Journal of Statistical Planning and Inference 143 (2013) 321–333

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Central Limit Theorem approximations for the number of runs in Markov-dependent binary sequences

George C. Mytalas, Michael A. Zazanis *

Department of Statistics, Athens University of Economics and Business, Athens 10434, Greece

ARTICLE INFO

Article history:
Received 21 October 2010
Received in revised form
15 July 2012
Accepted 18 July 2012
Available online 25 July 2012

Keywords:
Runs
Markov chains
Potential matrix
Central Limit Theorem for runs

ABSTRACT

We consider Markov-dependent binary sequences and study various types of success runs (overlapping, non-overlapping, exact, etc.) by examining additive functionals based on state visits and transitions in an appropriate Markov chain. We establish a multivariate Central Limit Theorem for the number of these types of runs and obtain its covariance matrix by means of the recurrent potential matrix of the Markov chain. Explicit expressions for the covariance matrix are given in the Bernoulli and a simple Markov-dependent case by expressing the recurrent potential matrix in terms of the stationary distribution and the mean transition times in the chain. We also obtain a multivariate Central Limit Theorem for the joint number of non-overlapping runs of various sizes and give its covariance matrix in explicit form for Markov dependent trials.

1. Introduction

In a sequence of binary trials, $\{\xi_i\}$ with $\xi_i \in \{0,1\}$, $i=1,2,\ldots,$ (success=1 or failure=0) a success run of length k is the occurrence of k consecutive successes. We will not discuss here the vast array of applications of the analysis of runs and patterns in Statistics and Applied Probability; for this we refer the reader to Balakrishnan and Koutras (2002) and Fu and Lou (2003).

Given a realization of n trials there are several different ways of counting the number of success runs of length k depending, among other considerations, on whether overlapping in counting is allowed or not. A success run of length k occurs at position $m \geq k$ of the binary string $\xi_1,\xi_2,\ldots,\xi_n$ if

$$\xi_{m-k+1} = \xi_{m-k+2} = \cdots = \xi_{m-1} = \xi_m = 1. \tag{1}$$

Counting all the positions $m=k,k=1,\ldots,n$ for which the above condition holds gives the count $M_{n,k}$ of the number of success runs with overlapping allowed. A *non-overlapping success run of length* k occurs at position m if (1) holds and no non-overlapping run of length k has already occurred in positions $m-k+1$, $m-k+2,\ldots,m-1$. The number of non overlapping runs of length k in a string of n trials is denoted by $N_{n,k}$. Runs of exact size k are also of interest, i.e. k consecutive successes flanked on the left and right by failures. An *exact* run of size k occurs at position $m \geq k$ if, in addition to (1), $\xi_{m-k}=0$ and $\xi_{m+1}=0$. We denote the number of exact runs of size k in a string of n trials by $J_{n,k}$. Note that, when the values of the ξ_i's are revealed sequentially, an exact run that occurs at position m will be counted when the value of ξ_{m+1} becomes known. Finally, we say that a run of size *greater than or equal to* k occurs at position m if (1) holds and $\xi_{m-k}=0$.

* Corresponding author. Tel.: +30 2108203523.
E-mail address: zazanis@aueb.gr (M.A. Zazanis).

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.07.014

$G_{n,k}$, denotes the number of success runs of size greater than or equal to k in a string of n trials. In a string of consecutive successes of length greater than k only one run of size greater than k is counted according to this definition. See Fu and Lou (2003) for further clarifications regarding these definitions. To illustrate, if $n=13$ and $k=2$, in the binary string 0111101110110 we have the following success run counts: $N_{13,2}=4$, $G_{13,2}=3$, $M_{13,2}=6$, and $J_{13,2}=1$.

In this paper we consider Markov dependent binary trials. We investigate the asymptotic form of the *joint distribution* of $(N_{n,k},M_{n,k},G_{n,k},J_{n,k})$, as $n\to\infty$ and k is fixed, and show that it obeys a multivariate Central Limit Theorem (CLT). For this purpose we consider an appropriate Markov chain and we express the number of success runs of the kind mentioned as additive functionals based on state visits and state transitions for this chain. The multivariate CLT follows then from standard results for such functionals. The covariance matrix of the limiting normal distribution is expressed in terms of the stationary distribution of the Markov chain and its *recurrent potential matrix* (also known as the fundamental matrix). This allows for the efficient numerical computation of the covariance matrix for quite general types of Markovian dependence of the binary trial sequence. In the special case of simple two-state Markovian dependence (and of course in the case of Bernoulli trials) we take advantage of the connection that exists between the potential matrix of the chain and the mean transition times between the states of the chain. By straightforward, if somewhat involved, calculations we are able to obtain explicit expressions for the potential matrix, and therefore for the covariance matrix itself, in terms of the parameters of the model. The same technique is used to obtain a multivariate CLT for the number of non-overlapping runs of different sizes, $(N_{n,k_1},\ldots,N_{n,k_v})$, in Markov dependent trials and to compute the corresponding covariance matrix. Besides its intrinsic interest and applications in areas including quality control, randomness tests, and reliability, we indicate applications of this result to the analysis of manufacturing systems with random yield.

Many results exist on the exact distribution of these types of runs for Bernoulli trials and, in some cases, also for Markov-dependent trials. Important unified approaches based on Markov chain techniques include Stefanov and Pakes (1997) where exact results are obtained for runs and more general patterns and Fu and Koutras (1994) who give the distributions of the run statistics $N_{n,k},M_{n,k},G_{n,k},J_{n,k}$, together with the size of the longest run for non-homogeneous Bernoulli trials. There are also approximations based on limit theorems that establish convergence to Poisson or compound Poisson limits in certain asymptotic regimes and which are especially important in applications. For an overview of these approximations see Barbour and Chryssaphinou (2001).

CLT approximations for the number of runs have a long history. Despite their limited accuracy they remain an important theoretical and practical tool in the analysis of runs. Feller (1968) using arguments based on the Central Limit Theorem for renewal processes, gave an normal approximation for the number of non-overlapping success runs in i.i.d. trials. Setting

$$\mu=\frac{1-p^k}{qp^k}\quad\text{and}\quad\sigma^2=\frac{qp^k-(2k+1)(qp^k)^2-qp^{3k+1}}{(1-p^k)^3},\tag{2}$$

he shows that $(N_{n,k}-n\mu)/\sigma\sqrt{n}\xrightarrow{d}\mathcal{N}(0,1)$. The same approach is essentially used in Fu et al. (2002) in order to obtain the limiting distribution of the number of successes in success runs of length greater than or equal k in a sequence of Markov dependent binary trials. Fu and Lou (2007) obtain in the same fashion a CLT approximation for the number of non-overlapping occurrences of a simple or compound pattern in i.i.d. multi-type trials.

A different approach towards establishing the asymptotic normality of the number of runs that has been widely used is via the Hoeffding–Robbins Central Limit Theorem for k-dependent random variables. A CLT for $M_{n,k}$ with i.i.d. Bernoulli trials was obtained along these lines by Godbole (1992) expressing $M_{n,k}$ as a sum of stationary $(k-1)$-dependent indicators. Using a similar approach Hirano et al. (1991) gave explicit results establishing that $(M_{n,k}-(n-k+1)p^k)/\sigma\sqrt{n}\xrightarrow{d}N(0,1)$ where

$$\sigma^2=-p^k(1-p^k)-2kp^{2k}+\frac{2p^k(1-p^k)}{q}.\tag{3}$$

Jennen-Steinmetz and Gasser (1986) also use the CLT for k-dependent random variables in order to obtain a multivariate limiting normal distribution for success and failure runs of various lengths with Bernoulli trials that do not necessarily have the same probability of success but which obey certain asymptotic conditions. The same approach is used in Fu and Lou (2007) in order to obtain normal approximations for the number of overlapping occurrences of simple patterns in multi-type i.i.d. trials and in Makri and Psillakis (2011) which obtains a CLT approximation for $J_{n,k}$.

An alternative and far-reaching approach based on exponential families of random variables has been pioneered by Stefanov (1995) for the analysis of the number of occurrences of runs and patterns in binary trials. By essentially constructing an exponential martingale from the transitions of an appropriate Markov chain, and a stopping time corresponding to the completion of the pattern in question, he is able to derive both exact distributional results and CLT approximations for the joint number of success runs of various sizes and to determine the corresponding limiting covariance matrix. We refer the reader also to Stefanov (2000) and Stefanov and Pakes (1997) for further details.

A very rich literature exists on the wider problem of the number of occurrences of patterns in strings of multi-type trials, typically independent or with Markovian dependence (see Reinert et al., 2000 for a review). In the study of the joint distribution of pattern frequencies in strings of multi-type trials in Rukhin (2007) the potential matrix of a Markov chain whose states are words of a given length is used explicitly in order to obtain the asymptotic covariance matrix in the corresponding CLT. The potential matrix is determined in that case via the *pattern correlation matrix* (as opposed to the results of this paper where it is determined using explicit computations for the mean transition times between states). We