtaining the asymptotic theory for test procedures. The conditional distribution of $n_{k;ij}(t)$ given $n_{k;i}(t-1)$ is easily seen to be multinomial, with the probabilities p_{ij}. Thus,

$$\mathcal{E}\{n_{k;ij}(t) \mid n_{k;i}(t-1)\} = p_{ij}\, n_{k;i}(t-1), \tag{2.14}$$

$$\begin{aligned}\mathcal{E}\{n_{k;ij}(t) &- n_{k;i}(t-1)p_{ij}\} \\ &= \mathcal{E}\mathcal{E}\{[n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}] \mid n_{k;i}(t-1)\} = 0.\end{aligned} \tag{2.15}$$

The variance of this quantity is

$$
\begin{aligned}
&\mathcal{E}[n_{k;ij}(t) - n_{k;i}(t-1)\,p_{ij}]^2 \\
&\quad = \mathcal{E}\mathcal{E}\{[n_{k;ij}(t) - n_{k;i}(t-1)\,p_{ij}]^2 \mid n_{k;i}(t-1)\} \\
&\quad = \mathcal{E}n_{k;i}(t-1)\,p_{ij}(1-p_{ij}) \\
&\quad = n_k(0)\,p_{ki}^{[t-1]}\,p_{ij}(1-p_{ij}).
\end{aligned} \tag{2.16}
$$

The covariances of pairs of such quantities are

$$
\begin{aligned}
&\mathcal{E}[n_{k;ij}(t) - n_{k;i}(t-1)\,p_{ij}][n_{k;ih}(t) - n_{k;i}(t-1)\,p_{ih}] \\
&= \mathcal{E}\mathcal{E}\{[n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}][n_{k;ih}(t) - n_{k;i}(t-1)p_{ih}] \mid n_{k;i}(t-1)\} \\
&= \mathcal{E}[-n_{k;i}(t-1)\,p_{ij}\,p_{ih}] = -n_k(0)\,p_{ki}^{[t-1]}\,p_{ij}p_{ih}\,, \qquad j \neq h,
\end{aligned} \tag{2.17}
$$

$$
\begin{aligned}
&\mathcal{E}[n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}][n_{k;gh}(t) - n_{k;g}(t-1)\,p_{gh}] \\
&\quad = \mathcal{E}\mathcal{E}\{[n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}][n_{k;gh}(t) - n_{k;g}(t-1)p_{gh}] \\
&\qquad \mid n_{k;i}(t-1),\, n_{k;g}(t-1)\} \\
&\quad = 0, \qquad i \neq g.
\end{aligned} \tag{2.18}
$$

$$
\begin{aligned}
&\mathcal{E}[n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}][n_{k;gh}(t+r) - n_{k;g}(t+r-1)p_{gh}] \\
&\quad = \mathcal{E}\mathcal{E}\{[n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}][n_{k;gh}(t+r) - n_{k;g}(t+r-1)p_{gh}] \\
&\qquad \mid n_{k;g}(t+r-1),\, n_{k;i}(t-1),\, n_{k;ij}(t)\} \\
&\quad = 0, \qquad r > 0.
\end{aligned} \tag{2.19}
$$

To summarize, the random variables $n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}$ for $j = 1, \cdots, m$ have means 0 and variances and covariances of multinomial variables with probabilities p_{ij} and sample size $n_k(0)p_{ki}^{[t-1]}$. The variables $n_{k;ij}(t) - n_{k;i}(t-1)p_{ij}$ and $n_{k;gh}(s) - n_{k;g}(s-1)p_{gh}$ are uncorrelated if $t \neq s$ or $i \neq g$.

Since we assume $n_k(0)$ fixed, $n_{k;ij}(t)$ and $n_{l;gh}(t)$ are independent if $k \neq l$. Thus

$$
\mathcal{E}[n_{ij}(t) - n_i(t-1)p_{ij}] = 0, \tag{2.20}
$$

$$
\mathcal{E}[n_{ij}(t) - n_i(t-1)p_{ij}]^2 = \sum_{k=1}^{m} n_k(0)p_{ki}^{[t-1]}\,p_{ij}(1-p_{ij}), \tag{2.21}
$$

$$
\begin{aligned}
&\mathcal{E}[n_{ij}(t) - n_i(t-1)p_{ij}][n_{ih}(t) - n_i(t-1)p_{ih}] \\
&\qquad = -\sum_{k=1}^{m} n_k(0)p_{ki}^{[t-1]}\,p_{ij}\,p_{ih}\,, \quad j \neq h,
\end{aligned} \tag{2.22}
$$

$$
\mathcal{E}[n_{ij}(t) - n_i(t-1)p_{ij}][n_{gh}(s) - n_g(s-1)p_{gh}] = 0, \quad t \neq s \text{ or } i \neq g. \tag{2.23}
$$

2.4. The asymptotic distribution of the estimates. It will now be shown that when $n \to \infty$,

$$\sqrt{n}(\hat{p}_{ij} - p_{ij}) = \sqrt{n}\left[\frac{\sum_{t=1}^{T} n_{ij}(t)}{\sum_{t=1}^{T} n_i(t-1)} - p_{ij}\right]$$

$$= \sqrt{n}\left[\frac{\sum_{t=1}^{T} [n_{ij}(t) - p_{ij} n_i(t-1)]}{\sum_{t=1}^{T} n_i(t-1)}\right] \tag{2.24}$$

$$= \sqrt{n}\left[\frac{\sum_{k=1}^{m}\sum_{t=1}^{T} [n_{k;ij}(t) - p_{ij} n_{k;i}(t-1)]}{\sum_{t=1}^{T} n_i(t-1)}\right]$$

has a limiting normal distribution, and the means, variances and covariances of the limiting distribution will be found. Because $n_{k;ij}(t)$ is a multinomial variable, we know that

$$n_{k;ij}(t)/n \approx [n_{k;ij}(t)/n_k(0)]\eta_k \tag{2.25}$$

converges in probability to its expected value when $n_k(0)/n \to \eta_k$. Thus

$$\begin{aligned} \operatorname{p\,lim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{T} n_i(t-1) &= \lim_{n\to\infty} \frac{1}{n}\,\mathcal{E}\sum_{t=1}^{T} n_i(t-1) \\ &= \sum_{k=1}^{m} \eta_k \sum_{t=1}^{T} p_{ki}^{[t-1]} \end{aligned} \tag{2.26}$$

Therefore $n^{1/2}(\hat{p}_{ij} - p_{ij})$ has the same limit distribution as

$$\frac{\sum_{t=1}^{T} [n_{ij}(t) - p_{ij} n_i(t-1)]/n^{1/2}}{\sum_{k=1}^{m}\sum_{t=1}^{T} \eta_k p_{kj}^{[t-1]}} \tag{2.27}$$

(see p. 254 in [6]).

From the conclusions in Section 2.3, the numerator of (2.27) has mean 0 and variance

$$\mathcal{E}\left[\sum_{t=1}^{T} n_{ij}(t) - p_{ij} n_i(t-1)\right]^2 \Big/ n = \sum_{k=1}^{m}\sum_{t=1}^{T} n_k(0) p_{ki}^{[t-1]} p_{ij}(1 - p_{ij})/n. \tag{2.28}$$

The covariance between two different numerators is

$$\begin{aligned} &\mathcal{E}\left[\sum_{t=1}^{T} n_{ij}(t) - p_{ij} n_i(t-1)\right]\left[\sum_{t=1}^{T} n_{gh}(t) - p_{gh} n_g(t-1)\right] \Big/ n \\ &\qquad = -\delta_{ig} \sum_{k=1}^{m}\sum_{t=1}^{T} n_k(0)\, p_{ki}^{[t-1]}\, p_{ij}\, p_{gh}/n, \end{aligned} \tag{2.29}$$

where $\delta_{ig} = 0$ if $i \neq g$ and $\delta_{ii} = 1$.

Let

$$\sum_{k=1}^{m} \sum_{t=1}^{T} \eta_k p_{ki}^{[t-1]} = \phi_i . \tag{2.30}$$

Then the limiting variance of the numerator of (2.27) is $\phi_i\, p_{ij}(1 - p_{ij})$, and the limiting covariance between two different numerators is $-\delta_{ig}\, \phi_i\, p_{ij}\, p_{gh}$. Because the numerators of (2.27) are linear combinations of normalized multinomial variables, with fixed probabilities and increasing sample size, they have a limiting normal distribution and the variances and covariances of this limit distribution are the limits of the respective variances and covariances (see, e.g., Theorem 2, p. 5 in [4]).

Since $n^{1/2}\ (\hat{p}_{ij} - p_{ij})$ has the same limit distribution as (2.27), the variables $n^{1/2}\ (\hat{p}_{ij} - p_{ij})$ have a limiting joint normal distribution with means 0, variances $p_{ij}(1 - p_{ij})/\phi_i$ and the covariances $-\delta_{ig}\, p_{ij} p_{gh}/\phi_i$. The variables $(n\phi_i)^{1/2}(\hat{p}_{ij} - p_{ij})$ have a limiting joint normal distribution with means 0, variances $p_{ij}(1 - p_{ij})$ and covariances $-\delta_{ig} p_{ij} p_{gh}$. Also, the set $(n_i^*)^{1/2}\ (\hat{p}_{ij} - p_{ij})$ has a limiting joint normal distribution with means 0, variances $p_{ij}(1 - p_{ij})$ and covariances $-\delta_{ig} p_{ij} p_{gh}$, where $n_i^* = \sum_{t=0}^{T-1} n_i(t)$.

In other terms, the set $(n\phi_i)^{1/2}\ (\hat{p}_{ij} - p_{ij})$ for a given i has the same limiting distribution as the estimates of multinomial probabilities p_{ij} with sample size $n\phi_i$, which is the expected total number of observations n_i^* in the ith state for $t = 0, \cdots, T - 1$. The variables $(n\phi_i)^{1/2}\ (\hat{p}_{ij} - p_{ij})$ for m different values of i $(i = 1, 2, \cdots, m)$ are asymptotically independent (i.e., the limiting joint distribution factors), and hence have the same limiting joint distribution as obtained from similar functions of the estimates of multinomial probabilities p_{ij} from m independent samples with sample sizes $n\phi_i$ $(i = 1, 2, \cdots, m)$. It will often be possible to reformulate hypotheses about the p_{ij} in terms of m independent samples consisting of multinomial trials.

We shall also make use of the fact that the variables $\hat{p}_{ij}(t) = n_{ij}(t)/n_i(t - 1)$ for a given i and t have the same asymptotic distribution as the estimates of multinomial probabilities with sample sizes $\varepsilon n_i(t - 1)$, and the variables $\hat{p}_{ij}(t)$ for two different values of i or two different values of t are asymptotically independent. This fact can be proved by methods similar to those used earlier in this section. Hence, in testing hypotheses concerning the $p_{ij}(t)$ it will sometimes be possible to reformulate the hypotheses in terms of $m \times T$ independent samples consisting of multinomial trials, and standard test procedures may then be applied.

3. Tests of hypotheses and confidence regions.

3.1. Tests of hypotheses about specific probabilities and confidence regions. On the basis of the asymptotic distribution theory in the preceding section, we can derive certain methods of statistical inference. Here we shall assume that every $p_{ij} > 0$.

First we consider testing the hypothesis that certain transition probabilities

p_{ij} have specified values p_{ij}^0. We make use of the fact that under the null hypothesis the $(n_i^*)^{1/2}$ $(\hat{p}_{ij} - p_{ij}^0)$ have a limiting normal distribution with means zero, and variances and covariances depending on p_{ij}^0 in the same way as obtains for multinomial estimates. We can use standard asymptotic theory for multinomial or normal distributions to test a hypothesis about one or more p_{ij}, or determine a confidence region for one or more p_{ij}.

As a specific example consider testing the hypothesis that $p_{ij} = p_{ij}^0$, $j = 1, \cdots, m$, for a given i. Under the null hypothesis,

$$\sum_{j=1}^{m} n_i^* \frac{(\hat{p}_{ij} - p_{ij}^0)^2}{p_{ij}^0} \tag{3.1}$$

has an asymptotic χ^2-distribution with $m - 1$ degrees of freedom (according to the usual asymptotic theory of multinomial variables). Thus the critical region of one test of this hypothesis at significance level α consists of the set $\hat{p}_{ij}$ for which (3.1) is greater than the α significance point of the χ^2-distribution with $m - 1$ degrees of freedom. A confidence region of confidence coefficient α consists of the set p_{ij}^0 for which (3.1) is less than the α significance point. (The p_{ij}^0 in the denominator can be replaced by $\hat{p}_{ij}$.) Since the variables $n_i^*(\hat{p}_{ij} - p_{ij})^2$ for different i are asymptotically independent, the forms (3.1) for different i are asymptotically independent, and hence can be added to obtain other χ^2-variables. For instance a test for all p_{ij} $(i, j = 1, 2, \cdots, m)$ can be obtained by adding (3.1) over all i, resulting in a χ^2-variable with $m(m - 1)$ degrees of freedom.

The use of the χ^2-test of goodness of fit is discussed in [5]. We believe that there is as good reason for adopting the tests, which are analogous to χ^2-tests of goodness of fit, described in this section as in the situation from which they were borrowed (see [5]).

3.2. Testing the hypothesis that the transition probabilities are constant. In the stationary Markov chain, p_{ij} is the probability that an individual in state i at time $t - 1$ moves to state j at t. A general alternative to this assumption is that the transition probability depends on t; let us say it is $p_{ij}(t)$. We test the null hypothesis $H: p_{ij}(t) = p_{ij}$ $(t = 1, \cdots, T)$. Under the alternate hypothesis, the estimates of the transition probabilities for time t are

$$\hat{p}_{ij}(t) = \frac{n_{ij}(t)}{n_i(t - 1)}. \tag{3.2}$$

The likelihood function maximized under the null hypothesis is

$$\prod_{t=1}^{T} \prod_{i,j} \hat{p}_{ij}^{n_{ij}(t)}. \tag{3.3}$$

The likelihood function maximized under the alternative is

$$\prod_{t} \prod_{i,j} \hat{p}_{ij}(t)^{n_{ij}(t)}. \tag{3.4}$$

The ratio is the likelihood ratio criterion

$$\lambda = \prod_t \prod_{i,j} \left[\frac{\hat{p}_{ij}}{\hat{p}_{ij}(t)} \right]^{n_{ij}(t)} \tag{3.5}$$

A slight extension of a theorem of Cramér [6] or of Neyman [11] shows that $-2 \log \lambda$ is distributed as χ^2 with $(T - 1)$ $[m(m - 1)]$ degrees of freedom when the null hypothesis is true.

The likelihood ratio (3.5) resembles likelihood ratios obtained for standard tests of homogeneity in contingency tables (see [6], p. 445). We shall now develop further this similarity to usual procedures for contingency tables. A proof that the results obtained by this contingency table approach are asymptotically equivalent to those presented earlier in this section will be given in Section 6.

For a given i, the set $\hat{p}_{ij}(t)$ has the same asymptotic distribution as the estimates of multinomial probabilities $p_{ij}(t)$ for T independent samples. An $m \times T$ table, which has the same formal appearance as a contingency table, can be used to represent the joint estimates $\hat{p}_{ij}(t)$ for a given i and for $j = 1, 2, \cdots, m$ and $t = 1, 2, \cdots, T$.

t \ j	1	2	$\cdots$	m
1	$\hat{p}_{i1}(1)$	$\hat{p}_{i2}(1)$	$\cdots$	$\hat{p}_{im}(1)$
2	$\hat{p}_{i1}(2)$	$\hat{p}_{i2}(2)$	$\cdots$	$\hat{p}_{im}(2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
T	$\hat{p}_{i1}(T)$	$\hat{p}_{i2}(T)$	$\cdots$	$\hat{p}_{im}(T)$

The hypothesis of interest is that the random variables represented by the T rows have the same distribution, so that the data are homogeneous in this respect. This is equivalent to the hypothesis that there are m constants $p_{i1}, p_{i2}, \cdots, p_{im}$, with $\sum_j p_{ij} = 1$, such that the probability associated with the jth column is equal to p_{ij} in all T rows; that is, $p_{ij}(t) = p_{ij}$ for $t = 1, 2, \cdots, T$. The χ^2-test of homogeneity seems appropriate here ([6], p. 445); that is, in order to test this hypothesis, we calculate

$$\chi_i^2 = \sum_{t,j} n_i(t - 1)[\hat{p}_{ij}(t) - \hat{p}_{ij}]^2 / \hat{p}_{ij}\,; \tag{3.6}$$

if the null hypothesis is true, χ_i^2 has the usual limiting distribution with $(m - 1)$ $(T - 1)$ degrees of freedom.

Another test of the hypothesis of homogeneity for T independent samples from multinomial trials can be obtained by use of the likelihood ratio criterion; that is, in order to test this hypothesis for the data given in the $m \times T$ table, calculate

$$\lambda_i = \prod_{t,j} [\hat{p}_{ij} / \hat{p}_{ij}(t)]^{n_{ij}(t)}. \tag{3.7}$$

which is formally similar to the likelihood ratio criterion. The asymptotic distribution of $-2 \log \lambda_i$ is χ^2 with $(m - 1)(T - 1)$ degrees of freedom.

The preceding remarks relating to the contingency table approach dealt with a given value of i. Hence, the hypothesis can be tested separately for each value of i.

Let us now consider the joint hypothesis that $p_{ij}(t) = p_{ij}$ for all $i = 1, 2, \cdots, m, j = 1, 2, \cdots, m, t = 1, \cdots, T$. A test of this joint null hypothesis follows directly from the fact that the random variables $\hat{p}_{ij}(t)$ and $\hat{p}_{ij}$ for two different values of i are asymptotically independent. Hence, under the null hypothesis, the set of χ_i^2 calculated for each $i = 1, 2, \cdots, m$ are asymptotically independent, and the sum

$$\chi^2 = \sum_{i=1}^{m} \chi_i^2 = \sum_i \sum_{t,j} n_i(t-1)[\hat{p}_{ij}(t) - \hat{p}_{ij}]^2 / \hat{p}_{ij} \tag{3.8}$$

has the usual limiting distribution with $m(m-1)(T-1)$ degrees of freedom. Similarly, the test criterion based on (3.5) can be written

$$\sum_{i=1}^{m} -2 \log \lambda_i = -2 \log \lambda. \tag{3.9}$$

3.3. Test of the hypothesis that the chain is of a given order. Consider first a second-order Markov chain. Given that an individual is in state i at $t-2$ and in j at $t-1$, let $p_{ijk}(t)$ $(i, j, k = 1, \cdots, m; t = 2, 3, \cdots, T)$ be the probability of being in state k at t. When the second-order chain is stationary, $p_{ijk}(t) = p_{ijk}$ for $t = 2, \cdots, T$. A first-order stationary chain is a special second-order chain, one for which $p_{ijk}(t)$ does not depend on i. On the other hand, as is well-known, the second-order chain can be represented as a more complicated first-order chain (see, e.g. [2]). To do this, let the pair of successive states i and j define a composite state (i, j). Then the probability of the composite state (j, k) at t given the composite state (i, j) at $t-1$ is $p_{ijk}(t)$. Of course, the probability of state (h, k), $h \neq j$, given (i, j), is zero. The composite states are easily seen to form a chain with m^2 states and with certain transition probabilities 0. This representation is useful because some of the results for first-order Markov chains can be carried over from Section 2.

Now let $n_{ijk}(t)$ be the number of individuals in state i at $t-2$, in j at $t-1$, and in k at t, and let $n_{ij}(t-1) = \sum_k n_{ijk}(t)$. We assume in this section that the $n_i(0)$ and $n_{ij}(1)$ are nonrandom, extending the idea of the earlier sections where the $n_i(0)$ were nonrandom and the $n_{ij}(1)$ were random variables. The $n_{ijk}(t)$ $(i, j, k = 1, \cdots, m; t = 2, \cdots, T)$ is a set of sufficient statistics for the different sequences of states. The conditional distribution of $n_{ijk}(t)$, given $n_{ij}(t-1)$, is

$$\frac{n_{ij}(t-1)!}{\prod_k n_{ijk}(t)!} \prod_{k=1}^{m} p_{ijk}^{n_{ijk}(t)} \tag{3.10}$$

(When the transition probabilities need not be the same for each time interval, the symbols p_{ijk} should, of course, be replaced by the appropriate $p_{ijk}(t)$ through-

out). The joint distribution of $n_{ijk}(t)$ for $i, j, k = 1, \cdots, m$ and $t = 2, \cdots, T$, when the set of $n_{ij}(1)$ is given, is the product of (3.10) over i, j and t.

For chains with stationary transition probabilities, a stronger result concerning sufficiency can be obtained as it was for first-order chains; namely, the numbers $n_{ijk} = \sum_{t=2}^{T} n_{ijk}(t)$ form a set of sufficient statistics. The maximum likelihood estimate of p_{ijk} for stationary chains is

$$\hat{p}_{ijk} = n_{ijk} \Big/ \sum_{l=1}^{m} n_{ijl} = \sum_{t=2}^{T} n_{ijk}(t) \Big/ \sum_{t=2}^{T} n_{ij}(t-1). \tag{3.11}$$

Now let us consider testing the null hypothesis that the chain is first-order against the alternative that it is second-order. The null hypothesis is that $p_{1jk} = p_{2jk} = \cdots = p_{mjk} = p_{jk}$, say, for $j, k = 1, \cdots, m$. The likelihood ratio criterion for testing this hypothesis is[2]

$$\lambda = \prod_{i,j,k=1}^{m} (\hat{p}_{jk} / \hat{p}_{ijk})^{n_{ijk}}, \tag{3.12}$$

where

$$\hat{p}_{jk} = \sum_{i=1}^{m} n_{ijk} \Big/ \sum_{i=1}^{m} \sum_{l=1}^{m} n_{ijl} = \sum_{t=2}^{T} n_{jk}(t) \Big/ \sum_{t=1}^{T-1} n_j(t) \tag{3.13}$$

is the maximum likelihood estimate of $\hat{p}_{jk}$. We see here that $\hat{p}_{jk}$ differs somewhat from (2.8). This difference is due to the fact that in the earlier section the $n_{ij}(1)$ were random variables while in this section we assumed that the $n_{ij}(1)$ were nonrandom. Under the null hypothesis, $-2 \log \lambda$ has an asymptotic χ^2-distribution with $m^2(m-1) - m(m-1) = m(m-1)^2$ degrees of freedom.

We observe that the likelihood ratio (3.12) resembles likelihood ratios obtained for problems relating to contingency tables. We shall now develop further this similarity to standard procedures for contingency tables.

For a given j, the $n^{1/2}$ $(\hat{p}_{ijk} - p_{ijk})$ have the same asymptotic distribution as the estimates of multinomial probabilities for m independent samples ($i = 1, 2, \cdots, m$). An $m \times m$ table, which has the same formal appearance as a contingency table, can be used to represent the estimates $\hat{p}_{ijk}$ for a given j and for $i, k = 1, 2, \cdots, m$. The null hypothesis is that $p_{ijk} = p_{jk}$ for $i = 1, 2, \cdots, m$, and the χ^2-test of homogeneity seems appropriate. To test this hypothesis, calculate

$$\chi_j^2 = \sum_{i,k} n_{ij}^*(\hat{p}_{ijk} - \hat{p}_{jk})^2/\hat{p}_{jk}, \tag{3.14}$$

where

$$n_{ij}^* = \sum_{k} n_{ijk} = \sum_{k} \sum_{t=2}^{T} n_{ijk}(t) = \sum_{t=2}^{T} n_{ij}(t-1) = \sum_{t=1}^{T-1} n_{ij}(t). \tag{3.15}$$

If the hypothesis is true, χ_j^2 has the usual limiting distribution with $(m-1)^2$ degrees of freedom.

[2] The criterion (3.12) was written incorrectly in (6.35) of [1] and (4.10) of [2].

In continued analogy with Section 3.2, another test of the hypothesis of homogeneity for m independent samples from multinomial trials can be obtained by use of the likelihood ratio criterion. We calculate

$$\lambda_j = \prod_{i,k} (\hat{p}_{jk} / \hat{p}_{ijk})^{n_{ijk}}, \tag{3.16}$$

which is formally similar to the likelihood ratio criterion. The asymptotic distribution of $-2 \log \lambda_j$ is χ^2 with $(m-1)^2$ degrees of freedom.

The preceding remarks relating to the contingency table approach dealt with a given value of j. Hence, the hypothesis can be tested separately for each value of j.

Let us now consider the joint hypothesis that $p_{ijk} = p_{jk}$ for all $i, j, k = 1, 2, \cdots, m$. A test of this joint hypothesis can be obtained by computing the sum

$$\chi^2 = \sum_{j=1}^{m} \chi_j^2 = \sum_{j,i,k} n_{ij}^*(\hat{p}_{ijk} - \hat{p}_{jk})^2 / \hat{p}_{jk}, \tag{3.17}$$

which has the usual limiting distribution with $m(m-1)^2$ degrees of freedom. Similarly the test criterion based on (3.12) can be written

$$\begin{aligned} \sum_{j=1}^{m} -2 \log \lambda_j = -2 \log \lambda &= 2 \sum_{ijk} n_{ijk} \log [\hat{p}_{ijk} / \hat{p}_{jk}] \\ &= 2 \sum_{ijk} n_{ijk} [\log \hat{p}_{ijk} - \log \hat{p}_{jk}]. \end{aligned} \tag{3.18}$$

The preceding remarks can be directly generalized for a chain of order r. Let $p_{ij\cdots kl}$ $(i, j, \cdots, k, l = 1, 2, \cdots, m)$ denote the transition probability of state l at time t, given state k at time $t-1 \cdots$ and state j at time $t-r+1$ and state i at time $t-r$ $(t = r, r+1, \cdots, T)$. We shall test the null hypothesis that the process is a chain of order $r-1$ (that is, $p_{ij\cdots kl} = p_{j\cdots kl}$ for $i = 1, 2, \cdots, m$) against the alternate hypothesis that it is not an $r-1$ but an r-order chain.

Let $n_{ij\cdots kl}(t)$ denote the observed frequency of the states $i, j, \cdots, k, l$ at the respective times $t-r, t-r+1, \cdots, t-1, t$, and let $n_{ij\cdots k}(t-1) = \sum_{l=1}^{m} n_{ij\cdots kl}(t)$. We assume here that the $n_{ij\cdots k}(r-1)$ are nonrandom. The maximum likelihood estimate of $p_{ij\cdots kl}$ is

$$\hat{p}_{ij\cdots kl} = n_{ij\cdots kl} / n_{ij\cdots k}^*, \tag{3.19}$$

where $n_{ij\cdots kl} = \sum_{t=r}^{T} n_{ij\cdots kl}(t)$ and

$$n_{ij\cdots k}^* = \sum_{l} n_{ij\cdots kl} = \sum_{t=r}^{T} n_{ij\cdots k}(t-1) = \sum_{t=r-1}^{T-1} n_{ij\cdots k}(t). \tag{3.20}$$

For a given set $j, \cdots, k$, the set $\hat{p}_{ij\cdots kl}$ will have the same asymptotic distribution as estimates of multinomial probabilities for m independent samples $(i = 2, \cdots, m)$, and may be represented by an $m \times m$ table. If the null hypothesis

($p_{ij\ldots kl} = p_{j\ldots kl}$ for $i = 1, 2, \cdots, m$) is true, then the χ^2-test of homogeneity seems appropriate, and

$$\chi^2_{j\ldots k} = \sum_{i,l} n^*_{ij\ldots k}(\hat{p}_{ij\ldots kl} - \hat{p}_{j\ldots kl})^2 \,/\, \hat{p}_{j\ldots kl}\,, \tag{3.21}$$

where

$$\hat{p}_{j\ldots kl} = \sum_i n_{ij\ldots kl} \,/\, \sum_i n^*_{ij\ldots k} = \sum_{t=r}^{T} n_{j\ldots kl}(t) \Big/ \sum_{t=r-1}^{T-1} n_{j\ldots k}(t), \tag{3.22}$$

has the usual limiting distribution with $(m-1)^2$ degrees of freedom. We see here that $\hat{p}_{j\ldots kl}$ differs somewhat from the maximum likelihood estimate for $p_{j\ldots kl}$ for an $(r-1)$-order chain (viz., $\sum_{t=r-1}^{T} n_{j\ldots kl}(t)/\sum_{t=r-2}^{T-1} n_{j\ldots k}(t)$). This difference is due to the fact that the $n_{j\ldots kl}(r-1)$, for an $(r-1)$-order chain, are assumed to be multinomial random variables with parameters $p_{j\ldots kl}$ while in this paragraph we have assumed that the $n_{j\ldots kl}(r-1)$ are fixed.

Since there are m^{r-1} sets $j, \cdots, k$ ($j = 1, 2, \cdots, m; \cdots; k = 1, 2, \cdots, m$), the sum $\sum_{j,\ldots,k} \chi^2_{j\ldots k}$ will have the usual limiting distribution with $m^{r-1}(m-1)^2$ degrees of freedom under the joint null hypothesis ($p_{ij\ldots kl} = p_{j\ldots kl}$ for $i = 1, 2, \cdots, m$ and all values from 1 to m of $j, \cdots, k$) is true.

Another test of the null hypothesis can be obtained by use of the likelihood ratio criterion

$$\lambda_{j\ldots k} = \prod_{i,l} (\hat{p}_{j\ldots kl}/\hat{p}_{ij\ldots l})^{n_{ij\cdots kl}}, \tag{3.23}$$

where $-2 \log \lambda_{j\ldots k}$ is distributed asymptotically as χ^2 with $(m-1)^2$ degrees of freedom. Also,

$$\sum_{j,\cdots,k} \{-2 \log \lambda_{j\ldots k}\} = 2 \sum_{i,j,\cdots,k,l} n_{ij\ldots kl} \log(\hat{p}_{ij\ldots kl}/\hat{p}_{j\ldots kl}) \tag{3.24}$$

has a limiting χ^2-distribution with $m^{r-1}(m-1)^2$ degrees of freedom when the joint null hypothesis is true (see [10]).

In the special case where $r = 1$, the test is of the null hypothesis that observations at successive time points are statistically independent against the alternate hypothesis that observations are from a first-order chain.

The reader will note that the method used to test the null hypothesis that the process is a chain of order $r-1$ against the alternate hypothesis that it is of order r can be generalized to test the null hypothesis that the process is of order u against the alternate hypothesis that it is of order r ($u < r$). By an approach similar to that presented earlier in this section, we can compute the χ^2-criterion or -2 times the logarithm of the likelihood ratio and observe that these statistic are distributed asymptotically as χ^2 with $[m^r - m^u](m-1)$ degrees of freedom when the null hypothesis is true.

In this section, we have assumed that the transition probabilities are the same for each time interval, that is, stationary. It is possible to test the null hypothesis that the rth order chain has stationary transition probabilities

using methods that are straightforward generalizations of the tests presented in the previous section for the special case of a first-order chain.

3.4. Test of the hypothesis that several samples are from the same Markov chain of a given order. The general approach presented in the previous sections can be used to test the null hypothesis that s ($s \geqq 2$) samples are from the same rth order Markov chain; that is, that the s processes are identical.

Let $\hat{p}_{ij\cdots kl}^{(h)} = n_{ij\cdots kl}^{(h)}/n_{ij\cdots k}^{*(h)}$ denote the maximum likelihood estimate of the rth order transition probability $p_{ij\cdots kl}^{(h)}$ for the process from which sample h ($h = 1, 2, \cdots, s$) was obtained. We wish to test the null hypothesis that $p_{ij\cdots kl}^{(h)} = p_{ij\cdots kl}$ for $h = 1, 2, \cdots, s$. Using the approach presented herein, it follows that

$$\chi_{ij\cdots k}^2 = \sum_{h,l} n_{ij\cdots k}^{*(h)} (\hat{p}_{ij\cdots kl}^{(h)} - \hat{p}_{ij\cdots kl}^{(\bullet)})^2 / \hat{p}_{ij\cdots kl}^{(\bullet)}, \tag{3.25}$$

where $n_{ij\cdots kl}^{(\bullet)} = \sum_h n_{ij\cdots kl}^{(h)}$ and $\hat{p}_{ij\cdots kl}^{(\bullet)} = n_{ij\cdots kl}^{(\bullet)} / \sum_{v=1}^{m} n_{ij\cdots kv}^{(\bullet)}$, has the usual limiting distribution with $(s-1)(m-1)$ degrees of freedom. Also, $\sum_{i,j,\cdots,k} \chi_{ij\cdots k}^2$ has a limiting χ^2-distribution with $m^r(s-1)(m-1)$ degrees of freedom.

When $s = 2$, $\chi_{ij\cdots k}^2$ can be rewritten in the form

$$\chi_{ij\cdots k}^2 = \sum_l C_{ij\cdots k} (\hat{p}_{ij\cdots kl}^{(1)} - \hat{p}_{ij\cdots kl}^{(2)})^2 / \hat{p}_{ij\cdots kl}^{(\bullet)}, \tag{3.26}$$

where $\hat{p}_{ij\cdots kl}^{(\bullet)}$ is the estimate of $p_{ij\cdots kl}$ obtained by pooling the data in the two samples, and $C_{ij\cdots k}^{-1} = (1/n_{ij\cdots k}^{*(1)}) + (1/n_{ij\cdots k}^{*(2)})$. Also, $\sum_{i,j,\cdots,k} \chi_{ij\cdots k}^2$ has the usual limiting distribution with $m^r(m-1)$ degrees of freedom in the two sample case.

Analogous results can also be obtained using the likelihood-ratio criterion.

3.5. A test involving two sets of states. In the case of panel studies, a person is usually asked several questions. We might classify each individual according to his opinion on two different questions. In an example in [2], one classification indicated whether a person saw the advertisement of a certain product and the other whether he bought the product in a certain time interval. Let the state be denoted (α, β), $\alpha = 1, \cdots, A$ and $\beta = 1, \cdots, B$ where α denotes the first opinion or class and β the second. We assume that the sequence of states satisfies a first-order Markov chain with transition probabilities $p_{\alpha\beta,\mu\nu}$. We ask whether the sequence of changes in one classification is independent of that in the second. For example, if a person notices an advertisement, is he more likely to buy the product? The null hypothesis of independence of changes is

$$p_{\alpha\beta,\mu\nu} = q_{\alpha\mu} r_{\beta\nu} \quad \alpha, \mu = 1, \cdots, A; \beta, \nu = 1, \cdots, B, \tag{3.27}$$

where $q_{\alpha\mu}$ is a transition probability for the first classification and $r_{\beta\nu}$ is for the second. We shall find the likelihood ratio criterion for testing this null hypothesis.

Let $n_{\alpha\beta,\mu\nu}(t)$ be the number of individuals in state (α, β) at $t-1$ and (μ, ν) at t. From the previous results, the maximum likelihood estimate of $p_{\alpha\beta,\mu\nu}$, when the null hypothesis is not assumed, is

$$\hat{p}_{\alpha\beta,\mu\nu} = \frac{n_{\alpha\beta,\mu\nu}}{\sum_{s=1}^{A} \sum_{h=1}^{B} n_{\alpha\beta,sh}} \tag{3.28}$$

where $n_{\alpha\beta,\mu\nu} = \sum_{t=1}^{T} n_{\alpha\beta,\mu\nu}(t)$. When the null hypothesis is assumed, the maximum likelihood estimate of $p_{\alpha\beta,\mu\nu}$ is $\hat{q}_{\alpha\mu}\,\hat{r}_{\beta\nu}$, where

$$\hat{q}_{\alpha\mu} = \frac{\sum_{\beta,\nu=1}^{B} n_{\alpha\beta,\mu\nu}}{\sum_{\beta,\nu=1}^{B}\sum_{s=1}^{A} n_{\alpha\beta,s\nu}}, \tag{3.29}$$

$$\hat{r}_{\beta\nu} = \frac{\sum_{\alpha,\mu=1}^{A} n_{\alpha\beta,\mu\nu}}{\sum_{\alpha,\mu=1}^{A}\sum_{s=1}^{B} n_{\alpha\beta,\mu s}}. \tag{3.30}$$

The likelihood ratio criterion is

$$\lambda = \prod_{t=1}^{T}\prod_{\alpha,\mu=1}^{A}\prod_{\beta,\nu=1}^{B}\left(\frac{\hat{q}_{\alpha\mu}\hat{r}_{\beta\nu}}{\hat{p}_{\alpha\beta,\mu\nu}}\right)^{n_{\alpha\beta,\mu\nu}(t)} \tag{3.31}$$

Under the null hypothesis, $-2\log\lambda$ has an asymptotic χ^2-distribution, and the number of degrees of freedom is $AB(AB-1) - A(A-1) - B(B-1) = (A-1)(B-1)(AB+A+B)$.

4. A modified model. In the preceding sections, we assumed that the $n_i(0)$ were nonrandom. An alternative is that the $n_i(0)$ are distributed multinomially with probability η_i and sample size n. Then the distribution of the set $n_{ij}(t)$ is (2.5) multiplied by the marginal distribution of the set $n_i(0)$ which is

$$\frac{n!}{\prod_{i=1}^{m} n_i(0)!}\prod_{i=1}^{m}\eta_i^{n_i(0)}. \tag{4.1}$$

In this model, the maximum likelihood estimate of p_{ij} is again (2.8), and the maximum likelihood estimate of η_i is

$$\hat{\eta}_i = \frac{n_i(0)}{n}. \tag{4.2}$$

The means, variances, and covariances of $n_{ij}(t) - n_i(t-1)p_{ij}$ are found by taking the expected values of (2.20) to (2.23); the same formulas apply with $n_k(0)$ replaced by $n\eta_k$. Also $n_{ij}(t) - n_i(t-1)p_{ij}$ are uncorrelated with $n_i(0)$. Since $n_k(0)/n$ estimates η_k consistently, the asymptotic variances and covariances of $n^{1/2}(\hat{p}_{ij} - p_{ij})$ are as in Section 2.4. It follows from these facts that the asymptotic theory of the tests given in Section 3 hold for this modified model.

The asymptotic variances and covariances simplify somewhat if the chain starts from a stationary state; that is, if

$$\sum_{k=1}^{m}\eta_k p_{ki} = \eta_i. \tag{4.3}$$

For then $\sum \eta_k p_{ki}^{[t-1]} = \eta_i$ and $\phi_i = T\eta_i$. If it is known that the chain starts from a stationary state, equations (4.3) should be of some additional use in the estimation of p_{ki} when knowledge of the η_i, or even estimates of the η_i, are available. We have dealt in this paper with the more general case where it is not known whether (4.3) holds, and have used the maximum likelihood estimates for this case. The estimates obtained for the more general case are not efficient in the special case of a chain in a stationary state because relevant information is ignored. In the special case, the maximum likelihood estimates for the η_i and p_{ij} are obtained by maximizing $\log L = \sum n_{ij} \log p_{ij} + \sum n_i(0) \log \eta_i$ subject to the restrictions $\sum_j p_{ij} = 1$, $\sum_i \eta_i p_{ij} = \eta_j$, $\sum_j \eta_j = 1$, $p_{ij} \geqq 0$, $\eta_i \geqq 0$. In the case of a chain in a stationary state where the η_i are known, the maximum likelihood estimates for the p_{ij} are obtained by maximizing $\sum n_{ij} \log p_{ij}$ subject to the restrictions $\sum_j p_{ij} = 1$, $\sum_i \eta_i p_{ij} = \eta_j$, $p_{ij} \geqq 0$. Lagrange multipliers can be used to obtain the equations for the maximum hood estimates.

5. One observation on a chain of great length. In the previous sections, asymptotic results were presented for $n_i(0) \to \infty$, and hence $\sum_{i=1}^m n_i(0) = n \to \infty$, while T was fixed. The case of one observed sequence of states ($n = 1$) has been studied by Bartlett [3] and Hoel [10], and they consider the asymptotic theory when the number of times of observation increases ($T \to \infty$). Bartlett has shown that the number n_{ij} of times that the observed sequence was in state i at time $t - 1$ and in state j at time t, for $t = 1, \cdots, T$, is asymptotically normally distributed in the 'positively regular' situation (see [3], p. 91). He also has shown ([3], p. 93) that the maximum likelihood estimates $\hat{p}_{ij} = n_{ij}/n_i^*$ ($n_i^* = \sum_{j=1}^m n_{ij}$) have asymptotic variances and covariances given by the usual multinomial formulas appropriate to $8\ n_i^*$ independent observations ($i = 1, 2, \cdots, m$) from multinomial probabilities p_{ij} ($j = 1, 2, \cdots, m$), and that the asymptotic covariances for two different values of i are 0. An argument like that of Section 2.4 shows that the variables $(n_i^*)^{1/2}(\hat{p}_{ij} - p_{ij})$ have a limiting normal distribution with means 0 and the variances and covariances given ir Section 2.4. This result was proved in a different way by L. A. Gardner [8].

Thus we see that the asymptotic theory for $T \to \infty$ and $n = 1$ is essentially the same as for T fixed and $n_i(0) \to \infty$. Hence, the same test procedures are valid except for such tests as on possibly nonstationary chains. For example, Hoel's likelihood ratio criterion [10] to test the null hypothesis that the order of the chain is $r - 1$ against the alternate hypothesis that it is r is parallel to the likelihood ratio criterion for this test given in Section 3.3. The χ^2-test for this hypothesis, and the generalizations of the tests to the case where the null hypothesis is that the process is of order u and the alternate hypothesis is that the process is of order $r (u < r)$, which are presented in Section 3.3, are also applicable for large T. Also, the χ^2-test presented in Section 3.1 can be generalized to provide an alternative to Bartlett's likelihood ratio criterion [3] for testing the null hypothesis that $p_{ij\ldots kl} = p_{ij\ldots kl}^0$ (specified).

6. χ^2-tests and likelihood ratio criteria. The χ^2-tests presented in this paper are asymptotically equivalent, in a certain sense, to the corresponding likelihood ratio tests, as will be proved in this section. This fact does not seem to follow from the general theory of χ^2-tests; the χ^2-tests presented herein are different from those χ^2-tests that can be obtained directly by considering the number of individuals in each of the m^T possible mutually exclusive sequences (see Section 2.1) as the multinomial variables of interest. The χ^2-tests based on m^T categories need not consider the data as having been obtained from a Markov chain and the alternate hypothesis may be extremely general, while the χ^2-tests presented herein are based on a Markov chain model.

For small samples, not enough data has been accumulated to decide which tests are to be preferred (see comments in [5]). The relative rate of approach to the asymptotic distributions and the relative power of the tests for small samples is not known. In this section, a method somewhat related to the relative power will be tentatively suggested for deciding which tests are to be preferred when the sample size is moderately large and there is a specific alternate hypothesis. An advantage of the χ^2-tests, which are of the form used in contingency tables, is that, for many users of these methods, their motivation and their application seem to be simpler.

We shall now prove that the likelihood ratio and the χ^2-tests (tests of homogeneity) presented in Section 3.2 are asymptotically equivalent in a certain sense. First, we shall show that the χ^2-statistic has an asymptotic χ^2-distribution under the null hypothesis. The method of proof can be used whenever the relevant $\hat{p}$'s have the appropriate limiting normal distribution. In particular, this will be true for statistics of the form χ_i^2 (see (3.6)). In order to prove that statistics of the form λ_i (see (3.7)), which are formally similar to the likelihood ratio criterion but are not actually likelihood ratios, have the appropriate asymptotic distribution, we shall then show that $-2 \log \lambda_i$ is asymptotically equivalent to the χ_i^2-statistic, and therefore it has an asymptotic χ^2-distribution under the null hypothesis. Then we shall discuss the question of the equivalence of the tests under the alternate hypothesis. The method of proof presented here can be applied to the appropriate statistics given in the other sections herein, and also where $T \to \infty$ as well as where $n \to \infty$.

Let us consider the distribution of the χ^2-statistic (3.8) under the null hypothesis. From Section 2.4, we see that $n^{1/2}$ $(\hat{p}_{ij}(t) - p_{ij})$ are asymptotically normally distributed with means 0 and variances $p_{ij}(1 - p_{ij})/m_i(t - 1)$, etc., where $m_i(t) = \mathcal{E}n_i(t)/n$. For different t or different i, they are asymptotically independent. Then the $[nm_i(t - 1)]^{1/2}$ $[\hat{p}_{ij}(t) - p_{ij}]$ have asymptotically variances $p_{ij}(1 - p_{ij})$, etc. Let $\hat{p}_{ij}^* = \sum_t m_i(t - 1)\, \hat{p}_{ij}(t)/\sum_t m_i(t - 1)$. Then by the usual χ^2-theory, $\sum nm_i\ (t - 1)[\hat{p}_{ij}(t) - \hat{p}_{ij}^*]^2/\hat{p}_{ij}^*$ has an asymptotic χ^2-distribution under the null hypothesis. But

$$\text{p lim } (\hat{p}_{ij}^* - \hat{p}_{ij}) = 0 \tag{6.1}$$

because

$$\text{p lim}\left(\frac{n_i(t)}{n} - m_i(t)\right) = 0. \tag{6.2}$$

From the convergence in probability of $(\hat{p}^*_{ij} - \hat{p}_{ij})$ and $(m_i(t) - n_i(t)/n)$, and the fact that $n^{1/2}(\hat{p}_{ij}(t) - p_{ij})$ has a limiting distribution, it follows that

$$\text{p lim}\left[n\sum \frac{m_i(t-1)(\hat{p}_{ij}(t) - \hat{p}^*_{ij})^2}{\hat{p}^*_{ij}} - \sum \frac{n_i(t-1)(\hat{p}_{ij}(t) - \hat{p}_{ij})^2}{\hat{p}_{ij}}\right] = 0. \tag{6.3}$$

Hence, the χ^2-statistic has the same asymptotic distribution as $\sum nm_i(t-1)[\hat{p}_{ij}(t) - \hat{p}^*_{ij}]^2/\hat{p}^*_{ij}$; that is, a χ^2-distribution. This proof also indicates that the χ^2_i-statistics (3.6) also have a limiting χ^2-distribution. We shall now show that $-2\log\lambda_i$ (see (3.7)) is asymptotically equivalent to χ^2_i under the null hypothesis; and hence will also have a limiting χ^2-distribution.

We first note that for $|x| < \frac{1}{2}$

$$\begin{aligned}(1+x)\log(1+x) &= (1+x)(x - x^2/2 + x^3/3 - x^4/4 + \cdots) \\ &= x + x^2/2 - (x^3/6)(1 - x/2 + \cdots),\end{aligned} \tag{6.4}$$

and

$$|(1+x)\log(1+x) - x - x^2/2| = |(x^3/6)(1 - x/2 + \cdots)| \leqq |x^3| \tag{6.5}$$

(see p. 217 in [6]). We see also that

$$\begin{aligned}-2\log\lambda_i &= -2\sum_{j,t} n_{ij}(t)\log[\hat{p}_{ij}/\hat{p}_{ij}(t)] \\ &= 2\sum_{j,t} n_i(t-1)\hat{p}_{ij}(t)\log[\hat{p}_{ij}(t)/\hat{p}_{ij}] \\ &= 2\sum_{j,t} n_i(t-1)\hat{p}_{ij}[1 + x_{ij}(t)]\log[1 + x_{ij}(t)],\end{aligned} \tag{6.6}$$

where $x_{ij}(t) = [\hat{p}_{ij}(t) - \hat{p}_{ij}]/\hat{p}_{ij}$. The difference Δ between $-2\log\lambda_i$ and the χ^2_i-statistic is

$$\begin{aligned}\Delta &= -2\log\lambda_i - \chi^2_i \\ &= 2\textstyle\sum_{j,t} n_i(t-1)\hat{p}_{ij}\{[1 + x_{ij}(t)]\log[1 + x_{ij}(t)] - [x_{ij}(t)]^2/2\}.\end{aligned} \tag{6.7}$$

Since $\sum_{j=1}^m \hat{p}_{ij}x_{ij}(t) = 0$,

$$\Delta = 2\sum_{j,t} n_i(t-1)\hat{p}_{ij}\{[1 + x_{ij}(t)]\log[1 + x_{ij}(t)] - x_{ij}(t) - [x_{ij}(t)]^2/2\}. \tag{6.8}$$

We shall show that Δ converges to 0 in probability; i.e. for any $\epsilon > 0$, the probability of the relation $|\Delta| < \epsilon$, under the null hypothesis, tends to unity as $n = \sum_i n_i(t) \to \infty$. The probability satisfies the relation

$$\begin{aligned}Pr\{|\Delta| < \epsilon\} &\geqq Pr\{|\Delta| < \epsilon \text{ and } |x_{ij}(t)| < \tfrac{1}{2}\} \\ &\geqq Pr\{|2\textstyle\sum_{j,t} n_i(t-1)\hat{p}_{ij}[x_{ij}(t)]^3| < \epsilon \text{ and } |x_{ij}(t)| < \tfrac{1}{2}\} \\ &\geqq Pr\{2n\textstyle\sum_{j,t} |x_{ij}(t)|^3 < \epsilon \text{ and } |x_{ij}(t)| < \tfrac{1}{2}\}.\end{aligned} \tag{6.9}$$

It is therefore necessary only to prove that $n[x_{ij}(t)]^3$ converges to 0 in probability. Since $x_{ij}(t) = [\hat{p}_{ij}(t) - \hat{p}_{ij}]/\hat{p}_{ij}$ converges to zero in probability under the null hypothesis, and

$$\sqrt{x_{ij}(t)n}\, x_{ij}(t) = \sqrt{x_{ij}(t)n}\left\{\left[\frac{\hat{p}_{ij}(t) - p_{ij}}{\hat{p}_{ij}}\right] - \left[\frac{\hat{p}_{ij} - p_{ij}}{\hat{p}_{ij}}\right]\right\}, \tag{6.10}$$

it follows that

$$n[x_{ij}(t)]^3 = [(x_{ij}(t)n)^{1/2}\, x_{ij}(t)]^2 \tag{6.11}$$

converges to zero in probability when the null hypothesis is true. Q.E.D.

Since the χ^2-statistic has a limiting χ^2-distribution under the null hypothesis, and $\Delta = -2 \log \lambda_i - \chi_i^2$ converges in probability to zero, $-2 \log \lambda_i = \chi_i^2 + \Delta$ has a limiting χ^2-distribution under the null hypothesis.

The method presented herein for showing the asymptotic equivalence of $-2 \log \lambda_i$ and χ_i^2 could also be used to show the asymptotic equivalence of statistics of the form $-2 \log \lambda$ and χ^2. It was proved in Section 3.2 that, under the null hypothesis, $-2 \log \lambda$ has a limiting χ^2-distribution with $m(m-1)(T-1)$ degrees of freedom. (The proof in Section 3.2 applied to λ, a likelihood ratio criterion, but would not apply to λ_i since they are not actually likelihood ratios.) Hence, we have another proof that the χ^2-statistic has the same limiting distribution as the likelihood ratio criterion under the null hypothesis.

The previous remarks refer to the case where the null hypothesis is true. Now suppose the alternate hypothesis is true; that is, $p_{ij}(t) \neq p_{ij}(s)$ for some t, s, i, j. It is easy to see that both the χ^2-test and the likelihood ratio test are consistent under any alternate hypothesis. In other words, if the values of $p_{ij}(t)$ for the alternate hypothesis and the significance level are kept fixed, then as n increases, the power of each test tends to 1 (see [5] and [11]).

In order to examine the situation in which the power is not close to 1 in large samples and also to make comparisons between tests, the alternate hypothesis may be moved closer to the null hypothesis as n increases. If the values of $p_{ij}(t)$ for the alternate hypothesis are not fixed but move closer to the null hypothesis, it can be seen that the two tests are again asymptotically equivalent. This can be deduced by a slight modification of the proof of asymptotic equivalence under the null hypothesis given in this section (see also [5], p. 323).

We shall now suggest another approach to the comparison of these tests when the alternate hypothesis is kept fixed. Since the null hypothesis is rejected when an appropriate statistic (χ^2 or $-2 \log \lambda$) exceeds a specified critical value, we might decide that the χ^2-test is to be preferred to the likelihood ratio test if the statistic χ^2 is in some sense (stochastically) larger than $-2 \log \lambda$ under the alternate hypothesis.

Since $n_i(t)$ is a linear combination of multinomial variables, we see that $n_i(t)/n$ converges in probability to its expected value $\mathcal{E}[n_i(t)/n] = m_i(t)$. Hence, χ^2/n converges in probability to

$$\sum_{i,j,t} m_i(t-1)[p_{ij}(t) - \bar{p}_{ij}]^2/\bar{p}_{ij}\,, \tag{6.12}$$

and $(-2 \log \lambda)/n$ converges in probability to

$$2 \sum_{i,j,t} m_i(t-1) p_{ij}(t) \log [p_{ij}(t)/\bar{p}_{ij}], \tag{6.13}$$

where

$$\bar{p}_{ij} = \sum_t p_{ij}(t)\, m_i(t-1) / \sum_t m_i(t-1) = \operatorname*{p\,lim}_{n\to\infty} \hat{p}_{ij}\,. \tag{6.14}$$

The difference between (6.12) and (6.13) is approximately

$$\sum m_i\,(t-1)[p_{ij}(t) - \bar{p}_{ij}]^3/(3\bar{p}_{ij}^2). \tag{6.15}$$

Under the alternate hypothesis, these two stochastic limits differ from 0, and computation of them suggests which test is better. If $(p_{ij}(t) - \bar{p}_{ij})/\bar{p}_{ij}$ is small, then there will be only a small difference between the two limits. When the alternative is some composite hypothesis, as is usually the case when χ^2-tests are applied, then these stochastic limits can be computed and compared for the simple alternatives that are included in the alternate hypothesis.

This method for comparing tests is somewhat related to Cochran's comment (see p. 323 in [5]) that either (a) the significance probability can be made to decrease as n increases, thus reducing the chance of an error of type I, or (b) the alternate hypothesis can be moved steadily closer to the null hypothesis. Method (b) was discussed in [3]. If method (a) is used, then the critical value of the statistic (χ^2 or $-\log \lambda$) will increase as n increases. When the critical value has the form cn, where c is a constant (there may be some question as to whether this form for the critical value is really suitable), we see from the remarks in the preceding paragraph that the power of a test will tend to 1 if c is less than the stochastic limit and it will tend to 0 if c is greater than the stochastic limit. Hence, by this approach we find that the power of the χ^2-test can be quite different from the power of the likelihood ratio test, and some approximate computations can suggest which test is to be preferred.

However, a more appealing approach is to vary the significance level so the ratio of significance level to the probability of some particular Type II error approaches a limit (or at least it seems that desirable sequences of significance points lie between c' and cn). While the usual asymptotic theory does not give enough information to handle this problem, the comparison of stochastic limits may suggest a comparison of powers.

The methods of comparison discussed herein can also be used in the study of the χ^2 and likelihood ratio methods for ordinary contingency tables. We have seen that, in a certain sense, the χ^2 and likelihood ratio methods are not equivalent when the alternate hypothesis is true and fixed, and we have suggested a method for determining which test is to be preferred.

REFERENCES

[1] T. W. Anderson, "Probability models for analyzing time changes in attitudes," RAND Research Memorandum No. 455, 1951.

[2] T. W. ANDERSON, "Probability models for analyzing time changes in attitudes," *Mathematical Thinking in the Social Sciences*, edited by Paul F. Lazarsfeld, The Free Press, Glencoe, Illinois, 1954.

[3] M. S. BARTLETT, "The frequency goodness of fit test for probability chains," *Proc. Cambridge Philos. Soc.*, Vol. 47 (1951), pp. 86–95.

[4] H. CHERNOFF, "Large-sample theory: parametric case," *Ann. Math. Stat.* Vol. 27 (1956), pp. 1–22.

[5] W. G. COCHRAN, "The χ^2-test of goodness of fit," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 315–345.

[6] H. CRAMÈR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[7] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. 1, John Wiley and Sons, New York, 1950.

[8] L. A. GARDNER, JR., "Some estimation and distribution problems in information theory," Master's Essay, Columbia University Library, 1954.

[9] L. A. GOODMAN, "On the statistical analysis of Markov chains" (abstract), *Ann. Math. Stat.*, Vol. 26 (1955), p. 771.

[10] P. G. HOEL, "A test for Markoff chains," *Biometrika*, Vol. 41 (1954), pp. 430–433.

[11] J. NEYMAN, "Contribution to the theory of the χ^2-test," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1949, pp. 239–274.