Arapis *et al. Journal of Statistical Distributions and Applications* (2017) 4:26
DOI 10.1186/s40488-017-0080-5

Journal of Statistical Distributions and Applications

RESEARCH Open Access

Joint distribution of *k*-tuple statistics in zero-one sequences of Markov-dependent trials

Anastasios N. Arapis[1], Frosso S. Makri[1*] and Zaharias M. Psillakis[2]

*Correspondence:
makri@math.upatras.gr
[1]Department of Mathematics, University of Patras, 26500 Patras, Greece
Full list of author information is available at the end of the article

Abstract

We consider a sequence of n, $n \geq 3$, zero (0) - one (1) Markov-dependent trials. We focus on k-tuples of 1s; i.e. runs of 1s of length at least equal to a fixed integer number k, $1 \leq k \leq n$. The statistics denoting the number of k-tuples of 1s, the number of 1s in them and the distance between the first and the last k-tuple of 1s in the sequence, are defined. The work provides, in a closed form, the exact conditional joint distribution of these statistics given that the number of k-tuples of 1s in the sequence is at least two. The case of independent and identical $0-1$ trials is also covered in the study. A numerical example illustrates further the theoretical results.

AMS Subject Classification: Primary 60E05, 62E15; Secondary 60J10, 60C05

Keywords: Exact Distributions, Runs, Binary trials, Markov chain

1 Introduction

Run counting statistics defined on a sequence of binary (zero (0) - one (1)) random variables (RVs), along with their exact and approximate distributions, have been extensively studied in the literature. Their popularity is due to the fact that such statistics appear as useful theoretical models in many research areas including statistics (e.g. hypothesis testing), engineering (e.g. system reliability and quality control), biology (e.g. population genetics and DNA sequence analysis), computer science (e.g. encoding/decoding/transmission of digital information) and financial engineering (e.g. insurance and risk analysis).

In such applications, a key point is the understanding how 1s and 0s are distributed and combined as elements of a $0-1$ sequence (finite or infinite, memoryless or not) and eventually forming runs of 1s or 0s which are enumerated according to certain counting schemes. Each scheme defines how runs of same symbols or strings (patterns) of both symbols are formed and consequently are enumerated. A counting scheme may depend on, among other considerations, whether overlapping counting is allowed or not as well as if the counting starts or not from scratch when a run/string of a certain size has been so far enumerated.

The counting scheme as well as the intrinsic uncertainty of a $0-1$ sequence are often suggested by the applications. Probabilistic models, in common use, for the internal structure of a $0-1$ sequence include the model of a sequence with elements independent of each other or a model for which it is assumed some kind of dependence among the elements of it. The methods used to derive exact/approximating, marginal/joint probability distributions include combinatorial analysis, generating functions, finite Markov chain imbedding technique, recursive schemes as well as normal, Poisson and large deviation approximations.

For extensive reviews of the recent literature on the distribution theory of runs and patterns we refer to Balakrishnan and Koutras (2002) and Fu and Lou (2003). Current works on the subject include, among others, those of Antzoulakos and Chadjiconstantinidis (2001); Eryilmaz (2006, 2015, 2016, 2017); Eryilmaz and Yalcin (2011); Johnson and Fu (2014); Koutras (2003); Koutras et al. (2016); Makri and Psillakis (2015); Makri et al. (2013) and Mytalas and Zazanis (2013, 2014).

In this article we derive expressions for a conditional distribution of a trivariate statistic. Its components denote the number of runs of 1s of length exceeding a fixed threshold number, the number of 1s in such runs of 1s and the length of the minimum sequence's segment in which these runs are concentrated. The study is developed on a sequence of two-state $(0-1)$ Markov-dependent trials. The runs are enumerated according to Mood's (1940) counting scheme.

More specifically, the manuscript is organized as follows. In Section 2 we present some preliminary material, including notation and definitions, necessary to develop our results which are obtained in Section 4. In Section 3 we give a motivation along with a statement of the aim of the work. A numerical example, showed in Section 5, clarifies the theoretical results of Section 4. A discussion on the results as well as a note on a future work are given in Section 6.

Throughout the article, for integers, n, m, $\binom{n}{m}$ denotes the extended binomial coefficient (see, Feller (1968), pp. 50, 63), $\lfloor x \rfloor$ stands for the greatest integer less than or equal to x and δ_{ij} denotes the Kronecker delta fuction of the integer arguments i and j. Further, for $\alpha > \beta$, we apply the conventions $\sum_{i=\alpha}^{\beta} y_i = 0$, $\prod_{i=\alpha}^{\beta} y_i = 1$, $\sum_{i=\alpha}^{\beta} \mathbf{Y}^{(i)} = \mathbf{O} \equiv \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, $\prod_{i=\alpha}^{\beta} \mathbf{Y}^{(i)} = \mathbf{I} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, where y_i and $\mathbf{Y}^{(i)}$ are scalars and 2×2 matrices, respectively.

2 Preliminaries

2.1 Run counting statistics

Let $\{X_t\}_{t=1}^{n}$, $n \geq 1$, be the first n trials of a binary $(0-1)$ sequence of RVs, $X_t = x_t \in \{0, 1\}$. A run of 1s, is a (sub)sequence of $\{X_t\}_{t=1}^{n}$ consisting of consecutive 1s, the number of which is referred to as its length, preceded and succeeded by 0s or by nothing.

Given a fixed integer k, $1 \leq k \leq n$, a k-tuple of 1s is a run of 1s of length k or more. In the paper we will deal with the following statistics defined on a $0-1$ $\{X_t\}_{t=1}^{n}$. For details see, e.g. Makri et al. (2015) and the references therein.

(I) $G_{n,k}$ denoting the number of k-tuples of 1s, $1 \leq k \leq n$. In particular, $G_{n,1}$ denotes the number of 1-tuples of 1s, i.e. it represents the number $R_n \equiv G_{n,1}$ of all runs of 1s in the sequence. Using the convention $X_0 = X_{n+1} \equiv 0$, we can define $G_{n,k}$ as

$$G_{n,k} = \sum_{i=k}^{n} E_{n,i}, \; 1 \leq k \leq n, \tag{1}$$

where

$$E_{n,i} = \sum_{j=i}^{n} J_j,\ J_j = \left(1 - X_{j-i}\right)\left(1 - X_{j+1}\right) \prod_{r=j-i+1}^{j} X_r.$$

(II) $S_{n,k}$ denoting the number of 1s in the $G_{n,k}$ k-tuples of 1s; i.e. $S_{n,k}$ represents the sum of lengths of the $G_{n,k}$ k-tuples of 1s, $1 \le k \le n$. In particular $S_{n,1}$ represents the number of all 1s in the sequence; hence, the number of 0s, Y_n, in the sequence is $Y_n = n - S_{n,1}$. $S_{n,k}$ is formally defined as

$$S_{n,k} = \sum_{i=k}^{n} iE_{n,i},\ 1 \le k \le n. \tag{2}$$

Readily, $kG_{n,k} \le S_{n,k}$.

(III) L_n, $n \ge 1$, denoting the length of the longest run of 1s in the sequence. By setting

$$\Lambda_n = \{i : G_{n,i} > 0, 1 \le i \le n\},$$

we have that

$$L_n = \max\{k : k \in \Lambda_n\},\ \text{if } \Lambda_n \ne \emptyset;\ 0,\ \text{otherwise}. \tag{3}$$

Readily $L_n < k$ iff $G_{n,k} < 1$.

(IV) For $G_{n,k} \ge 1$, $1 \le k \le n$, $D_{n,k}$ denotes the distance (number of trials) between and including the first 1 of the first k-tuple of 1s and the last 1 of the last k-tuple of 1s in the sequence. If there is only one k-tuple of 1s in the sequence then $D_{n,k}$ denotes its length. That is, $D_{n,k}$ represents the size (length) of the minimum (sub)sequence of $\{X_t\}_{t=1}^{n}$ in which all $G_{n,k}$ k-tuple of 1s are concentrated. In particular, $D_{n,1}$ represents the length of the minimum segment of the sequence containing all R_n runs of 1s or in other words all $S_{n,1}$ 1s appearing in the sequence. For $G_{n,k} \ge 1$, $1 \le k \le n$, $D_{n,k}$ can be formally defined as

$$D_{n,k} = U_{n,k}^{(2)} - U_{n,k}^{(1)} + 1, \tag{4}$$

where

$$U_{n,k}^{(1)} = \min\{j : I_j = 1, 1 \le j \le n - k + 1\},$$

$$U_{n,k}^{(2)} = \max\{j : I_{j-k+1} = 1, k \le j \le n\},$$

$$I_j = \prod_{r=j}^{j+k-1} X_r,\ 1 \le j \le n - k + 1.$$

Readily, $D_{n,k} = S_{n,k} = L_n$, if $G_{n,k} = 1$ and $D_{n,k} > S_{n,k} > L_n$, if $G_{n,k} > 1$.

(V) For $G_{n,k} \ge 1$, $1 \le k \le n$, set $\mathbf{V}_{n,k} = (D_{n,k}, G_{n,k}, S_{n,k})$. This is the RV we focus on in the article.

Example: By way of illustration consider the trials 1110001100010001010011101 111001001001001 numbered from 1 to 40. Then, $L_{40} = 4$ and $\mathbf{V}_{40,1} = (40, 11, 19)$, $\mathbf{V}_{40,2} = (28, 4, 12)$, $\mathbf{V}_{40,3} = (28, 3, 10)$, $\mathbf{V}_{40,4} = (4, 1, 4)$.

2.2 Internal structure's models

A general enough model for the internal structure of a $0-1$ $\{X_t\}_{t=1}^{n}$, $n \ge 2$, is that of the first n trials of a homogeneous $0-1$ Markov chain of first order (HMC1). On such a model we will develop our results. Accordingly, we next state the necessary notation/definitions.

Let $\{X_t\}_{t\geq 1}$ be a HMC1 with state space $\mathcal{A} = \{0,1\}$, one step transition probability matrix

$$\mathbf{P} = (p_{ij}) = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix},$$

with

$$p_{ij} = P\left(X_t = j \mid X_{t-1} = i\right),\ i,j \in \mathcal{A},\ \sum_{j\in\mathcal{A}} p_{ij} = 1,\ i \in \mathcal{A},\ t \geq 2 \tag{5}$$

and probability distribution vector at time t

$$\mathbf{p}^{(t)} = \left(p_0^{(t)}, p_1^{(t)}\right),$$

with

$$p_i^{(t)} = P(X_t = i),\ i \in \mathcal{A},\ \sum_{i\in\mathcal{A}} p_i^{(t)} = 1,\ t \geq 1. \tag{6}$$

Readily, because of the homogeneity of $\{X_t\}_{t\geq 1}$, it holds

$$\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)}\mathbf{P} = \mathbf{p}^{(1)}\mathbf{P}^{t-1},\ t \geq 2;\ \mathbf{p}^{(1)},\ t = 1 \text{ and } \mathbf{P}^{t-1} = \left(p_{ij}^{(t-1)}\right),\ t \geq 2,$$

with

$$p_i^{(t)} = \mathbf{p}^{(t)}\mathbf{e}_i^{'},\ i \in \mathcal{A},\ t \geq 1,$$

$$p_{ij}^{(t-1)} = P(X_{t-1+m} = j \mid X_m = i) = \mathbf{e}_i\mathbf{P}^{t-1}\mathbf{e}_j^{'},\ i,j \in \mathcal{A},\ t \geq 2,\ m \geq 1, \tag{7}$$

where $\mathbf{e}_i^{'}$ is the transpose (i.e. the column vector) of the row vector $\mathbf{e}_i$, $i \in \mathcal{A}$, with $\mathbf{e}_0 = (1,0)$ and $\mathbf{e}_1 = (0,1)$.

In particular, for $p_{01} + p_{10} \neq 0$, i.e. $\mathbf{P} \neq \mathbf{I}$, it holds

$$\mathbf{P}^{t-1} = (p_{01} + p_{10})^{-1}\left\{\begin{pmatrix} p_{10} & p_{01} \\ p_{10} & p_{01} \end{pmatrix} + (1 - p_{01} - p_{10})^{t-1}\begin{pmatrix} p_{01} & -p_{01} \\ -p_{10} & p_{10} \end{pmatrix}\right\},\ t \geq 2, \tag{8}$$

$$p_0^{(t)} = p_0^{(1)}(1 - p_{01} - p_{10})^{t-1} + p_{10}(p_{01} + p_{10})^{-1}\left[1 - (1 - p_{01} - p_{10})^{t-1}\right],\ t \geq 1. \tag{9}$$

The setup of a $0-1$ HMC1 $\{X_t\}_{t=1}^n$, $n \geq 2$, covers the case of a $0-1$ sequence of independent and identically distributed (IID) RVs, too. This is so, because a $0-1$ $\{X_t\}_{t=1}^n$, $n \geq 2$, IID sequence with

$$P(X_t = 1) = 1 - P(X_t = 0) = p_1,\ 1 \leq t \leq n, \tag{10}$$

is a particular HMC1 with

$$p_{ij} = 1 - p_1,\ j = 0;\ p_1, j = 1,\ i \in \mathcal{A},\ p_{ij}^{(t-1)} = p_{ij},\ i,j \in \mathcal{A},\ t \geq 2,$$

$$p_1^{(t)} = p_1 = 1 - p_0^{(t)},\ 1 \leq t \leq n. \tag{11}$$

2.3 A combinatorial result

In combinatorial analysis which will be used in Section 4, the following result, recalled from Makri et al. (2007), is useful. The coefficient

$$H_m(\alpha,r,k) = \sum_{j=0}^{\left\lfloor \frac{\alpha}{k+1} \right\rfloor} (-1)^j \binom{m}{j} \binom{\alpha-(k+1)j+r-1}{\alpha-(k+1)j}, \tag{12}$$

represents the number of allocations of α indistinguishable balls into r distinguishable cells where each of the m, $0 \leq m \leq r$, specified cells is occupied by at most k balls. Equivalently, it gives the number of nonnegative integer solutions of the linear equation $x_1 + x_2 + \ldots + x_r = \alpha$ with the restrictions, for $m \geq 1$, $0 \leq x_{i_j} \leq k$, $1 \leq j \leq m$, for some specific m-combination $\{i_1, i_2, \ldots, i_m\}$ of $\{1, 2, \ldots, r\}$, and no restrictions on x_js, $1 \leq j \leq r$, for $m = 0$.

Moreover, $H_r(\alpha, r, k)$ is Riordan's (1964, p. 104) coefficient

$$C(\alpha,r,k) = \sum_{j=0}^{\left\lfloor \frac{\alpha}{k+1} \right\rfloor} (-1)^j \binom{r}{j} \binom{\alpha-(k+1)j+r-1}{\alpha-(k+1)j}. \tag{13}$$

3 Motivation and aim of the work

In a study of a $0-1$ sequence $\{X_t\}_{t=1}^n$, $n \geq 3$, it is reasonable for one to be interested in the probabilistic behavior of RV $\mathbf{V}_{n,k} = (D_{n,k}, G_{n,k}, S_{n,k})$. This happens because jointly its components provide a more refined view of the internal clustering structure of the sequence than the information extracted by each one alone.

Interpreting a k-tuple of 1s as a cluster of consecutive 1s of size at least k, $D_{n,k}$ represents the size of the minimum segment of $\{X_t\}_{t=1}^n$ in which $G_{n,k}$ clusters of size at least k and at most L_n are concentrated. The overall density of $G_{n,k}$ clusters, with respect to the number of 1s in them, as well as of the minimum concentration segment is evaluated by $S_{n,k}$. Large values of $D_{n,k}$ suggest that these $G_{n,k}$ clusters spread over the interval between the left and the right side of the sequence whereas small values of $D_{n,k}$ indicate rather that the clusters are concentrated in a segment of the sequence of small size leaving the rest part(s) of the sequence empty of such clusters.

In addition to this information, a large value of $S_{n,k}$ paired with a small value of $G_{n,k}$ indicates the existence of clusters of 1s of a large size and therefore a trend whereas the same value of $S_{n,k}$ paired with a large value of $G_{n,k}$ indicates rather a distribution of clusters of small size in the (sub)sequence in which they are concentrated.

Therefore, based on the former interpretation, the motivation for the study as well as the usefulness of the statistic $\mathbf{V}_{n,k} = (D_{n,k}, G_{n,k}, S_{n,k})$ is apparent. In the sequel, we assume that $G_{n,k} \geq 2$ in order to have at least two k-tuples of 1s in the sequence and accordingly the distance $D_{n,k}$ is not a degenerate one. Moreover, this assumption is a common one in an application area of $D_{n,k}$; e.g., in detecting pattern (tandem or non-tandem direct) repeats in DNA sequences (Benson 1999).

For $1 \leq k \leq n$, set

$$\mathcal{M}_{n,k} = \{G_{n,k} \geq 2\},\ \alpha_{n,k} = P\left(\mathcal{M}_{n,k}\right) \tag{14}$$

and for $n \geq 3$, $1 \leq k \leq \lfloor (n-1)/2 \rfloor$, define

$$\Omega_{n,k} = \{(d,m,s) : 2k+1 \le d \le n, 2k \le s \le d-1, 2 \le m \le \min\left(\lfloor s/k \rfloor, d-s+1\right)\} \tag{15}$$

and for $(d,m,s) \in \Omega_{n,k}$,

$$h_{n,k}(d,m,s) = P\left(\mathbf{V}_{n,k} = (d,m,s), \mathcal{M}_{n,k}\right),$$

$$v_{n,k}(d,m,s) = P\left(\mathbf{V}_{n,k} = (d,m,s) \mid \mathcal{M}_{n,k}\right) = h_{n,k}(d,m,s)/\alpha_{n,k}. \tag{16}$$

The paper provides exact closed form expressions for $\alpha_{n,k}$, $h_{n,k}(d,m,s)$ and eventually for $v_{n,k}(d,m,s)$ when $\mathbf{V}_{n,k}$ is defined on a $0-1$ HMC1/IID. The expressions are obtained via combinatorial analysis.

More specifically, closed formulae are established for the first time for $h_{n,k}(d,m,s)$, $1 \le k \le \lfloor (n-1)/2 \rfloor$, when $\mathbf{V}_{n,k}$ is defined on a $0-1$ HMC1 with given $\mathbf{P}$ and $\mathbf{p}^{(1)}$. Since, the general frame of HMC1 covers as a particular case IID sequences, the so implied expressions for $v_{n,k}(d,m,s)$ are alternative to those obtained for $v_{n,k}(d,m,s)$, $1 \le k \le \lfloor (n-1)/2 \rfloor$, by Makri et al. (2015) for IID sequences.

Moreover, for $n \ge 3$, $1 \le k \le \lfloor (n-1)/2 \rfloor$, $2k+1 \le d \le n$, let

$$f_{n,k}(d) = P\left(D_{n,k} = d \mid \mathcal{M}_{n,k}\right).$$

Therefore, since

$$f_{n,k}(d) = \sum_{s=2k}^{d-1} \sum_{m=2}^{\min(\lfloor s/k \rfloor, d-s+1)} v_{n,k}(d,m,s) = \alpha_{n,k}^{-1} \sum_{s=2k}^{d-1} \sum_{m=2}^{\min(\lfloor s/k \rfloor, d-s+1)} h_{n,k}(d,m,s), \tag{17}$$

hence, the work provides closed form expressions for determining $f_{n,k}(d)$ for HMC1 and IID $0-1$ $\{X_t\}_{t=1}^n$. These expressions are alternative to those derived, for IID sequences, by Makri et al. (2015) for $1 \le k \le \lfloor (n-1)/2 \rfloor$ as well as to those obtained, for HMC1, by Arapis et al. (2016) for $k=1$ and by Arapis et al. (2017) for $1 \le k \le \lfloor (n-1)/2 \rfloor$.

4 Results

In a $0-1$ sequence $\{X_t\}_{t=1}^n$, $n \ge 2$, for $0 \le y \le n$, $0 \le r \le \lfloor (n+1)/2 \rfloor$ and $(i,j) \in \{0,1\}^2$, define

$$B_n^{(i,j)}(y,r) = \{X_1 = i, X_n = j, Y_n = y, G_{n,1} = r\},$$

$$\pi_n^{(i,j)}(y,r) = P(B_n^{(i,j)}(y,r)).$$

Accordingly, for a HMC1 $\{X_t\}_{t=1}^n$, $n \ge 2$, with given $\mathbf{P}$ and $\mathbf{p}^{(1)}$, it holds

$$\pi_n^{(i,j)}(y,r) = \left(p_1^{(1)}\right)^i \left(1-p_1^{(1)}\right)^{1-i} p_{00}^{y-r-1+i+j} (1-p_{00})^{r-i} (1-p_{11})^{r-j} p_{11}^{n-y-r}, \tag{18}$$

for $2-(i+j) \le y \le n-(i+j)$, $1-\delta_{y,0}-\delta_{y,n}+\delta_{i+j,2} \le r \le \min\{n-y, y-1+i+j\}$ and $\pi_n^{(i,j)}(y,r) = 0$, otherwise.

Consequently, $\pi_n^{(i,j)}(y,r)$, for a $0-1$ IID sequence, reduces to

$$\pi_n^{(i,j)}(y,r) = \pi_n(y) = p_1^{n-y}(1-p_1)^y,\ 0 \le y \le n. \tag{19}$$

Theorem 1 *For $n \geq 3$, $(d,m,s) \in \Omega_{n,1}$, $0 < p_1^{(1)} < 1$, it holds*

$$h_{n,1}(d,m,s) = \binom{s-1}{m-1}\binom{d-s-1}{m-2}\pi_d^{(1,1)}(d-s,m)\varepsilon_n(d) \tag{20}$$

where $\varepsilon_n(d) = 1$, if $n = d$; $p_{00}^{n-d-2}\left\{p_{10}p_{00} + p_0^{(1)}(p_1^{(1)})^{-1}p_{01}\left[(n-d-1)p_{10} + p_{00}\right]\right\}$, if $n \geq d+1$.

Proof For $d = 3,\ldots,n-2$, $i = 2,3,\ldots,n-d$, $s = 2,3,\ldots,d-1$, $m = 2,3,\ldots,\min\{s,d-s+1\}$ an element of the event $\Gamma_{i,d,m,s} = \{U_{n,1}^{(1)} = i, D_{n,1} = d, R_n = m, S_{n,1} = s\}$ is a $0-1$ sequence of length n with probability

$$p_0^{(1)}p_{00}^{i-2}p_{01}\left[\pi_d^{(1,1)}(d-s,m)\left(p_1^{(1)}\right)^{-1}\right]p_{10}p_{00}^{n-i-d}.$$

Fix i. Then the number of elements of the event $\Gamma_{i,d,m,s}$ is $\binom{s-1}{m-1}\binom{d-s-1}{m-2}$, since the number of allocations of s 1s in m runs of 1s is $\binom{s-1}{m-1}$ and the number of allocations of $d-s$ 0s in $m-1$ runs of 0s is $\binom{d-s-1}{m-2}$, so that

$$P\left(\Gamma_{i,d,m,s}\right) = \binom{s-1}{m-1}\binom{d-s-1}{m-2}p_0^{(1)}p_{01}\left[\pi_d^{(1,1)}(d-s,m)\left(p_1^{(1)}\right)^{-1}\right]p_{10}p_{00}^{n-d-2}.$$

We use similar reasoning for the rest cases. Then summing with respect to i we get the result. □

For a sequence $\{X_t\}_{t=1}^n$ of $0-1$ IID RVs, $h_{n,1}(d,m,s)$ reduces to the explicit formula given in the next Corollary.

Corollary 1 *For $n \geq 3$, $(d,m,s) \in \Omega_{n,1}$, $0 < p_1 < 1$, it is true that*

$$h_{n,1}(d,m,s) = (n-d+1)\binom{s-1}{m-1}\binom{d-s-1}{m-2}p_1^s(1-p_1)^{n-s}. \quad \diamondsuit \tag{21}$$

In order to derive for HMC1, in the forthcoming Theorem 2, $h_{n,k}(d,m,s)$, $5 \leq 2k+1 \leq n$, we next recall, in Lemma 1, a result from (Makri et al.: On the concentration of runs of ones of length exceeding a threshold in a Markov chain, submitted).

Lemma 1 *For $(i,j) \in \{0,1\}^2$, $n \geq 2$, set $\lambda_{n,k}^{(i,j)}(x) = P(G_{n,k} = x, X_1 = i, X_n = j)$, $x = 0,1$. Then, it holds that:*

(I) For $2 \leq k \leq n-2+i+j$,

$$\lambda_{n,k}^{(i,j)}(0) = \sum_{y=1}^{n-(i+j)}\sum_{r=i+j}^{y-1+i+j}\binom{y-1}{r-i-j}C(n-y-r,r,k-2)\pi_n^{(i,j)}(y,r),$$

$$\lambda_{n,k}^{(i,j)}(1) = \pi_n^{(i,j)}(0,1)\delta_{2,i+j}+\sum_{y=1}^{n-k}\sum_{r=1}^{y-1+i+j}r\binom{y-1}{r-i-j}H_{r-1}(n-y-r-k+1,r,k-2)\pi_n^{(i,j)}(y,r). \tag{22}$$

(II) For $k > n-2+i+j$,

$$\lambda_{n,k}^{(i,j)}(0) = \left(p_1^{(1)}\right)^i\left(1-p_1^{(1)}\right)^{1-i}p_{ij}^{(n-1)},$$

$$\lambda_{n,k}^{(i,j)}(1) = 0. \tag{23}$$

Theorem 2 *For* $n \geq 5$, $2 \leq k \leq \lfloor (n-1)/2 \rfloor$, $(d,m,s) \in \Omega_{n,k}$, $0 < p_1^{(1)} < 1$, *it holds*

$$h_{n,k}(d,m,s) = p_{11}^{2k-2}\left(p_1^{(1)}\right)^{-1} \sum_{i=1}^{n-d+1} \ell^{(\alpha)}_{i-1,k}\ell^{(\beta)}_{n-d-i+1,k} \sum_{r=m}^{m+\left\lfloor \frac{d-s-m+1}{2} \right\rfloor} \sum_{y=r-1}^{d-s-r+m} \gamma_{d,m,s}(y,r)\pi^{(1,1)}_{d-2k+2}(y,r), \tag{24}$$

where

$$\ell^{(\alpha)}_{n,k} = p_1^{(1)}, \textit{for}\, n=0; \quad p_0^{(n)}p_{01}, \textit{for}\, 1 \leq n \leq k; \quad p_{01}\left[\lambda^{(0,0)}_{n,k}(0) + \lambda^{(1,0)}_{n,k}(0)\right], \textit{for}\, n \geq k+1,$$

$$\ell^{(\beta)}_{n,k} = 1, \textit{for}\, n=0; \quad p_{10}, \textit{for}\, 1 \leq n \leq k; \quad p_{10}(p_0^{(1)})^{-1}\left[\lambda^{(0,0)}_{n,k}(0) + \lambda^{(0,1)}_{n,k}(0)\right], \textit{for}\, n \geq k+1 \tag{25}$$

and

$$\gamma_{d,m,s}(y,r) = \binom{y-1}{r-2}\binom{r-2}{m-2}\binom{s-mk+m-1}{m-1}C(d-y-s-r+m, r-m, k-2). \tag{26}$$

Proof For $1 \leq r_1 \leq r_2 \leq n$ let Y_{r_1,r_2}, R_{r_1,r_2}, L_{r_1,r_2}, $S_{r_1,r_2,k}$, $D_{r_1,r_2,k}$, $G_{r_1,r_2,k}$ be RVs defined on the subsequence $X_{r_1}, X_{r_1+1}, \ldots, X_{r_2}$ of $\{X_t\}_{t=1}^n$. For $m \geq 2$ define the event

$$\begin{aligned}&\Delta_{r_1,r_1+d-1}(d,s,m,y,r)\\ &\quad = \{D_{r_1,r_1+d-1,k} = d, G_{r_1,r_1+d-1,k} = m, S_{r_1,r_1+d-1,k} = s, Y_{r_1,r_1+d-1} = y, R_{r_1,r_1+d-1} = r\}.\end{aligned}$$

An element of this event is a 0 - 1 sequence of length d, starting and ending with a 1, for which y_j's and z_j's, representing the lengths of the failure and success runs, respectively, satisfy the conditions:

(a) $y_1 + y_2 + \ldots + y_{r-1} = y, y_j \geq 1, 1 \leq j \leq r-1$.
(b) $z_1 + z_{i_1} + z_{i_2} + \ldots + z_{i_{m-2}} + z_r = s, z_j \geq k, j \in \{1, i_1, i_2, \ldots, i_{m-2}, r\}$, for some specific combination $\{1, i_1, i_2, \ldots, i_{m-2}, r\}$ of $\{1, 2, \ldots, r-1, r\}$ among the $\binom{r-2}{m-2}$ ones.
(c) $z_{i_{m-1}} + z_{i_m} + \ldots + z_{i_{r-2}} = d-y-s, 1 \leq z_{i_j} \leq k-1, m-1 \leq j \leq r-2$, for $\{i_{m-1}, \ldots, i_{r-2}\} \in \{1, 2, \ldots, r\} - \{1, i_1, i_2, \ldots, i_{m-2}, r\}$.

Fix $i_1, i_2, \ldots, i_{m-2}$. Then the number of such sequences, i.e. the number of solutions of the system (a)-(c), is

$$\binom{y-1}{r-2}C(d-y-s-r+m, r-m, k-2)\binom{s-mk+m-1}{m-1}$$

and each such sequence has probability

$$p_1^{(1)}p_{11}^{k-1}(p_1^{(1)})^{-1}\pi^{(1,1)}_{d-2k+2}(y,r)p_{11}^{k-1} = p_{11}^{2k-2}\pi^{(1,1)}_{d-2k+2}(y,r).$$

Hence,

$$\begin{aligned}P(\Delta_{r_1,r_1+d-1}(d,s,m,y,r)) &= p_{11}^{2k-2}\pi^{(1,1)}_{d-2k+2}(y,r)\binom{r-2}{m-2}\binom{y-1}{r-2}\binom{s-mk+m-1}{m-1}\\ &\quad \times C(d-y-s-r+m, r-m, k-2).\end{aligned}$$

For $k+2 \le i \le n-k-d$, $m \ge 2$, we have that

$$\begin{aligned}
&P\left(U^{(1)}_{n,k}=i, D_{n,k}=d, G_{n,k}=m, S_{n,k}=s, Y_{i,i+d-1}=y, R_{i,i+d-1}=r\right)\\
&= P\Big\{\left[(L_{1,i-1}<k, X_{i-1}=0)\cap[(X_1=0)\cup(X_1=1)]\right]\cap\Delta_{i,i+d-1}(d,s,m,y,r)\\
&\quad\cap\left[(L_{i+d,n}<k, X_{i+d}=0)\cap[(X_n=0)\cup(X_n=1)]\right]\Big\}\\
&= \left[\lambda^{(0,0)}_{i-1,k}(0)+\lambda^{(1,0)}_{i-1,k}(0)\right]p_{01}\\
&\quad\times\left(p_1^{(1)}\right)^{-1}P\left(\Delta_{i,i+d-1}(d,s,m,y,r)\right)p_{10}\left[\lambda^{(0,0)}_{n-i-d+1,k}(0)+\lambda^{(0,1)}_{n-i-d+1,k}(0)\right]/p_0^{(1)}\\
&= \left[\lambda^{(0,0)}_{i-1,k}(0)+\lambda^{(1,0)}_{i-1,k}(0)\right]p_{01}\left(p_1^{(1)}\right)^{-1}p_{11}^{2k-2}\pi^{(1,1)}_{d-2k+2}(y,r)\\
&\quad\times\binom{r-2}{m-2}\binom{y-1}{r-2}\binom{s-mk+m-1}{m-1}\\
&\quad\times C(d-y-s-r+m, r-m, k-2)p_{10}\left(p_0^{(1)}\right)^{-1}\left[\lambda^{(0,0)}_{n-i-d+1,k}(0)+\lambda^{(0,1)}_{n-i-d+1,k}(0)\right].
\end{aligned}$$

By similar reasoning we get the remaining cases of i, i.e. $1 \le i \le k+1$ and $n-d+1-k \le i \le n-d+1$. Then summing with respect to i, y and r we get the result. □

Having found $h_{n,k}(d,m,s)$, we next proceed to obtain $v_{n,k}(d,m,s)$. In accomplishing it, the required probabilities $\alpha_{n,k}$ for HMC1 are recalled, in Lemma 2, from Arapis et al. (2016) for $k=1$, and they are computed via Lemma 1 for $2 \le k \le \lfloor (n-1)/2 \rfloor$.

Lemma 2 *For $n \ge k \ge 1$, the probability $\alpha_{n,k}$, for HMC1, is computed via the expressions:*
(I) For $k=1$,

$$\alpha_{n,1} = 1-p_{00}^{n-3}\left\{p_{00}\left(1+(n-2)p_{01}\right)+\frac{(n-1)(n-2)}{2}p_0^{(1)}p_{01}^2\right\}, \quad \textit{if} \quad p_{00}=p_{11}$$

and

$$\begin{aligned}
\alpha_{n,1} &= 1-p_0^{(1)}p_{00}^{n-1}-p_{11}^{n-2}\left(p_1^{(1)}+p_0^{(1)}p_{01}\right)-p_{00}\left(p_0^{(1)}p_{01}+p_1^{(1)}p_{10}\right)\frac{p_{11}^{n-2}-p_{00}^{n-2}}{p_{11}-p_{00}}\\
&\quad -p_0^{(1)}p_{01}p_{10}\frac{p_{11}^{n-1}-p_{00}^{n-2}\left[p_{11}+(n-2)\left(p_{11}-p_{00}\right)\right]}{\left(p_{11}-p_{00}\right)^2}, \quad \textit{if} \quad p_{00}\ne p_{11}. \qquad (27)
\end{aligned}$$

(II) For $2 \le k \le n$,

$$\alpha_{n,k} = 1-\sum_{(i,j)\in\{0,1\}^2}\left[\lambda^{(i,j)}_{n,k}(0)+\lambda^{(i,j)}_{n,k}(1)\right]. \qquad (28)$$

Theorem 3 *For $n \ge 3$, $1 \le k \le \lfloor (n-1)/2 \rfloor$, $(d,m,s) \in \Omega_{n,k}$, $0 < p_1^{(1)} < 1$, the PMF $v_{n,k}(d,m,s)$ for a HMC1, with given $\boldsymbol{P}$ and $\boldsymbol{p}^{(1)}$, is calculated by*

$$v_{n,k}(d,m,s) = \alpha_{n,k}^{-1}h_{n,k}(d,m,s), \qquad (29)$$

where $\alpha_{n,k}$ and $h_{n,k}(d,m,s)$ are provided by Lemma 2 and Theorems 1 (for $k=1$) and 2 (for $2 \le k \le \lfloor (n-1)/2 \rfloor$), respectively.

Remark 1 *For IID sequences, in implementing Theorem 3, one has to take into consideration Eqs. (10) - (11), (19) and (21). Moreover, for speeding up calculations, one has to set $\pi_n(y)$ in front of the inner summation in (22).*

5 A numerical example

In this example we compute some indicative numerics concerning two model (i.e. HMC1 and IID) $0-1$ sequences $\{X_t\}_{t=1}^{n}$ which are considered in the paper. The common length of these was taken small, i.e. $n=8$, so that the required computations can also be carried out by a hand/pocket calculator and thus it is possible to gain insight in the formulae developed in Section Results, and also because of space limitations. The sequences that have been used are as follows. Table 1: An IID sequence with $p_1=0.5$. Table 2: A HMC1 sequence with $p_{00}=p_{11}=0.9$, $p_1^{(1)}=0.5$.

Both tables depict for $k=1,2,3$, $v_{8,k}(d,m,s)$, $(d,m,s)\in\Omega_{8,k}$ and$f_{8,k}(d)$, $2k+1\le d\le 8$ illustrating the numeric values of the involved probabilities. $v_{8,k}(d,m,s)$ and $f_{8,k}(d)$ were computed via Eqs. (29) and (17), respectively.

6 Discussion and further study

In this article we have derived exact closed form expressions for PMF $v_{n,k}(d,m,s)$, $n\ge 3$, $1\le k\le\lfloor(n-1)/2\rfloor$, $(d,m,s)\in\Omega_{n,k}$, of the RV $\mathbf{V}_{n,k}\mid\mathcal{M}_{n,k}$ defined on a $0-1$ sequence of homogeneous Markov-dependent trials. The method used is a combinatorial one relied on results exploiting the internal structure of such a sequence.

As it is noticed in the Introduction the application domain of runs contains a diverse range of fields. Indicative potential ones are next discussed.

Table 1 $0-1$ IID sequence with $p_1=0.5$

s	m	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$
				$v_{8,1}(d,m,s)$			
2	2	0.02739726	0.02283105	0.01826484	0.01369863	0.00913242	0.00456621
3	2		0.04566210	0.03652968	0.02739726	0.01826484	0.00913242
	3			0.01826484	0.02739726	0.02739726	0.01826484
4	2			0.05479452	0.04109589	0.02739726	0.01369863
	3				0.04109589	0.05479452	0.04109589
	4					0.00913242	0.01369863
5	2				0.05479452	0.03652968	0.01826484
	3					0.05479452	0.05479452
	4						0.01826484
6	2					0.04566210	0.02283105
	3						0.04566210
7	2						0.02739726
$f_{8,1}(d)$		0.02739726	0.06849315	0.12785388	0.20547945	0.28310503	0.28767123
				$v_{8,2}(d,m,s)$			
4	2			0.18518519	0.09259259	0.07407407	0.05555556
5	2				0.18518519	0.07407407	0.07407407
6	2					0.11111111	0.05555556
	3						0.01851852
7	2						0.07407407
$f_{8,2}(d)$				0.18518519	0.27777778	0.25925925	0.27777778
				$v_{8,3}(d,m,s)$			
6	2					0.40000000	0.20000000
7	2						0.40000000
$f_{8,3}(d)$						0.40000000	0.60000000

Table 2 0 − 1 HMC1 sequence with $p_{00} = p_{11} = 0.9, p_1^{(1)} = 0.5$

s	m	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$
				$v_{8,1}(d, m, s)$			
2	2	0.00914441	0.00872875	0.00831310	0.00789744	0.00748179	0.03366804
3	2		0.01745750	0.01662619	0.01579488	0.01496357	0.06733609
	3			0.00010263	0.00019500	0.00027710	0.00166262
4	2			0.02493929	0.02369233	0.02244536	0.10100413
	3				0.00029250	0.00055421	0.00374089
	4					0.00000114	0.00001539
5	2				0.03158977	0.02992715	0.13467217
	3					0.00055421	0.00498786
	4						0.00002053
6	2					0.03740894	0.16834021
	3						0.00415655
7	2						0.20200826
$f_{8,1}(d)$		0.00914441	0.02618626	0.04998121	0.07946192	0.11361346	0.72161274
				$v_{8,2}(d, m, s)$			
4	2			0.02225160	0.02081956	0.01806565	0.08228685
5	2				0.04163913	0.03569068	0.16259088
6	2					0.05353602	0.24091210
	3						0.00099141
7	2						0.32121613
$f_{8,2}(d)$				0.02225160	0.06245869	0.10729236	0.80799735
				$v_{8,3}(d, m, s)$			
6	2					0.06896552	0.31034483
7	2						0.62068966
$f_{8,3}(d)$						0.06896552	0.93103448

Encoding, compression and transmission of digital information calls for the understanding the distributions of runs of 1s or 0s. Such a knowledge helps in analyzing, and also in comparing, several techniques used in communication networks. In such networks 0 − 1 data ranging from a few kilobytes (e.g. e-mails) to many gigabytes of greedy multimedia applications (e.g. video on demand) are highly encoded, decoded and eventually proceeded under security. For details, see e.g., Sinha and Sinha (2009), Makri and Psillakis (2011a) and Tabatabaei and Zivic (2015).

An area where the study of runs of 1s and 0s has become increasingly useful is the field of bioinformatics or computational biology. For instance, molecular biologists design similarity tests between two DNA sequences where a 1 is interpreted as a match of the sequences at a given position and everything else as a 0. Moreover, the probabilistic analysis of such sequences according to the form, the length and the number of detected patterns as well as of the positions and the lengths of the segments of the sequence in which they are concentrated, probably suggests a functional reason for the internal structure of the examined sequence. The latter facts might be useful in suggesting a further investigation of the underline sequence(s) by biologists. See, e.g. Avery and Henderson (1999), Benson (1999) and Nuel et al. (2010).

Another active area where run statistics, in particular $G_{n,k}$ and $S_{n,k}$, have interesting statistical applications is that connected to hypothesis testing; e.g., in tests of randomness.

For a systematic study of such a topic, we refer among others, the works of Koutras and Alexandrou (1997) and Antzoulakos et al. (2003).

Accordingly, it is reasonable for one to use the exact expressions obtained for $v_{n,k}(d,m,s)$ in applications like the ones mentioned above. This is so, because this distribution, as a joint one, is more flexible than each one of its marginals which have been used in such applications. See, e.g. Lou (2003), Makri and Psillakis (2011b) and Arapis et al. (2016).

Moreover, in handling 0 - 1 sequences of a large length, with dependent or not elements, a Monte - Carlo simulation, based on Eqs. (1) - (4) would be a useful tool in obtaining approximate values for $v_{n,k}(d,m,s)$. In addition, the general approximating methods, suggested by Johnson and Fu (2014), might be helpful in deriving approximate values for $f_{n,k}(d)$.

Acknowledgements
The authors wish to thank the Editor for the thorough reading, and the anonymous reviewers for useful comments and suggestions which improved the article.

Authors' contributions
The authors, ANA, FSM and ZMP with the consultation of each other carried out this work and drafted the manuscript together. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Mathematics, University of Patras, 26500 Patras, Greece. [2]Department of Physics, University of Patras, 26500 Patras, Greece.

Received: 29 March 2017 Accepted: 18 October 2017
Published online: 15 November 2017

References

Antzoulakos, DL, Bersimis, S, Koutras, MV: On the distribution of the total number of run lengths. Ann. Inst. Statist. Math. **55**, 865–884 (2003)

Antzoulakos, DL, Chadjiconstantinidis, S: Distributions of numbers of success runs of fixed length in Markov dependent trials. Ann. Inst. Statist. Math. **53**, 559–619 (2001)

Arapis, AN, Makri, FS, Psillakis, ZM: On the length and the position of the minimum sequence containing all runs of ones in a Markovian binary sequence. Statist. Probab. Lett. **116**, 45–54 (2016)

Arapis, AN, Makri, FS, Psillakis, ZM: Distribution of statistics describing concentration of runs in non homogeneous Markov-dependent trials. Commun. Statist. Theor. Meth. (2017). doi:10.1080/03610926.2017.1337144

Avery, PJ, Henderson, D: Fiting Markov chain models to discrete state series such as DNA sequences. Appl. Statist. **48** (Part 1), 53–61 (1999)

Balakrishnan, N, Koutras, MV: Runs and Scans with Applications. Wiley, New York (2002)

Benson, G: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**, 573–580 (1999)

Eryilmaz, S: Some results associated with the longest run statistic in a sequence of Markov dependent trials. Appl. Math. Comput. **175**, 119–130 (2006)

Eryilmaz, S: Discrete time shock models involving runs. Statist. Probab. Lett. **107**, 93–100 (2015)

Eryilmaz, S: Generalized waiting time distributions associated with runs. Metrika. **79**, 357–368 (2016)

Eryilmaz, S: The concept of weak exchangeability and its applications. Metrika. **80**, 259–271 (2017)

Eryilmaz, S, Yalcin, F: Distribution of run statistics in partially exchangeable processes. Metrika. **73**, 293–304 (2011)

Feller, W: An Introduction to Probability Theory and Its Applications. 3rd Ed., Vol. I. Wiley, New York (1968)

Fu, JC, Lou, WYW: Distribution Theory of Runs and Patterns and Its Applications: A finite Markov chain imbedding approach. World Scientific, River Edge (2003)

Johnson, BC, Fu, JC: Approximating the distributions of runs and patterns. J. Stat. Distrib. Appl. **1:5**, 1–15 (2014)

Koutras, MV: Applications of Markov chains to the distribution of runs and patterns. In: Shanbhag, DN, Rao, CR (eds.) Handbook of Statistics, pp. 431–472. Elsevier, North-Holland, (2003)

Koutras, MV, Alexandrou, V: Non-parametric randomness tests based on success runs of fixed length. Statist. Probab. Lett. **32**, 393–404 (1997)

Koutras, VM, Koutras, MV, Yalcin, F: A simple compound scan statistic useful for modeling insurance and risk management problems. Insur. Math. Econ. **69**, 202–209 (2016)

Lou, WYW: The exact distribution of the k-tuple statistic for sequence homology. Statist. Probab. Lett. **61**, 51–59 (2003)

Makri, FS, Philippou, AN, Psillakis, ZM: Success run statistics defined on an urn model. Adv. Appl. Prob. **39**, 991–1019 (2007)

Makri, FS, Psillakis, ZM: On success runs of a fixed length in Bernoulli sequences: Exact and asymptotic results. Comput. Math. Appl. **61**, 761–772 (2011a)

Makri, FS, Psillakis, ZM: On runs of length exceeding a threshold: normal approximation. Stat. Papers. **52**, 531–551 (2011b)

Makri, FS, Psillakis, ZM: On ℓ-overlapping runs of ones of length k in sequences of independent binary random variables. Commun. Statist. Theor. Meth. **44**, 3865–3884 (2015)

Makri, FS, Psillakis, ZM, Arapis, AN: Counting runs of ones with overlapping parts in binary strings ordered linearly and circularly. Intern. J. Statist. Probab. **2**, 50–60 (2013)

Makri, FS, Psillakis, ZM, Arapis, AN: Length of the minimum sequence containing repeats of success runs. Statist. Probab. Lett. **96**, 28–37 (2015)

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

Mytalas, GC, Zazanis, MA: Central limit theorem approximations for the number of runs in Markov-dependent binary sequences. J. Statist. Plann. Infer. **143**, 321–333 (2013)

Mytalas, GC, Zazanis, MA: Central limit theorem approximations for the number of runs in Markov-dependent multi-type sequences. Commun. Statist. Theor. Meth. **43**, 1340–1350 (2014)

Nuel, G, Regad, L, Martin, J, Camproux, A-C: Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. Algorithm Mol. Biol. **5**, 1–18 (2010)

Riordan, AM: An Introduction to Combinatorial Analysis. Second Ed. John Wiley, New York (1964)

Sinha, K, Sinha, BP: On the distribution of runs of ones in binary trials. Comput. Math. Appl. **58**, 1816–1829 (2009)

Tabatabaei, SAH, Zivic, N: A review of approximate message authentication codes. In: Zivic, N (ed.) Robust Image Authentication in the Presence of Noise, pp. 106–127. Springer International Publishing AG, Cham (ZG), Switzerland, (2015)