Ann. Inst. Statist. Math.
Vol. 47, No. 4, 743-766 (1995)

RUNS, SCANS AND URN MODEL DISTRIBUTIONS: A UNIFIED MARKOV CHAIN APPROACH

M. V. KOUTRAS AND V. A. ALEXANDROU

Department of Mathematics, University of Athens, Panepistemiopolis, Athens 157 84, Greece

(Received June 6, 1994; revised December 2, 1994)

Abstract. This paper presents a unified approach for the study of the exact distribution (probability mass function, mean, generating functions) of three types of random variables: (a) variables related to success runs in a sequence of Bernoulli trials (b) scan statistics, i.e. variables enumerating the moving windows in a linearly ordered sequence of binary outcomes (success or failure) which contain prescribed number of successes and (c) success run statistics related to several well known urn models. Our approach is based on a Markov chain imbedding which permits the construction of probability vectors satisfying triangular recurrence relations. The results presented here cover not only the case of identical and independently distributed Bernoulli variables, but the non-identical case as well. An extension to models exhibiting Markov dependence among the successive trials is also discussed in brief.

Key words and phrases: Success runs, scan statistics, urn models, Markov chains, triangular multidimensional recurrence relations, distributions of order k.

1. Introduction

It is quite common for a statistician to face problems involving experimental trials with two possible outcomes. An educational psychologist evaluates subject's or material's efficiency by examining patterns of successes or failures in a learning process. An ecologist studies the spread of a specific disease by observing the patterns of infected or non-infected plants in a transect through a field. An acceptance sampling specialist develops plans based on sequences of acceptable or non-acceptable lots. A physician studies success and failure of treatments in therapeutic trials.

The statistical analysis of such phenomena seeks criteria for detecting changes in the underlying process generating the outcomes. Intuitively, the heavy congestion of outcomes of a specific type (for example "success") signals the occurrence of a change in the observed process.

A reasonable and intuitively appealing criterion for the analysis of the above mentioned situations, is the one based on the concept of (success) runs. In early forties it was used by Mood (1940) in the area of statistical hypothesis testing and by Mosteller (1941) and Wolfowitz (1943) in statistical control problems. Recently it has been successfully employed in a lot of diverse areas such as reliability (see Chao *et al.* (1995) or Papastavridis and Koutras (1994)), DNA sequencing (Arratia and Waterman (1985), Goldstein (1990)), psychology, ecology, radar astronomy (Schwager (1983)) etc.

There are various ways of counting runs. Consider a sequence of n Bernoulli trials $Z_1, Z_2, \ldots, Z_n$ with success (S) probabilities p_t and failure (F) probabilities q_t, $t = 1, 2, \ldots, n$. The number of non-overlapping and recurrent success runs of length k (k is a positive integer) was first introduced by Feller (1968) and is usually denoted by $N_{n,k}$. Another counting scheme proposed by Ling (1988) gives rise to the number $M_{n,k}$ of overlapping success runs of length k. Finally, of great statistical importance is also the number $G_{n,k}$ of success runs of length at least k (see Gibbons (1971)). Instead of giving the mathematical definition of the above mentioned variables, we mention the following illustrative example: if in a sequence of $n = 12$ trials, the outcomes were $SFSSSSFSSSFS$ then $N_{12,2} = 3$, $M_{12,2} = 5$, $G_{12,2} = 2$, $N_{12,3} = 2$, $M_{12,3} = 3$, $G_{12,3} = 2$.

The distributions of the random variables $N_{n,k}$, $M_{n,k}$ are known as *binomial distributions of order* k and have been studied extensively by Hirano *et al.* (1984), Aki (1985), Aki and Hirano (1988), Chryssaphinou *et al.* (1993), Godbole (1990*a*, 1991), Hirano *et al.* (1991), Philippou and Makri (1986) etc. Manifestly, the binomial distributions of order $k = 1$ coincide with the usual binomial probability, this fact being responsible for the *order* k nomenclature.

A natural generalization of the success-run criteria arises by interpreting as evidence of lack of randomness, the appearance of many k-tuples of consecutive trials containing among them large number (say greater than or equal to r) of successes. The respective random variables will be called *binomial scan statistics* or simply *scan statistics*. Problems leading to scan statistics may arise in the following practical context. Suppose data on the output of an assembly line is to be used for determining whether the production of defectives is a "contagious" phenomenon. A sample of n-units is examined and each defective (non-defective) item is marked as success (failure). Checking the sequence of outcomes for evidence of contagion amounts to making test for non-random clustering of S's relative to F's. A criterion that suggests itself in this context is the following: scan the sequence with an interval (window) of length k and mark all the windows containing at least r successes (defectives). If the total number of marked windows is "too large" reject the hypothesis that production of defectives is not contagious. Other applications of scan-statistics analysis pertain to phenomena such as clusters of disease in time, generalized birthday proximities and the nearest neighbour problems (see e.g. Dembo and Karlin (1992), Glaz (1989), Saperstein (1972, 1975)).

It is apparent that, several counting processes could be considered, leading to different statistics. For example, generalising the notion of overlapping success runs, we may denote by $M_{n,k,r}$ the number of overlapping k-tuples $\{i, i+1, \ldots, i+k-1\}$, $i = 1, 2, \ldots, n-k+1$ which contain at least r successes. There are also two non-overlapping analogues for $M_{n,k,r}$. The first of them, to be denoted by $N^{(1)}_{n,k,r}$, is computed by counting from scratch each time we encounter r successes placed within a "window" of length at most k. Alternatively, one could start counting anew only after the completion of a k-tuple of successive trials with at least r successes. This gives birth to a variable $N^{(2)}_{n,k,r}$. It is evident that $N^{(1)}_{n,k,k} = N^{(2)}_{n,k,k} = N_{n,k}$, $M_{n,k,k} = M_{k,k}$. To make the previous definitions clear and transparent, we mention in passing that in the sequence of outcomes *SFSFFSSSFFSFFSSF* we have $N^{(1)}_{16,4,2} = 4$, $N^{(2)}_{16,4,2} = 3$, $M_{16,4,2} = 9$.

Since the variables $N^{(1)}_{n,k,r}$, $N^{(2)}_{n,k,r}$, $M_{n,k,r}$ are created by a counting process performed in a scanning (moving) window, we use for them the name (*discrete*) *Scan Statistics* (see also Glaz and Naus (1991) and Wallenstein *et al.* (1994)). Several problems related to the continuous analogue of Scan Statistics can be found in Huntington (1978), Naus (1982) and references therein.

Currently, except for a few special cases, the exact distributions of the statistics $N^{(1)}_{n,k,r}$, $N^{(2)}_{n,k,r}$, $M_{n,k,r}$ are mainly unknown, especially for non-identical Bernoulli trials. The probabilities $\Pr(N^{(1)}_{n,k,r} = 0) = \Pr(N^{(2)}_{n,k,r} = 0) = \Pr(M_{n,k,r} = 0)$ are related to the well known generalized birthday problem (see Saperstein (1972), Naus (1974, 1982)), and certain quality control, queuing and reliability models (Greenberg (1970), Saperstein (1973), Chao *et al.* (1995)). We mention also that Karlin and Macken (1991) and Dembo and Karlin (1992), motivated by the study of inhomogeneities in long DNA sequences, developed certain Poisson approximations (through the Chen-Stein method) for a class of general scan statistics; their approximations are also applicable for the case of Bernoulli trials.

When studying finite populations, the development of randomness tests for dichotomous characteristics, calls for the study of without-replacement sampling schemes. Consider an urn that contains a white and b black balls. Assume that n balls are randomly drawn, one at a time without replacement. The distribution of the number $N^*_{n,k}$ of occurrences of non-overlapping consecutive k-tuples of white balls is called *hypergeometric distribution of order k* and has been studied by Panaretos and Xekalaki (1986), Aki and Hirano (1988) and Godbole (1990*b*). If in the above sampling scheme, each ball is returned to the urn together with c balls of the same colour before the next drawing, the resulting distribution is called *the Polya distribution of order k*. (The special case $c = 1$ is usually referred as *negative hypergeometric distribution of order k*.) An analogous random variable might be defined by considering *Friedman's urn scheme* (see Friedman (1949) or Freedman (1965)) in which besides the c balls of the same colour, we add to the urn d balls of the opposite colour.

Recently, Fu and Koutras (1994) taking a completely different approach to the problem of evaluating the probability mass function of the run-statistics $N_{n,k}$, $M_{n,k}$, $G_{n,k}$, used proper finite Markov chains and expressed the distribution of the variables of interest in terms of transition probability matrices products. A similar approach was also used by Fu (1994*a*, 1994*b*) for the study of the number of successions in a random permutation and patterns in a sequence of multistate trials respectively.

The purpose of the present paper is to develop a general workable framework for the study of all statistics mentioned before. The basic tool for our approach is a Markov chain imbedding technique. In Section 2 we introduce the concept of

Markov chain imbeddable variables of Binomial type (MVB) and provide methods for evaluating their distribution functions and generating functions. For independent and identically distributed (iid) MVB's, the double generating function is expressed through a matrix inverse, and certain simple matrix formulae are given for the mean and the generating function of the means.

In Sections 3 and 4, we show how the run statistics $M_{n,k}$, $N_{n,k}$, $G_{n,k}$ and scan statistics $M_{n,k,r}$, $N^{(1)}_{n,k,r}$, $N^{(2)}_{n,k,r}$ can be viewed as MVB's; as a consequence certain properties of them are explored through the general results presented in Section 2. Section 5 deals with the Markov chain imbedding of the urn model associated variable $N^*_{n,k}$. In Section 6 we work out some illustrative examples providing numerical results for the exact distribution function of certain scan statistics. Finally, in Section 7 we present several concluding remarks on our approach and discuss possible extensions to Markov dependent random variables and waiting time problems.

2. Markov chain imbeddable variables of binomial type

Recently, Fu and Koutras (1994) studied the distribution of the most common run statistics by establishing an imbedding into a finite Markov chain and expressed the probability distribution function of them via products of proper transition probability matrices. The motivation of the present paper stems from the observation that in most of the cases studied there, the transition probability matrix can be viewed as a bidiagonal matrix with non-zero blocks appearing only on the main diagonal and on the diagonal next to it. As a consequence, the introduction of proper ***probability vectors*** describing the *overall state formulation* of the observed Markovian structure at time t, would naturally lead to certain triangular (multidimensional) recurrence relations. Let us first introduce the notion of a *Markov chain imbeddable variable*, which is similar to the one used by Fu and Koutras (1994).

Let X_n (n a non-negative integer) be an integer valued random variable and denote by $\ell_n = \max\{x : \Pr(X_n = x) > 0\}$ its upper end point.

DEFINITION 1. The random variable X_n will be called Markov chain imbeddable variable if

(i) there exists a Markov chain $\{Y_t : t \geq 0\}$ defined on a state space Ω,

(ii) there exists a partition $\{C_x, x = 0, 1, \dots\}$ on Ω,

(iii) for every $x = 0, 1, \dots, \ell_n$ the probabilities $\Pr(X_n = x)$ can be deduced by considering the projection of the probability space of Y_n onto C_x i.e.

$$\Pr(X_n = x) = \Pr(Y_n \in C_x), \qquad x = 0, 1, \dots, \ell_n. \tag{2.1}$$

In order to proceed to the mathematical formulation of our model, let us introduce some additional notations and definitions. Assume first that the sets (state subspaces) C_x of the partition $\{C_x, x = 0, 1, \dots\}$ have the same cardinality $s = |C_x|$, $x = 0, 1, \dots$, more specifically

$$C_x = \{c_{x0}, c_{x1}, \dots, c_{x,s-1}\}.$$

This can be done without loss of generality, since one can always expand the cardinalities of non-maximal C_x's by incorporating into them additional hypothetical states. In most of the cases these states are inaccessible and their behaviour does not affect the chain at all. Next, we introduce the ***probability*** (row) ***vectors***

$$\text{(2.2)} \qquad \mathbf{f}_t(x) = (\Pr(Y_t \in c_{x0}), \Pr(Y_t \in c_{x1}), \dots, \Pr(Y_t \in c_{x,s-1})), \qquad 0 \le t \le n$$

displaying the marginal probabilities in which $\Pr(Y_t \in C_x)$ can be decomposed. From now on we shall be using the index t for the t-th step of the Markov chain and n for its final stage, where the distribution of X_n is attained (through Y_n).

We are now ready to define the basic notion of our presentation which is the Markov chain imbeddable variable of Binomial type (MVB).

DEFINITION 2. A non-negative integer random variable X_n will be called MVB if

(i) X_n can be imbedded into a Markov chain as in Definition 1,
(ii) $\Pr(Y_t \in c_{yj} \mid Y_{t-1} \in c_{xi}) = 0$ for all $y \neq x, x+1$.

For any MVB we introduce the next two $s \times s$ transition probability matrices

$$A_t(x) = (\Pr(Y_t \in c_{xj} \mid Y_{t-1} \in c_{xi})), \qquad B_t(x) = (\Pr(Y_t \in c_{x+1,j} \mid Y_{t-1} \in c_{xi})).$$

In order to illuminate the reasoning hidden in the above definitions, let the term *state* x refer to the collection $C_x = \{c_{x0}, c_{x1}, \dots, c_{x,s-1}\}$ and *substate of* x refer to the elements c_{xi} of C_x. Then, roughly speaking, the process described by a MVB cannot move backwards or jump directly to a higher state, without visiting first its next state. Regarding the matrices $A_t(x)$ and $B_t(x)$ we may state the following.

a. The entries of $A_t(x)$ control the *within state* one-step transitions i.e. the transitions of the Markov chain from a substate c_{xi} to another substate c_{xj} of the same state x.

b. The entries of $B_t(x)$ control the *between states* one-step transitions i.e. the transitions from a substate c_{xi} to a substate $c_{x+1,j}$.

c. The sum $A_t(x) + B_t(x)$ is a stochastic matrix.

Definition 2 provides a fairly broad framework, wide enough to accommodate a lot of diverse probability applications (for more details see next Sections). On the other hand it permits the derivation of a number of general results which can be subsequently applied to specific problems, providing new results and alternative ways of proving well known results.

Let $\boldsymbol{\pi}_x$ denote the initial probabilities of the Markov chain $\{Y_t : t \ge 0\}$, i.e.

$$\boldsymbol{\pi}_x = (\Pr(Y_0 \in c_{x0}), \Pr(Y_0 \in c_{x1}), \dots, \Pr(Y_0 \in c_{x,s-1})), \qquad x \ge 0$$

and $\mathbf{1} = (1, 1, \dots, 1)$ the (row) vector of R^s with all its entries being 1. The next theorem provides a method for the evaluation of the distribution function of a MVB.

THEOREM 2.1. *The double sequence of vectors* $\mathbf{f}_t(x)$, $0 \le x \le \ell_n$, $1 \le t \le n$ *satisfies the recurrence relations*

$$\begin{aligned} &(2.3\text{a}) \qquad \mathbf{f}_t(0) = \mathbf{f}_{t-1}(0)A_t(0), \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t = 1, 2, \ldots, n \\ &(2.3\text{b}) \qquad \mathbf{f}_t(x) = \mathbf{f}_{t-1}(x)A_t(x) + \mathbf{f}_{t-1}(x-1)B_t(x-1), \qquad 1 \le x \le \ell_n \end{aligned}$$

with initial conditions $\mathbf{f}_0(x) = \boldsymbol{\pi}_x$, $0 \le x \le \ell_n$. *In addition the probability distribution function of the MVB* X_n *is given by*

$$\Pr(X_n = x) = \mathbf{f}_n(x)\mathbf{1}', \qquad x = 0, 1, \ldots, \ell_n.$$

PROOF. The recurrences (2.3) are immediate consequences of the total probability theorem (or Chapman-Kolmogorov equations), Definition 2 and the form of the matrices $A_t(x)$ and $B_t(x)$. The proof of the theorem is completed by observing that

$$\Pr(X_n = x) = \Pr(Y_n \in C_x) = \sum_{j=0}^{s-1} \Pr(Y_n \in c_{xj}). \tag{2.4}$$

The use of the nomenclature "*Binomial Type*" is justified by the apparent similarity of recurrences (2.3) to the following relations, satisfied by the binomial distribution $b(n,p;x) = \binom{n}{x}p^x q^{n-x}$,

$$\begin{aligned} b(t,p;0) &= b(t-1,p;0)q, \\ & \qquad\qquad\qquad\qquad\qquad\qquad 1 \le t \le n \\ b(t,p;x) &= b(t-1,p;x)q + b(t-1,p;x-1)p. \end{aligned}$$

The generating function

$$\varphi_n(z) = \sum_{x=0}^{\ell_n} \Pr(X_n = x)z^x$$

of a *MVB* X_n, in view of (2.4), takes the form

$$\varphi_n(z) = \sum_{j=0}^{s-1} \left(\sum_{x=0}^{\ell_n} \Pr(Y_n \in c_{xj})z^x \right)$$

which, on introducing the ***vector generating functions***

$$\boldsymbol{\varphi}_n(z) = \sum_{x=0}^{\ell_n} \mathbf{f}_n(x)z^x, \tag{2.5}$$

can be expressed as

$$\varphi_n(z) = \boldsymbol{\varphi}_n(z)\mathbf{1}'. \tag{2.6}$$

In most of the applications presented here, the matrices $A_t(x)$ and $B_t(x)$ appearing in recurrences (2.3) do not depend on x. In this case the vector generating function $\boldsymbol{\varphi}_n(z)$ can be expressed as a product in the following way

THEOREM 2.2. *If $A_t(x) = A_t$, $B_t(x) = B_t$ for all $x = 0, 1, \dots$, then the vector generating function of the MVB X_n is given by*

$$\boldsymbol{\varphi}_n(z) = \boldsymbol{\varphi}_0(z) \prod_{t=1}^{n} (A_t + zB_t)$$

where

$$\boldsymbol{\varphi}_0(z) = \sum_{x=0}^{\ell_0} \boldsymbol{\pi}_x z^x$$

is the vector generating function of the initial probabilities $\boldsymbol{\pi}_x$.

PROOF. Multiplying both sides of (2.3b) by z^x, summing up for all $x = 1, 2, \dots, \ell_t$ and adding (2.3a) we obtain, for $t \geq 1$

$$\boldsymbol{\varphi}_t(z) = \left(\sum_{x=0}^{\ell_t} \mathbf{f}_{t-1}(x) z^x\right) A_t + z \left(\sum_{x=0}^{\ell_t - 1} \mathbf{f}_{t-1}(x) z^x\right) B_t.$$

Condition (ii) of Definition 2 implies that $\ell_t - \ell_{t-1} \in \{0, 1\}$. If $\ell_t = \ell_{t-1}$ we have

$$\boldsymbol{\varphi}_t(z) = \boldsymbol{\varphi}_{t-1}(z)(A_t + zB_t) - z^{\ell_{t-1}+1} \mathbf{f}_{t-1}(\ell_{t-1}) B_t.$$

Considering this equality for $z = 1$, post multiplying by $\mathbf{1}'$ and taking into account that

$$\boldsymbol{\varphi}_t(1)\mathbf{1}' = \boldsymbol{\varphi}_{t-1}(1)\mathbf{1}' = 1, \qquad (A_t + B_t)\mathbf{1}' = \mathbf{1}'$$

we get $\mathbf{f}_{t-1}(\ell_{t-1}) B_t = \mathbf{0}$. If $\ell_t = \ell_{t-1} + 1$ we may write

$$\boldsymbol{\varphi}_t(z) = \boldsymbol{\varphi}_{t-1}(z)(A_t + zB_t) + z^{\ell_t} \mathbf{f}_{t-1}(\ell_t) A_t$$

and the last term is easily checked to vanish by the same argument as before. Therefore, in both cases

$$\boldsymbol{\varphi}_t(z) = \boldsymbol{\varphi}_{t-1}(z)(A_t + zB_t), \qquad t \geq 1$$

and the proof of the theorem follows immediately.

We recall that for the generalized binomial distribution (number of successes in a sequence of n non-identical independent Bernoulli trials), a similar formula holds true for the respective (1-dimensional) generating function, namely

$$\varphi_n(z) = \varphi_0(z) \prod_{t=1}^{n} (q_t + zp_t), \qquad \varphi_0(z) = 1.$$

It's worth mentioning that in most of the applications we have

$$\boldsymbol{\pi}_0 = (1, 0, \ldots, 0) = \mathbf{e}_1, \qquad \boldsymbol{\pi}_x = \mathbf{0} = (0, 0, \ldots, 0) \qquad \text{for all} \quad x \geq 1$$

which implies that $\boldsymbol{\varphi}_0(z) = \boldsymbol{\pi}_0 = \mathbf{e}_1$.

The rest of this section will be devoted to the presentation of some results for the special case of homogeneous *MVB*, i.e. if $A_t(x) = A$, $B_t(x) = B$ for all $t \geq 1$ and $x \geq 0$.

THEOREM 2.3. *The double vector generating function*

$$\boldsymbol{\Phi}(z, w) = \sum_{n=0}^{\infty} \boldsymbol{\varphi}_n(z) w^n$$

of an homogeneous MVB X_n is given by

$$\boldsymbol{\Phi}(z, w) = \boldsymbol{\varphi}_0(z)[I - w(A + zB)]^{-1}, \qquad 0 < w < 1$$

where I is the identity $s \times s$ matrix.

PROOF. Making use of Theorem 2.2 we may write

$$\boldsymbol{\Phi}(z, w) = \boldsymbol{\varphi}_0(z) \sum_{n=0}^{\infty} [w(A + zB)]^n$$

and under proper conditions for the series to converge (e.g. if the elements of the matrix $A + zB$ lie in the closed interval $[0, 1]$, a condition which is usually satisfied) we are immediately led the desired conclusion.

Notice that the sum of the entries of $\boldsymbol{\Phi}(z, w)$ gives the double generating function $\Phi(z, w)$ of the probabilities $\Pr(X_n = x)$ i.e.

$$\Phi(z, w) = \sum_{n=0}^{\infty} \sum_{x=0}^{\ell_n} \Pr(X_n = x) z^x w^n = \boldsymbol{\Phi}(z, w)\mathbf{1}'. \tag{2.7}$$

For an homogeneous *MVB* X_n let $m_n = E(X_n)$ denote the mean of X_n, and

$$M(x) = \sum_{n=1}^{\infty} m_n w^n \tag{2.8}$$

its generating function. Then we have

THEOREM 2.4. *The means m_n and their generating function $M(w)$ are given by*

$$m_n = \boldsymbol{\varphi}_0(1) \left\{ \sum_{i=1}^{n} (A + B)^{i-1} \right\} B\mathbf{1}'$$

$$M(x) = \frac{w}{1 - w} \boldsymbol{\varphi}_0(1)[I - w(A + B)]^{-1} B\mathbf{1}'.$$

PROOF. Since

$$m_n = E(X_n) = \frac{d}{dz}[\varphi_n(z)\mathbf{1}']_{z=1}$$

and (see for example Pham (1962), p. 75)

$$\frac{d}{dz}(A+zB)^n = \sum_{i=1}^{n}(A+zB)^{i-1}B(A+zB)^{n-i}$$

we get, in virtue of Theorem 2.2 and relation (2.6)

$$m_n = \varphi_0(1)\sum_{i=1}^{n}(A+B)^{i-1}B(A+B)^{n-i}\mathbf{1}'.$$

The result follows easily by taking into account that $(A+B)\mathbf{1}' = \mathbf{1}'$. The second conclusion of the theorem is derived immediately by substituting m_n in (2.8), interchanging the order of summation and making use of the identity $\sum_{i=1}^{\infty}[w(A+B)]^{i-1} = [I - w(A+B)]^{-1}$.

3. Distribution of success runs

Consider a sequence of Bernoulli trials $Z_1, Z_2, \dots$ with success (S) probabilities $p_t = \Pr(Z_t = 1)$, and failure (F) probabilities $q_t = \Pr(Z_t = 0) = 1 - p_t$, $t \geq 1$. If k is a positive integer, let $W_t = \prod_{j=t}^{t+k-1} Z_j$, $t = 1, 2, \dots, n-k+1$ and

$$\hat{W}_t = \begin{cases} W_t & \text{if } \sum_{i=1}^{k-1} \hat{W}_{t-i} = 0, \qquad t = 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

(convention: $\hat{W}_t = 0$ for $t \leq 0$). Then, the three success run statistics described in Section 1 can be formally defined by

$$N_{n,k} = \sum_{t=1}^{n-k+1} \hat{W}_t, \qquad M_{n,k} = \sum_{t=1}^{n-k+1} W_t,$$

$$G_{n,k} = \sum_{t=1}^{n-k+1} (1 - Z_{t-1})W_t \qquad (\text{Convention} : Z_0 = 0).$$

Adopting Fu and Koutras' (1994) approach, let us denote by x the number of success runs (non-overlapping, overlapping or greater than) in a sequence of Bernoulli trials and by m the number of trailing successes i.e. the number of last consecutive successes counting backwards. For more details and illustrative examples the reader is referred to Fu and Koutras (1994).

3.a *Non-overlapping success runs*

Let $\ell_n = [n/k]$ and define $C_x = \{c_{x0}, c_{x1}, \ldots, c_{x,k-1}\}$, $x = 0, 1, \ldots, \ell_n$ where

$$c_{xi} = \{(x,i)\}, \qquad 0 \le i \le k-1, \quad x = 0, 1, \ldots, \ell_n. \tag{3.1}$$

To introduce a proper Markov chain $\{Y_t : t \ge 0\}$, we define $Y_t \in c_{xi}$ (or equivalently $Y_t = (x,i)$) if in the first t outcomes, say $SFFS \cdots F\overbrace{SS\cdots S}^{m}$, there exist x non-overlapping success runs and $m = i \pmod k$. With this set up, the random variable $N_{n,k}$ becomes a MVB, with

$$A_t(x) = A_t = \left[\begin{array}{ccccc} (\cdot,0) & (\cdot,1) & (\cdot,2) & & (\cdot,k-1) \\ q_t & p_t & 0 & \cdot & 0 \\ q_t & 0 & p_t & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ q_t & 0 & 0 & \cdot & p_t \\ q_t & 0 & 0 & \cdot & 0 \end{array}\right]_{k\times k},$$

$$B_t(x) = B_t = \left[\begin{array}{ccccc} (\cdot,0) & (\cdot,1) & (\cdot,2) & & (\cdot,k-1) \\ 0 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ p_t & 0 & 0 & \cdot & 0 \end{array}\right]_{k\times k}.$$

Therefore the probability mass function of the variable $N_{t,k}$ could be successively evaluated for all $t = k, k+1, \ldots, n$ by making use of Theorem 2.1. This provides an alternative computation scheme to the one proposed by Fu and Koutras (1994); its advantage lies in the fact that instead of multiplying matrices of order $(\ell_n + 1)k \times (\ell_n + 1)k$ we have to deal with vector recurrences involving multiplications of $k \times k$ matrices. Another interesting feature of our approach is that in the iid case ($p_t = p$, $A_t = A$, $B_t = B$ for all $t \ge 1$) it provides easy to apply formulae for the evaluation of means and generating functions. For example, making use to Theorem 2.4, we may write for $m_n = E(N_{n,k})$,

$$M(w) = \sum_{n=1}^{\infty} m_n w^n = \frac{w}{1-w}\varphi_0(1)[I - w(A+B)]^{-1}B\mathbf{1}'$$

with $\varphi_0(1) = \pi_0 = (1, 0, \ldots, 0)$, $B\mathbf{1}' = (0, 0, \ldots, 0, p)$ and calculating the $(1,k)$ element of $(I - w(A+B))^{-1}$ as

$$\frac{(pw)^{k-1}}{\det(I - w(A+B))} = \frac{(pw)^{k-1}}{(1-w)(1-(pw)^k)(1-pw)^{-1}}$$

we finally obtain

$$M(w) = \frac{(pw)^k(1-pw)}{(1-w)^2(1-(pw)^k)}.$$

By the same reasoning, working with Theorem 2.3 and formula (2.7) we deduce that

$$\Phi(z,w) = (1,0,\dots,0)[I - w(A+zB)]^{-1}\mathbf{1}' = \frac{1-(pw)^k}{(1-w)+(pw)^k[qw - z(1-pw)]}.$$

Finally, we mention that, employing Theorem 2.2 and formula (2.6), one could easily capture Aki and Hirano's (1988) recurrences for the generating function $\varphi_n(z)$ by observing that

$$\varphi_n(z) = \varphi_{n-1}(z)(A+zB)\mathbf{1}'.$$

Analogous recurrences could also be established for the non-iid case.

3.b *Overlapping success runs*

To imbed the random variable $M_{n,k}$ into a Markov chain, we set $\ell_n = n-k+1$, expand state x by incorporating into C_x an additional substate $c_{x,-1} = \{(x,-1)\}$ and define $Y_t = (x,m)$ if $m \le k-1$ and $Y_t = (x,-1)$ if $m > k$. Note that the hypothetical state $c_{0,-1} = \{(0,-1)\}$, which is in fact inaccessible by the system is used only for increasing c_0's cardinality from k to $|c_x| = k+1$, $x \ge 1$. Definition 2 is manifestly fulfilled and

$$A_t(x) = A_t = \left[\begin{array}{ccccc|c} (\cdot,0) & (\cdot,1) & (\cdot,2) & \cdot & (\cdot,k-1) & (\cdot,-1) \\ q_t & p_t & 0 & \cdot & 0 & 0 \\ q_t & 0 & p_t & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ q_t & 0 & 0 & \cdot & p_t & 0 \\ q_t & 0 & 0 & \cdot & 0 & 0 \\ \hline q_t & 0 & 0 & \cdot & 0 & 0 \end{array}\right]_{(k+1)\times(k+1)},$$

$$B_t(x) = B_t = \left[\begin{array}{ccccc|c} (\cdot,0) & (\cdot,1) & (\cdot,2) & \cdot & (\cdot,k-1) & (\cdot,-1) \\ & & & & & 0 \\ & & & & & 0 \\ & & \mathbf{0}_{k\times k} & & & \cdot \\ & & & & & 0 \\ & & & & & p_t \\ \hline 0 & 0 & 0 & \cdot & 0 & p_t \end{array}\right]_{(k+1)\times(k+1)}.$$

The double generating function of the probabilities $\Pr(M_{n,k} = x)$ (in the iid case) can be now immediately derived through Theorem 2.3 and relation (2.7).

3.c *Success runs of length at least k*

Using the same notation as in the overlapping case, with $\ell_n = [n + 1/k + 1]$ instead of $\ell_n = n-k+1$, we introduce the Markov chain $\{Y_t : t \ge 0\}$ as follows: For $0 \le m \le k-1$, we define $Y_t = (x,m)$ when there exist exactly $x \ge 0$ success runs of length at least k before the last $m+1$ outcomes. If $m \ge k$ and there exist $x-1 \ge 0$ success runs before the last $m+1$ outcomes, we define $Y_t = (x,-1)$.

An hypothetical (inaccessible) state, labelled as $(0, -1)$, is also added in order to make the cardinality of C_0 equal to $|C_x|$, $x \geq 1$. The transition matrices $A_t(x)$, $B_t(x)$ are given by

$$A_t(x) = A_t = \left[\begin{array}{ccccc|c} (\cdot,0) & (\cdot,1) & (\cdot,2) & \cdot & (\cdot,k-1) & (\cdot,-1) \\ q_t & p_t & 0 & \cdot & 0 & 0 \\ q_t & 0 & p_t & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ q_t & 0 & 0 & \cdot & p_t & 0 \\ q_t & 0 & 0 & \cdot & 0 & 0 \\ \hline q_t & 0 & 0 & \cdot & 0 & p_t \end{array}\right]_{(k+1)\times(k+1)},$$

$$B_t(x) = B_t = \left[\begin{array}{ccccc|c} (\cdot,0) & (\cdot,1) & (\cdot,2) & \cdot & (\cdot,k-1) & (\cdot,-1) \\ & & & & & 0 \\ & & & & & 0 \\ & & \mathbf{0}_{k\times k} & & & \cdot \\ & & & & & 0 \\ & & & & & p_t \\ \hline 0 & 0 & 0 & \cdot & 0 & 0 \end{array}\right]_{(k+1)\times(k+1)}.$$

Hence, one could benefit from the general theorems presented in Section 2 to derive results for the random variable $G_{n,k}$. For example, Theorems 2.1 and 2.2 highlight easy to program numerical methods for obtaining the probability distribution and generating function of $G_{n,k}$. Moreover, some routine algebraic manipulations on Theorems' 2.3 and 2.4 formulae immediately yield the double generating function $\Phi(z, w)$ and the means generating function $M(w)$.

4. Scan statistics

Let $Z_1, Z_2, \ldots$ be a sequence of independent Bernoulli trials as in Section 3, and $r \leq k$ two positive integers. Introducing the auxiliary variables

$$W_t = \begin{cases} 1 & \text{if } \sum_{j=t}^{t+k-1} Z_j \geq r \\ 0 & \text{otherwise} \end{cases}, \qquad \hat{W}_t = \begin{cases} W_t & \text{if } \sum_{i=1}^{k-1} \hat{W}_{t-i} = 0 \\ 0 & \text{otherwise} \end{cases},$$

for $t = 1, 2, \ldots, n-k+1$ (convention: $\hat{W}_t = 0$ for $t \leq 0$) and

$$r_t = \begin{cases} \min\left\{ \alpha : \sum_{j=t}^{t+\alpha-1} Z_j \geq r \right\}, & \text{if } \sum_{j=t}^{t+k-1} Z_j \geq r \\ 0, & \text{if } \sum_{j=t}^{t+k-1} Z_j < r \end{cases},$$

$$\tilde{W}_t = \begin{cases} 1 & \text{if } r_t > 0 \text{ and } \sum_{i<t:i+r_i>t} \tilde{W}_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

$t = 1, 2, \dots, n-r+1$, we may define the three scan statistics mentioned in Section 1 as follows

$$N^{(1)}_{n,k,r} = \sum_{t=1}^{n-r+1} \tilde{W}_t, \qquad N^{(2)}_{n,k,r} = \sum_{t=1}^{n-k+1} \hat{W}_t, \qquad M_{n,k,r} = \sum_{t=1}^{n-k+1} W_t.$$

For the establishment of a proper Markov structure we are going to employ pairs $(x; \mathbf{j})$ which keep track of the number (x) of scan counts till the t-th trial and the stage of formation of the next appearance $(\mathbf{j})$. The vector $\mathbf{j}$ is an m-tuple $\mathbf{j} = (j_1, j_2, \dots, j_m)$, $m \le k$ defined by

$$j_i = \begin{cases} 1 & \text{if the } (t-m+i)\text{-th trial is success } (S) \\ 0 & \text{if the } (t-m+i)\text{-th trial is failure } (F). \end{cases}$$

As we shall see later on in the treatment of specific counting processes, a substantial number of $\mathbf{j}$-combinations can be ruled out, a fact leading to a reduction of the states' cardinality.

4.a *Non-overlapping scans* $N^{(1)}_{n,k,r}$

Let $\ell_n = [n/r]$. The typical element of the state space Ω will be represented by a pair $(x; \mathbf{j})$ where $x \ge 0$ and $\mathbf{j} = (j_1, j_2, \dots, j_k)$ with $j_i \in \{0, 1\}$, $\sum_{i=1}^{k} j_i < r$. The event $Y_t = (x; \mathbf{j})$ means that

(i) in the sequence of outcomes $Z_1, Z_2, \dots, Z_{t-k}$ there appear x non-overlapping windows of length at most k, containing exactly r successes, i.e. $N^{(1)}_{t-k,k,r} = x$.

(ii) $Z_{t-k+i} = j_i$ for $i = 1, 2, \dots, k$. If $t - k + i \le 0$ or trial $t - k + i$ falls within a scan window that has already been counted we assume that $j_i = 0$.

It is not difficult to check that all the requirements of Definitions 1 and 2 are met if we define $c_{x,\mathbf{j}} = \{(x; \mathbf{j})\}$, $x \ge 0$ and

$$C_x = \left\{ c_{x,\mathbf{j}} : \mathbf{j} = (j_1, \dots, j_k) \text{ with } j_i \in \{0,1\} \text{ and } \sum_{i=1}^{k} j_i < r \right\}.$$

Obviously, the cardinality of each state C_x equals

$$s = |C_x| = \sum_{i=0}^{r-1} \binom{k}{i}.$$

The transition probabilities of the Markov chain $\{Y_t : t \ge 0\}$ are given as follows: If $j_2 + \dots + j_k < r - 1$ then

$$\Pr(Y_t = (x; j_2, \dots, j_k, j) \mid Y_{t-1} = (x; j_1, \dots, j_k)) = p_t^j (1 - p_t)^{1-j}, \qquad j = 0, 1 \tag{4.1}$$

whereas for $j_2 + \dots + j_k = r - 1$ we have

$$\Pr(Y_t = (x; j_2, \dots, j_k, 0) \mid Y_{t-1} = (x; j_1, \dots, j_k)) = q_t \tag{4.2}$$

$$\Pr(Y_t = (x+1; 0, \dots, 0) \mid Y_{t-1} = (x; j_1, \dots, j_k)) = p_t. \tag{4.3}$$

To construct the $s \times s$ matrix $A_t(x) = A_t$, it suffices to fill in the transition probabilities deduced by formulae (4.1) and (4.2). Finally, formula (4.3) provides the only non-zero transition probabilities for matrix $B_t(x) = B_t$.

As an illustration, for the random variable $N^{(1)}_{n,3,2}$ we obtain (the 3-tuples above the first row represent the values of the vector $\mathbf{j}$)

$$A_t = \left[\begin{array}{cccc} (0,0,0) & (0,0,1) & (0,1,0) & (1,0,0) \\ q_t & p_t & 0 & 0 \\ 0 & 0 & q_t & 0 \\ 0 & 0 & 0 & q_t \\ q_t & p_t & 0 & 0 \end{array}\right],$$

$$B_t = \left[\begin{array}{cccc} (0,0,0) & (0,0,1) & (0,1,0) & (1,0,0) \\ 0 & 0 & 0 & 0 \\ p_t & 0 & 0 & 0 \\ p_t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] = [(I - A_t)\mathbf{1}', \mathbf{0}_{4\times 3}].$$

Repeated application of (2.3a), (2.3b) (with initial conditions $\mathbf{f}_0(0) = \boldsymbol{\pi}_0 = \mathbf{e}_1$, $\mathbf{f}_0(x) = \boldsymbol{\pi}_x = \mathbf{0}$ for all $x \geq 1$), yields the sequence of probability vectors $\mathbf{f}_t(x)$, $t = 1, 2, \ldots, n$ and the distribution of $N^{(1)}_{n,3,2}$ will be given by (see Theorem 2.1)

$$\Pr(N^{(1)}_{n,3,2} = x) = \mathbf{f}_n(x)\mathbf{1}', \qquad x = 0, 1, \ldots, \ell_n. \tag{4.4}$$

Moreover, employing Theorems 2.3 and 2.4, we can easily deduce (in the iid case) the double generating function of the probability function (4.4) and the generating function of the means $m_n = E(N^{(1)}_{n,3,2})$ as

$$\Phi(z, w) = \sum_{n=0}^{\infty} \sum_{x=0}^{\ell_n} \Pr(N^{(1)}_{n,3,2} = x) z^x w^n = \frac{1 + pw + pqw^2}{1 - qw - pq^2w^3 - (pw)^2(1 + qw)z},$$

$$M(w) = \sum_{n=1}^{\infty} m_n w^n = \frac{(pw)^2}{1 - w} \cdot \frac{1 + qw}{1 - qw - pq^2w^3 - (pw)^2(1 + qw)}.$$

4.b *Non-overlapping scans* $N^{(2)}_{n,k,r}$

For the study of the random variable $N^{(2)}_{n,k,r}$ we use $\ell_n = [n/k]$ and

$$C_x = \left\{ (x; \mathbf{j}) : \mathbf{j} = (j_1, \ldots, j_m) \text{ with } 1 \leq m \leq k \text{ and } \sum_{i=1}^{m} j_i < r \right\}$$
$$\cup \{(x; *)\} \cup \{(x; -1, m) : r \leq m \leq k - 1\}.$$

The meaning of the event $Y_t = (x; \mathbf{j})$ is the same as in the treatment of $N^{(1)}_{n,k,r}$, the only difference being that in the beginning and after each count, we are keeping the full description of a gradually increasing window of length $m = 1, 2, \ldots, k$

containing at most $r-1$ successes. An additional number of $k-r+1$ states is also employed to denote

(i) the completion of a scan count at the t-th trial (state $(x;*)$),

(ii) the appearance of at least r successes in a window of length $m = r, r+1, \dots, k-1$ (states $(x;-1,m)$).

Obviously, the cardinality of each state C_x, $x \geq 0$ equals

$$|C_x| = (k-r+1) + \sum_{i=1}^{k} \sum_{j=0}^{\min(r-1,i)} \binom{i}{j}.$$

Instead of specifying the transition formulae for the general case (as a matter of fact a number of additional transition probabilities are attached to (4.1)–(4.3), to handle the new states of the chain) we prefer to provide the form of the matrices $A_t(x) = A_t$, $B_t(x) = B_t$ for the special cases $N^{(2)}_{n,3,1}$ and $N^{(2)}_{n,3,2}$. A careful investigation of those matrices reveals the essence of our approach for the study of the variables $N^{(2)}_{n,k,r}$.

- Special case $k=3$, $r=1$. The matrices $A_t(x) = A_t$, $B_t(x) = B_t$ are given by

$$A_t = \left[\begin{array}{cccccc}
(*) & (0) & (0,0) & (0,0,0) & (-1,1) & (-1,2) \\
0 & q_t & 0 & 0 & p_t & 0 \\
0 & 0 & q_t & 0 & 0 & p_t \\
0 & 0 & 0 & q_t & 0 & 0 \\
0 & 0 & 0 & q_t & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{array}\right],$$

$$B_t = [(I - A_t)\mathbf{1}', \mathbf{0}_{6\times 5}].$$

In the iid case, making use of Theorem 2.4, we can easily compute the generating function $M(w)$ of $m_n = E(N^{(2)}_{n,3,1})$ as

$$M(w) = \sum_{n=1}^{\infty} m_n w^n = \frac{pw^3[p^2 - 3p + 3 + (-p^2 + 3p - 2)w]}{(1-w)^2[1 + pw + pw^2 + (-p^3 + 3p^2 - 2p)w^3]}.$$

- Special case $k=3$, $r=2$. The matrices $A_t(x) = A_t$, $B_t(x) = B_t$ are given by

$$A_t = \left[\begin{array}{ccccccccccc}
(*) & (0) & (1) & (0,0) & (0,1) & (1,0) & (0,0,0) & (0,0,1) & (0,1,0) & (1,0,0) & (-1,2) \\
0 & q_t & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & q_t & p_t & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & q_t & 0 & 0 & 0 & 0 & p_t \\
0 & 0 & 0 & 0 & 0 & 0 & q_t & p_t & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & q_t & p_t & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q_t & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & q_t & p_t & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right],$$

$$B_t = [(I - A_t)\mathbf{1}', \mathbf{0}_{11\times 10}].$$

4.c *Overlapping scans* $M_{n,k,r}$

Let $\ell_n = n - k + 1$ and $\Omega = \bigcup_{x \geq 0} C_x$ where

$$C_x = \{(x;\mathbf{j}) : \mathbf{j} = (j_1, \dots, j_k) \text{ with } j_i \in \{0,1\}, 1 \leq i \leq k\}.$$

The occurrence of the event $Y_t = (x;\mathbf{j})$, $t > k$ means that

(i) in the sequence of outcomes $1, 2, \dots, t$ there exist x windows of length k including at least r successes (i.e. $M_{t,k,r} = x$),

(ii) $Z_{t-k+i} = j_i$ for $i = 1, 2, \dots, k$.

It is not difficult to verify that, for the transition probabilities of the Markov chain $\{Y_t : t \geq k\}$, formulae (4.1), (4.2) are still valid. In addition we have

$$\begin{aligned} &\Pr(Y_t = (x+1; j_2, \dots, j_k, 1) \mid Y_{t-1} = (x; j_1, \dots, j_k)) = p_t \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad j_2 + \cdots + j_k = r - 1, \\ &\Pr(Y_t = (x+1; j_2, \dots, j_k, 0) \mid Y_{t-1} = (x; j_1, \dots, j_k)) = q_t \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad j_2 + \cdots + j_k = r. \end{aligned} \tag{4.5}$$

Matrix $A_t(x) = A_t$ contains transitions of type (4.1) and (4.2), while B_t's non zero entries are provided by (4.5). The computation of the probability distribution function of $M_{n,k,r}$ can now be easily performed by repeated application of recurrences (2.3a), (2.3b) for $t = k+1, \dots, n$. The initial conditions required i.e. $\mathbf{f}_k(x)$, $x = 0, 1, \dots, \ell_n$, depend on the relationship between r and k. For example, in the special case $k = 3$, $r = 2$ we have

$$\begin{aligned} \mathbf{f}_3(0) &= (q_1q_2q_3, q_1q_2p_3, q_1p_2q_3, p_1q_2q_3, 0, 0, 0, 0), \\ \mathbf{f}_3(1) &= (0, 0, 0, 0, q_1p_2p_3, p_1q_2p_3, p_1p_2q_3, p_1p_2p_3), \\ \mathbf{f}_3(x) &= 0 \qquad \text{for} \quad x > 1 \end{aligned}$$

and

$$A_t = \left[\begin{array}{cccccccc} (0,0,0) & (0,0,1) & (0,1,0) & (1,0,0) & (0,1,1) & (1,0,1) & (1,1,0) & (1,1,1) \\ q_t & p_t & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & q_t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & q_t & 0 & 0 & 0 & 0 \\ q_t & p_t & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & q_t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & q_t & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

$$B_t = \left[\begin{array}{cccccccc} (0,0,0) & (0,0,1) & (0,1,0) & (1,0,0) & (0,1,1) & (1,0,1) & (1,1,0) & (1,1,1) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_t & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & q_t & p_t \\ 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & p_t & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & q_t & p_t \end{array}\right].$$

Note that for the study of $M_{n,k,r}$ one could use an alternative Markov chain $\{Y_t : t \geq 0\}$ defined on the extended state space

$$\Omega^* = \{(x,*)\} \cup \Omega \cup \left(\bigcup_{x \geq 0} \{(x; j_1, \dots, j_m) : 1 \leq m \leq k-1\} \right).$$

The additional states $(x,*)$ and $(x; j_1, \dots, j_m)$, $1 \leq m \leq k-1$ have an interpretation analogous to the one used for the study of the non-overlapping scans $N^{(2)}_{n,k,r}$ in Subsection 4.b. Their mission is to take care of the first k steps of the Markov chain and guarantee the validity of recurrence relations (2.3a), (2.3b) for the whole range $t = 1, 2, \dots$ (instead for $t \geq k$ which was the case for our first method).

5. Urn models

Consider a random sample of n balls drawn without replacement from an urn containing a white and b black balls. The number $N^*_{n,k}$ of (non-overlapping) runs of white balls of length k in the sample, follows an hypergeometric distribution of order k. To achieve a Markov chain description for $N^*_{n,k}$ along the lines of Section 2, let us consider $\ell_n = [n/k]$ and

$$C_x = \{(x; j, y) : 0 \leq j \leq k-1, 0 \leq j + y \leq a\}, \qquad x \geq 0.$$

We define $Y_t = (x; j, y)$ if and only if in the sequence of the first t draws there exist x runs of white balls of length k, j trailing white balls and y white balls not involved in any run 'or in the "current trail". It is rather straightforward that the transition probabilities of the Markov chain $\{Y_t : t \geq 0\}$ are given by

$$\Pr(Y_t = (x; 0, y+j) \mid Y_{t-1} = (x; j, y)) = \frac{b - (t - 1 - xk - y - j)}{a + b - (t-1)}, \qquad 0 \leq j \leq k-1, \tag{5.1}$$

$$\Pr(Y_t = (x; j+1, y) \mid Y_{t-1} = (x; j, y)) = \frac{a - (xk + y + j)}{a + b - (t-1)}, \qquad 0 \leq j < k-1, \tag{5.2}$$

$$\Pr(Y_t = (x+1; 0, y) \mid Y_{t-1} = (x; j, y)) = \frac{a - (xk + y + j)}{a + b - (t-1)}, \qquad j = k-1. \tag{5.3}$$

Note that, in this case, the transition probability matrices $A_t(x)$, $B_t(x)$ depend on both t and x. As an illustration, consider the special case $a = 4$, $k = 2$. Then

$$A_t(x)=\frac{1}{\gamma_t}\left[\begin{array}{ccccccccc}
(x;0,0) & (x;0,1) & (x;0,2) & (x;0,3) & (x;0,4) & (x;1,0) & (x;1,1) & (x;1,2) & (x;1,3)\\
\beta_t(x) & 0 & 0 & 0 & 0 & \delta(x) & 0 & 0 & 0\\
0 & \beta_t(x)+1 & 0 & 0 & 0 & 0 & \delta(x)-1 & 0 & 0\\
0 & 0 & \beta_t(x)+2 & 0 & 0 & 0 & 0 & \delta(x)-2 & 0\\
0 & 0 & 0 & \beta_t(x)+3 & 0 & 0 & 0 & 0 & \delta(x)-3\\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\
0 & \beta_t(x)+1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & \beta_t(x)+2 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \beta_t(x)+3 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0
\end{array}\right].$$

$$B_t(x)=\frac{1}{\gamma_t}\left[\begin{array}{ccccccccc}
(x+1;0,0) & (x+1;0,1) & (x+1;0,2) & (x+1;0,3) & (x+1;0,4) & (x+1;1,0) & (x+1;1,1) & (x+1;1,2) & (x+1;1,3)\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\delta(x)-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & \delta(x)-2 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & \delta(x)-3 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]$$

where $\beta_t(x)=b-t+1+2x$, $\gamma_t=5+b-t$, $\delta(x)=4-2x$.

The Polya distribution of order k can be embedded into a Markov chain by a proper modification of the state space (the third entry of the state triples $(x;j,y)$ varies now from 0 to $(c+1)a-j)$ and the transition probabilities (5.1)-(5.3) (the terms in parentheses are multiplied by $(c+1)$ instead of -1).

Finally we mention that in order to cover Friedman's urn model, only a slight modification of Polya's transition probabilities is needed whereas the state space remains unchanged. The details are left to the reader.

6. Numerical calculations

In view of the recurrences (2.3), the memory space requirements for the numerical evaluation of a MVB's distribution function through our method, depend mainly on the dimension s of the $\mathbf{f}_t(\cdot)$ vectors. Due to the special form of (2.3), the transition from $\mathbf{f}_{t-1}(\cdot)$ to $\mathbf{f}_t(\cdot)$ can be performed by the use of a single vector with $(\ell_n+1)\cdot s$ coordinates. Therefore, should our memory availability be enough to register the $(\ell_n+1)\cdot s$ entries of the $\mathbf{f}_t(\cdot)$ vector, we can proceed to the evaluation of X_n's distribution function. This gives a rough idea how could one estimate the range of the parameters where our method works.

As an application of the approach presented in the previous paragraphs, we provide some numerical results for the scan statistics $N^{(1)}_{n,k,r}$, $N^{(2)}_{n,k,r}$ and $M_{n,k,r}$ in a sequence of Bernoulli trials. Two special cases are treated. For the first one ($n=5$, $k=3$, $r=2$) the reader can easily repeat the calculations by hand (see Section 4 for the form of the matrices $A_t(x)$, $B_t(x)$), a fact that will help him to get a better grip on the underlying mechanism of our approach. The second, ($n=15$, $k=3$, $r=2$), provides a more realistic example revealing the combinatorial complexity of the problem under consideration. Nevertheless, our approach easily succeeds; compared to the first special case, it only requires 10 additional repetitions of the same recursive scheme. This remark highlights the following essential feature of the proposed method: during the computation of the distribution function of a specific scan statistic, say $N^{(1)}_{n,k,r}$, the distribution functions of all $N^{(1)}_{t,k,r}$, $t \leq n$ are also derived as by-products.

Three different choices for the success probabilities p_t of the t-th trial were considered

$$\begin{array}{lll} \text{I.} & p_t = 1/(1+t), & t \geq 1 \\ \text{II.} & p_t = 1 - 2^{-t}, & t \geq 1 \\ \text{III.} & p_t = 0.90, & t \geq 1 \text{ (iid case).} \end{array}$$

Table 1. Exact distribution of the scan statistics $N^{(1)}_{5,3,2}$, $N^{(2)}_{5,3,2}$ and $M_{5,3,2}$.

	$p_t = (1+t)^{-1}$			$p_t = 1 - 2^{-t}$			$p_t = 0.9$		
x	$N^{(1)}_{5,3,2}$	$N^{(2)}_{5,3,2}$	$M_{5,3,2}$	$N^{(1)}_{5,3,2}$	$N^{(2)}_{5,3,2}$	$M_{5,3,2}$	$N^{(1)}_{5,3,2}$	$N^{(2)}_{5,3,2}$	$M_{5,3,2}$
0	0.63889	0.63889	0.63889	0.00601	0.00601	0.00601	0.00289	0.00289	0.00289
1	0.33889	0.36111	0.20416	0.22659	0.99398	0.03952	0.07857	0.99711	0.01863
2	0.02222		0.10833	0.76739		0.15438	0.91854		0.03807
3			0.04861			0.80008			0.94041
mean	0.38333	0.36111	0.56666	1.76138	0.99398	2.74856	1.91565	0.99711	2.91600

7. Concluding remarks

The Markov chain imbedding technique merits a great potential. It provides a proper framework for developing the exact distribution of the most success run and scan statistics encountered in the study of randomness tests and other statistical problems related to sequences of binary outcomes. Certain run statistics arising from well-known urn models can also be accommodated in the same set-up.

A disadvantage of Fu and Koutras' approach is that, should one wish to work with large sequences of Bernoulli trials (or large samples in urn models), he would be forced to work with incredibly big matrices; as a matter of fact the dimension of the matrices used, tends to infinity as the number n of the trials increases. This handicap is incurred here by the consideration and study of proper probability vectors whose dimension is independent of n; the evaluation of the target distribution is then easily performed recursively, working on (triangular) matrix recurrence relations. Our approach takes advantage of the underlying sequential nature of the model under consideration and exhibits the following useful features:

a. Computational efficiency in deriving numerical results for the exact distribution of a lot of significant statistics in both iid and non-iid cases (sequences of independent Bernoulli trials, with not necessarily common success probabilities).

Table 2. Exact distribution of the scan statistics $N^{(1)}_{15,3,2}$, $N^{(2)}_{15,3,2}$ and $M_{15,3,2}$.

	$p_t = (1+t)^{-1}$			$p_t = 1 - 2^{-t}$			$p_t = 0.9$		
x	$N^{(1)}_{15,3,2}$	$N^{(2)}_{15,3,2}$	$M_{15,3,2}$	$N^{(1)}_{15,3,2}$	$N^{(2)}_{15,3,2}$	$M_{15,3,2}$	$N^{(1)}_{15,3,2}$	$N^{(2)}_{15,3,2}$	$M_{15,3,2}$
0	0.53805	0.53805	0.53805	$5.94 \cdot 10^{-23}$	$5.94 \cdot 10^{-23}$	$5.94 \cdot 10^{-23}$	$2.19 \cdot 10^{-9}$	$2.19 \cdot 10^{-9}$	$2.19 \cdot 10^{-9}$
1	0.38188	0.39114	0.20052	$2.64 \cdot 10^{-17}$	$7.77 \cdot 10^{-15}$	$5.2 \cdot 10^{-20}$	$2.63 \cdot 10^{-7}$	$4.38 \cdot 10^{-7}$	$2.26 \cdot 10^{-8}$
2	0.07430	0.06685	0.15010	$1.87 \cdot 10^{-12}$	$1.55 \cdot 10^{-8}$	$2.47 \cdot 10^{-17}$	$1.4 \cdot 10^{-5}$	$4.77 \cdot 10^{-5}$	$1.59 \cdot 10^{-7}$
3	0.00559	0.00388	0.07099	$1.82 \cdot 10^{-8}$	0.00045	$8.06 \cdot 10^{-15}$	0.0004	0.00333	$9.89 \cdot 10^{-7}$
4	0.00017	0.00007	0.02690	$2.87 \cdot 10^{-5}$	0.18981	$1.72 \cdot 10^{-12}$	0.0069	0.12899	$5.46 \cdot 10^{-6}$
5	$2.22 \cdot 10^{-6}$	$1.95 \cdot 10^{-7}$	0.00946	0.00755	0.80973	$2.02 \cdot 10^{-10}$	0.07026	0.86762	0.00002
6	$8.65 \cdot 10^{-9}$		0.00293	0.23962		$1.63 \cdot 10^{-8}$	0.37337		0.00012
7	$5.8 \cdot 10^{-12}$		0.00079	0.75279		$8.64 \cdot 10^{-7}$	0.54904		0.00052
8			0.00019			0.00002			0.00199
9			$4.22 \cdot 10^{-5}$			0.00047			0.00646
10			$8.21 \cdot 10^{-6}$			0.00614			0.02439
11			$1.41 \cdot 10^{-6}$			0.04089			0.08222
12			$2.11 \cdot 10^{-7}$			0.15371			0.08631
13			$2.39 \cdot 10^{-8}$			0.79875			0.79793
mean	0.54797	0.53678	0.89375	6.74519	4.80927	12.74405	6.46370	4.86419	12.636

b. Remarkable potential in developing results of theoretical interest such as the derivation of simple and mathematically manageable formulae for generating functions, means and generating functions for means. It is of great importance that the evaluation of these quantities is based on the computation of certain elements of a matrix inverse, which nowadays can be easily achieved by computer packages performing symbolic algebra manipulations (e.g. Mathematica, Mathcad etc.).

It is very essential that the Markov chain approach offers total control over the stage of formulation of the patterns we are interested in. This provides a powerful tool for capturing the easy way the distribution of some additional variables. As an example we mention the waiting time problems encountered when we look for the first (or m-th) occurrence of a specific pattern. For the waiting time distribution of a success run, Fu and Koutras (1994) provided a formula based on their Markov chain approach. This formula could be easily restated in terms of our triangular binomial vector probabilities. Further results, pertaining to waiting times for a specific scan configuration, or success run in urn models, will be presented in a forthcoming paper.

Another interesting feature of the techniques used here is that the established theory extends routinely to the case where the random variables $Z_1, Z_2, \dots$ are generated in a Markov dependent manner. Some trivial modifications on the transition matrices of Section 3 are enough to transfer the Bernoulli model description to the respective Markovian structures. For example, let us consider a time-homogeneous Markov chain $\{Z_n, n \geq 0\}$ with states labelled as 1 (Success) and 0 (Failure) and assume that the one-step transition probabilities $p_{ij}(t) = \Pr(Z_t = j \mid Z_{t-1} = i)$ are given by

$$(7.1) \quad p_{10}(t) = \alpha_t, \quad p_{11}(t) = 1 - \alpha_t, \quad p_{01}(t) = \beta_t, \quad p_{00}(t) = 1 - \beta_t, \qquad t \geq 0.$$

Then, the distribution of the number of non-overlapping success runs of length k can be analysed by making use of a Markov chain similar to the one employed in Subsection 3.a and respective matrices

$$A_t(x) = A_t = \left[\begin{array}{ccccc} (\cdot,0) & (\cdot,1) & (\cdot,2) & & (\cdot,k-1) \\ 1-\beta_t & \beta_t & 0 & \cdot & 0 \\ \alpha_t & 0 & 1-\alpha_t & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_t & 0 & 0 & \cdot & 1-\alpha_t \\ \alpha_t & 0 & 0 & \cdot & 0 \end{array}\right]_{k\times k},$$

$$B_t(x) = B_t = \left[\begin{array}{ccccc} (\cdot,0) & (\cdot,1) & (\cdot,2) & & (\cdot,k-1) \\ 0 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & 0 \\ 1-\alpha_t & 0 & 0 & \cdot & 0 \end{array}\right]_{k\times k}.$$

The special case $\alpha_t = \alpha$, $\beta_t = \beta$ for all $t \geq 0$, was considered by Rajarshi (1974) and Hirano and Aki (1993). Quite a few of the results presented there can be easily derived by making use of our approach and Theorems 2.1–2.3. We mention that Lou (1995) studied the conditional distribution of success runs given the total number of successes in n trials, under the assumption that the trials are Markov dependent. It goes without saying that, for Markov dependent trials, the scan statistics of Section 4 can also be defined and subsequently studied as MVB's. As an illustration we mention that, for the sequence described by (7.1), the transition probability matrices of the scan statistic $N^{(1)}_{n,3,2}$ become

$$A_t = \begin{bmatrix} (0,0,0) & (0,0,1) & (0,1,0) & (1,0,0) \\ 1-\beta_t & \beta_t & 0 & 0 \\ 0 & 0 & \alpha_t & 0 \\ 0 & 0 & 0 & 1-\beta_t \\ 1-\beta_t & \beta_t & 0 & 0 \end{bmatrix},$$

$$B_t = \begin{bmatrix} (0,0,0) & (0,0,1) & (0,1,0) & (1,0,0) \\ 0 & 0 & 0 & 0 \\ 1-\alpha_t & 0 & 0 & 0 \\ \beta_t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Finally it's worth noticing that, one could make the definitions and techniques given in this paper work for variables arising from a sequence of trials with more than two outcomes in each trial (see Schwager (1983)).

Acknowledgements

The authors would like to thank the referees for their valuable comments. The second author has been partially supported by the Greek Scholarship Foundation.

References

Aki, S. (1985). Discrete distributions of order k on a binary sequence, *Ann. Inst. Statist. Math.*, **37**, 205–224.

Aki, S. and Hirano, K. (1988). Some characteristics of the binomial distribution of order k and related distributions, *Statistical Theory and Data Analysis II* (ed. K. Matusita), 211–222, North-Holland.

Arratia, R. and Waterman, M. S. (1985). Critical phenomena in sequence matching, *Ann. Probab.*, **13**, 1236–1249.

Chao, M. T., Fu, J. C. and Koutras, M. V. (1995). Survey of reliability studies of consecutive-k-out-of-n: F and related systems, *IEEE Transactions on Reliability*, **44**, 120–127.

Chryssaphinou, O., Papastavridis, S. and Tsapelas, T. (1993). On the number of overlapping success runs in a sequence of independent Bernoulli trials, *Applications of Fibonacci Numbers*, **5**, 103–112.

Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes, *Annals of Applied Probability*, **2**, 329–357.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., Wiley, New York.

Freedman, D. (1965). Bernard Friedman's urn, *Ann. Math. Statist.*, **36**, 956–970.

Freidman, B. (1949). A simple urn model, *Comm. Pure Appl. Math.*, **2**, 59–70.

Fu, J. C. (1994*a*). Exact and limiting distributions of the number of successions in a random permutation (preprint).

Fu, J. C. (1994*b*). Distribution theory of runs and patterns associated with a sequence of multi-state trials, Tech. Report, Department of Statistics, University of Manitoba, Canada.

Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach, *J. Amer. Statist. Assoc.*, **89**, 1050–1058.

Gibbons, J. D. (1971). *Non Parametric Statistical Inference*, McGraw-Hill, New York.

Glaz, J. (1989). Approximations and bounds for the distribution of the scan statistic, *J. Amer. Statist. Assoc.*, **84**, 560–566.

Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data, *Annals of Applied Probability*, **1**, 306–318.

Godbole, A. P. (1990*a*). Specific formulae for some success run distributions, *Statist. Probab. Lett.*, **10**, 119–124.

Godbole, A. P. (1990*b*). On hypergeometric and related distributions of order k, *Comm. Statist. Theory Method*, **19**, 1291–1301.

Godbole, A. P. (1991). Poisson approximations for runs and patterns of rare events, *Adv. in Appl. Probab.*, **23**, 851–865.

Goldstein, L. (1990). Poisson approximation in DNA sequence matching, *Comm. Statist. Theory Method*, **19** (11), 4167–4179.

Greenberg, I. (1970). The first occurrence of n successes in N trials, *Technometrics*, **12**, 627–634.

Hirano, K. and Aki, S. (1993). On the number of occurrences of success runs of specified length in a two-state Markov chain, *Statistica Sinica*, **3**, 313–320.

Hirano, K., Kuboki, H., Aki, S. and Kuribayashi, A. (1984). Figures of probability density functions in statistics II—discrete univariate case, *Comput. Sci. Monographs*, No. 20, The Institute of Statistical Mathematics, Tokyo.

Hirano, K., Aki, S., Kashiwagi, N. and Kuboki, H. (1991). On Ling's binomial and negative binomial distributions of order k, *Statist. Probab. Lett.*, **11**, 503–509.

Huntington, R. (1978). Distribution of the minimum number of points in a scanning interval on the line, *Stochastic Process Appl.*, **7**, 73–77.

Karlin, S. and MacKen, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data, *J. Amer. Statist. Assoc.*, **86**, 27–35.

Ling, K. D. (1988). On binomial distributions of order k, *Statist. Probab. Lett.*, **6**, 247–250.

Lou, W. W. (1995). On runs tests for independence of binary longitudinal data using the method of finite Markov chain imbedding, Ph.D. thesis, Deptartment of Community Health, University of Toronto.

Mood, A. M. (1940). The distribution theory of runs, *Ann. Math. Statist.*, **11**, 367–392.

Mosteller, F. (1941). Note on an application of runs to quality control charts, *Ann. Math. Statist.*, **12**, 228–232.

Naus, J. (1974). Probabilities for a generalized birthday problem, *J. Amer. Statist. Assoc.*, **69**, 810–815.

Naus, J. (1982). Approximations for distributions of scan statistics, *J. Amer. Statist. Assoc.*, **77**, 177–183.

Panaretos, J. and Xekalaki, E. (1986). On some distributions arising from certain generalized sampling schemes, *Comm. Statist. Theory Method*, **15**, 873–891.

Papastavridis, S. G. and Koutras, M. V. (1994). Consecutive-k-out-of-n systems, *New Trends in System Reliability Evaluation* (ed. K. B. Misra), 228–248, Elsevier, Amsterdam.

Pham, D. (1962), *Techniques du Calcul Matriciel*, Dunod, Paris.

Philippou, A. N. and Makri, F. S. (1986). Successes, runs and longest runs, *Statist. Probab. Lett.*, **4**, 211–215.

Rajarshi, M. B. (1974). Success runs in a two-state Markov chain, *J. Appl. Probab.*, **11**, 190–194.

Saperstein, B. (1972). The generalized birthday problem, *J. Amer. Statist. Assoc.*, **67**, 425–428.

Saperstein, B. (1973). On the occurrence of n successes within N Bernoulli trials, *Technometrics*, **15**, 809–818.

Saperstein, B. (1975). Note on a clustering problem, *J. Appl. Probab.*, **12**, 629–632.

Schwager, S. (1983). Run probabilities in sequences of Markov dependent trials, *J. Amer. Statist. Assoc.*, **78**, 168–175.

Wallenstein, S., Naus, J. and Glaz, J. (1994). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence, *Biometrika*, **81**, 595–601.

Wolfowitz, J. (1943). On the theory of runs with some applications to quality control, *Ann. Math. Statist.*, **14**, 280–288.