Journal of Scientific Research & Reports
8(4): 1-6, 2015; Article no.JSRR.19642
ISSN: 2320-0227

SCIENCEDOMAIN *international*
www.sciencedomain.org

SCIENCEDOMAIN

Testing the Transition Probabilities in Square Contingency Tables

Serpil Aktaş[1*]

[1]*Department of Statistics, Hacettepe University, Beytepe, Ankara, Turkey.*

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JSRR/2015/19642
Editor(s):
(1) Janusz Brzdek, Department of Mathematics, Pedagogical University, Poland.
Reviewers:
(1) Jesus Soria-Ruiz, National Institute of Research for Forestry, México.
(2) S. K. Srivatsa, Computer Science and Engineering, Prathusha Institute of Technology and Management, India.
Complete Peer review History: http://sciencedomain.org/review-history/10325

Original Research Article

Received 21st June 2015
Accepted 18th July 2015
Published 25th July 2015

ABSTRACT

Repeated responses may be obtained from different time points in longitudinal studies. When modelling such data, transition models like Markov type models concentrate on changes between the consecutive time points. Markov model consists of all possible states of a randomly changing system where it is assumed that next states depend only on the current state. For categorical data, Markov models help us to summarize the data and parameter estimation in contingency table form. Square contingency tables having the same row and column categories occur for the repeated observations on the response variable. In this paper, a computer program written in C# is developed to test whether the stationary probabilities are constant for several square contingency tables. It is also shown that if the transition probabilities are the same for each time interval, a single transition matrix may be estimated from the aggregrated tables. Limiting behavior of Markov chains as $n \to \infty$ is also calculated.

Keywords: Markov chain; Markov models; stationary; transition probabilities; square contingency tables.

1. INTRODUCTION

Dependency of two categorical data arises in many situations, such as, the same subjects are surveyed at different points in time (panel studies, voting); matched pairs are surveyed (father and son, husband and wife); two people rate the same object (ranking brands, pathalogist classifications). These data are generally

**Corresponding author: Email: spxl@hacettepe.edu.tr;*

collected in the form of RxR square contingency tables. Markov chain models are used in various applied fields such as longitudinal studies because we can display the Markov chain data in a contingency table form. Behavior of a Markov chain depends on the transition matrix which contains transitional probabilities. A stationarity test on Markov chain models is proposed by Sirdari et al [1]. Weissbach and Walter [2] studied the time-stationarity of rating transitions. In most practical studies the transition matrix is unknown and estimated from the emprical distribution [3].

One of the most important tests on Markov chain models is stationarity of transition probabilities [4]. The transition probabilities may be the same for each time interval. Square contingency tables where observations are cross-classified by two variables with the same categories arise frequently in social, behavioral and the medical studies, and display changes in state from one period of time to another [5]. We assume that a square RxR contingency table constructed for different time points. When the state space is categorical and observations ocur at a discrete set of times has discrete state space and discrete time. For example the children were examined annualy at ages 9 through 12 and classified according to presence and absence of wheeze [6].

Potential voters are asked their party or candidate preference from May to October. It is of interest to know the a voter's intention is contant over time. In sociology, social mobility researches are of great importance. One may wish to study the population changes as regards some certain traits from generation to generation. Analogously, in marketing researches, brand loyality can be expressed through the stationary of the process. If the transisiton probabilities on the main diagonal are close to one, we will assume that the process is time independent and the customers are loyal to the brand. Markov chain models are widely used in demography as well. In such surveys, we extensively wish to know that the transition probabilities for the *R*x*R* tables are the same over time. The aim of the paper is to develop computer codes to test the stationary probabilities are constant or not for several square contingency tables. The codes are written in *C#* language to analyze such data. The implementation of the program is displayed on a Danish Subjective health data is given in the third section. Throughout our paper, we will assume that we know the state or category for member of sample at every point in time and we will use the observed transitions.

2. MATERIALS AND METHODS

2.1 Testing the Hypothesis that the Transition Probablities are Constant

A Markov chain is a stochastic process that for given sequence of random v ariables, $X_0,\ldots,X_t$. The distribution of X_{t+1} is identical to the contitional distribution of X_{t+1} for given X_t [6].

Let p_{ij} denotes the probability that an individual in state *i* at time *t*-1 moves to state *j* at time *t* [5].

Maximum likelihood estimates for p_{ij}'s are

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i} = \frac{\sum_{t=1}^{T} n_{ij(t)}}{\sum_{t=0}^{T-1} n_{i(t)}}. \qquad (1)$$

Where T is the number of tables observed at t and t-1 times,
$n_{ij}(t)$ denotes the number of individuals in state *i* at *t*-1 and *j* at *t*,
n_i (t) is the row totals of each time point *t*, $n_i(t)=\sum_j n_{ij}(t)$,
n_{ij} is the cell frequencies over *t*, $n_{ij}=\sum_j n_{ij}(t)$,
and n_i is the row totals over *t*, $n_i=\sum_t n_i(t)$.
For instance, the transition matrix for a binary case is defined by,

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}.$$

This matrix shows the probabilities of making the transition from one state to the other or to the same state [7].

The hypothesis of interest is that the random variables for T have the same distribution. A Markov process is stationary if p_{ij} (t) is independent of time *t*.

We test the null hypothesis

$$H_0: P_{ij}(t)=P_{ij} \qquad t=1,\ldots,T \qquad (2)$$

under the alternative hypothesis the estimation of the transition probabilities for time *t* are

$$\hat{P}_{ij(t)} = \frac{n_{ij(t)}}{n_{i(t-1)}} \ . \qquad (3)$$

Likelihood ratio statistic for testing the null hypothesis of stationarity as follows

$$LR = 2\sum_t \sum_i \sum_j n_{ij}(t)\left\{ln\, n_{ij}(t) - ln\left(n_{ij}\, n_i(t)\right)/n_i\right\} \quad (4)$$

The likelihood ratio has chi-squared distribution with $R(R\text{-}1)(T\text{-}1)$ degrees of freedom under H_0 [8].

The null hypothesis of T independent samples fom multionomial trials alternatively can be tested by the likelihood ratio criterion

$$\lambda = \prod_t \prod_{i,j}\left(\frac{\hat{p}_{ij}}{\hat{p}_{ij(t)}}\right)^{n_{ij(t)}}. \quad (5)$$

-2logλ is similarly distributed as chi-square distribution with $R(R\text{-}1)(T\text{-}1)$ degrees of freedom when the null hypothesis is true [6].

<u>Theorem:</u> *If a Markov chain is irreducible, aperiodic, and positive recurrent, then, for every i, j $\in$ S(state, space), the chain has a limiting distribution* $lim_{n\to\infty} P_{ij}^n = \pi_j$.

$\pi_{1,\ldots,}\ \pi_n$ are the uniqe solutions to

$$\pi_j = \sum_{i\in S}\pi_i P_{ij}.$$

We suppose that the pattern of change is constant or not over time. This means that we have a first order Markov chain and we wish to test for stationary. This analysis actually equivalent to certain contingency table analysis based on log linear models, but unlike the marginal models, this modeling is conditional on the previous response [6]. In Markov models transitions from one state to another time point is investigated.

If there exists a unique stationary distribution, the distribution of the Markov chain converges to the stationary distribution starting from any initial state is of interest.

2.2 Numerical Example

Data are based on Danish longitudinal study of subjective health and given in Table 4 were taken directly from Andersen [8]. A population of 570 Danish elderly people were asked to report their subjective health every third year on a three-category scale over a period of 9 years. Table 1, Table 2 and Table 3 give data. Categories denote: A:Good health, B:Neither good nor bad health, C:Bad health.

Table 1. Subjective health from 1962 to 1965

		1965			
		A	**B**	**C**	**Total**
1962	A	168	51	9	228
	B	42	73	23	138
	C	5	17	23	45
	Total	215	141	55	411

Table 2. Subjective Health from 1965 to 1968

		1968			
		A	**B**	**C**	**Total**
1965	A	178	32	4	214
	B	55	71	15	141
	C	7	30	18	55
	Total	240	133	37	410

Table 3. Subjective Health from 1968 to 1971

		1971			
		A	**B**	**C**	**Total**
1968	A	170	56	13	239
	B	32	79	22	133
	C	4	14	18	36
	Total	206	149	53	408

Transition matrix that is usually of great interest is the time until a subject takes the worse state. it is assumed that the transition matrix will be estimated from longitudinal data.

The likelihood ratio statistic for stationarity hypothesis is found as 24.1863 with 12 degrees of freedom. This value is significant at the %1 level and we do not reject the hypothesis of constant transition probabilities, which can be interpreted that people's subjective health do not change over time.

As the transition probabilities are stationary, then the estimated transition probabilities will be based on the aggregrated tables as in Table 4.

Table 4. Estimated transition frequencies and probabilities for frequencies in Tables 1-3

		***i*+1**			
		A	**B**	**C**	**Total**
i	A	516 (0.7577)	139 (0.204)	26 (0.038)	681
	B	129 (0.313)	223 (0.541)	60 (0.145)	412
	C	16 (0.117)	61 (0.448)	59 (0.433)	136

It is clear that the proportions in Table 4 relatively differ from each other.

3. RESULTS

A C# program was developed to test the transition probabilities are constant and to find at which point the process the equabrilium. The program is organized as first reading the data as matrices of successive event transitions. The program is available from the author upon request.

Matrix entry window for Tables 1-3 is given in Fig. 1.

Each matrix can be entried after definition of dimension and number of matrices. Here we enter three matrices at the same time.

Output window is displayed in Fig. 2. After definition the matrices seperately, the results of chi-square, associated degrees of freedom and P-value can be easly obtained for testing the null hypothesis given in Equation (2). As it is said earlier, this value is significant at the %1 level and we do not reject the hypothesis of constant transition probabilities. This means that people's subjective health do not change over time.

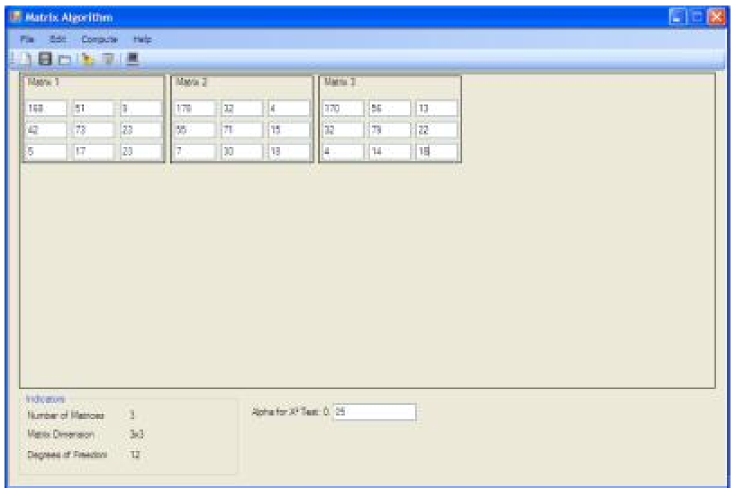

Fig. 1. Illustration of data entry window

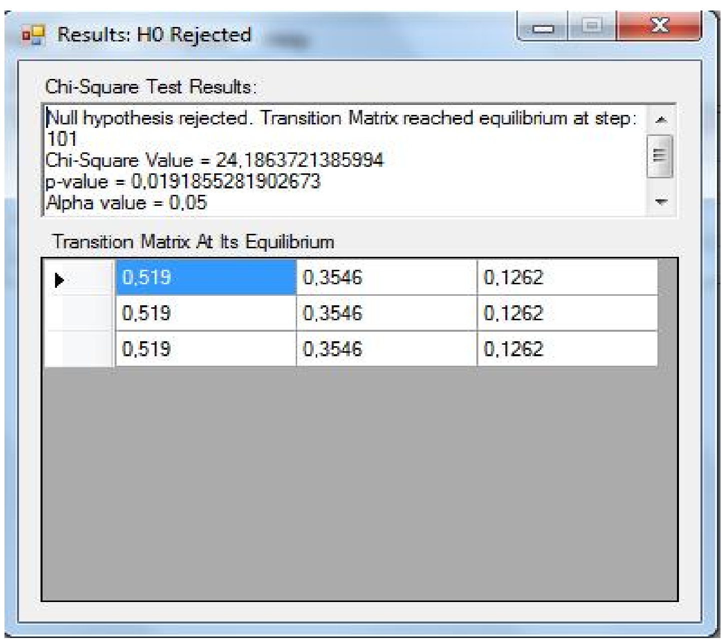

Fig. 2. Output window

Program also investigates at which step the population will reach the equilibrium by the following transition matrices.

Let π_j denotes the long run proportion of time that the chain stays in state *j*. Hence we can find the limit distribution using $\pi = (\pi_1\ \pi_2\ \pi_3)$.

We obtained the probabilities as, π_1=0.5190, π_2=0.3546, π_3=0.1262.

where, π_1=Probability of good health ; π_2=Probability of neither good nor bad health and π_3=Probability of bad health for $T \to \infty$.

We also investigated at which step the population will reach the equilibrium by the following transition matrices. We have seen from the results that the population reached the equilibrium at almost 101st step.

$$\underline{P} = \begin{bmatrix} 0.7577 & 0.2041 & 0.0382 \\ 0.3131 & 0.5413 & 0.1456 \\ 0.1212 & 0.4621 & 0.4167 \end{bmatrix}, \quad \underline{P}^2 = \begin{bmatrix} 0.6457 & 0.2958 & 0.0760 \\ 0.4244 & 0.4273 & 0.1514 \\ 0.2870 & 0.4686 & 0.2455 \end{bmatrix},$$

$$\underline{P}^3 = \begin{bmatrix} 0.5911 & 0.3335 & 0.0994 \\ 0.4736 & 0.3921 & 0.1415 \\ 0.3939 & 0.4285 & 0.1815 \end{bmatrix} \cdots \underline{P}^{101} = \begin{bmatrix} 0.5190 & 0.3546 & 0.1262 \\ 0.5190 & 0.3546 & 0.1262 \\ 0.5190 & 0.3546 & 0.1262 \end{bmatrix}.$$

We can see from the results that the population reached the equilibrium at the 101st step as 0.52, 0.35 and 0.13.

For instance, only for Table 1, we get (π_1, π_2, π_3)=(0.4813, 0,3578, 0.1610) as given in Fig. 3.

Output shows the population is in equilibrium as regards subjective health for data in Table 1. It seems that the marginal distributions are related to the marginal homogeneity model in a square contingency table.

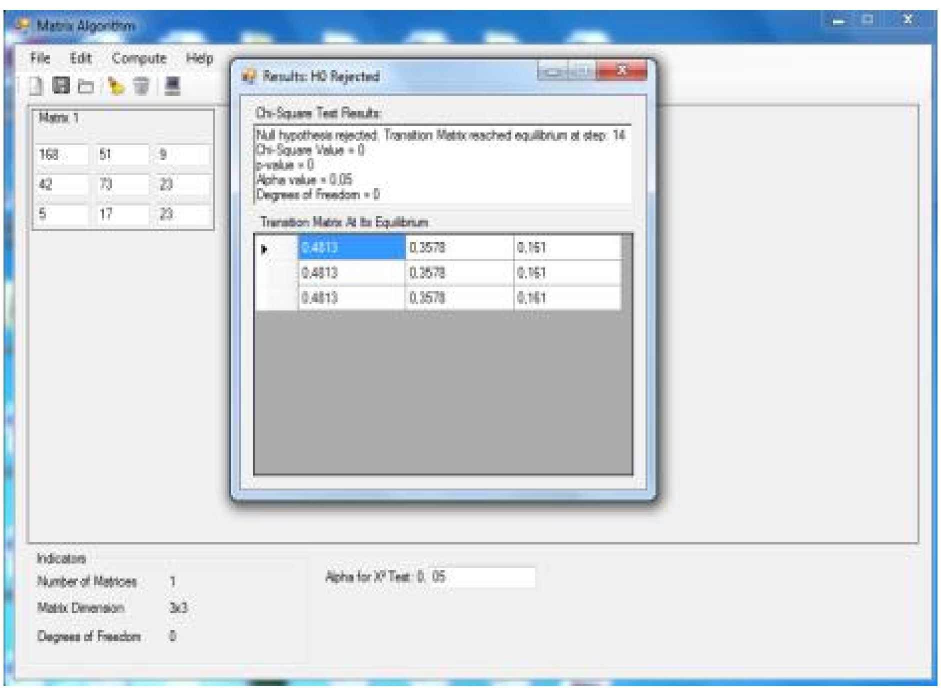

Fig. 3. Equilibrium Output

4. CONCLUSIONS

Markov chain models are used in many applied field such as time series analysis, longitudinal studies.

In time time dependent studies, individuals are measured over time. We frequently deal with random variables that depend on time in some social, demographic or health applications. In such cases, fundemental concept is the probability of changing from one state in time *t*, to another state in *t*+1. For instance, political opinion of the voters, consumer behaviours, health status of the patients and so on. In such cases, transition probabilities over time are of interest. Markov chain models were developed based on marginal probabilities by considering repeated measures data [8]. We test the hypothesis the matrix of transition probabilities is constant over time. A fundamental question in this context is whether the transition probabilities can be assumed to be constant in time or not. A Markov process is stationary if the transition probabilities $p_{ij}(t)$ is independent of *t*. In order to collect data for Markov chain, we should observe the population at equally spaced time points.

A first approach to analyze the time-stability of transition probabilities is to compare the estimated transition probabilities per period for *T* periods with estimates from pooled data. A Markov chain with constant transition matrices is a homogenous chain. When the transition probabilities are homogenous, we can aggregate the multiway table into a single table. Using this single table, the transition probabilities, the limit distribution and the parameter estimates can be estimated. Markov models are recently appropriate for the analysis of longitudinal studies [9-12]. Relevant inferences can be made such as for brand loyality, social mobility, health status and i.e surveys.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Sirdari MZ,Islam MA, Awang N. A stationary test on Markov chain models based on marginal distribution. Proceedings of the 6th IMT-GT Conference on Mathematics, Statistics and its Applications, Kuala Lumper, Malaysia; 2010.
2. Weissbach R, Walter R. A likelihood ratio test for stationarity of rating transitions. Journal of Econometrics. 2010;155(2): 188-194.
3. Karlin S, Taylor HM. A first course in stochastic processes. Academic Press, second edition; 1975.
4. Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis: Theory and practice. MIT Press, Cambridge; 1975.
5. Aktas S, Inal C. Discrete time Markov chains for square contingency tables. International Journal of Statistics and Economics. 2010;4(S10):74-83.
6. Agresti A. Categorical data analysis. John Wiley&Sons, New-York; 2002.
7. Anderson T, Goodman L. Statistical ınference about Markov chains. Annals of Mathematical Statistics. 1957;28:89-110.
8. Andersen E. Discrete statistical models with social science applications. North Holland Publishing Com., Amsterdam; 1980.
9. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics 2002;58:342–351.
10. Albert PS. A Markov model for sequences of ordinal data from a relapsing-remitting disease. Biometrics. 1994;50:51–60.
11. Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. Journal of the American Statistical Association. 1985;80:863–871.
12. Mandel M, Gauthier SA, Guttmann CR, Weiner HL, Betensky RA. Estimating time to event from longitudinal categorical data: An analysis of multiple sclerosis progression. J Am Stat Assoc. 2007;102(480):1254-1266.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/10325