PSI-SQUARED STATISTICS FOR MARKOV SEQUENCES AND THEIR RELATIVE EFFICIENCIES

By P. V. Krishna Iyer* and P. Samarasimhudu

Defence Science Laboratory, Delhi

Abstract

This paper gives the first two asymptotic cumulants for different states and transitions between successive observations of the first-order Markov chains. Using these results, the first two asymptotic cumulants of the statistics ψ_1^2 and ψ_2^2 have been obtained and their relative efficiencies have been discussed for large n and for some values of k (where n is the number of observations and k is the number of different states). It is pointed out that ψ_2^2 can be used for testing the homogeneity of k samples.

Introduction

A sequence of observations $x_1, x_2, \ldots, x_n$ is Markovian if the probability for x_i depends on the previous observation x_{i-1} and is governed by the conditional probability matrix (p_{ij}), where p_{ij} is the probability for an observation to be in state j when its preceding one is in state i. Studies of such sequences enable us to develop methods of analysis suited for long-term series in which neighbouring observations are correlated. They also provide methods for examining the relative efficiency for various statistical procedures for testing two or more samples with respect to the alternate hypothesis wherein the observations are not independent. It may be observed in this connection that most of the studies undertaken, so far, are based on the assumption that all the observations are independent of each other. The efficiency of these tests, when the observations are not independent, is not known.

In view of the basic nature of such studies, a good deal of work has been carried out by a number of workers on Markov sequences

* The author is now in the Department of Mathematics, Punjab University, Chandigarh.

and chains. Recently, an excellent review of the investigations on this topic has been given by Billingsley.[3] Good[4] and Bhat[2] have tried to generalise some of the results obtained by Bartlett,[1] Whittle[6] and others. The results given by them, besides being approximate, are not in a form that can be readily used for practical purposes. We have given in this paper the first two cumulants for (i) the number of states; (ii) the number of first order transitions; (iii) ψ^2 statistics for a first order Markov chain. The results are utilised for testing the relative efficiency of Psi-squared statistics.

II. Cumulants for First Order Chains

(1) *Different States.*—Let X_i be the number of observations in state i in the sequence. Then

$$X_i = x_1 + x_2 + \ldots + x_n$$

where

$$x_r \begin{cases} = 1 \text{ if } x_r \text{ is in state } i \\ = 0 \text{ otherwise.} \end{cases}$$

$$E(X_i) = E(x_1 + x_2, \ldots + x_n)$$
$$= n\,E(x_r)$$
$$= np_i \text{ where } p_i \text{ is the asymptotic probability for an observation to be in state } i. \qquad (2.1.1)$$

$$E(X_i^2) = nE(x_r^2) + 2\sum_{s=1}^{n-1}(n-s)\,E(x_r x_{r+s}),$$

$E(x_r x_{r+s})$ = The probability for the r-th and $(r+s)$-th observations to be in state i, when nothing is known of the observations between r-th and $(r+s)$-th observations.

$$= p_i p_{ii}^{(s)},$$ where $p_{ii}^{(s)}$ is the conditional probability for $(r+s)$-th observation to be in state i when the r-th observation is in state i.

$$\kappa_2(X_i) = np_i(1-p_i) + 2\sum_{s=1}^{n-1}(n-s)(p_{ii}^{(s)} - p_i)$$

$$= np_i\left[q_i + 2\sum_{s=1}^{n-1}\left(1 - \frac{s}{n}\right)P_{ii}^{(s)}\right] \qquad (2.1.2)$$

where

$$P_{ii}^{(s)} = (p_{ii}^{(s)} - p_i)$$

As $n \to \infty$ this reduces to the form

$$\kappa_2(X_i) \approx np_i\left[q_i + 2\frac{1-E^{n-1}}{1-E}P_{ii}^{(1)}\right] \tag{2.1.3}$$

where

$$E^s P_{ii}^{(1)} = P_{ii}^{(s+1)}$$

i.e., E refers to the usual operator in finite differences.

$$\kappa_{11}(X_i, X_j) \approx -np_ip_j + n(p_i\alpha_{ij}' + p_{jj}\alpha'_{ij}) \tag{2.1.4}$$

where

$$\alpha_{ij}' = \sum_{s=1}^{n-1}\left(1 - \frac{s}{n}\right)(p_{ij}^{(s)} - p_j) \text{ for } i, j = a, b, \ldots, k$$

where $a, b, \ldots, k$ refer to different states.

(2) *Transitions like* (ii) *and* (ij) *between successive observations.*— Let X_{ii} be the number of (ii) transitions between successive observations in the sequence.

Then

$$X_{ii} = x_{12} + x_{23} + \ldots + x_{n-1,\ n},$$

where

$x_{r,\ r+1}$ represents the transition between r and $r+1$-th observations, and

$$x_{r,\ r+1} \begin{cases} = 1 \text{ if the } r\text{-th and } r+1\text{-th observations are in state } i. \\ = 0 \text{ otherwise.} \end{cases}$$

$$\begin{aligned} E(X_{ii}) &= (n-1)E(x_{r,\ r+1}) \\ &= (n-1)p_ip_{ii} \\ &\approx nR_{ii}, \qquad \text{where} \quad R_{ii} = p_ip_{ii} \end{aligned} \tag{2.2.1}$$

Evaluating similarly, we get

$$\kappa_2(X_{ii}) \approx nR_{ii}\left[1 - 3R_{ii} + 2p_{ii} + 2p_{ii}\sum_{s=1}^{n-3}\left(1 - \frac{s}{n}\right)P_{ii}^{(s)}\right] \tag{2.2.2}$$

$$\kappa_1(X_{ij}) = nR_{ij}, \qquad \text{where} \quad R_{ij} = p_ip_{ij} \tag{2.2.3}$$

and

$$\kappa_2(X_{ij}) \approx nR_{ij}\left[1 - 3R_{ij} + 2p_{ij}\sum_{s=1}^{n-3}\left(1 - \frac{s}{n}\right)P_{jj}^{(s)}\right] \quad (2.2.4)$$

$$\kappa_{11}(X_{ii}, X_{jj}) \approx -3nR_{ii}R_{jj} + np_{ii}p_{jj}(p_i\alpha_{ij} + p_j\alpha_{ji}) \quad (2.2.5)$$

where

$$\alpha_{ij} = \sum_{s=1}^{n-3}\left(1 - \frac{s}{n}\right)(p_{ij}^{(s)} - p_j) \quad \text{for} \quad i, j = a, b, \ldots, k.$$

$$\kappa_{11}(X_{ii}, X_{ij}) \approx -3nR_{ii}R_{ij} + nR_{ii}p_{ij} + np_{ii}p_{ij}(p_i\alpha_{ii} + p_i\alpha_{ji}). \quad (2.2.6)$$

$$\kappa_{11}(X_{ii}, X_{jm}) \approx -3nR_{ii}R_{jm} + np_{ii}p_{jm}(p_i\alpha_{ij} + p_j\alpha_{mi}) \quad (2.2.7)$$

$$\kappa_{11}(X_{ij}, X_{jm}) \approx -3nR_{ij}R_{jm} + nR_{ij}p_{jm} + np_{ij}p_{jm}(p_i\alpha_{jj} + p_j\alpha_{mi}) \quad (2.2.8)$$

$$\kappa_{11}(X_{ij}, X_{lm}) \approx -3nR_{ij}R_{lm} + np_{ij}p_{lm}(p_i\alpha_{jl} + p_i\alpha_{mi}) \quad (2.2.9)$$

$$\kappa_{11}(X_{ij}, X_{ji}) \approx -3nR_{ii}R_{ji} + np_{ij}p_{ji}[p_i(1 + \alpha_{jj}) + p_j(1 + \alpha_{ii})] \quad (2.2.10)$$

$$\kappa_{11}(X_{ii}, X_{ji}) \approx -3nR_{ii}R_{ji} + nR_{ji}p_{ii} + np_{ii}p_{ji} \times (p_i\alpha_{ij} + p_j\alpha_{ii}) \quad (2.2.11)$$

3. Psi-Squared Statistics for First Order Chains

Using the various cumulants given earlier, the first two asymptotic moments of ψ_1^2 and ψ_2^2 defined by

$$\psi_1^2 = \sum_i \frac{[X_i - E(X_i)]^2}{E(X_1)} \qquad i = a, b, \ldots, k$$

and

$$\psi_2^2 = \sum_{i,j} \frac{(X_{ij} - E(X_{ij})]^2}{E(X_{ij})}, \quad i, j = a, b, \ldots k$$

have been calculated and given below.

$$\kappa_1(\psi_1^2) \approx \left[(k-1) + 2\sum_i \sum_{s=1}^{n-1}\left(1 - \frac{s}{n}\right)P_{ii}^{(s)}\right] = A_1 \quad \text{say} \quad (3.1)$$

$$\kappa_2(\psi_1^2) \approx 2\left[(k-1) + 4\sum_i \sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)P_{ii}^{(s)} + 4\sum_i \sum_{s_1,\,s_2=1}^{n-1}\left(1-\frac{s_1}{n}\right)\left(1-\frac{s_2}{n}\right)P_{ii}^{(s_1+s_2)}\right] = 2B_1 \quad \text{say} \tag{3.2}$$

$$\kappa_1(\psi_2^2) \approx (k^2-3) + 2\sum_i p_{ii} + 2\sum_i \sum_{s=1}^{n-3}\left(1-\frac{s}{n}\right)P_{ii}^{(s+1)} = A_2 \quad \text{say} \tag{3.3}$$

$$\kappa_2(\psi_2^2) \approx 2\left[k^2+2k-9+4\sum_i p_{ii} + 2\sum_{i,j} p_{ij}p_{ji} + 4\sum_i \sum_{s=1}^{n-3}\left(1-\frac{s}{n}\right)\{P_{ii}^{(s)} + P_{ii}^{(s+1)} + P_{ii}^{(s+2)}\} + 4\sum_i \sum_{s_1,\,s_2=1}^{n-3}\left(1-\frac{s_1}{n}\right)\left(1-\frac{s_2}{n}\right)P_{ii}^{s_1+s_2+2}\right] = 2B_2 \quad \text{say} \qquad i,j = a, b, \ldots k. \tag{3.4}$$

By examining the first two cumulants of ψ_1^2 and ψ_2^2 with those modified χ^2, discussed by Patankar,[5] we find that the asymptotic distributions of $A_i\psi_i^2/B_i$ $(i = 1, 2)$ can be approximated to the χ^2 distribution with A_i^2/B_i degrees of freedom.

4. Relative Efficiency of ψ_1^2 and ψ_2^2

The statistics ψ_1^2 and ψ_2^2 are such that they can be used for testing the hypothesis of randomness regarding a sequence of observations. To decide which of these is more suited for different situations, the relative powers of ψ_1^2 and ψ_2^2 for different H_0 and H_1, have been calculated on the basis of 5% level of significance by comparing the corresponding ordinary χ^2's with degrees of freedom $(k-1)$ and (k^2-1) respectively and are given in Table 1. Thus, for example, in the case of $k = 6$, power of ψ_1^2 is calculated as folllows,

5% value of $\chi^2_{(k-1)}$ for the hypothesis $H_0 = 11{\cdot}070$. Then we calculate the values of u and p to enable us to use Pearson's table[7] in evaluating the area $I(u, p)$,

$$u \text{ for } \psi_1^2 = \frac{11{\cdot}070}{\sqrt{2B_1}} = 2{\cdot}7139$$

and $p = (\frac{1}{2}$ of degrees of freedom$) - 1$

$$= \frac{A_1^2}{2B_1} - 1 = 1{\cdot}4695.$$

Then $I(u, p)$ comes out to be $0{\cdot}8744$ and $\{1 - I(u, p)\}$ gives the power of the statistics ψ_1^2. Similarly, for ψ_2^2 also, by replacing $\chi^2_{(k-1)}$ with $\chi^2_{(k^2-1)}$ and A_1, B_1 with A_2, B_2 and proceeding in the same way, we get the power of ψ_2^2 to be $0{\cdot}1040$.

It will be seen from Tables I and II that the statistics ψ_1^2 is more powerful than ψ_2^2 for testing H_0 against H_1 when $\sum_i p_{ii} > 1$, and when $\sum_i p_{ii} \leqslant 1$, ψ_2^2 is more powerful than ψ_1^2. Therefore it should be decided on the basis of $\sum_i p_{ii}$ as to which of the statistics ψ_1^2 and ψ_2^2 should be used for testing randomness. Further investigations are required to decide about the relative efficiency of ψ_3^2, ψ_4^2, etc., which are on hand and will be published in due course.

5. Large Sample Test for Testing Homogeneity of k Samples

Before concluding, it may be pointed out that the statistic ψ_2^2 can be used for testing homogeneity of k samples by extending Wald and Wolfowitz's procedure for testing two samples.

Suppose there are k samples $a, b, c, \ldots k$ consisting of $n_1, n_2, \ldots n_k$ observations respectively. The k samples are pooled together and arranged in ascending or descending order, noting down the sample to which each of them belongs. Thus we will be obtaining a sequence of the form (acdekca....). The above sequence can be tested for randomness for the characters a, b, c, etc., with the help of ψ_2^2. For the null case, it will be assumed that

$$p_a = n_1/S, \quad p_b = n_2/S \ldots, \quad \text{where} \quad S = n_1 + n_2 + \ldots + n_k.$$

It may be noted that for this assumption the expected value and variance of ψ_2^2 can be approximated to χ^2 of a $k \times k$ contingency table showing the frequencies for various transitions, *i.e.*, $(k-1)^2$ and $2(k-1)^2$ respectively. Instead of assuming $p_r = n_r/S$ $(r = a, b, \ldots k)$, $p_a, p_b, \ldots,$

TABLE I*

Powers of ψ_1^2 and ψ_2^2 for some alternatives and for large values of n.

$\alpha = 0{\cdot}05 \quad \Sigma p_{ii} > 1$

k	Hypothesis	Power	
		ψ_1^2	ψ_2^2
3	$H_0: P_i = P_j, \ \sum_i p_{ii} = 1. \quad i, j = 1, 2, 3$		
	$H_1:$ $\begin{pmatrix} 0{\cdot}625 & 0{\cdot}250 & 0{\cdot}250 \\ 0{\cdot}250 & 0{\cdot}500 & 0{\cdot}375 \\ 0{\cdot}125 & 0{\cdot}250 & 0{\cdot}375 \end{pmatrix}$	0·1785	0·1486
4	$H_0: P_i = P_j, \quad \sum_i p_{ii} = 1. \quad i, j = 1, 2, 3, 4$		
	$H_1:$ $\begin{pmatrix} 0{\cdot}500 & 0{\cdot}200 & 0{\cdot}250 & 0{\cdot}300 \\ 0{\cdot}250 & 0{\cdot}450 & 0{\cdot}200 & 0{\cdot}125 \\ 0{\cdot}125 & 0{\cdot}250 & 0{\cdot}300 & 0{\cdot}200 \\ 0{\cdot}125 & 0{\cdot}100 & 0{\cdot}250 & 0{\cdot}375 \end{pmatrix}$	0·1743	0·1349
5	$H_0: P_i = P_j, \ \sum_i p_{ii} = 1. \quad i, j = 1, 2, 3, 4, 5.$		
	$H_1:$ $\begin{pmatrix} 0{\cdot}60 & 0{\cdot}10 & 0{\cdot}20 & 0{\cdot}25 & 0{\cdot}21 \\ 0{\cdot}05 & 0{\cdot}50 & 0{\cdot}15 & 0{\cdot}11 & 0{\cdot}18 \\ 0{\cdot}10 & 0{\cdot}15 & 0{\cdot}45 & 0{\cdot}17 & 0{\cdot}13 \\ 0{\cdot}16 & 0{\cdot}10 & 0{\cdot}06 & 0{\cdot}40 & 0{\cdot}18 \\ 0{\cdot}09 & 0{\cdot}15 & 0{\cdot}14 & 0{\cdot}07 & 0{\cdot}30 \end{pmatrix}$	0·3054	0·1892
6	$H_0: P_i = P_j, \ \sum_i p_{ii} = 1. \ i, j = 1, 2, \ldots 6$		
	$\begin{pmatrix} 0{\cdot}20 & 0{\cdot}15 & 0{\cdot}10 & 0{\cdot}09 & 0{\cdot}16 & 0{\cdot}14 \\ 0{\cdot}15 & 0{\cdot}20 & 0{\cdot}15 & 0{\cdot}10 & 0{\cdot}14 & 0{\cdot}12 \\ 0{\cdot}15 & 0{\cdot}16 & 0{\cdot}32 & 0{\cdot}14 & 0{\cdot}15 & 0{\cdot}18 \\ 0{\cdot}15 & 0{\cdot}11 & 0{\cdot}13 & 0{\cdot}25 & 0{\cdot}10 & 0{\cdot}16 \\ 0{\cdot}16 & 0{\cdot}21 & 0{\cdot}14 & 0{\cdot}22 & 0{\cdot}33 & 0{\cdot}10 \\ 0{\cdot}19 & 0{\cdot}17 & 0{\cdot}16 & 0{\cdot}20 & 0{\cdot}12 & 0{\cdot}30 \end{pmatrix}$	0·1256	0·1040

TABLE II*

Powers of ψ_1^2 and ψ_2^2 for some alternatives and for large values of n

$\alpha = 0{\cdot}05 \quad \sum_i p_{ii} \leqslant 1$

k	Hypothesis	Power	
		ψ_1^2	ψ_2^2
3	H_0: $P_i = P_j$, $\sum_i p_{ii} = 1$. $i, j = 1, 2, 3$		
	H_1: $\begin{pmatrix} 0{\cdot}40 & 0{\cdot}50 & 0{\cdot}45 \\ 0{\cdot}35 & 0{\cdot}20 & 0{\cdot}45 \\ 0{\cdot}25 & 0{\cdot}30 & 0{\cdot}10 \end{pmatrix}$	0·0190	0·0583
3	H_0: $p_i = p_j$ $\sum_i p_{ii} = 1$, $i, j = 1, 2, 3$		
	H_1: $\begin{pmatrix} 0{\cdot}40 & 0{\cdot}25 & 0{\cdot}40 \\ 0{\cdot}30 & 0{\cdot}35 & 0{\cdot}35 \\ 0{\cdot}30 & 0{\cdot}40 & 0{\cdot}25 \end{pmatrix}$	0·0539	0·0812
4	H_0: $p_i = p_j$ $\sum_i p_{ii} = 1$, $i, j = 1, 2, 3, 4$		
	H_1: $\begin{pmatrix} 0{\cdot}20 & 0{\cdot}25 & 0{\cdot}30 & 0{\cdot}40 \\ 0{\cdot}15 & 0{\cdot}20 & 0{\cdot}25 & 0{\cdot}30 \\ 0{\cdot}30 & 0{\cdot}35 & 0{\cdot}15 & 0{\cdot}20 \\ 0{\cdot}35 & 0{\cdot}20 & 0{\cdot}30 & 0{\cdot}10 \end{pmatrix}$	0·0224	0·0591
5	H_0: $p_i = p_j$ $\sum_i p_{ii} = 1$, $i, j = 1, 2, 3, 4, 5$		
	$\begin{pmatrix} 0{\cdot}20 & 0{\cdot}25 & 0{\cdot}15 & 0{\cdot}18 & 0{\cdot}25 \\ 0{\cdot}16 & 0{\cdot}15 & 0{\cdot}30 & 0{\cdot}25 & 0{\cdot}15 \\ 0{\cdot}15 & 0{\cdot}20 & 0{\cdot}20 & 0{\cdot}30 & 0{\cdot}15 \\ 0{\cdot}30 & 0{\cdot}15 & 0{\cdot}10 & 0{\cdot}12 & 0{\cdot}35 \\ 0{\cdot}19 & 0{\cdot}25 & 0{\cdot}25 & 0{\cdot}15 & 0{\cdot}10 \end{pmatrix}$	0·0285	0·0670

* In calculating the cumulants, terms upto $\sum_i P_{ii}^{(4)}$ have been used, since the terms beyond these are of no consequence in these cases.

may be estimated from the matrix $[\delta_{ij} - p_{ij}]$, where $p_{ij} = n_{ij}/S$, n_i^f being the number of ij transitions and

$$\delta_{ij} \begin{cases} = 1 & \text{if } i = j \\ = 0 & \text{otherwise.} \end{cases}$$

In that case $p_a = \Delta_{aa}/\Sigma\Delta_{rr}$, where Δ_{aa} is the first minor of $(1 - p_{aa})$ of the determinant $|\delta_{ij} - p_{ij}|$. If we use the above values of p_a, p_b ..., the expected value of ψ_2^2 and its variance can be approximated to

$$\kappa_1(\psi_2^2) = k(k-1)$$

$$\kappa_2(\psi_2^2) = 2(k+1)(k-1)$$

and the test may be carried out by using these values for the null case. Further work is necessary to decide its power in relation to usual variance ratio test for testing the homogeneity of k samples.

6. Acknowledgements

Our sincere thanks are due to Prof. Bartlett, F.R.S., for pointing out some of the discrepancies in the earlier draft of this paper.

7. References

1. Bartlett, M. S. .. *Proc. Camb. Phil. Soc.*, 1951, **47**, 86–95.
2. Bhat, B. R. .. *Ann. Math. Stat.*, 1961, **32**, 59.
3. Billingsley, P. .. *Ibid.*, 1961, **32**, 12.
4. Good, I. J. .. *Ibid.*, 1961, **32**, 41.
5. Patankar, V. N. .. *Biometrika*, 1954, **41**, 450.
6. Whittle, P. .. *J. Roy. Stat. Soc.*, 1955, **17 A**, 235.
7. Pearson, K. .. Tables of the Incomplete Γ-Function, Re-issue, 1951.