Stat Papers (2010) 51:959–973
DOI 10.1007/s00362-008-0191-7

REGULAR ARTICLE

Run statistics in a sequence of arbitrarily dependent binary trials

Sevcan Demir · Serkan Eryılmaz

Received: 25 July 2008 / Revised: 24 November 2008 / Published online: 9 December 2008

Abstract Let $\{Z_i\}_{i\geq 1}$ be an arbitrary sequence of trials with two possible outcomes either success (1) or failure (0). General expressions for the exact distributions of runs, both success and failure, in $Z_1, \ldots, Z_n$ are presented. Our method is based on the use of joint distribution of success and failure run lengths and unifies the results on distribution of runs. As a special case of our results we obtain the distributions of runs for various binary sequences. As illustrated in the paper the results enable us to derive the distribution of runs for binary trials arising in urn models.

Keywords Binary trials · Exchangeable trials · Markov dependent trials · Records · Runs · Urn model

1 Introduction

Runs based on a sequence of binary trials have attracted much attention in the literature because of the wide range of applications in many areas including computer science, molecular biology, statistical reliability and quality, and statistical hypothesis testing. Past and current developments on the topic are well documented in Balakrishnan and Koutras (2002) as well as in Fu and Lou (2003). Recent discussions on the topic appear in the works of Eryılmaz (2005), Makri and Philippou (2005), Kong (2006), Makri et al. (2007a), Makri et al. (2007b), Eryılmaz and Demir (2007).

S. Demir
Department of Statistics, Ege University, 35100 Bornova, Izmir, Turkey
e-mail: sevcan.demir@ege.edu.tr

S. Eryılmaz (✉)
Department of Mathematics, Izmir University of Economics, 35330 Balcova, Izmir, Turkey
e-mail: serkan.eryilmaz@ieu.edu.tr

Let $\{Z_i\}_{i\geq 1}$ be a sequence of trials with two possible outcomes either success (1) or failure (0). The main problem in run theory is to obtain the distributions of run statistics under the various types of dependencies among the elements of $\{Z_i\}_{i\geq 1}$. The problem has been extensively studied in the literature whenever the elements of $\{Z_i\}_{i\geq 1}$ are independent (identical/nonidentical) or exchangeable or dependent in a Markovian fashion (homogeneous/nonhomogeneous). However, in many cases, the elements of $\{Z_i\}_{i\geq 1}$ may not be independent but dependent in a form different from Markov dependence.

The distribution of runs under particular assumptions on $\{Z_i\}_{i\geq 1}$ can be obtained by a simple unified combinatorial approach as shown in the present paper (see Corollary 2).

Total number of successes (1s), to be denoted by S_n, among $Z_1, Z_2, \ldots, Z_n$ can be seen as the simplest run statistic. The distribution of $S_n = \sum_{i=1}^{n} Z_i$ has been widely studied in the literature under the various types of possible dependencies among $Z_1, Z_2, \ldots, Z_n$. Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. Then, for $k = 0, 1, \ldots, n$,

$$P\{S_n = k\} = \sum_{j=k}^{n} (-1)^{j-k} \binom{j}{k} T_j,$$

where

$$T_j = \sum_{1\leq i_1<i_2<\cdots<i_j\leq n} P\left\{Z_{i_1} = 1, Z_{i_2} = 1, \ldots, Z_{i_j} = 1\right\},$$

(see, e.g. Blom et al. 1994, p. 30).

The problem of testing randomness of observations arises in many fields. The tests based on runs, in particular the total number of runs and the longest run, are best known and easiest to apply for testing randomness in a sequence of observations. The hypothesis of randomness implies that one is considering a sequence of binary trials which are independent and identically distributed (i.i.d.). Thus the main interest in this problem is to test the null hypothesis H_0 of i.i.d. against the alternative hypothesis H_a of dependence. To compute the powers of the test it is necessary to derive the distributions of test (run) statistics under H_a of dependence. This is not an easy task and derivations heavily depend on the dependence among $Z_1, Z_2, \ldots, Z_n$.

In the present paper, we provide some general expressions for the distributions of runs based on a sequence of arbitrarily dependent binary trials. That is, we do not suppose either the trials are identically distributed or independent. The results enable us to compute the distribution of runs for various dependence structures among $Z_1, Z_2, \ldots, Z_n$. In the second section, we provide general expressions for the distribution of runs. In Sect. 3, we illustrate our results for various binary sequences including exchangeable binary trials, binary trials arising in urn models, homogeneous Markov dependent binary trials, binary trials arising in record sequences.

2 Distribution of runs

Let $\{Z_i\}_{i\geq 1}$ be an arbitrary sequence of binary trials. Denote by $R_n^{(1)}$ and $R_n^{(0)}$ the number of success runs and failure runs in $Z_1, Z_2, \ldots, Z_n$, respectively. Let $\theta_i^{(j)}$ denote the length of the ith run of type j $(j = 0, 1)$ in $Z_1, Z_2, \ldots, Z_n$. For example in a sequence of ten trials 1011111001, we have $R_n^{(0)} = 2$, $R_n^{(1)} = 3$, and $\theta_1^{(1)} = 1$, $\theta_2^{(1)} = 5$, $\theta_3^{(1)} = 1$, $\theta_1^{(0)} = 1$, $\theta_2^{(0)} = 2$. The distribution of any run statistic defined on a sequence $Z_1, Z_2, \ldots, Z_n$ can be evaluated using the distribution of the vector $(\theta_1^{(1)}, \ldots, \theta_{r_1}^{(1)}, \theta_1^{(0)}, \ldots, \theta_{r_2}^{(0)}, R_n^{(1)} = r_1, R_n^{(0)} = r_2)$. The method which is based on the use of the joint distribution of the run lengths and the number of runs has also been used in a recent work of Eryılmaz (2008a).

In this section, we provide general expressions for the distribution of runs without making any assumption on a binary sequence $\{Z_i\}_{i\geq 1}$. Let us start our discussion considering the probability of the event

$$E_n(\vec{i}_{r_1}, \vec{j}_{r_2}): \left\{\theta_1^{(1)} = i_1, \ldots, \theta_{r_1}^{(1)} = i_{r_1}, \theta_1^{(0)} = j_1, \ldots, \theta_{r_2}^{(0)} = j_{r_2}, R_n^{(1)} = r_1, R_n^{(0)} = r_2\right\},$$

where $\vec{i}_{r_1} = (i_1, \ldots, i_{r_1})$ and $\vec{j}_{r_2} = (j_1, \ldots, j_{r_2})$. The sequence $Z_1, \ldots, Z_n$ has one of the following four forms for the occurrence of the event $E_n(\vec{i}_{r_1}, \vec{j}_{r_2})$:

$$A_n(\vec{i}_{r_1}, \vec{j}_{r_2}): \quad \overbrace{0\ldots0}^{j_1} \mid \overbrace{1\ldots1}^{i_1} \mid \overbrace{0\ldots0}^{j_2} \mid \overbrace{1\ldots1}^{i_2} \mid \ldots \mid \overbrace{0\ldots0}^{j_{r_2}} \mid \overbrace{1\ldots1}^{i_{r_1}} \tag{1}$$

$$B_n(\vec{i}_{r_1}, \vec{j}_{r_2}): \quad \overbrace{0\ldots0}^{j_1} \mid \overbrace{1\ldots1}^{i_1} \mid \overbrace{0\ldots0}^{j_2} \mid \overbrace{1\ldots1}^{i_2} \mid \ldots \mid \overbrace{0\ldots0}^{j_{r_2-1}} \mid \overbrace{1\ldots1}^{i_{r_1}} \mid \overbrace{0\ldots0}^{j_{r_2}} \tag{2}$$

$$C_n(\vec{i}_{r_1}, \vec{j}_{r_2}): \quad \overbrace{1\ldots1}^{i_1} \mid \overbrace{0\ldots0}^{j_1} \mid \overbrace{1\ldots1}^{i_2} \mid \overbrace{0\ldots0}^{j_2} \mid \ldots \mid \overbrace{0\ldots0}^{j_{r_2}} \mid \overbrace{1\ldots1}^{i_{r_1}} \tag{3}$$

$$D_n(\vec{i}_{r_1}, \vec{j}_{r_2}): \quad \overbrace{1\ldots1}^{i_1} \mid \overbrace{0\ldots0}^{j_1} \mid \overbrace{1\ldots1}^{i_2} \mid \overbrace{0\ldots0}^{j_2} \mid \ldots \mid \overbrace{0\ldots0}^{j_{r_2-1}} \mid \overbrace{1\ldots1}^{i_{r_1}} \mid \overbrace{0\ldots0}^{j_{r_2}}, \tag{4}$$

It should be noted that the definitions of (1)–(4) are based on the arguments given in the proof of Theorem 2.1 of Gibbons and Chakraborti (2003, p. 78). It is clear that for the first and fourth forms total number of success and failure runs are both equal $(r_1 = r_2)$ and we have $r_2 = r_1 + 1$ and $r_1 = r_2 + 1$ for the second and third forms, respectively. Thus we proved the following.

Lemma 1 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. Then*

$$P\left\{E_n(\vec{i}_{r_1}, \vec{j}_{r_2})\right\} = \begin{cases} P\left(A_n(\vec{i}_{r_1}, \vec{j}_{r_2})\right) + P\left(D_n(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \text{if } r_1 = r_2 \\ P\left(B_n(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \text{if } r_2 = r_1 + 1 \\ P\left(C_n(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \text{if } r_1 = r_2 + 1. \end{cases}$$

We have the following relationships:

$$P\left(B_n(\vec{i}_{r_1}, \vec{j}_{r_2})\right) = P\left(A_{n-j_{r_2}}(\vec{i}_{r_1}, \vec{j}_{r_2-1}), Z_{n-j_{r_2}+1} = 0, \ldots, Z_n = 0\right),$$

and

$$P\left(D_n(\vec{i}_{r_1}, \vec{j}_{r_2})\right) = P\left(C_{n-j_{r_2}}(\vec{i}_{r_1}, \vec{j}_{r_2-1}), Z_{n-j_{r_2}+1} = 0, \ldots, Z_n = 0\right).$$

If S_n denotes the total number of successes in $Z_1, Z_2, \ldots, Z_n$ then we also have the following.

$$P\left\{E_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right\} = P\left\{E_n(\vec{i}_{r_1}, \vec{j}_{r_2}), S_n = n_1\right\}$$

$$= P\left\{\theta_1^{(1)} = i_1, \ldots, \theta_{r_1}^{(1)} = i_{r_1}, \theta_1^{(0)} = j_1, \ldots, \theta_{r_2}^{(0)} = j_{r_2}, R_n^{(1)} = r_1, R_n^{(0)} = r_2, S_n = n_1\right\}$$

$$= \begin{cases} P\left(A_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) + P\left(D_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \text{if } r_1 = r_2 \\ P\left(B_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \text{if } r_2 = r_1 + 1 \\ P\left(C_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \text{if } r_1 = r_2 + 1, \end{cases}$$

where $\sum_{j=1}^{r_1} i_j = n_1, \sum_{i=1}^{r_2} j_i = n - n_1$ and the events $A, B, C, D,$ and E with superscript n_1 represent the corresponding forms such that the sequence contains n_1 successes.

Lemma 2 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. Then for $r_1, r_2 > 0$*

$$P\left\{R_n^{(1)} = r_1, R_n^{(0)} = r_2\right\}$$
$$= \begin{cases} \sum_{n_1=r_1}^{n-r_2} \sum_{\vec{i}_{r_1} \in I} \sum_{\vec{j}_{r_2} \in J} \left[P\left(A_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) + P\left(D_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right)\right] & \textit{if } r_1 = r_2 \\ \sum_{n_1=r_1}^{n-r_2} \sum_{\vec{i}_{r_1} \in I} \sum_{\vec{j}_{r_2} \in J} P\left(B_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \textit{if } r_2 = r_1 + 1 \\ \sum_{n_1=r_1}^{n-r_2} \sum_{\vec{i}_{r_1} \in I} \sum_{\vec{j}_{r_2} \in J} P\left(C_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) & \textit{if } r_1 = r_2 + 1, \end{cases}$$

where

$$I = \left\{(i_1, \ldots, i_{r_1}) : i_1 + \cdots + i_{r_1} = n_1; i_k > 0, k = 1, \ldots, r_1\right\}$$

and

$$J = \left\{(j_1, \ldots, j_{r_2}) : j_1 + \cdots + j_{r_2} = n - n_1; j_k > 0, k = 1, \ldots, r_2\right\}.$$

Proof Conditioning on the total number of successes and noting that $\sum_{j=1}^{r_1} i_j = n_1$ we have

$$
\begin{aligned}
&P\left\{R_n^{(1)} = r_1, R_n^{(0)} = r_2\right\} \\
&= \sum_{n_1=r_1}^{n-r_2} \underset{i_1+\cdots+i_{r_1}=n_1}{\sum_{i_1}\cdots\sum_{i_{r_1}}} \underset{j_1+\cdots+j_{r_2}=n-n_1}{\sum_{j_1}\cdots\sum_{j_{r_2}}} P\left\{\theta_1^{(1)} = i_1, \ldots, \theta_{r_1}^{(1)} = i_{r_1},\right. \\
&\qquad\qquad \left.\theta_1^{(0)} = j_1, \ldots, \theta_{r_2}^{(0)} = j_{r_2}, R_n^{(1)} = r_1, R_n^{(0)} = r_2, S_n = n_1\right\}.
\end{aligned}
$$

The result now follows considering the cases $r_1 = r_2, r_2 = r_1 \pm 1$. □

Remark 1 The sums over the sets I and J contain $\binom{n_1-1}{r_1-1}$ and $\binom{n-n_1-1}{r_2-1}$ terms, respectively. Therefore there are $\sum_{n_1=r_1}^{n-r_2} \binom{n_1-1}{r_1-1}\binom{n-n_1-1}{r_2-1}$ terms contributing to the sum giving $P\{R_n^{(1)} = r_1, R_n^{(0)} = r_2\}$.

Corollary 1 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. If the probabilities associated with the forms (1)–(4) depend on $\vec{i}_{r_1}$ and $\vec{j}_{r_2}$ only through the values of $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$ (this is the case whenever $Z_1, Z_2, \ldots, Z_n$ are exchangeable or homogeneous Markov dependent and hence i.i.d.) then we have*

$$
\begin{aligned}
&P\left\{R_n^{(1)} = r_1, R_n^{(0)} = r_2\right\} \\
&= \begin{cases}
\sum_{n_1=r_1}^{n-r_2} \binom{n_1-1}{r_1-1}\binom{n-n_1-1}{r_2-1}\left[P\left(A_n^{n_1}(r_1, r_2)\right) + P\left(D_n^{n_1}(r_1, r_2)\right)\right] & \text{if } r_1 = r_2 \\
\sum_{n_1=r_1}^{n-r_2} \binom{n_1-1}{r_1-1}\binom{n-n_1-1}{r_2-1} P\left(B_n^{n_1}(r_1, r_2)\right) & \text{if } r_2 = r_1 + 1 \\
\sum_{n_1=r_1}^{n-r_2} \binom{n_1-1}{r_1-1}\binom{n-n_1-1}{r_2-1} P\left(C_n^{n_1}(r_1, r_2)\right) & \text{if } r_1 = r_2 + 1.
\end{cases}
\end{aligned}
$$

Note that since the probabilities of (1)–(4) depend on $\vec{i}_{r_1}$ and $\vec{j}_{r_2}$ only through the values of $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$ in Corollary 1 we use $P(A_n^{n_1}(r_1, r_2))$ instead of $P(A_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2}))$.

Run statistics can be expressed as a function of run lengths and the total number of runs. That is, given the total number of success runs $R_n^{(1)} = r_1$, a run statistic associated with successes can be viewed mathematically as

$$
X_n^{(1)} = \phi\left(\theta_1^{(1)}, \ldots, \theta_{r_1}^{(1)}\right), \tag{5}
$$

Similarly a run statistic associated with failures given the total number of failure runs $R_n^{(0)} = r_2$ can be represented as

$$
X_n^{(0)} = \psi\left(\theta_1^{(0)}, \ldots, \theta_{r_2}^{(0)}\right), \tag{6}
$$

where ϕ and ψ are Borel measurable functions.

The following theorem provides the joint distribution of $X_n^{(1)}$ and $X_n^{(0)}$.

Theorem 1 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials and $X_n^{(1)}$ and $X_n^{(0)}$ denote the run statistics associated with successes and failures in $Z_1, Z_2, \ldots, Z_n$, respectively. Then*

$$P\left\{X_n^{(1)} \in B_1, X_n^{(0)} \in B_2\right\} = \sum_{r_1}\sum_{r_2}\sum_{n_1}\sum_{\vec{i}_{r_1} \in I(B_1)}\sum_{\vec{j}_{r_2} \in J(B_2)} P\left(E_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right),$$

where B_1 and B_2 are Borel sets and

$$I(B_1) = \left\{\vec{i}_{r_1} : i_1 + \cdots + i_{r_1} = n_1; \phi(i_1, \ldots, i_{r_1}) \in B_1; i_j > 0, j = 1, \ldots, r_1\right\}$$

and

$$J(B_2) = \left\{\vec{j}_{r_2} : j_1 + \cdots + j_{r_2} = n - n_1; \psi(j_1, \ldots, j_{r_2}) \in B_2; j_i > 0, i = 1, \ldots, r_2\right\}.$$

Proof Using the representations given in (5) and (6) and conditioning on the total number of runs we have

$$P\left\{X_n^{(1)} \in B_1, X_n^{(0)} \in B_2\right\} = \sum_{r_1}\sum_{r_2} P\left\{\phi(\theta_1^{(1)}, \ldots, \theta_{r_1}^{(1)}) \in B_1, \right.$$
$$\left. \psi(\theta_1^{(0)}, \ldots, \theta_{r_2}^{(0)}) \in B_2, R_n^{(1)} = r_1, R_n^{(0)} = r_2\right\}.$$

Now conditioning on the total number of successes one obtains

$$P\left\{X_n^{(1)} \in B_1, X_n^{(0)} \in B_2\right\}$$
$$= \sum_{r_1}\sum_{r_2}\sum_{n_1}\sum_{\substack{i_1+\cdots+i_{r_1}=n_1 \\ \phi(i_1,\ldots,i_{r_1})\in B_1}}\cdots\sum\;\sum_{\substack{j_1+\cdots+j_{r_2}=n-n_1 \\ \psi(j_1,\ldots,j_{r_2})\in B_2}}\cdots\sum P\left\{\theta_1^{(1)} = i_1, \ldots, \theta_{r_1}^{(1)} = i_{r_1}, \right.$$
$$\left. \theta_1^{(0)} = j_1, \ldots, \theta_{r_2}^{(0)} = j_{r_2}, R_n^{(1)} = r_1, R_n^{(0)} = r_2, S_n = n_1\right\}.$$

Thus the proof is completed. □

We readily get the following result for the case whenever the probabilities associated with the forms (1)–(4) depend on $\vec{i}_{r_1}$ and $\vec{j}_{r_2}$ only through the values of $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$. This is the case for i.i.d., exchangeable, and homogeneous Markov dependent binary trials as it will be illustrated in Sect. 3.

Corollary 2 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. If the probabilities associated with the forms (1)–(4) depend on $\vec{i}_{r_1}$ and $\vec{j}_{r_2}$ only through the values of $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$ then we have*

$$P\left\{X_n^{(1)} \in B_1, X_n^{(0)} \in B_2\right\} = \sum_{r_1}\sum_{r_2}\sum_{n_1} |I(B_1)|\,|J(B_2)|\, P\left(E_n^{n_1}(r_1, r_2)\right),$$

where $|A|$ shows the cardinality of the set A.

Note that the problem of finding the cardinalities of $|I(B_1)|$ and $|J(B_2)|$ is combinatorial one. For example $|I(B_1)|$ corresponds to the total number of integer solutions to the equation $i_1 + i_2 + \cdots + i_{r_1} = n_1$ s.t. $\phi(i_1, \ldots, i_{r_1}) \in B_1; i_j > 0, j = 1, \ldots, r_1$.

Theorem 1 and Corollary 2 enable us to get the distribution of various run statistics for particular selections of the functions ϕ and ψ. Below we obtain the distributions of some well known run statistics.

Let $L_n^{(1)}$ and $L_n^{(0)}$ denote the longest run of successes and failures in $Z_1, Z_2, \ldots, Z_n$, respectively. We can express the random variables $L_n^{(1)}$ and $L_n^{(0)}$ as

$$L_n^{(1)} = \max_{1\le i\le R_n^{(1)}} \theta_i^{(1)} \quad \text{and} \quad L_n^{(0)} = \max_{1\le i\le R_n^{(0)}} \theta_i^{(0)}.$$

The proof of the following result readily follows taking $\phi(x_1, \ldots, x_r) = \max(x_1, \ldots, x_r)$, $\psi(x_1, \ldots, x_r) = \max(x_1, \ldots, x_r)$ and $B_1 = (0, k_1)$, $B_2 = (0, k_2)$ in Theorem 1.

Corollary 3 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. Then*

$$P\left\{L_n^{(1)} < k_1, L_n^{(0)} < k_2\right\} = \sum_{r_1}\sum_{r_2}\sum_{n_1}\sum_{\vec{i}_{r_1}\in I(k_1)}\sum_{\vec{j}_{r_2}\in J(k_2)} P\left(E_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right),$$

where

$$I(k_1) = \left\{\left(i_1, \ldots, i_{r_1}\right) : i_1 + \cdots + i_{r_1} = n_1; 0 < i_j < k_1, j = 1, \ldots, r_1\right\}$$

and

$$J(k_2) = \left\{\left(j_1, \ldots, j_{r_2}\right) : j_1 + \cdots + j_{r_2} = n - n_1; 0 < j_i < k_2, i = 1, \ldots, r_2\right\}.$$

Remark 2 The sums over the sets $I(k_1)$ and $J(k_2)$ contain $N(r_1, k_1, n_1)$ and $N(r_2, k_2, n-n_1)$ terms, respectively, where $N(a, b, c)$ denotes the total number of integer solutions to the equation $x_1 + x_2 + \cdots + x_a = c$, s.t. $0 < x_i < b$, $i = 1, 2, \ldots, a$, and is given by

$$N(a, b, c) = \sum_{j=0}^{a} (-1)^j \binom{a}{j}\binom{c - j(b-1) - 1}{a-1}.$$

See, e.g. Charalambides (2002).

Remark 3 The marginal distribution of $L_n^{(1)}$ ($L_n^{(0)}$) can be obtained taking $k_2 = n + 1$ ($k_1 = n + 1$) in Corollary 3.

Remark 4 The distribution of the longest run of any type, denoted by $L_n = \max(L_n^{(1)}, L_n^{(0)})$, follows from Corollary 3 since

$$P\{L_n < k\} = P\left\{L_n^{(1)} < k, L_n^{(0)} < k\right\}.$$

Corollary 4 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. If the probabilities associated with the forms (1)–(4) depend on $\vec{i}_{r_1}$ and $\vec{j}_{r_2}$ only through the values of $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$ then we have*

$$P\left\{L_n^{(1)} < k_1, L_n^{(0)} < k_2\right\} = \sum_{r_1}\sum_{r_2}\sum_{n_1} N(r_1, k_1, n_1) N(r_2, k_2, n - n_1) P\left(E_n^{n_1}(r_1, r_2)\right).$$

Another popular run statistic is the total number of runs of length at least k. Let $G_{n,k_1}^{(1)}$ ($G_{n,k_2}^{(0)}$) denote the total number of success (failure) runs of length at least k_1 (k_2) in $Z_1, Z_2, \ldots, Z_n$. These statistics can be expressed as

$$G_{n,k_1}^{(1)} = \sum_{i=1}^{R_n^{(1)}} I\left(\theta_i^{(1)} \geq k_1\right) \quad \text{and} \quad G_{n,k_2}^{(0)} = \sum_{i=1}^{R_n^{(0)}} I\left(\theta_i^{(0)} \geq k_2\right).$$

Note that the event $\{G_{n,k_1}^{(1)} \geq x\}$ is equivalent to $\left\{\theta^{(1)}_{R_n^{(1)}-x+1:R_n^{(1)}} \geq k_1\right\}$, where $\theta^{(1)}_{m:R_n^{(1)}}$ denotes the mth smallest among $\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta^{(1)}_{R_n^{(1)}}$. Now using Theorem 1 with $B_1 = [x, n_1)$ and $B_2 = [y, n - n_1)$ we have the following corollary.

Corollary 5 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. Then*

$$P\left\{G_{n,k_1}^{(1)} \geq x, G_{n,k_2}^{(0)} \geq y\right\} = \sum_{r_1}\sum_{r_2}\sum_{n_1}\sum_{\vec{i}_{r_1}\in I_x(k_1)}\sum_{\vec{j}_{r_2}\in J_y(k_2)} P\left(E_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right),$$

where

$$I_x(k_1) = \left\{\left(i_1, \ldots, i_{r_1}\right) : i_1 + \cdots + i_{r_1} = n_1; i_{r_1-x+1:r_1} \geq k_1\right\}$$

and

$$J_y(k_2) = \left\{\left(j_1, \ldots, j_{r_2}\right) : j_1 + \cdots + j_{r_2} = n - n_1; j_{r_2-y+1:r_2} \geq k_2\right\},$$

where $i_{m:r_1}$ ($j_{m:r_2}$) shows the mth smallest among $i_1, i_2, \ldots, i_{r_1}$ ($j_1, j_2, \ldots, j_{r_2}$).

It should be noted that, using Theorem 1 we can find the joint distributions not only the same type of runs for successes and failures but also the different types. That is, the forms of the functions ϕ and ψ can be chosen different from each other. For example choosing

$$\phi(x_1,\ldots,x_r) = \max(x_1,\ldots,x_r),$$
$$\psi(x_1,\ldots,x_r) = \sum_{m=1}^{r} x_m$$

we get the joint distribution of the longest success run and the total number of successes (failures) as provided in the following corollary.

Corollary 6 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. Then*

$$P\left\{L_n^{(1)} < k, S_n = n_1\right\} = \sum_{r_1}\sum_{r_2}\sum_{\vec{i}_{r_1}\in I(k)}\sum_{\vec{j}_{r_2}\in J} P\left(E_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right),$$

where

$$I(k) = \left\{\left(i_1,\ldots,i_{r_1}\right) : i_1 + \cdots + i_{r_1} = n_1; 0 < i_j < k, j = 1,\ldots,r_1\right\}$$

and

$$J = \left\{\left(j_1,\ldots,j_{r_2}\right) : j_1 + \cdots + j_{r_2} = n - n_1; j_i > 0, i = 1,\ldots,r_2\right\}.$$

Corollary 7 *Let $Z_1, Z_2, \ldots, Z_n$ be an arbitrary sequence of binary trials. If the probabilities associated with the forms (1)–(4) depend on $\vec{i}_{r_1}$ and $\vec{j}_{r_2}$ only through the values of $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$ then we have*

$$P\left\{L_n^{(1)} < k, S_n = n_1\right\} = \sum_{r_1}\sum_{r_2}\binom{n-n_1-1}{r_2-1} N(r_1,k,n_1) P\left(E_n^{n_1}(r_1,r_2)\right).$$

3 Particular cases

As it seen from the previous section, it is enough to compute the probabilities of the forms (1)–(4) for finding the distributions of runs. This can be done for a given structure of a binary sequence. Below we provide some examples for various type of binary trials.

3.1 Exchangeable binary trials

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of exchangeable binary trials. That is, the joint distribution of $Z_1, Z_2, \ldots, Z_n$ is invariant under permutation of its arguments. In this case $P(A_n(\vec{i}_{r_1}, \vec{j}_{r_2})) = P(D_n(\vec{i}_{r_1}, \vec{j}_{r_2})) = g(n, \sum_{j=1}^{r_1} i_j)$ if $r_1 = r_2$; 0 otherwise, $P(B_n(\vec{i}_{r_1}, \vec{j}_{r_2})) = g(n, \sum_{j=1}^{r_1} i_j)$ if $r_2 = r_1 + 1$; 0 otherwise, and $P(C_n(\vec{i}_{r_1}, \vec{j}_{r_2})) = g(n, \sum_{j=1}^{r_1} i_j)$ if $r_1 = r_2 + 1$; 0 otherwise, where $g(n, x)$ denotes the probability of getting x successes in $Z_1, Z_2, \ldots, Z_n$. Since the corresponding probabilities depend on $\vec{i}_{r_1}$ and $\vec{j}_{r_2}$ only through the values of $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$ the usage of the material presented in Sect. 2 with

$$\begin{aligned}P\left(A_n^{n_1}(r_1, r_2 = r_1)\right) &= P\left(B_n^{n_1}(r_1, r_2 = r_1 + 1)\right)\\ &= P\left(C_n^{n_1}(r_1, r_2 = r_1 - 1)\right) = P\left(D_n^{n_1}(r_1, r_2 = r_1)\right)\\ &= P\left\{Z_1 = 1, \ldots, Z_{n_1} = 1, Z_{n_1+1} = 0, \ldots, Z_n = 0\right\}\\ &= g(n, n_1)\end{aligned}$$

provides the distribution of runs for a sequence of exchangeable binary trials. Specifically using Corollary 4 the joint distribution of the longest success and longest failure run is

$$P\left\{L_n^{(1)} < k_1, L_n^{(0)} < k_2\right\} = \sum_{r_1}\sum_{r_2}\sum_{n_1} N(r_1, k_1, n_1)N(r_2, k_2, n - n_1)P\left(E_n^{n_1}(r_1, r_2)\right), \tag{7}$$

where $P(E_n^{n_1}(r_1, r_2)) = 2g(n, n_1)$ if $r_1 = r_2$ and $P(E_n^{n_1}(r_1, r_2)) = g(n, n_1)$ if $r_2 = r_1 \pm 1$. Equation (7) is consistent with the Corollary 5 of Eryılmaz (2008a).

For an illustration we compute the distribution of the longest success run for exchangeable trials arising in a record threshold model. Let $\{Y_i\}_{i\geq 1}$ be a sequence of i.i.d. random variables with continuous distribution function F. The random variable Y_j is called a record if $Y_j > Y_i$ for all $i = 1, 2, \ldots, j - 1$, where by convention Y_1 is a record. Let $u(k)$ be the time (index) of the kth record, $k = 1, 2, \ldots$ Then $u(1) = 1, u(k) = \min\{j : Y_j > Y_{u(k-1)}\}, k > 1$. Then $Y_{u(1)}, Y_{u(2)}, \ldots$ denote the record values associated with $\{Y_i\}_{i\geq 1}$. The probability density function of the rth record value ($Y_{u(r)}$) is given by

$$f_r(x) = \frac{1}{(r-1)!}\left(-\ln(\bar{F}(x))\right)^{r-1} f(x), \quad r > 1,$$

where $\bar{F}(x) = 1 - F(x)$ and $f(x)$ is the probability density function associated with $F(x)$ (Nevzorov 2001, p. 69). Let $Y_1', Y_2', \ldots, Y_n'$ be i.i.d random variables with continuous distribution function G and independent of $\{Y_i\}_{i\geq 1}$. Define

$$Z_i = \begin{cases} 1, & \text{if } Y_i' > Y_{u(r)} \\ 0, & \text{otherwise} \end{cases} \quad i = 1, 2, \ldots, n.$$

The random variables $Z_1, Z_2, \ldots, Z_n$ are exchangeable and under the hypothesis $\mathrm{H}_0 : F = G$ we have

Table 1 Distribution and expectation of the longest success run for the record threshold model

r	k	$P\{L_n^{(1)} < k\}$	r	k	$P\{L_n^{(1)} < k\}$
2	1	0.2745	3	1	0.4853
	2	0.6348		2	0.8338
	3	0.8082		3	0.9371
	4	0.8920		4	0.9726
	5	0.9354		5	0.9868
	6	0.9605		6	0.9932
	7	0.9745		7	0.9963
	8	0.9830		8	0.9979
	9	0.9883		9	0.9988
	10	0.9917		10	0.9992
$E(L_n^{(1)})$		1.5572			0.8736

$$
\begin{aligned}
g(n, n_1) &= P\left\{Y_1' > Y_{u(r)}, \ldots, Y_{n_1}' > Y_{u(r)}, Y_{n_1+1}' \leq Y_{u(r)}, \ldots, Y_n' \leq Y_{u(r)}\right\} \\
&= \int_{-\infty}^{\infty} (\bar{F}(x))^{n_1} (F(x))^{n-n_1} \frac{1}{(r-1)!} \left(-\ln \bar{F}(x)\right)^{r-1} dF(x) \\
&= \frac{1}{(r-1)!} \int_0^1 u^{n_1} (1-u)^{n-n_1} (-\ln u)^{r-1} du \\
&= \sum_{i=0}^{n-n_1} (-1)^i \binom{n-n_1}{i} \frac{1}{(n_1+i+1)^r}, \qquad n_1 \geq 0.
\end{aligned}
$$

The latter equation is obtained using the binomial expansion for the term $(1-u)^{n-n_1}$. Table 1 contains the distribution of the longest success run (longest run of exceedances) for $n = 10$ and $r = 2, 3$. We observe that an increase in r leads to a decrease in the mean length of the longest success run.

3.2 Binary trials arising in urn models

Urn models have been a popular topic in probability and statistics. A class of these models is the Pólya urn which was introduced by Eggenberger and Pólya (1923). For a detailed and lucid review of urn models we refer to Johnson and Kotz (1977). A two-color Pólya urn is an urn which initially contains m_i balls of color i, $i = 1, 2$. At each stage a ball is drawn from the urn and its color is noted. If a ball of color i is drawn at stage, a_{ij} balls of color j, $j = 1, 2$ are added to the urn. This scheme is described by 2×2 addition matrix $(a_{ij}), i, j = 1, 2$ whose rows are indexed by the color of the ball selected and whose columns are indexed by the color of the ball added. Let the urn initially contains m_1 black and m_2 white balls and define

$$Z_i = \begin{cases} 1 & \text{if the } i\text{th ball selected is black} \\ 0 & \text{if the } i\text{th ball selected is white} \end{cases}, \quad i = 1, 2, \ldots$$

The resulting binary trials $Z_1, Z_2, \ldots$ are generally dependent and the type of this dependence is determined by the structure of the addition matrix. Distribution of success runs in $Z_1, Z_2, \ldots$ have been investigated in the literature for the diagonal addition matrix whose entries are $a_{11} = a_{22} = a$ and $a_{12} = a_{21} = 0$ (Sen et al. 2002; Makri et al. 2007b,c; Eryılmaz 2008b). For this scheme a ball is drawn form the urn and then replaced together with a balls of the same color. Thus this scheme generates an exchangeable binary sequence with the probability of getting n_1 successes-black balls- in $Z_1, Z_2, \ldots, Z_n$ given by

$$g(n, n_1) = \frac{\prod_{j=0}^{n_1-1}(m_1 + j \cdot a) \prod_{j=0}^{n-n_1-1}(m_2 + j \cdot a)}{\prod_{j=0}^{n-1}(m_1 + m_2 + j \cdot a)}, \quad 0 \le n_1 \le n.$$

The usage of $g(n, n_1)$ in the formulas presented in Sect. 2 provides the joint distribution of runs for this exchangeable urn scheme. These results extend the results of Sen et al. (2002), Eryılmaz and Demir (2007), Makri et al. (2007b,c) since they enable us to obtain the joint distributions not only the same type of runs for successes and failures but also the different types.

The random trials $Z_1, Z_2, \ldots$ may not be exchangeable for particular selections of addition matrix. That is, different schemes might generate a sequence whose elements are dependent but not exchangeable. For example consider the case $a_{11} = -1$, $a_{12} = 1$, $a_{21} = a_{22} = 0$ (Styve 1965), i.e. any black ball is replaced by a white one, whereas white balls are returned to the urn. This is a useful model for inspection of items in a lot, in which items found to be defective (black) are immediately replaced by non-defective (white) items. The resulting binary trials $Z_1, Z_2, \ldots$ are no longer exchangeable. Distribution of runs under this scheme can be obtained computing the probabilities of the forms (1)–(4). Under this scheme we have

$$P\left(A_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) = \binom{m_1}{n_1} \frac{n_1!}{(m_1+m_2)^n} \prod_{s=1}^{r_2} (m_2+i_0+\cdots+i_{s-1})^{j_s}, \quad i_0 \equiv 0, r_1 = r_2,$$

$$P\left(B_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) = \binom{m_1}{n_1} \frac{n_1!}{(m_1+m_2)^n} \prod_{s=1}^{r_2} (m_2+i_0+\cdots+i_{s-1})^{j_s}, \quad i_0 \equiv 0, r_2 = r_1 + 1,$$

$$P\left(C_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) = \binom{m_1}{n_1} \frac{n_1!}{(m_1+m_2)^n} \prod_{s=1}^{r_2} (m_2+i_1+\cdots+i_s)^{j_s}, \quad r_1 = r_2 + 1,$$

$$P\left(D_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) = \binom{m_1}{n_1} \frac{n_1!}{(m_1+m_2)^n} \prod_{s=1}^{r_2} (m_2+i_1+\cdots+i_s)^{j_s}, \quad r_1 = r_2.$$

3.3 Markov-dependent binary trials

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of homogeneous Markov dependent binary trials with transition probabilities $p_{00}, p_{01}, p_{10}, p_{11}$ and initial probabilities $P\{Z_1 = 1\} = p_1$, $P\{Z_1 = 0\} = p_0 = 1 - p_1$. Then we have

$$P\left(A_n^{n_1}(r_1, r_2)\right) = p_{11}^{n_1 - r_1} p_{01}^{r_1} p_{10}^{r_1 - 1} p_{00}^{n - n_1 - r_2} p_0,$$

$$P\left(B_n^{n_1}(r_1, r_2)\right) = p_{11}^{n_1 - r_1} p_{01}^{r_1} p_{10}^{r_1} p_{00}^{n - n_1 - r_2} p_0,$$

$$P\left(C_n^{n_1}(r_1, r_2)\right) = p_{11}^{n_1 - r_1} p_{01}^{r_2} p_{10}^{r_1 - 1} p_{00}^{n - n_1 - r_2} p_1,$$

$$P\left(D_n^{n_1}(r_1, r_2)\right) = p_{11}^{n_1 - r_1} p_{01}^{r_2 - 1} p_{10}^{r_1} p_{00}^{n - n_1 - r_2} p_1$$

As it seen the probabilities of the forms (1)–(4) depend on $\sum_{j=1}^{r_1} i_j = n_1$ and $\sum_{i=1}^{r_2} j_i = n - n_1$. Therefore Eq. (7) also holds for a sequence of homogeneous Markov-dependent binary trials with

$$P\left\{E_n^{n_1}(r_1, r_2)\right\} = \begin{cases} p_{11}^{n_1 - r_1} p_{01}^{r_1} p_{10}^{r_1 - 1} p_{00}^{n - n_1 - r_2} p_0 + p_{11}^{n_1 - r_1} p_{01}^{r_2 - 1} p_{10}^{r_1} p_{00}^{n - n_1 - r_2} p_1 & \text{if } r_1 = r_2 \\ p_{11}^{n_1 - r_1} p_{01}^{r_1} p_{10}^{r_1} p_{00}^{n - n_1 - r_2} p_0 & \text{if } r_2 = r_1 + 1 \\ p_{11}^{n_1 - r_1} p_{01}^{r_2} p_{10}^{r_1 - 1} p_{00}^{n - n_1 - r_2} p_1 & \text{if } r_1 = r_2 + 1. \end{cases}$$

3.4 Consecutive records (independent nonidentical binary trials)

Let $\{Y_i\}_{i \geq 1}$ be a sequence of random variables. Define

$$Z_i = \begin{cases} 1, & \text{if } Y_i \text{ is record} \\ 0, & \text{otherwise} \end{cases} \quad i = 1, 2, \ldots \tag{8}$$

The random variables defined by (8) are known as record indicators and if $\{Y_i\}_{i \geq 1}$ is a sequence of i.i.d. random variables with a common absolutely continuous distribution function then the record indicators are independent and $P\{Z_i = 1\} = 1 - P\{Z_i = 0\} = \frac{1}{i}$, $i \geq 1$ (Nevzorov 2001, pp. 57–58). Chern et al. (2000) and Chern and Hwang (2005) studied the distribution of the number of consecutive records, i.e. the runs in $\{Z_i\}_{i \geq 1}$. Runs associated with $\{Z_i\}_{i \geq 1}$ can be studied using the formulas presented in this paper. We only need to consider the probabilities of the forms (3) and (4) since $P\{Z_1 = 1\} = 1$. Using the independence of record indicators we have

$$P\left(C_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right) = \frac{1}{i_1!} \prod_{s=1}^{r_2 = r_1 - 1} \frac{\left(n_1 - \sum_{m=s+1}^{r_1} i_m + \sum_{m=1}^{s} j_m\right)!}{\left(n_1 - \sum_{m=s+2}^{r_1} i_m + \sum_{m=1}^{s} j_m\right)!} \frac{\left(n_1 - \sum_{m=s+1}^{r_1} i_m + \sum_{m=0}^{s-1} j_m\right)}{\left(n_1 - \sum_{m=s+1}^{r_1} i_m + \sum_{m=0}^{s} j_m\right)},$$

Table 2 Waiting time probabilities for the first k consecutive records

x	$P\{W_2 = x\}$	$P\{W_3 = x\}$
2	0.5000	
3	0.0000	0.1667
4	0.0417	0.0000
5	0.0167	0.0083
6	0.0125	0.0056
7	0.0087	0.0030
8	0.0066	0.0020
9	0.0051	0.0014
10	0.0041	0.0010

$$P\left(D_n^{n_1}(\vec{i}_{r_1}, \vec{j}_{r_2})\right)$$
$$= \frac{1}{i_1!} \prod_{s=1}^{r_1-1=r_2-1} \frac{\left(n_1 - \sum_{m=s+1}^{r_1} i_m + \sum_{m=1}^{s} j_m\right)!}{\left(n_1 - \sum_{m=s+2}^{r_1} i_m + \sum_{m=1}^{s} j_m\right)!} \frac{\left(n_1 - \sum_{m=s+1}^{r_1} i_m + \sum_{m=0}^{s-1} j_m\right)}{\left(n_1 - \sum_{m=s+1}^{r_1} i_m + \sum_{m=0}^{s} j_m\right)}$$
$$\times \frac{n - j_{r_2}}{n},$$

where $j_0 \equiv 0$, and $\sum_{i=a}^{b} = 0$ for $a > b$.

Table 2 gives the waiting-time probabilities for the first k consecutive records (success run of length k). If W_k denotes the waiting time for the first success run length of k then the distribution of W_k can be computed from $P\{W_k = x\} = P\{L_{x-1}^{(1)} < k\} - P\{L_x^{(1)} < k\}$.

Acknowledgments The authors would like to express gratitude to the referee for a thorough review and valuable comments that led to some improvements in this paper.

References

Balakrishnan N, Koutras MV (2002) Runs and scans with applications. Wiley Series in probability and statistics, New York

Blom G, Holst L, Sandell D (1994) Problems and snapshots from the world of probability. Springer, New York

Charalambides CA (2002) Enumerative combinatorics. Chapman and Hall/CRC, London

Chern H-H, Hwang H-K, Yeh Y-N (2000) Distribution of the number of consecutive records. Random Struct Algorithms 17:169–196

Chern H-H, Hwang H-K (2005) Limit distribution of the number of consecutive records. Random Struct Algorithms 26:404–417

Eggenberger F, Pólya G (1923) Über die Statistik verketteter Vorgänge. Z. für Ang. Math. und Mech. 3:279–289

Eryılmaz S (2005) On the distribution and expectation of success runs in nonhomogeneous Markov dependent trials. Stat Papers 46:117–128

Eryılmaz S, Demir S (2007) Success runs in a sequence of exchangeable binary trials. J Stat Planning Inference 137:2954–2963

Eryılmaz S (2008a) Distribution of runs in a sequence of exchangeable multistate trials. Stat Probab Lett 78:1505–1513

Eryılmaz S (2008b) Run statistics defined on the multicolor urn model. J Appl Probab (to appear)
Fu JC, Lou WYW (2003) Distribution theory of runs and patterns and its applications. A finite markov chain imbedding technique. World Scientific Pub., USA
Gibbons JD, Chakraborti S (2003) Nonparametric statistical inference, 4th edn. Marcel Dekker, New York
Johnson N, Kotz S (1977) Urn models and their application: an approach to modern discrete probability theory. Wiley, New York
Kong Y (2006) Distribution of runs and longest runs: a new generating function approach. J Am Stat Assoc 101:1253–1263
Makri FS, Philippou AN (2005) On binomial and circular binomial distributions of order k for l-overlapping success runs of length k. Stat Papers 46:411–432
Makri FS, Philippou AN, Psillakis ZM (2007a) Shortest and longest length of success runs in binary sequences. J Stat Planning Inference 137:2226–2239
Makri FS, Philippou AN, Psillakis ZM (2007b) Success run statistics defined on an urn model. Adv Appl Probab 39:991–1019
Makri FS, Philippou AN, Psillakis ZM (2007c) Pólya, inverse Pólya, and circular Pólya distributions of order k for l -overlapping success runs. Commun Stat Theory Methods 36:657–668
Nevzorov VB (2001) Records: mathematical theory. American Mathematical Society, Rhode Island
Sen K, Agarwal ML, Chakraborty S (2002) Lengths of runs and waiting time distributions by using Pólya–Eggenberger sampling scheme. Studia Sci Math Hungar 39:309–332
Styve B (1965) A variant of Pólya urn model. Nordisk Mat Tidsk 13:147–152