Hypothesis Tests for Markov Process Models Estimated from Aggregate Frequency Data

W. David Kelton
Department of Industrial & Operations Engineering
The University of Michigan

Christina M. L. Kelton
Department of Economics
Wayne State University

Technical Report 84-2

January, 1984

Hypothesis Tests for Markov Process Models Estimated from Aggregate Frequency Data

W. David Kelton
Department of Industrial and Operations Engineering
The University of Michigan

Christina M. L. Kelton
Department of Economics
Wayne State University

ABSTRACT

When the only data available for estimating the transition probabilities of a Markov chain are state occupation frequencies (rather than interstate transition frequencies), a least-squares estimation technique and an accompanying hypothesis-testing methodology are proposed. This general hypothesis-testing procedure is used to develop three tests for adequacy of the simple stationary model. Null hypotheses of a zero-order process, stationarity, and homogeneity are considered. The distributions of the test statistics are investigated in a factorially designed Monte Carlo study. In general, it is found that treating the test statistics as having F distributions with appropriate degrees of freedom under the null hypothesis of interest leads to rejection proportions close to the desired levels. Additional Monte Carlo results indicate favorable power of the proposed tests.

KEY WORDS: Markov processes; Estimation and testing; Aggregate frequency data.

This research was partially supported by a Faculty Development Grant from the Standard Oil Company (Ohio) to the Graduate School of Management at Kent State University. We appreciate the input of Professor John Geweke, Department of Economics, University of Wisconsin-Madison. Further, we are grateful for the suggestions of an anonymous referee. We also thank the Madison Academic Computing Center of the University of Wisconsin for supplying the program QUADMP used in this study.

1. INTRODUCTION

Many socioeconomic phenomena are most appropriately modeled by stochastic processes; examples include population migration, consumer brand shift, firm size change, voter preference behavior, and occupational mobility. A Markov chain model is frequently chosen since it provides a simple and useful stochastic framework. The usefulness of this type of modeling, however, has been generally restricted by the difficulty (in the case of "micro" data) or inability (in the case of "macro" data) to perform statistical inference (e.g., hypothesis tests) from the results. In this paper, we propose a general methodology for devising hypothesis tests when only macro data (see below) are available. This methodology is applied to the case of testing some of the basic assumptions (first-order dependence, parameter stationarity, and entity homogeneity) underlying the Markov chain model. Monte Carlo studies for these three tests provide empirical support for the proposed methodology.

If data are available for individual interstate transitions ("micro" data), suitable estimation techniques and hypothesis tests have long been available. (See, for example, Anderson and Goodman [1957], Billingsley [1961a, 1961b], and Kullback, Kupperman, and Ku [1962]. Further references to papers dealing with specialized tests may be found in Lee, Judge, and Zellner [1977].) Let N be the number of entities (i.e., individuals), and assume that we can observe the sequence of states occupied by each entity at the discrete time points $t = 0,1, \ldots,T$. Let n_{ij} be the number of times any of the entities moves from state i to state j in any one-step transition. Then the maximum-likelihood estimator of the true (stationary) transition probability p_{ij} is simply

$$\tilde{p}_{ij} = n_{ij} / \sum_{k=1}^{R} n_{ik},$$

where the states are numbered 1,2,...,R. These estimators have been shown to be consistent, asymptotically unbiased, and asymptotically normal. Furthermore, Anderson and Goodman (1957) developed likelihood-ratio and chi-square statistics for testing parameter stationarity, specified process order, and specified transition probability values.

In many applications, however, such micro data cannot be obtained. Often, one has access only to aggregate time-series "macro" data on state occupation frequencies, $n_i(t)$ = the number of the N entities occupying state i at time t. Thus, we do not observe individual movements from state to state, and know only how many entities are present in each state at each time. Least-squares (LS) estimators for the transition probabilities p_{ij} were proposed initially by Miller (1952) and Madansky (1959) and refined and summarized by Lee, Judge, and Zellner (1977) and MacRae (1977); see Section 2 for a brief discussion. Bedall (1978) proposed chi-square tests for goodness of fit and for equality of Markov chains, based on analogy to frequency table analysis techniques. His Monte Carlo results indicated good performance for the latter hypothesis test, and that a modification of his technique was suitable for the former hypothesis test. However, sampling properties of the LS estimators are, in general, unknown, and (other than Bedall's study) hypothesis-testing procedures have not as yet been developed. Our purpose in this paper is to propose a general hypothesis-testing framework based on analogy to linear regression theory and to apply this framework to develop three specific tests aimed at evaluating the adequacy of a simple Markov chain model in a given situation. Validity and favorable power properties of these three tests are supported by the results of a substantial Monte Carlo study.

In Section 2, we review the LS estimation procedure, give the general form of our test statistics, and develop statistics for the three specific tests mentioned above. Section 3 contains a discussion of a Monte Carlo study

designed to evaluate the performance of the statistics, and Section 4 presents and discusses the empirical results. Finally, we draw some conclusions in Section 5.

2. TRANSITION PROBABILITY ESTIMATORS AND TEST STATISTICS

In this section we review the LS estimation of the matrix of transition probabilities from macro data and establish a general procedure for constructing test statistics for a wide class of null hypotheses. The procedure is used to form test statistics for three hypotheses related to adequacy of the stationary discrete-time Markov chain model.

2.1 Constrained LS Estimation

As in Section 1, let the state space be $\{1,2,\ldots,R\}$, the time instants of observation be $0,1,\ldots,T$ (so that each entity experiences T transitions), and N be the number of entities. For each state i and each time instant t, let

$n_i(t)$ = the number of entities in state i at time t,

and

$y_i(t) = n_i(t)/N$ = the proportion of entities in state i at time t.

Suppose that the true transition probability matrix P has (i,j)th element p_{ij}.

Define

$\underline{y} = [y_1(1),\ldots,y_1(T),\ldots,y_{R-1}(1),\ldots,y_{R-1}(T)]'$,

$\underline{p} = (p_{11},\ldots,p_{R1},\ldots,p_{1,R-1},\ldots,p_{R,R-1})'$,

X^* = a T x R matrix with (t+1,i)th element $y_i(t)$ for $t = 0,\ldots,T-1$, and $i = 1,\ldots,R$, and

X = a T(R-1) x R(R-1) block-diagonal matrix with R-1 copies of X^* along the diagonal.

Let $\pi_i(t)$ be the true probability (under the assumed stationary Markov chain model) that an entity will occupy state i at time t, and let

$\pi(t) = [\pi_1(t),\pi_2(t),\ldots,\pi_R(t)]$.

One of the basic properties of the model is the recursion

$$\pi(t) = \pi(t-1)\ P.$$

As $y_i(t)$ is an unbiased estimator of $\pi_i(t)$, we would expect that $X\underline{p}$ should approximate $\underline{y}$. The (constrained) LS estimator of $\underline{p}$ is thus the solution $\hat{\underline{p}}$ to the quadratic programming (QP) problem

$$\min_{\underline{p}} (\underline{y} - X\underline{p})' (\underline{y} - X\underline{p})$$

subject to

$$p_{ij} \geq 0 \quad \text{for all } i \text{ and } j, \text{ and } \sum_{j=1}^{R-1} p_{ij} \leq 1 \text{ for all } i.$$

Since the sum of each row of a transition probability matrix must equal 1, we take

$$\hat{p}_{iR} = 1 - \sum_{j=1}^{R-1} \hat{p}_{ij}.$$

Thus, the last column of P was omitted from the above formulation. The above QP problem may be solved by a simplex-like pivotal algorithm, such as Lemke's (1968). It has been shown (see Kelton [1981]) that this QP approach yields estimates which are usually identical to those obtained from a more complicated algorithm due to MacRae (1977) for which consistency properties have been established.

2.2 General Test Statistic Construction

Following Chow (1960), Fisher (1970), and Theil (1971), the idea in testing a desired null hypothesis H_0 on the underlying Markov model is to compare an "unrestricted" sum of squared residuals SSR_U to the sum of squared residuals SSR_R from a QP problem which forces the restrictions required by H_0. The methods of computing SSR_U and SSR_R depend on the particular H_0 of interest; three examples appear in Section 2.3.

The general formula for the test statistic is

$$F_{q,v} = \frac{(SSR_R - SSR_U)/q}{SSR_U/v},$$

where

q = the number of (additional) restrictions imposed by H_0,

and

v = the degrees of freedom (number of independent observations minus number of parameters estimated) in the unrestricted QP problem.

We propose that $F_{q,v}$ be treated as having an F distribution with (q,v) degrees of freedom (d.f.), under H_0. However, we are violating many assumptions necessary for this to be true, e.g., that $\underline{y} - X\underline{p}$ is a vector of uncorrelated, homoscedastic, normally distributed error terms. Thus, we investigated in a Monte Carlo study the actual distribution of $F_{q,v}$; see Sections 3 and 4.

2.3 Test Statistics for Some Specific Null Hypotheses

First, we test the null hypothesis that all rows of the transition probability matrix are identical: i.e.,

$$H_0: \; p_{ij} = p_j \text{ for all } i.$$

This implies that the process is zero-order; i.e., S_t is independent of S_{t-1}, where S_t is the state of the system at time t. Also, the distribution of S_t for each $t \geq 1$ would be the same as the steady-state distribution, under H_0. The test distinguishes between independence and some dependence (memory), but cannot be used to establish the length of memory of the process.

In order to implement this test and to calculate $F_{q,v}$, we let

SSR_U = the sum of squared residuals from the QP problem detailed in Section 2.1,

$q = (R-1)^2$,

SSR_R = the sum of squared residuals from a QP problem, QP(1), similar to that of Section 2.1, but with $\underline{p}$ and X^* replaced by

$\underline{p}(1) = (p_1, p_2, \ldots, p_{R-1})'$ and

$X^*(1) = (1, 1, \ldots, 1)'$,

respectively, and X(1) defined to be block-diagonal on $X^*(1)$,

and

$$v = T(R-1) - R(R-1).$$

To illustrate how $\underline{p}(1)$ and $X^*(1)$ were found, note first that $\underline{p}(1)$ simply contains the parameters to be estimated, under H_0. Next, since

$$\sum_{i=1}^{R} y_i(t) = 1,$$

we have (under H_0) that $y_j(t)$ has expectation p_j. Thus, we would expect that $X(1)\underline{p}(1)$ should approximate $\underline{y}$. Hence, the LS estimator of $\underline{p}(1)$ is the solution to QP(1):

$$\min_{\underline{p}(1)} [\underline{y} - X(1)\underline{p}(1)]'[\underline{y} - X(1)\underline{p}(1)]$$

$$\text{subject to } p_j \geq 0 \text{ and } \sum_{j=1}^{R-1} p_j \leq 1.$$

The restriction that $p_{ij} = p_j$ for all i is thus embedded in QP(1). (Note that only R-1 transition probabilities are estimated in the restricted model.) Under the conditions for the standard general linear model (which we do not meet here), Lemma 2.3 of Fisher (1970) would imply that $F_{q,v}$ has an F distribution with (q,v) d.f., under H_0. Our Monte Carlo study below investigates the actual null distribution of $F_{q,v}$ in our case.

Second, we test the null hypothesis that the transition probabilities are stationary over time:

$$H_0: \; p_{ij}(e) = p_{ij}(f) \text{ for all } i,j,$$

where e refers to the "early" time instants t = 1,2,...,T/2, and f refers to the "later" time instants t = T/2+1,...,T. The Markov chain observation

period is divided into two equal-length subperiods, and transition probabilities are estimated for each subperiod; this is the "unrestricted" model in this case. For the later time instants, the proportions at "time t=0" are actually the observed proportions at time T/2. (The proposed approach is also easily extended to test equality of transition probabilities across multiple subperiods.)

In this case, let

SSR_U = $SSR_U(e) + SSR_U(f)$, where $SSR_U(e)$ is the sum of squared residuals from the QP problem detailed in Section 2.1 for t = 1,...,T/2, and where $SSR_U(f)$ is the sum of squared residuals from the QP problem detailed in Section 2.1 for t = T/2+1,...,T,

$q = R(R-1)$,

SSR_R = the sum of squared residuals from the QP problem of Section 2.1 for t = 1,2,...,T,

and

$v = T(R-1) - 2R(R-1)$;

$F_{q,v}$ is defined as before.

Finally, we test the null hypothesis that one group of entities has the same transition probabilities as a second group (again, the problem can be generalized to multiple groups): i.e.,

H_0: $p_{ij}(g) = p_{ij}(h)$ for all i,j,

where g and h refer to two distinct groups of entities. This is the homogeneity assumption which requires that all individuals follow the same Markov chain. Let N(g) and N(h) be the number of entities in group g and group h, respectively. (N(g) + N(h) = N.) If

$n_i(t,g)$ = the number of group-g entities in state i at time t

and

$n_i(t,h)$ = the number of group-h entities in state i at time t,

then $y_i(t,g) = n_i(t,g)/N(g)$ and $y_i(t,h) = n_i(t,h)/N(h)$ are entity proportions

in state i at time t for groups g and h, respectively. The "unrestricted" model (as for the stationarity test above) consists of two transition-probability estimations, one for group g and one for group h.

Here,

SSR_U = $SSR_U(g) + SSR_U(h)$, where $SSR_U(g)$ is the sum of squared residuals from the QP problem detailed in Section 2.1 with the N(g) entities of group g, and where $SSR_U(h)$ is the sum of squared residuals from Section 2.1's QP with the N(h) entities of group h,

$q = R(R-1)$,

SSR_R = the sum of squared residuals from a QP problem, QP(3), similar to that of Section 2.1, but with $\underline{y}$ and X^* replaced by

$\underline{y}(3) = (\underline{y}_1 \mid \underline{y}_2 \mid \ldots \mid \underline{y}_{R-1})'$, where

$\underline{y}_j = [y_j(1,g),\ldots,y_j(T,g),y_j(1,h),\ldots,y_j(T,h)]$

for $j = 1,\ldots,R-1$, and

$X^*(3)$ = a $2T \times R$ matrix with, for $i = 1,\ldots,R$, $(t+1,i)$th element $y_i(t,g)$ for $t = 0,\ldots,T-1$, and $y_i(t-T,h)$ for $t = T,\ldots,2T-1$,

respectively, and X(3) defined to be block-diagonal on $X^*(3)$,

and

$v = 2T(R-1) - 2R(R-1)$.

$F_{q,v}$ is defined as before.

3. EXPERIMENTAL PROCEDURES AND DESIGN

The test statistics defined in Section 2 would have F distributions with (q,v) d.f., under their respective H_0's, if the assumptions for the general linear regression model held. Since several of these assumptions will, in general, be violated, we undertook an extensive Monte Carlo study to determine whether this F distributional assumption can be made in practice. In this section, we describe the data generation, the criteria used in evaluating the proposed tests, and the experimental design.

For each of the three null hypotheses considered, values for R, N, T, $\pi(0)$, and P were specified as called for by an experimental design discussed below; for a validity investigation, these process parameters and the data generation must conform to the requirements of the H_0 under consideration. Given this underlying process structure, N independent realizations (i.e., entities) were generated by first generating the initial state, S_0, according to the probabilities in $\pi(0)$. S_1 was then generated according to the probabilities in row S_0 of P; S_2 was generated according to row S_1 of P, etc. The proportions $y_i(t)$ were tallied as the data generation progressed. The random-number generator used is the multiplicative congruential generator developed by Lewis, Goodman, and Miller (1969) in the FORTRAN implementation of Schrage (1979). All data generated across design points and different null hypotheses were made independent by using nonoverlapping random number streams.

Rather than specifying the underlying process parameters in a few combinations and levels chosen essentially arbitrarily, we view R, N, T, $\pi(0)$, and P as five factors in an experimental design context. In this way, we can examine the effects of these factors on the performance of the tests in an organized fashion. For each null hypothesis, we chose two levels for each of the five factors (see Section 4 for the particular values), and used them in 16 combinations called for by a resolution V, 2^{5-1} fractional factorial design. The design was constructed by writing a full 2^4 factorial design in the first four factors, and taking the level (sign) for P to be the positive product of the signs of the levels of the other four factors (see Box, Hunter, and Hunter [1978]). For each of the 16 design points, 200 independent replications of the entire estimation and testing procedure were made, resulting in 200 independent observations on the test statistic $F_{q,v}$.

We selected a number of criteria to evaluate the performance of the tests

proposed. The 200 test statistic values were used in chi-square (χ^2) and Kolmogorov-Smirnov (KS) goodness-of-fit tests in comparison with the F distribution with (q,v) d.f. Since robustness of these tests more importantly depends on upper-tail properties, we computed the percentage of the 200 values which fell above the upper 10%, 5%, and 1% critical values of the F distribution with (q,v) d.f. Finally, we noted the absolute differences between these observed rejection percentages and their corresponding "desired" values. These criteria are the "responses" for the experimental designs.

In addition to our designed robustness experiments, we evaluated the power of the three hypothesis tests, at various degrees of departure from the respective null hypotheses, at the mean values (see below) of four of the design parameters: i.e., T = 38, N = 300, $\pi(0)$ = (0.52, 0.18, 0.15, 0.15), and P's elements averaged over their "-" and "+" values in the validity experiments. Power functions were estimated for both state space sizes (R = 2 and R = 4) with respect to a multinomial logit parameterization (see MacRae [1977]).

This parameterization sets

$$p_{ij} = \exp \theta_{ij} \; / \; (1 + \sum_{m=1}^{R-1} \exp \theta_{im}),$$

$$\text{for } i = 1,\ldots,R \text{ and } j = 1,\ldots,R-1.$$

Thus, the null hypothesis of zero-order dependence requires $\theta_{1j} = \theta_{2j} = \cdots = \theta_{Rj}$, for j = 1,...,R-1. To allow deviation from H_0, we set $\theta_{21} = \theta_{11} + \delta$, where δ varied from -3 to +3. (When $\delta = 0$, the null hypothesis is true.) Results are presented in Section 4. For the stationarity and homogeneity tests, the null hypotheses require, in turn, that $\theta_{ij}(e) = \theta_{ij}(f)$ and $\theta_{ij}(g) = \theta_{ij}(h)$, for i = 1,...,R, and j = 1,...,R-1. In these two cases, for R = 2, we estimated a power surface by allowing $\theta_{11}(f) = \theta_{11}(e) + \delta_1$ and $\theta_{21}(f) = \theta_{21}(e) + \delta_2$ (in the case of stationarity) and $\theta_{11}(h) = \theta_{11}(g) + \delta_1$ and $\theta_{21}(h) =$

$\theta_{21}(g) + \delta_2$ (in the case of homogeneity) and letting δ_1 and δ_2 vary from -2 to +2. For R = 4, we simply let $\theta_{21}(f) = \theta_{21}(e) + \delta$ and $\theta_{21}(h) = \theta_{21}(g) + \delta$ for stationarity and homogeneity, respectively, allowing δ to vary from -3 to +3. This departure from H_0 involves only one of the four rows of P(f) and P(h), maintaining H_0 for the other three rows; this should yield a worst case (lower bound) for power. Again, results are presented in Section 4. For further power results from a factorially designed study, see Kelton and Kelton (1983).

4. EMPIRICAL RESULTS

In this section we present the results of the Monte Carlo studies described in Section 3. For the validity investigations, the choices for the levels of the factors were made taking into account our experience with use of these models in practice. The "-" and "+" levels for R are 2 and 4, respectively, for each of the three null hypotheses. Except for the null hypothesis of stationarity, the "-" and "+" levels for T are 25 and 50; since T should be even for the stationarity H_0, we use 26 for the "-" level for T in this test. For N, we take 100 and 500 as the "-" and "+" levels, respectively. Since $\pi(0)$ and P may depend on the H_0 under consideration, their values will be given as we discuss each test; generally, we attempt to use one of the values of P to indicate reluctance to change states (i.e., large diagonal elements), and the other value to exhibit somewhat greater mobility. For each test, we give a table of means (over the 16 design points) and main effects estimates for each of our nine responses. For the χ^2 goodness-of-fit test, we report the p-value (probability of obtaining a value of χ^2 at least as large as the one observed) as well as the χ^2 value itself; 20 equiprobable intervals were used for each χ^2 test. For the KS test we report the adjusted test statistic $D' = D_n[(n)^{0.5} + 0.12 + 0.11/(n)^{0.5}]$, developed by Stephens (1974), instead of the usual KS statistic, D_n; this

allows the use of his very compact tables of critical values. Also, for each test, we report power results for a number of alternative hypotheses.

4.1 First-Order Dependence

The two levels for $\pi(0)$ here were taken to be the discrete uniform distribution on {1,2,...,R} at the "-" level, and a distribution degenerate on state 1 for the "+" level. Since this H_0 requires that each row of P be the same, we need specify only a single row. For the "-" level, we took this row to be uniform, and the "+" level row was taken as (0.79, 0.21) when R = 2 and (0.79, 0.11, 0.05, 0.05) when R = 4.

Table 1 contains, for each response, the overall mean and the main effects of the factors; C_{10}, C_5, and C_1 denote the observed rejection percentages for the (desired) 10%, 5%, and 1% levels, respectively. The average p-value of χ^2, 0.34, indicates generally good agreement between $F_{q,v}$ and the F distribution. For some factor combinations, however, the overall fit is not good (see Kelton and Kelton [1982b]). 7 of 16 values are significant at the 0.10 level; 6 of 16 values of D' are significant at the 0.10 level. A 90% confidence interval fails to cover the target rejection percentage in 14 out of 48 cases. The average observed rejection proportions, 7.7%, 4.2%, and 1.0%, are mildly conservative, but are quite close to the desired levels. The absolute deviations of the C_r's further indicate good performance of the proposed test; for example, the average deviation for a 5% test is 1.3 percentage points. We feel that these rejection proportions are the most relevant criteria for judging the proposed testing procedure, and their closeness to the desired levels in this case is encouraging.

The main effects in Table 1 and the two-way interactions not reported here (see Kelton and Kelton [1982b] for details) indicate that R appears to be the most important single factor affecting the overall fit, with the larger number

Table 1. Means and Main Effects for Testing First-Order Dependence

Response	Means	Main Effects				
		R	T	N	$\pi(0)$	P
χ^2	43.36	47.73	-5.63	-22.53	4.28	45.58
p-value of χ^2	0.34	-0.32	-0.14	0.27	0.02	-0.21
D'	1.50	1.61	-0.30	-0.26	0.15	1.06
C_{10}	7.66	-0.19	1.69	-0.56	-1.31	0.19
C_5	4.16	0.56	1.31	-0.19	-0.81	-0.56
C_1	1.03	0.31	0.44	0.06	-0.56	-0.19
$\lvert C_{10}-10\rvert$	2.78	-0.44	-1.06	-0.06	1.69	0.44
$\lvert C_5-5\rvert$	1.34	-0.06	-1.06	-0.31	0.56	0.06
$\lvert C_1-1\rvert$	0.41	0.31	0.19	-0.19	-0.31	-0.19

of states (R = 4) worsening the fit. Also, a better fit is obtained if the rows of P are uniform, as opposed to having a large probability of being in a particular state. Furthermore, a large positive interaction between R and P is consistent with their main effects. These three effects also appeared to be significant from a normal probability plot.

The rejection proportions, however, do not appear to be affected as much by any main effect or interaction. A probability plot of the effects on $|C_5-5|$ did not indicate any obviously significant effect. It appears that there is some tendency for a larger value of T to lead to better rejection proportions; thus, it may be advisable to obtain longer time records for the observations, if possible.

Table 2 shows the rejection percentages, each obtained from 200 experimental replications, for different values of δ (violations of H_0). For both R = 2 and R = 4, the power function behaves as expected (is unbiased), with power's rising with the extent of violation of H_0. For R = 2, power reaches 100% when δ = -2 or -3, rising again to 74% (10% test level) when δ = +3. Note that when δ = 0, type I error probabilities are obtained which do not differ greatly from the mean values in Table 1. As one would expect intuitively, a larger state space (R = 4) produces lower power results overall, with a high rejection percentage of 54.5 (at the 10% level). Recall, however, that our departure from H_0 in the case of R = 4 is only for row 2; higher power would be obtained for departures involving other rows as well.

4.2 Parameter Stationarity

In this case the "-" level for $\pi(0)$ is a uniform distribution, and the "+" level is (0.79, 0.21) when R = 2, and (0.79, 0.11, 0.05, 0.05) when R = 4. When R = 2, the "-" and "+" levels for P are

Table 2. Rejection Percentages for Power of Test for First-Order Dependence

δ	R = 2: Level of Test			R = 4: Level of Test		
	10%	5%	1%	10%	5%	1%
-3	100.00	100.00	100.00	50.00	37.00	21.00
-2	100.00	100.00	100.00	38.00	25.50	12.00
-1	47.50	34.50	9.00	18.50	12.00	4.00
0	8.50	2.50	1.00	10.00	4.50	1.50
+1	37.50	24.00	7.50	16.00	11.50	5.00
+2	61.50	51.50	30.00	42.00	30.00	14.00
+3	74.00	65.50	40.00	54.50	42.50	22.50

$$\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.6 & 0.4 \\ 0.5 & 0.5 \end{bmatrix}$$

respectively. When R = 4, the "-" and "+" levels for P are

$$\begin{bmatrix} 0.8 & 0.2 & 0.0 & 0.0 \\ 0.1 & 0.8 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.8 & 0.1 \\ 0.0 & 0.0 & 0.2 & 0.8 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.5 & 0.2 & 0.2 & 0.1 \\ 0.2 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.2 & 0.2 & 0.5 \end{bmatrix}$$

respectively.

From Table 3 (which follows the same format as Table 1), the χ^2 test yields an average p-value of 0.40, indicating a good overall fit; Further, χ^2 was significant at the 0.10 level in only 3 out of 16 cases, and the KS statistic was significant in 7 cases. The average observed rejection percentages, 9.9%, 5.5%, and 1.2%, are very close to the desired values. Individual rejection percentages differed from the target values at the 0.10 level of significance in only 8 out of 48 cases. Again, in Table 3, the absolute deviations $|C_r - r|$ are 1.9, 1.5, and 0.7 percentage points, further indicating reasonable rejection probability performance. Thus, this test for stationarity appears to be quite well-behaved with respect to size.

A probability plot of the main effects and two-way interactions for the overall fit statistics did not indicate that any were clearly significant, although large R again appears to worsen the overall fit. Also, it appears that large R causes the rejection percentages to deviate somewhat more from their respective target values.

The power values in Tables 4 and 5 were obtained from 100 independent replications and show quite high power (100% in many cases). These results are encouraging since even small one- or two-parameter departures from H_0 elicit high rejection levels. Again, power behaves generally as expected, with low rejection percentages obtained when H_0 is true.

able 3. Means and Main Effects for Testing Parameter Stationarity

Response	Means	Main Effects				
		R	T	N	$\pi(0)$	P
χ^2	22.70	9.25	-3.25	5.75	6.15	-0.10
p-value of χ^2	0.40	-0.33	0.04	-0.12	-0.15	-0.19
D'	1.28	0.38	0.00	0.19	0.41	-0.12
C_{10}	9.94	-0.13	0.13	-0.50	-0.50	-2.13
C_5	5.53	-0.06	-0.06	-0.94	0.44	-1.06
C_1	1.16	0.31	-0.19	-0.19	0.31	-1.19
$\lvert C_{10}-10\rvert$	1.88	0.50	-0.25	0.13	0.13	-1.25
$\lvert C_5-5\rvert$	1.53	0.19	-0.31	-0.44	0.19	0.19
$\lvert C_1-1\rvert$	0.66	0.31	0.06	0.06	-0.44	-0.44

Table 4. Rejection Percentages for Power of Test for Parameter Stationarity (R = 2)*

δ_2	δ_1 = -2	δ_1 = -1	δ_1 = 0	δ_1 = +1	δ_1 = +2
	100.00	100.00	100.00	100.00	57.00
-2	100.00	100.00	100.00	100.00	40.00
	100.00	100.00	100.00	97.00	16.00
	100.00	100.00	100.00	30.00	100.00
-1	100.00	100.00	100.00	20.00	100.00
	100.00	100.00	100.00	2.00	100.00
	100.00	100.00	10.00	100.00	100.00
0	100.00	100.00	3.00	100.00	100.00
	100.00	100.00	0.00	100.00	100.00
	100.00	50.00	100.00	100.00	100.00
+1	100.00	37.00	100.00	100.00	100.00
	100.00	17.00	100.00	100.00	100.00
	98.00	100.00	100.00	100.00	100.00
+2	93.00	100.00	100.00	100.00	100.00
	88.00	100.00	100.00	100.00	100.00

* Values are ordered for 10%, 5%, and 1% test levels, respectively.

Table 5. Rejection Percentages for Power of Test for Parameter Stationarity (R = 4)

	Level of Test		
δ	10%	5%	1%
-3	100.00	100.00	99.00
-2	100.00	100.00	93.00
-1	87.00	77.00	58.00
0	8.00	5.00	1.00
+1	100.00	100.00	98.00
+2	100.00	100.00	100.00
+3	100.00	100.00	100.00

4.3 Entity Homogeneity

The levels for $\pi(0)$ and P in this case are exactly the same as for the parameter stationarity test in Section 4.2. From Table 6, the average overall fit is again reasonably good (χ^2 was significant at the 0.10 level in 4 of 16 cases, and the KS statistic was significant in 5 cases.) 90% confidence interval coverage for target rejection percentages was not obtained in 12 out of 48 cases. The average rejection percentages, 12.3%, 7.1%, and 1.8%, are slightly higher than desired, but are still quite close. The values of $|C_r - r|$ indicate that the rejection percentages are, on average, within a few percentage points of the target values.

Again, the proposed tests behave better when the state space is smaller, and there is some evidence that a P matrix consistent with higher mobility leads to better test performance; these remarks are true both for the overall goodness-of-fit tests and for the rejection percentages. Furthermore, the main effects of T and N on the values of $|C_r - r|$ indicate the advisability of having longer time records and a larger sample of entities, if possible.

The estimated power percentages in Tables 7 and 8, again obtained from 100 replications, are even higher than those presented in Tables 4 and 5. The good power properties of the testing methodology are evidenced by the high power observed at all deviations from H_0.

5. CONCLUSIONS

The empirical application of stochastic processes has been inhibited by the lack of statistical inference techniques when aggregate frequency data alone are available. In this paper, we have developed a framework for devising suitable test statistics and distributions for examining various null hypotheses of interest, and have applied this framework to three specific

Table 6. Means and Main Effects for Testing Entity Homogeneity

		Main Effects				
Response	Means	R	T	N	$\pi(0)$	P
χ^2	28.55	16.10	-7.70	0.25	8.15	-16.70
p-value of χ^2	0.29	-0.23	0.12	0.15	0.07	0.26
D'	1.12	0.38	-0.28	0.11	0.41	-0.62
C_{10}	12.31	0.50	-1.38	-1.38	-0.50	-2.38
C_5	7.13	1.13	-0.63	-0.88	-0.25	-2.13
C_1	1.78	0.56	0.31	-0.56	-0.19	-1.31
$\lvert C_{10}-10\rvert$	2.56	-0.00	-1.38	-1.38	0.00	-1.88
$\lvert C_5-5\rvert$	2.25	0.88	-0.88	-0.63	0.00	-1.88
$\lvert C_1-1\rvert$	1.03	0.56	-0.19	-0.06	-0.19	-0.81

Table 7. Rejection Percentages for Power of Test for Entity Homogeneity (R = 2)*

δ_2	δ_1 = -2	δ_1 = -1	δ_1 = 0	δ_1 = +1	δ_1 = +2
	100.00	100.00	100.00	100.00	93.00
-2	100.00	100.00	100.00	100.00	87.00
	100.00	100.00	100.00	100.00	75.00
	100.00	100.00	100.00	58.00	100.00
-1	100.00	100.00	100.00	42.00	100.00
	100.00	100.00	100.00	19.00	100.00
	100.00	100.00	10.00	100.00	100.00
0	100.00	100.00	6.00	100.00	100.00
	100.00	100.00	1.00	100.00	100.00
	100.00	82.00	100.00	100.00	100.00
+1	100.00	68.00	100.00	100.00	100.00
	100.00	39.00	100.00	100.00	100.00
	100.00	100.00	100.00	100.00	100.00
+2	100.00	100.00	100.00	100.00	100.00
	99.00	100.00	100.00	100.00	100.00

* Values are ordered for 10%, 5%, and 1% test levels, respectively.

Table 8. Rejection Percentages for Power of Test for Entity Homogeneity (R = 4)

	Level of Test		
δ	10%	5%	1%
-3	100.00	100.00	100.00
-2	100.00	100.00	100.00
-1	95.00	91.00	78.00
0	13.00	7.00	2.00
+1	100.00	100.00	100.00
+2	100.00	100.00	100.00
+3	100.00	100.00	100.00

hypotheses for assessing the adequacy of a simple stationary Markov chain model. For a given problem, if these three null hypotheses cannot be rejected at some chosen level of significance, a Markov chain model could presumably be used for this process along with its predictive capability and its associated steady-state distribution. If the stationarity assumption is rejected, a nonstationary Markov process would be a more appropriate model. (For an application of a nonstationary Markov process to the brewing industry, see Kelton and Kelton [1982a].)

The Monte Carlo studies that we carried out for the three null hypotheses considered here indicate that treating the test statistics as having an F distribution with the appropriate d.f. under H_0 leads to tests having rejection probabilities which are quite close to the desired levels. Thus, we would anticipate that, in most applications, the tests would be valid. The experimental design of our Monte Carlo studies suggests that, among those factors which might be under the control of the investigator, enhanced test validity can be expected if the state space is kept small. There is also some evidence that long time records and large samples of entities would be desirable. In any case, however, the tests generally appear to be quite robust. Furthermore, the results of this study (as well as those of Kelton and Kelton [1983]) indicate that the three hypothesis tests are fairly powerful against various alternative hypotheses.

Additional research might include development and evaluation of more specialized tests which could be of interest in various applications. We are currently investigating the application of the general techniques of this paper to some particular hypothesis tests of a more specialized nature.

REFERENCES

ANDERSON, T.W., and GOODMAN, L.A. (1957), "Statistical Inference About Markov Chains," The Annals of Mathematical Statistics, 28, 89-110.

BEDALL, FRITZ K. (1978), "Test Statistics for Simple Markov Chains. A Monte Carlo Study," Biometrical Journal, 20, 41-49.

BILLINGSLEY, PATRICK (1961a), Statistical Inference for Markov Processes, Chicago, Illinois: The University of Chicago Press.

——— (1961b), "Statistical Methods in Markov Chains," The Annals of Mathematical Statistics, 32, 12-40.

BOX, GEORGE E.P., HUNTER, WILLIAM G., and HUNTER, J. STUART (1978), Statistics for Experimenters, New York: John Wiley & Sons, Inc.

CHOW, GREGORY C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," Econometrica, 28, 591-605.

FISHER, FRANKLIN M. (1970), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions: An Expository Note," Econometrica, 38, 361-366.

KELTON, CHRISTINA M.L. (1981), "Estimation of Time-Independent Markov Processes with Aggregate Data: A Comparison of Techniques," Econometrica, 49, 517-518.

KELTON, CHRISTINA M.L., and KELTON, W. DAVID (1982a), "Advertising and Intraindustry Brand Shift in the U. S. Brewing Industry," The Journal of Industrial Economics, 30, 293-303.

KELTON, W. DAVID, and KELTON, CHRISTINA M.L. (1982b), "Hypothesis Tests for Markov Process Models Estimated from Aggregate Frequency Data," Kent State University Department of Administrative Sciences, Working Paper WS-8202.

KELTON, CHRISTINA M.L., and KELTON, W. DAVID (1983), "Markov Process Models: A General Framework for Estimation and Inference in the Absence of State Transition Data," to appear in Proceedings of the Fourth International Conference on Mathematical Modeling, Oxford: Pergamon Press.

KULLBACK, S., KUPPERMAN, M., and KU, H.H. (1962), "Tests for Contingency Tables and Markov Chains," Technometrics, 4, 573-608.

LEE, T.C., JUDGE, G.G., and ZELLNER, A. (1977), Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data, 2nd ed., Amsterdam: North-Holland Publishing Company.

LEMKE, C.E. (1968), "On Complementary Pivot Theory," in Mathematics of the Decision Sciences, ed. by G. B. Dantzig and A. F. Veinott, Providence, Rhode Island: American Mathematical Society, 95-114.

LEWIS, P.A.W., GOODMAN, A.S., and MILLER, J.M. (1969), "A Pseudo-Random Number Generator for the System/360," IBM Systems Journal, 8, 136-146.

MACRAE, ELIZABETH CHASE (1977), "Estimation of Time-Varying Markov Processes with Aggregate Data," Econometrica, 45, 183-198.

MADANSKY, ALBERT (1959), "Least Squares Estimation in Finite Markov Processes," Psychometrika, 24, 137-144.

MILLER, GEORGE A. (1952), "Finite Markov Processes in Psychology," Psychometrika, 17, 149-167.

SCHRAGE, L. (1979), "A More Portable Fortran Random Number Generator," ACM Transactions on Mathematical Software, 5, 132-138.

STEPHENS, M.A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," Journal of the American Statistical Association, 69, 730-737.

THEIL, HENRI (1971), Principles of Econometrics, New York: John Wiley & Sons, Inc.