Computers and Mathematics with Applications 61 (2011) 761–772

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

# On success runs of a fixed length in Bernoulli sequences: Exact and asymptotic results

Frosso S. Makri [a,*], Zaharias M. Psillakis [b]

[a] Department of Mathematics, University of Patras, 26500 Patras, Greece

[b] Department of Physics, University of Patras, 26500 Patras, Greece

ARTICLE INFO

*Article history:*
Received 26 March 2010
Received in revised form 13 December 2010
Accepted 13 December 2010

*Keywords:*
Runs
Waiting time
Bernoulli trials
Exact distribution
Limiting distribution
Exchangeable trials

ABSTRACT

Consider a sequence of $n$ Bernoulli (Success–Failure or 1–0) trials. The exact and limiting distribution of the random variable $E_{n,k}$ denoting the number of success runs of a fixed length $k$, $1 \leq k \leq n$, is derived along with its mean and variance. An associated waiting time is examined as well. The exact distribution is given in terms of binomial coefficients and an extension of it covering exchangeable sequences is also discussed. Limiting distributions of $E_{n,k}$ are obtained using Poisson and normal approximations. The exact mean and variance of $E_{n,k}$ which are given in explicit forms are also used to derive bounds and an additional approximation of the distribution of $E_{n,k}$. Numbers, associated with $E_{n,k}$ and related random variables, counting binary strings and runs of 1's useful in applications of computer science are provided. The overall study is illustrated by an extensive numerical experimentation.



## 1. Introduction and preliminaries

Runs and related statistics, counting according to several enumerating schemes, defined on binary sequences of several internal structures are used in many diverse areas of applied research. Such areas include statistical hypothesis testing, reliability and quality control, molecular biology, financial engineering and computer science. Past and current works on runs/patterns literature are well documented in [1,2]. Recent studies on the topic are included among others in works [3–13]. In counting runs and patterns we often are faced with large sequences of trials. In such cases the exact distribution of the studied random variable is approximated by another, simpler in a computational sense, distribution; see, e.g. [14] for a comprehensive review.

Let $\{X_i\}_{i\geq 1}$ be an ordered sequence of binary trials resulting in either a success (denoted by $S$ or 1) or a failure (denoted by $F$ or 0). According to Mood's [15] enumeration scheme a success run is defined to be a sequence of consecutive successes preceded and succeeded by failures or by nothing. The number of successes in a success run is referred to as its length (or its size). Given a sequence of length $n$ and a run length $k$, $1 \leq k \leq n$, the random variable (RV) $E_{n,k}$, denoting the number of success runs of length exactly $k$ may be defined as

$$E_{n,k} = \sum_{j=k}^{n} U_j, \quad U_j = (1 - X_{j-k})(1 - X_{j+1}) \prod_{i=j-k+1}^{j} X_i, \ k \leq j \leq n \tag{1}$$

(using the convention that $X_0 = X_{n+1} = 0$). The support of the RV $E_{n,k}$ is the set $\mathcal{R}(E_{n,k}) = \left\{0, 1, \ldots, \left\lfloor \frac{n+1}{k+1} \right\rfloor\right\}$ where by $\lfloor x \rfloor$ we denote the greatest integer less than or equal to $x$. The setup (1) holds for any binary sequence and it is a useful

* Corresponding author.
*E-mail addresses:* makri@math.upatras.gr (F.S. Makri), psillaki@physics.upatras.gr (Z.M. Psillakis).

0898-1221/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2010.12.023

apparatus to derive expected value, variance and limiting distribution of $E_{n,k}$. Furthermore, it is helpful to determine numeric values of $E_{n,k}$ when it is used as a descriptive statistic in various applications that require arithmetic values (e.g. to compute experimental relative frequencies).

A RV related to $E_{n,k}$ is the waiting time $W_{r,k}$ until the $r$th, $r \geq 1$ occurrence of a success run of length exactly $k$. It is defined and related to $E_{n,k}$ as follows

$$W_{r,k} = \min\{n \geq r(k+1) - 1 : E_{n,k} = r\}; \quad W_{r,k} > n \text{ iff } E_{n,k} < r,\ r \in \mathcal{R}(E_{n,k}) - \{0\}. \tag{2}$$

Hence, via Eq. (2) it is offered an alternative way of obtaining results for the waiting time RV $W_{r,k}$ through formulae established for the run enumerative RV $E_{n,k}$ and vice versa. The RV $W_{1,k}$, the minimum number of binary trials needed to observe a sequence of (exactly) $k$ consecutive successes for the first time, has been studied extensively (see, e.g. [1, pp. 9–10]) from Moivre's era (see, e.g. [16, pp. 173–175]). The RVs $E_{n,k}$ and $W_{r,k}$ are fundamental in run literature and have been studied on binary sequences of several internal structures by many researchers who used various approaches. See, e.g. [7–9,15,17–20].

Definitions (1) and (2) are illustrated using the following example. Let the outcomes of the first 13 binary trials be *SFFSFSSSFFSSS*. Then, $W_{1,1} = 1$ since $E_{1,1} = 1$, $W_{1,2} = 7$ since $E_{7,2} = 1$, $W_{2,3} = 13$ since $E_{13,3} = 2$ and $W_{r,k} > 13$ for $k \geq 4$, $r \geq 1$ since $E_{13,k} = 0$ for $k \geq 4$.

In a recent interesting paper, Sinha and Sinha [13] addressed the usefulness of $E_{n,k}$ defined on Bernoulli sequences (that is, sequences of independent and identically distributed binary RVs with a common success probability $p$, $0 < p < 1$) in several areas of computer science including encoding, compression and transmission of digital information. For the particular case of equiprobable binary trials (i.e. Bernoulli trials with $p = 1/2$) they provided the exact distribution of $E_{n,k}$ using generating functions. Also, they presented lucid experimental results (relative frequencies) of $E_{n,k}$ defined on non-necessarily equiprobable Bernoulli sequences of small, moderate and large length.

In the present paper we study the RV $E_{n,k}$ defined on Bernoulli sequences of non-necessarily equiprobable binary trials. Specifically, our paper is organized as follows. In Theorems 2.1 and 2.2, Corollary 2.3 and Theorem 2.3 we establish the exact probability mass function (PMF) and the limiting distribution of $E_{n,k}$. Practical implementation of Corollary 2.3 and Theorem 2.3 are presented in Remark 2.3. In Proposition 2.1 we obtain the exact mean value and the variance of $E_{n,k}$. A simple combinatorial approach is used in Theorem 2.1 whereas the setup (1) is the main tool to derive Theorem 2.3 and Proposition 2.1. Furthermore, a number counting binary strings which is useful in applications is derived in Corollary 2.1. In Proposition 2.2 we present lower/upper bounds and an additional approximation of the probability distribution of $E_{n,k}$. In Section 3, first we offer an extension of Theorem 2.1 to exchangeable binary sequences and second we clarify the vast majority of the formulae given in the paper, by providing applications and numerical results helpful to a practical minded reader. Finally, in Section 4, further results concerning other important RVs which are related to $E_{n,k}$ and can be useful in engineering applications like the ones discussed in [13], are derived.

Next, for completeness and reader's convenience we restate some results useful in our study. For further details see e.g. [1, pp. 167–168], [16, pp. 94–96] and [21, pp. 508–509].

Let us consider two non-negative integer valued RVs $X$ and $Y$ with distributions $\mathcal{L}(X)$ and $\mathcal{L}(Y)$, respectively. If we want to approximate the distribution of one of the RVs with that of the other, a common measure of the accuracy of the approximation is the total variation distance $d(\mathcal{L}(X), \mathcal{L}(Y))$ or simply $d(X, Y) = \frac{1}{2}\sum_{x=0}^{\infty} |P(X = x) - P(Y = x)|$. Readily, $0 \leq d(X, Y) \leq 1$ and $d(X, Y) = 0$ iff $X$ and $Y$ have exactly the same distribution. Furthermore, let $0 \leq x_n = \max\{x : P(X = x) > 0\} < \infty$, i.e. $0 < |\mathcal{R}(X)| < \infty$, $0 \leq x_n = \max\{x : x \in \mathcal{R}(X)\}$, where $|\mathcal{R}(X)|$ denotes the cardinality of the range set $\mathcal{R}(X)$ of the non-negative integer valued RV $X$. If $Y$ is a Poisson RV with parameter (mean) $\lambda > 0$, i.e. $Y \sim \text{Po}(\lambda)$, $\mathcal{R}(Y) = \{0, 1, 2, \ldots\}$, then

$$d(X, Y) = \frac{1}{2}\sum_{x \in \mathcal{R}(X)} |P(X = x) - \text{po}(x; \lambda)| + \frac{1}{2}[1 - \text{Po}(x_n; \lambda)] \tag{3}$$

where $\text{po}(y; \lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^y}{y!}$ and $\text{Po}(y; \lambda) = \sum_{t=0}^{y} \text{po}(t; \lambda)$ are the PMF and the CDF (cumulative distribution function) of $Y$, respectively. By (3) the actual distance $d(X, Y)$ consists of two terms: the residual or tail term $d_{\text{tail}} = \frac{1}{2}[1 - \text{Po}(x_n; \lambda)]$ and the truncated (since it is computed in $\mathcal{R}(X)$) term $d_{\text{trc}} = \frac{1}{2}\sum_{x \in \mathcal{R}(X)} |P(X = x) - \text{po}(x; \lambda)|$. When the distribution of $X$ is well approximated by that of $Y$ $d_{\text{tail}}$ is negligible with respect to $d_{\text{trc}}$.

Another practical way of measuring the error between the theoretical distribution $f(x) = P(X = x)$ of a RV $X$ with $0 < |\mathcal{R}(X)| < \infty$ and an approximating (experimental or predicting) distribution $f^*(x)$ of a RV $Y$ with $\mathcal{R}(X) \subseteq \mathcal{R}(Y)$ is the root mean square error

$$\text{rmse}(f, f^*) = \left\{\frac{1}{|\mathcal{R}(X)|}\sum_{x \in \mathcal{R}(X)} [f(x) - f^*(x)]^2\right\}^{1/2}. \tag{4}$$

If, in addition, $X$ and $Y$ are non-negative integer valued RVs then $\text{rmse} \leq d$ since $|P(X = x) - P(Y = x)| \leq d(X, Y)$. It is clear that $d(\mathcal{L}(X), \mathcal{L}(Y))$ and $\text{rmse}(f, f^*)$ decrease as the approximate distribution of $Y$ to that of $X$ performs better.

**Lemma 1.1** ([22]). *For a probability function $H(x)$ of a RV $X$, it holds*

$$0 < L(x) \le H(x) \le U(x), \quad x \in \mathcal{R}(X). \tag{5}$$

*Then, an approximation $\hat{H}(x)$ of $H(x)$ and an upper bound $\hat{B}(x)$ of the relative error between $H(x)$ and $\hat{H}(x)$, $B(x) = |H(x) - \hat{H}(x)|/H(x)$, are given by*

$$\hat{H}(x) = [L(x) + U(x)]/2, \qquad \hat{B}(x) = [U(x) - L(x)]/[2L(x)]. \tag{6}$$

**Remark 1.1.** As $\hat{B}(x)$ does not assume any knowledge of the exact value of $H(x)$, the pair $(\hat{H}, \hat{B})$ gives an advantage in cases for which a formula for $H(x)$ does not exist or it exists but it is difficult to be implemented via the available computers.

Throughout the article, for integers $n$, $m$, $\binom{n}{m}$ denotes the extended binomial coefficient (see, e.g. [23, pp. 50,63]); $\lceil x \rceil$ denotes the least integer greater than or equal to $x$; $\Phi(x)$ stands for the standard normal CDF $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2/2} \mathrm{d}t$, $x \in R$; $F(x; n, k, p) = P(E_{n,k} \le x)$, $x \in \mathcal{R}(E_{n,k})$; $G(x) = F(x-1; n, k, p) = P(W_{x,k} > n)$, $x \in \mathcal{R}(E_{n,k}) - \{0\}$; and $\xrightarrow{d}$ denotes convergence in distribution.

## 2. Main results

In this section we consider a Bernoulli sequence $\{X_i\}_{i \ge 1}$ with success probability $p$, $p = P(X_i = 1) = 1 - P(X_i = 0) = 1 - q$, $0 < p < 1$, $i = 1, 2, \ldots$. First we give the exact PMF of $E_{n,k}$ using a result of [8] for the derivation of which the combinatorial method of distributing balls in cells (see, e.g. [24]) was employed.

**Theorem 2.1.** *For $x \in \mathcal{R}(E_{n,k})$, $1 \le k \le n$ it holds*

$$P(E_{n,k} = x) = \sum_{y=0}^{n-kx} p^{n-y} q^y \binom{y+1}{x} \sum_{j=0}^{\left\lfloor \frac{n-y-kx}{k} \right\rfloor} (-1)^j \binom{y+1-x}{j} \binom{n-(k+1)(x+j)}{n-y-k(x+j)}. \tag{7}$$

**Proof.** Consider a Polya–Eggenberger urn model, $PE(w, b, s)$, with initial urn composition of $w$ white and $b$ black balls (see, e.g. [25, pp. 176–178]) where the drawings of the balls are done with replacements, i.e. $s = 0$. Then, if we realize a drawing of a white ball as a success (1) and a drawing of a black ball as a failure (0), a sequence of $n$ drawings is a Bernoulli one with $p = w/(w + b)$. Then, setting $p_n(y) = p^{n-y} q^y$ in Theorem 3.1 of Makri et al. [8] the result is concluded. □

**Remark 2.1.** Note that, since $\mathcal{R}(E_{n,k}) = \{0, 1\}$ if $n \le 2k$, it holds $P(E_{n,k} = 0) = P(W_{1,k} > n) = 1 - E(E_{n,k})$ where the mean value $E(E_{n,k})$ is given by the forthcoming Proposition 2.1.

Using the PMF of $E_{n,k}$ we get the following two useful numbers in engineering applications.

**Corollary 2.1.** (a) *The number $N_{n;e_k}$ of binary strings which contain exactly $e_k$, $e_k = 0, 1, \ldots, \left\lfloor \frac{n+1}{k+1} \right\rfloor$, runs of 1's of length (exactly) equal to $k$ in all possible binary strings of length $n$, $1 \le k \le n$, is*

$$N_{n;e_k} = \sum_{y=0}^{n-ke_k} \binom{y+1}{e_k} \sum_{j=0}^{\left\lfloor \frac{n-y-ke_k}{k} \right\rfloor} (-1)^j \binom{y+1-e_k}{j} \binom{n-(k+1)(e_k+j)}{n-y-k(e_k+j)}. \tag{8}$$

(b) *The total number $R_{n,k}^{(e)}$ of occurrences of all runs of 1's of length (exactly) equal to $k$, in all possible binary strings of length $n$, $1 \le k \le n$, is*

$$R_{n,k}^{(e)} = \sum_{e_k=1}^{\left\lfloor \frac{n+1}{k+1} \right\rfloor} e_k N_{n;e_k}. \tag{9}$$

**Proof.** Setting $x = e_k$ and $p = 1/2$ in (7) we get that $P(E_{n,k} = e_k) = N_{n;e_k}/2^n$. Since the cardinality of the proper sample space is equal to $2^n$ and all sequences are equally likely to occur part (a) of the corollary is derived. Part (b) follows directly from the definitions of $R_{n,k}^{(e)}$ and $N_{n;e_k}$. □

**Remark 2.2.** The counting problem of determination of $N_{n;e_k}$ has been addressed by Sinha and Sinha [13] who provided an alternative formula using generating functions. Their expression (given by Eq. (3)) is more complicated than (8) since it contains one additional summation of binomial coefficients. Therefore, $N_{n;e_k}$ may be evaluated faster computationally via (8).

Using Theorem 2.1 and relation (2) we get the PMF and the mean value of $W_{r,k}$.

**Corollary 2.2.** (a) *The exact PMF of $W_{r,k}$, $r \geq 1$, is given by*

$$P(W_{r,k}=t)=\begin{cases} q^{r-1}p^{t-r+1}, & \text{if } t=r(k+1)-1 \\ \sum_{x=0}^{r-1}\left[P(E_{t-1,k}=x)-P(E_{t,k}=x)\right], & \text{if } t\geq r(k+1). \end{cases} \tag{10}$$

(b) *The mean value of the RV $W_{r,k}$ can be computed by*

$$E(W_{r,k})=r(k+1)-1+\sum_{t=r(k+1)-1}^{\infty}P(E_{t,k}<r)\simeq r(k+1)-1+\sum_{t=r(k+1)-1}^{t_\infty}\sum_{x=0}^{r-1}P(E_{t,k}=x) \tag{11}$$

*$t_\infty = t_\infty(\varepsilon; r, k)$ is a stopping time such that $\sum_{x=0}^{r-1}P(E_{t_\infty,k}=x)\leq\varepsilon\sum_{t=r(k+1)-1}^{t_\infty}\sum_{x=0}^{r-1}P(E_{t,k}=x)$ where $\varepsilon$ is a prespecified small positive number, defined by the demanded accuracy of the results.*

Next, two asymptotic results for the RV $E_{n,k}$ are established. The first one is a Poisson Limit Law derived via the Chen–Stein method (see, e.g. [26]) and the second one is a Central Limit Theorem obtained using a Theorem of [27]. To derive the results we employ two different setups for the parameters $n$, $k$ and $p$. The first setup, which implies the Poisson approximation, assumes that the success run length $k$ is fixed and the success probability $p$ tends to zero as the length of the sequence $n$ tends to infinity. The second setup, that provides a normal approximation, assumes that both $p$ and $k$ are fixed whereas $n$ tends to infinity. Accordingly, first we obtain (in Theorem 2.2) a total variation upper bound for the rate of convergence of $E_{n,k}$ to a suitable Poisson RV and then we state (in Corollary 2.3) a Poisson Limit Law. Second, we establish (in Theorem 2.3) a Central Limit Theorem for $E_{n,k}$.

**Theorem 2.2.** *Let $Y_\lambda$ be a Poisson RV with mean $\lambda>0$. Then, for fixed $k$ it holds*

$$d(E_{n,k},Y_\lambda)\leq\alpha(n,k,p)+\beta(n,k,p,\lambda) \tag{12}$$

*where $\alpha(n,k,p)=\{[(n-k-1)q+1]p+(2k+1)q+1\}p^k$, $\beta(n,k,p,\lambda)=\min\{|\lambda_n-\lambda|, |\lambda_n^{1/2}-\lambda^{1/2}|\}$ and $\lambda_n=(n-k)qp^k$.*

**Proof.** Let $Y_{\lambda_n}$ be a Poisson RV with mean $\lambda_n=(n-k)qp^k$ and $G_{n,k}$ a RV denoting the number of success runs of length at least $k$ in $n$ Bernoulli trials. Then, the total variation distance $d(E_{n,k},Y_{\lambda_n})$ between the distributions of $E_{n,k}$ and $Y_{\lambda_n}$ satisfies the condition $d(E_{n,k},Y_{\lambda_n})\leq\alpha(n,k,p)$ since $d(E_{n,k},Y_{\lambda_n})\leq d(E_{n,k},G_{n,k})+d(G_{n,k},Y_{\lambda_n})$, by the triangle inequality, $d(E_{n,k},G_{n,k})\leq P(E_{n,k}\neq G_{n,k})=P(E_{n,k}<G_{n,k})=P(G_{n,k+1}>0)\leq p^{k+1}+(n-k-1)qp^{k+1}$ (see, [11, Eq. (34)]) and $d(G_{n,k},Y_{\lambda_n})\leq[(2k+1)q+1]p^k$ (see, [1, Eq. (5.20)]). Also, from Theorem 2.1 of Yannaros [28] we obtain $d(Y_{\lambda_n},Y_\lambda)\leq\beta(n,k,p,\lambda)$. Using again the triangular inequality $d(E_{n,k},Y_\lambda)\leq d(E_{n,k},Y_{\lambda_n})+d(Y_{\lambda_n},Y_\lambda)$ the theorem follows. □

**Corollary 2.3.** *For fixed $k$ if $np^k\to\lambda>0$ and $p\to0$ as $n\to\infty$, then*

$$E_{n,k}\xrightarrow{d}Y_\lambda\sim\mathrm{Po}(\lambda),\quad \text{as } n\to\infty. \tag{13}$$

**Proof.** Since $k$ is fixed and $np^k\to\lambda$, $p\to0$, $q\to1$, as $n\to\infty$ it holds $\lambda_n=(n-k)qp^k\to\lambda$, i.e. $\beta(n,k,p,\lambda)\to0$ and $\alpha(n,k,p)\to0$. Therefore, $d(E_{n,k},Y_\lambda)\to0$, i.e. $E_{n,k}\xrightarrow{d}Y_\lambda$ as $n\to\infty$. □

**Theorem 2.3.** *For fixed $k$ and $p$ it holds*

$$\frac{E_{n,k}-n\mu}{\sigma\sqrt{n}}\xrightarrow{d}Z\sim N(0,1),\quad \text{as } n\to\infty \tag{14}$$

*where $\mu=q^2p^k$ and $\sigma^2=\mu\{1+\mu[2(p/q-k)-1]\}$.*

**Proof.** In the proof all limits are taken as $n\to\infty$. Let $U_j$, $j=k,k+1,\ldots,n$ be as in (1). For $j=1,2,\ldots,n-k-1$ set $I_j=U_{k+j}$. The RVs $I_j$, by their definition, are $k+1$-dependent. Also, since $X_i$'s are i.i.d. it is implied that the sequence $I_1,I_2,\ldots,$ is stationary. Hence, noting that $E(I_1^3)<\infty$ it follows that Theorem 2 of Hoeffding and Robbins [27] holds for the RV $V_n=E_{n,k}-U_k-U_n=\sum_{j=1}^{n-k-1}I_j$ with $E(V_n)=(n-k-1)q^2p^k$ and $\sigma^2=V(I_1)+2\sum_{j=2}^{k+2}\mathrm{Cov}(I_1,I_j)=q^2p^k(1-q^2p^k)-2kq^4p^{2k}+2q^3p^{2k}-2q^4p^{2k}=q^2p^k(1-q^2p^k)+2q^3p^{2k}(p-kq)$. That is, $Z_n=\frac{V_n-(n-k-1)q^2p^k}{\sigma\sqrt{n-k-1}}\xrightarrow{d}Z\sim N(0,1)$ or equivalently, $Z_n=\sqrt{\frac{n}{n-k-1}}\left(\frac{E_{n,k}-nq^2p^k}{\sigma\sqrt{n}}\right)-\frac{U_k+U_n}{\sigma\sqrt{n-k-1}}+\frac{(k+1)q^2p^k}{\sigma\sqrt{n-k-1}}\xrightarrow{d}Z\sim N(0,1)$. Then, since $c_n=\sqrt{\frac{n-k-1}{n}}\longrightarrow1$, $d_n=-\frac{(k+1)q^2p^k}{\sigma\sqrt{n}}\longrightarrow0$ and $\frac{U_k+U_n}{\sigma\sqrt{n}}\xrightarrow{d}0$, we first obtain that $c_nZ_n+d_n\longrightarrow Z\sim N(0,1)$ and finally, by Slutsky's Theorem $\frac{E_{n,k}-nq^2p^k}{\sigma\sqrt{n}}\xrightarrow{d}Z\sim N(0,1)$. □

**Remark 2.3.** A practical interpretation of (14) and (13) is that, for sufficient large $n$, i.e. $n \gg 1$ we have for the CDF $F(x; n, k, p)$, $x \in \mathcal{R}(E_{n,k})$, the approximations

$$F(x; n, k, p) \simeq \Phi\left(\frac{x + 0.5 - n\mu}{\sigma\sqrt{n}}\right) \tag{15}$$

with $\mu = q^2p^k$, $\sigma^2 = \mu\{1 + \mu[2(p/q - k) - 1]\}$ and

$$|F(x; n, k, p) - \mathrm{Po}(x; \lambda)| \le d_{\mathrm{TV}}(n, k, p), \tag{16}$$

where $\lambda = np^k$ and $d_{\mathrm{TV}}(n, k, p) = \alpha(n, k, p) + \beta(n, k, p, \lambda)$.

We note that (16) provides via $d_{\mathrm{TV}}(n, k, p)$ lower/upper bounds of the CDF of $E_{n,k}$, along with an upper bound (via Lemma 1.1) of the error committed by the approximation $\mathrm{Po}(x; \lambda)$ of $F(x; n, k, p)$. Moreover, it is a common practice to consider that the success probability $p$ is continuously varying with $n$ following the law $p = p_n = 1 - \mathrm{e}^{-(\lambda/n)^{1/k}} = (\lambda/n)^{1/k} + o(n^{-1/k})$. In this case, the condition $np_n^k \to \lambda$ as $n \to \infty$, implies that $p_n \to 0$ (and thus $q_n \to 1$) and secures the convergence (see e.g. [1, pp. 178–180]).

On the other hand, in order to use (15) we have to check first if the condition (see, e.g. [12])

$$\mu - \frac{3\sigma}{\sqrt{n}} > 0 \quad \text{and} \quad \mu + \frac{3\sigma}{\sqrt{n}} < \frac{\left\lfloor \frac{n+1}{k+1} \right\rfloor}{n} \tag{17}$$

is satisfied for the parameter vector $(n, k, p)$. Readily, $P(n\mu - 3\sigma\sqrt{n} \le E_{n,k} \le n\mu + 3\sigma\sqrt{n}) \simeq 0.9973$.

Next, we obtain the exact mean value and variance of $E_{n,k}$ using the setup (1).

**Proposition 2.1.** *Let $E(E_{n,k})$ and $V(E_{n,k})$ be the mean value and the variance of the RV $E_{n,k}$ for $0 < p < 1$. Then, for $n = k$, $E(E_{n,k}) = p^k$, $V(E_{n,k}) = p^k(1 - p^k)$ and for $n \ge k + 1$,*

$$\begin{aligned} &E(E_{n,k}) = qp^k[2 + (n - k - 1)q], \\ &V(E_{n,k}) = v_1, \text{ for } n \le 2k; \quad v_2, \text{ for } n = 2k + 1; \quad v_3, \text{ for } n \ge 2k + 2 \end{aligned} \tag{18}$$

*where $v_1 = 2qp^k - 4q^2p^{2k} + (n - k - 1)q^2p^k - (n - k - 1)^2q^4p^{2k} - 4(n - k - 1)q^3p^{2k}$, $v_2 = v_1 + 2qp^{2k}$ and $v_3 = 2qp^k + 2q^2p^{2k} + (n - k - 1)q^2p^k + 2(n - 4k - 4)q^3p^{2k} - [(n - k - 1)^2 - (n - 2k - 2)(n - 2k - 3)]q^4p^{2k}$.*

**Proof.** By means of Eq. (1) we have that $E(E_{n,k}) = \sum_{j=k}^{n} E(U_j)$ and $V(E_{n,k}) = \sum_{j=k}^{n} E(U_j)\big(1 - E(U_j)\big) + 2\sum_{k \le i < j \le n}\big(E(U_iU_j) - E(U_i)E(U_j)\big)$. It is clear that $E(U_k) = E(U_n) = qp^k$, $E(U_j) = q^2p^k$ for $j = k + 1, \ldots, n - 1$, $E(U_iU_j) = 0$ for $j - i \le k$, $E(U_kU_{2k+1}) = E(U_{n-k-1}U_n) = q^2p^{2k}$, $E(U_iU_{i+k+1}) = q^3p^{2k}$ for $i = k + 1, k + 2, \ldots, n - k - 2$ and $E(U_iU_j) = q^4p^{2k}$, for $j - i \ge k + 2$. The results then follow after some algebraic manipulations. □

For an alternative derivation of (18) see [15]. The expressions (18), in addition to their independent merit, can be used to derive additional bounds and approximations for the probability $G(x)$. Specifically, for large $n$ calculating the exact $G(x)$, $x \in \left\{1, 2, \ldots, \left\lfloor \frac{n+1}{k+1} \right\rfloor\right\}$ is often a hard task, because of the computation effort needed to calculate the sums of the binomial coefficients involved. A first solution to this problem was addressed in Remark 2.3. Another approach is presented next and it is of particular importance when condition (17) does not hold or the error bound of (16), $d_{\mathrm{TV}}(n, k, p)$, is not acceptable.

**Proposition 2.2.** *Let $m = E(E_{n,k})$ and $v^2 = V(E_{n,k})$ be as in Proposition 2.1. Then it holds*

$$G(x) \ge L_{MC}(x), \quad \text{for } x \ge m; \qquad G(x) \le U_C(x), \quad \text{for } x < m + 1 \tag{19}$$

*where $L_{MC}(x) = 0$, if $x = m$; $1 - m/x$, if $m < x \le m + v^2/m$; $1 - v^2/[v^2 + (x - m)^2]$, if $x > m + v^2/m$ and $U_C(x) = v^2/[v^2 + (1 + m - x)^2]$.*

**Proof.** Employing Markov's and one-sided Chebyshev's inequalities as well as similar arguments that have been used to derive Eqs. (34)–(36) of Makri and Psillakis [11] the bounds (19) are derived. □

**Remark 2.4.** For $x$, $n$, $k$ and $p$ such that both bounds (19) can be used, by setting $L(x) = L_{MC}(x)$ and $U(x) = U_C(x)$ in Lemma 1.1, we obtain the respective approximation $\hat{G}_{MC}(x)$ and the error estimate $\hat{B}_{MC}(x)$. These numbers have to be compared with the respective ones $\hat{G}_{\mathrm{Po}}(x)$, $\hat{B}_{\mathrm{Po}}(x)$, which are derived by Eq. (16) and Lemma 1.1, for $L(x) = L_{\mathrm{Po}}(x) = \mathrm{Po}(x - 1; \lambda) - d_{\mathrm{TV}}(n, k, p)$ and $U(x) = U_{\mathrm{Po}}(x) = \mathrm{Po}(x - 1; \lambda) + d_{\mathrm{TV}}(n, k, p)$.

## 3. Extensions, applications and numerics

In Section 3.1, we extend Theorem 2.1 for exchangeable binary sequences and then we consider two indicative examples of such sequences. In Section 3.2, we provide reasonable fits to the experimental data frequencies, presented in [13], using Theorem 2.1 for Bernoulli sequences of non-necessarily equiprobable binary trials. Finally, in Section 3.3 we present some possible scenarios referring to approximations and bounds which clarify further Remarks 2.3 and 2.4 and Propositions 2.1 and 2.2. Extensive numerical experimentation supports and illustrates the results.

**Table 1**
PMFs, means and variances of $E_{n,2}$ and $E_{n,4}$ for an RTM with $j = 2, 3, 5$.

| $n$ | 10 | | | 20 | | |
|---|---|---|---|---|---|---|
| $j$ | 2 | 3 | 5 | 2 | 3 | 5 |
| $x$ | $P(E_{n,2} = x)$ | | | | | |
| 0 | 0.765092 | 0.878494 | 0.977979 | 0.629561 | 0.796540 | 0.959267 |
| 1 | 0.193577 | 0.104871 | 0.020456 | 0.239498 | 0.145030 | 0.034158 |
| 2 | 0.037500 | 0.015441 | 0.001498 | 0.097566 | 0.045340 | 0.005541 |
| 3 | 0.003761 | 0.001194 | 0.000066 | 0.027659 | 0.011070 | 0.000912 |
| 4 | | | | 0.005107 | 0.001826 | 0.000113 |
| 5 | | | | 0.000574 | 0.000184 | 0.000009 |
| 6 | | | | 0.000033 | 0.000009 | $0.3\times10^{-6}$ |
| 7 | | | | $0.5\times10^{-6}$ | $0.1\times10^{-6}$ | $0.3\times10^{-8}$ |
| $E(E_{n,2})$ | 0.280000 | 0.139333 | 0.023625 | 0.541111 | 0.277204 | 0.048473 |
| $V(E_{n,2})$ | 0.299306 | 0.157962 | 0.026488 | 0.683188 | 0.383349 | 0.064217 |
| $x$ | $P(E_{n,4} = x)$ | | | | | |
| 0 | 0.953123 | 0.985347 | 0.999015 | 0.913049 | 0.970968 | 0.997872 |
| 1 | 0.045047 | 0.014283 | 0.000975 | 0.077274 | 0.026603 | 0.002041 |
| 2 | 0.001830 | 0.000369 | 0.000009 | 0.009084 | 0.002307 | 0.000084 |
| 3 | | | | 0.000580 | 0.000120 | 0.000003 |
| 4 | | | | 0.000013 | 0.000002 | $0.3\times10^{-7}$ |
| $E(E_{n,4})$ | 0.048707 | 0.015022 | 0.000994 | 0.097234 | 0.031584 | 0.002217 |
| $V(E_{n,4})$ | 0.049996 | 0.015534 | 0.001012 | 0.109583 | 0.035942 | 0.002397 |

### 3.1. Exchangeable binary sequences

Let $X_1, X_2, \ldots, X_n, \ldots$ be an exchangeable binary sequence. Exchangeability implies that all finite sequences with the same length $n$ and the same number of failures $y$ ($y = 0, 1, 2, \ldots, n$) are equally likely. Replacing $p^{n-y}q^y$ in (7) by $p_n(y) = P(X_1 = X_2 = \cdots = X_{n-y} = 1, X_{n-y+1} = X_{n-y+2} = \cdots = X_n = 0)$ we obtain the exact PMF of $E_{n,k}$ defined on an exchangeable binary sequence, i.e.

$$P(E_{n,k} = x) = \sum_{y=0}^{n-kx} p_n(y) \binom{y+1}{x} \sum_{j=0}^{\left\lfloor \frac{n-y-kx}{k} \right\rfloor} (-1)^j \binom{y+1-x}{j} \binom{n-(k+1)(x+j)}{n-y-k(x+j)}. \tag{20}$$

Using Theorem 2.1 of George and Bowman [29] the probability $p_n(y)$ may be expressed as $p_n(y) = \sum_{i=0}^{y}(-1)^i \binom{y}{i} \lambda_{n-y+i}$, $y = 0, 1, \ldots, n$ with $\lambda_i = P(X_1 = X_2 = \cdots = X_i = 1)$, $i = 1, 2, \ldots, n$ and $\lambda_0 = 1$. We mention that $p_n(y)$ (or $\lambda_i$) is explicitly determined by the considered exchangeable sequence. See, for instance [11,30].

A classical example of an exchangeable binary sequence is that derived according to the Polya–Eggenberger urn model. For a detailed study of the RV $E_{n,k}$ on such a model see [7–9,20].

Another interesting example is the exchangeable binary sequence derived according to the record threshold model (see, e.g. [10,30]). Let $\{Y_i\}_{i\geq 1}$ be a binary sequence of independent and identically distributed (i.i.d.) RVs with continuous distribution function $F_Y$. For such sequences we define record times $K_j$ and record values $R_j$ as follows (see, e.g. [31, pp. 56–57])

$$K_1 = 1, \qquad K_j = \min\{i > K_{j-1} : Y_i > \max(Y_1, Y_2, \ldots, Y_{i-1})\}, \quad j = 2, 3, \ldots \quad \text{and} \quad R_j = Y_{K_j}, \quad j = 1, 2, \ldots,$$

that is, $K_j$ is the index (or the position) of the $j$th record the value of which is $R_j$. By convention, $Y_1$ is a record (since $K_1 = 1$).

Let $Y_1', Y_2', \ldots, Y_n'$ be i.i.d. RVs with continuous distribution function $F_{Y'}$ and independent of $\{Y_i\}_{i\geq 1}$. If the $j$th record value $R_j$ is chosen as a random threshold, then the sequence associated with this record threshold model (RTM) defined by

$$X_i = 1, \quad \text{if } Y_i' > R_j; \qquad 0, \quad \text{otherwise}, \ i = 1, 2, \ldots, n, \ j > 1 \tag{21}$$

is exchangeable and under the hypothesis $H_0 : F_Y = F_{Y'}$ it holds (see [10])

$$p_n(y) = \sum_{i=0}^{y} (-1)^i \binom{y}{i} \frac{1}{(n-y+i+1)^j}, \quad j > 1, \ 0 \leq y \leq n. \tag{22}$$

The number of occurrences of success runs of length exactly $k$ defined on sequence (21) can be studied using (20) with $p_n(y)$ given by (22).

As an illustration of the study of $E_{n,k}$ on an RTM we provide Table 1. In this table we selected the indicative values $n = 10, 20$; $k = 2, 4$ and $j = 2, 3, 5$. We observe that an increase in $j$ (the order of the record used as a threshold) leads to an increase of $P(E_{n,k} = 0)$ as well as a decrease of the mean value and the variance of $E_{n,k}$.

**Table 2**
Approximate success probabilities $p$, rmse($p$; 8, $k$) and rmse(0.5; 8, $k$) for $1 \leq k \leq 8$ and for several data file types.

| File type | $k$ | $p$ | rmse($p$; 8, $k$) | rmse(0.5; 8, $k$) | File type | $k$ | $p$ | rmse($p$; 8, $k$) | rmse(0.5; 8, $k$) |
|---|---|---|---|---|---|---|---|---|---|
| binary | 1 | 0.22156 | 0.89200 | 1.90410 | text | 1 | 0.22791 | 2.39734 | 4.85674 |
| | 2 | 0.39835 | 0.15358 | 3.20841 | | 2 | 0.51416 | 2.23350 | 2.26233 |
| | 3 | 0.43066 | 0.17004 | 3.88278 | | 3 | 0.45306 | 0.18972 | 2.56225 |
| | 4 | 0.40905 | 0.00006 | 4.69550 | | 4 | 0.40969 | 0.00006 | 4.66660 |
| | 5 | 0.41466 | 0.00000 | 2.41230 | | 5 | 0.38186 | 0.00000 | 3.06330 |
| | 6 | 0.47486 | 0.00003 | 0.43273 | | 6 | 0.20368 | 0.00000 | 1.93722 |
| | 7 | 0.40290 | 0.00000 | 0.57545 | | 7 | 0.01563 | 0.00000 | 0.78125 |
| | 8 | 0.49654 | 0.00001 | 0.02112 | | 8 | 0.01563 | 0.00000 | 0.39063 |
| jpeg | 1 | 0.51001 | 0.22983 | 0.42301 | mpeg | 1 | 0.48621 | 0.05021 | 0.46052 |
| | 2 | 0.52539 | 0.18085 | 0.39868 | | 2 | 0.44379 | 0.22599 | 1.35044 |
| | 3 | 0.53983 | 0.11194 | 1.82725 | | 3 | 0.47233 | 0.04011 | 1.46117 |
| | 4 | 0.92413 | 0.00014 | 1.38950 | | 4 | 0.47152 | 0.00006 | 1.57100 |
| | 5 | 0.50609 | 0.00006 | 0.21180 | | 5 | 0.47849 | 0.00006 | 0.70720 |
| | 6 | 0.50508 | 0.00002 | 0.09687 | | 6 | 0.48556 | 0.00003 | 0.25793 |
| | 7 | 0.50455 | 0.00001 | 0.04355 | | 7 | 0.49331 | 0.00001 | 0.06075 |
| | 8 | 0.49697 | 0.00000 | 0.01852 | | 8 | 0.54662 | 0.00002 | 0.40637 |
| mp3 | 1 | 0.45959 | 0.33830 | 1.29144 | pdf | 1 | 0.53040 | 1.11262 | 1.59403 |
| | 2 | 0.49030 | 0.11961 | 0.20868 | | 2 | 0.52246 | 0.85132 | 0.95513 |
| | 3 | 0.46626 | 0.07802 | 1.79924 | | 3 | 0.47665 | 0.03464 | 1.22510 |
| | 4 | 0.46807 | 0.00009 | 1.75660 | | 4 | 0.93783 | 0.00012 | 0.42190 |
| | 5 | 0.47609 | 0.00003 | 0.78180 | | 5 | 0.48626 | 0.00003 | 0.45930 |
| | 6 | 0.47752 | 0.00003 | 0.39052 | | 6 | 0.49361 | 0.00000 | 0.11723 |
| | 7 | 0.48587 | 0.00001 | 0.12395 | | 7 | 0.50755 | 0.00001 | 0.07335 |
| | 8 | 0.55455 | 0.00001 | 0.50378 | | 8 | 0.56218 | 0.00003 | 0.60707 |

**Table 3**
Deviation in percentage values of $E_{n,k}$ in binary files from the theoretical values with $n = 8$.

| $k$ | Deviation in percentage values of $E_{n,k}$ | | | | |
|---|---|---|---|---|---|
| | $x = 0$ | $x = 1$ | $x = 2$ | $x = 3$ | $x = 4$ |
| 1 | 0.7006 | −1.5730 | −0.4146 | 0.7243 | 0.5627 |
| 2 | 0.0503 | 0.2204 | −0.1927 | −0.0780 | – |
| 3 | 0.1041 | 0.1357 | −0.2398 | – | – |
| 4 | −0.0001 | 0.0001 | – | – | – |
| 5 | 0.0000 | 0.0000 | – | – | – |
| 6 | 0.0000 | 0.0000 | – | – | – |
| 7 | 0.0000 | 0.0000 | – | – | – |
| 8 | 0.0000 | 0.0000 | – | – | – |

### 3.2. Fit of experimental data frequencies

In this example we will try to provide reasonable fits to the experimental data frequencies presented by Sinha and Sinha [13] which refer to a wide range of file types commonly encountered in real life computer applications. To achieve this goal we search for a success probability $p$, $0 < p < 1$, such that given the length of the sequence $n$ and the success run length $k$, the root mean square error rmse($p$; $n$, $k$) between the experimental relative frequencies $f^*(x)$-expressed in percentage values, and the theoretical values $f(x) = 100P(E_{n,k} = x)$ becomes approximately a (global) minimum. To solve the latter problem we used a bracketing procedure to determine local minima of rmse in the interval (0, 1) which then was accompanied by a bisection like scheme for every bracketing interval. See, e.g. [32, pp. 138–140].

To get a sense of our approach, let us consider $n = 8$ and the experimental relative frequencies of their Tables 5–10. Then, Table 2 provides for every concerned file type an approximate success probability $p$ proper for the corresponding values of $k$ and $n$. In the table the respective values of rmse($p$; 8, $k$) and rmse(0.5; 8, $k$) between $f(x)$ and $f^*(x)$, are also given. Comparing the latter values and observing that the majority of the entries of the columns of $p$ are not so close to 0.5 the usefulness of our approach is evident.

As an indicative example we present in Table 3 the deviations of the experimental relative frequencies expressed in percentage values of $E_{n,k}$, from the theoretically calculated frequencies (using the suggested values of $p$ in Table 2) for binary files with $n = 8$. The entries of Table 3 should be compared to their Table 11. The goodness of our fitted results with success probability not necessarily equal to 0.5 is clear.

Table 4 offers an overall picture of our results. We tabulate for $n = 8$ the root mean square deviation (that is rmse) in percentage values of the experimental frequencies from the theoretically obtained ones (using the probabilities $p$ given by Table 2) over all the studied file types for different values of $k$ and $x$. The entries of Table 4 compared to the respective entries of their Table 17 suggest a serious reduction, ranging from 48.0740% to 99.9975%, between the corresponding rmse. This improvement supports our fitting approach using non-necessarily equiprobable Bernoulli trials.

**Table 4**
rmse in percentage values of $E_{n,k}$ in application data from theoretical values with $n = 8$.

| $k$ | rmse in percentage values of $E_{n,k}$ | | | | |
|---|---|---|---|---|---|
| | $x = 0$ | $x = 1$ | $x = 2$ | $x = 3$ | $x = 4$ |
| 1 | 1.20247 | 2.13115 | 0.33127 | 0.64823 | 0.32839 |
| 2 | 0.81828 | 1.65479 | 0.66739 | 0.19142 | – |
| 3 | 0.06979 | 0.09922 | 0.16868 | – | – |
| 4 | 0.00009 | 0.00009 | – | – | – |
| 5 | 0.00004 | 0.00004 | – | – | – |
| 6 | 0.00002 | 0.00002 | – | – | – |
| 7 | 0.00001 | 0.00001 | – | – | – |
| 8 | 0.00001 | 0.00001 | – | – | – |

**Table 5**
rmse$(n, k, p)$, $d(n, k, p)$ and $d_{\mathrm{TV}}(n, k, p)$ for $\lambda = 1$, 10 and several values of $k$, $n$ and $p$.

| $\lambda$ | $k$ | $n$ | $p$ | $d_{\mathrm{TV}}(n, k, p)$ | $d(n, k, p)$ | rmse$(n, k, p)$ |
|---|---|---|---|---|---|---|
| 1 | 2 | 50 | 0.131877 | 0.269970 | 0.094311 | 0.025897 |
| | | 100 | 0.095163 | 0.182036 | 0.068608 | 0.013720 |
| | | 200 | 0.068269 | 0.123301 | 0.049486 | 0.007006 |
| | | 1 000 | 0.031128 | 0.051239 | – | – |
| | | 10 000 | 0.009950 | 0.015402 | – | – |
| | 10 | 100 | 0.467918 | 0.043730 | 0.035927 | 0.015815 |
| | | 200 | 0.444954 | 0.046943 | 0.041121 | 0.013085 |
| | | 1 000 | 0.394189 | 0.058928 | – | – |
| | | 10 000 | 0.328410 | 0.080402 | – | – |
| 10 | 10 | 100 | 0.548115 | 0.226177 | 0.171281 | 0.071667 |
| | | 200 | 0.523427 | 0.259540 | 0.199930 | 0.059673 |
| | | 1 000 | 0.467918 | 0.324762 | – | – |
| | | 10 000 | 0.394189 | 0.428692 | – | – |
| | 30 | 100 | 0.603912 | 0.000031 | 0.000024 | 0.000017 |
| | | 200 | 0.595444 | 0.000035 | 0.000030 | 0.000016 |
| | | 1 000 | 0.575862 | 0.000055 | – | – |
| | | 10 000 | 0.548115 | 0.000177 | – | – |

Readily, the same method could be used to get better fits for the experimental results of their Figs. 2–8. Finally, to fit the results of their Fig. 9 we could approximate the exact distribution of $E_{n,k}$ by Theorem 2.3 since they refer to a long sequence ($n = 2048$) the length of which might prevent (in some computers) the implementation of the involved binomial coefficients of Theorem 2.1; see also Section 3.3 (Table 7).

### 3.3. Approximations and bounds

(A) In order to illustrate Remark 2.3 about possible implementation of a Poisson or a normal approximation of $E_{n,k}$ defined on Bernoulli sequences we discuss some indicative cases along with the associated numerics. Accordingly, we might have the following cases.

Case I: Let $k$ be fixed, and $\lambda > 0$ be given or it can be estimated. We suppose that $p = p_n$ is changing continuously as $n$ increases following the law $p_n = 1 - \mathrm{e}^{-(\lambda/n)^{1/k}}$. Table 5 presents for $\lambda = 1$, 10 and for several values of $k$ and $n$ the upper bound, $d_{\mathrm{TV}}(n, k, p)$, of the total variation distance, $d(n, k, p)$. When the used values of $n$ allow the computation of the exact values of PMF of $E_{n,k}$ and consequently of the actual values of $d(n, k, p)$ as well as of the root mean square error, rmse$(n, k, p)$, the latter values are also provided. The entries of the table clarify the Poisson approximation of $E_{n,k}$.

Case II: Let a fixed pair $(p, k)$ be given. Then, we examine if $n$ is large enough so that condition (17) is satisfied so that $E_{n,k}$ be well approximated by a normal RV. Table 6 shows for $p = 0.1, 0.5$; $k = 1, 2, 5$ and for $n = 50, 100, 200, 500$ the respective computed $\mathrm{rmse}_N(n, k, p)$ of a normal approximation to $E_{n,k}$. The values of $n$ with stars $(*)$ are the minimum values of $n$ such that (17) holds for the depicted values of $p$ and $k$. In the table the $\mathrm{rmse}_{\mathrm{Po}}(n, k, p)$ as well as the $d(n, k, p)$ of a Poisson approximate RV with mean $\lambda = np^k$ are also included for comparison. The rmse entries of the table suggest that in the vast majority of the studied cases the results are in favor of the normal approximation. We note that although in some depicted cases the conservative condition (17) is not satisfied the corresponding $\mathrm{rmse}_N$ could be acceptable.

Case III: Usually in practice, a long enough sequence with length $n$ is given and someone wants to know if for a candidate pair $(p, k)$ the distribution of $E_{n,k}$ can be well approximated by that of a normal or Poisson RV. In such situations for which the exact distribution of $E_{n,k}$ cannot be computed because of the large value of $n$ we might proceed as follows: initially, we check if condition (17) is satisfied. If it is, we use a normal approximation of $E_{n,k}$ and if it is not satisfied we check if the error bound $d_{\mathrm{TV}}(n, k, p)$ is acceptable (in any case smaller than 1). If it is acceptable, depending on the demanding

**Table 6**
Condition (17), $\text{rmse}_\alpha$ ($\alpha = N$, Normal; $\alpha =$ Po, Poisson) and $d(n, k, p)$ for $p = 0.1, 0.5$ and several values of $k$ and $n$.

| $p$ | $k$ | $n$ | (17) | $\text{rmse}_N(n, k, p)$ | $\text{rmse}_{\text{Po}}(n, k, p)$ | $d(n, k, p)$ |
|---|---|---|---|---|---|---|
| 0.1 | 1 | 50 | F | 0.005728 | 0.023656 | 0.179521 |
| | | 87* | T | 0.002691 | 0.020762 | 0.242380 |
| | | 100 | T | 0.002266 | 0.020176 | 0.259895 |
| 0.5 | 1 | 50 | F | 0.003444 | 0.076555 | 0.990571 |
| | | 64* | T | 0.002469 | 0.063872 | 0.996620 |
| | | 100 | T | 0.001404 | 0.045882 | 0.999792 |
| | 2 | 100 | F | 0.003029 | 0.072891 | 0.991804 |
| | | 117* | T | 0.002448 | 0.064692 | 0.996012 |
| | | 200 | T | 0.001237 | 0.043437 | 0.999834 |
| | 5 | 100 | F | 0.026092 | 0.134856 | 0.643177 |
| | | 500 | F | 0.004982 | 0.050942 | 0.958555 |

**Table 7**
Critical $k$'s, condition (17) and upper bound $\delta(n, k, p)$ for several values of $p$, $k$ and $n = 2048, 8192$.

| $n$ | $p$ | $\lceil E(L_n) \rceil$ | $k$ | (17) | $\delta(n, k, p)$ | $n$ | $p$ | $\lceil E(L_n) \rceil$ | $k$ | (17) | $\delta(n, k, p)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2048 | 0.90 | 56 | 1 | T | 1.000000 | 8192 | 0.90 | 69 | 1 | T | 1.000000 |
| | | | 7* | T | 1.000000 | | | | 21* | T | 1.000000 |
| | | | 56 | F | 1.000000 | | | | 69 | F | 1.000000 |
| | | | 101** | F | 0.049021 | | | | 114** | F | 0.049437 |
| | 0.50 | 11 | 1 | T | 1.000000 | | 0.50 | 13 | 1 | T | 1.000000 |
| | | | 5* | T | 1.000000 | | | | 7* | T | 1.000000 |
| | | | 11 | F | 0.549678 | | | | 13 | F | 0.544858 |
| | | | 15** | F | 0.047501 | | | | | | |
| | 0.10 | 4 | 1 | T | 1.000000 | | | | 17** | F | 0.047050 |
| | | | 2* | T | 1.000000 | | 0.10 | 4 | 1 | T | 1.000000 |
| | | | 4** | F | 0.040147 | | | | 2* | T | 1.000000 |
| | 0.05 | 3 | 1* | T | 1.000000 | | | | 4 | F | 0.121259 |
| | | | 3** | F | 0.026255 | | | | 5** | F | 0.015714 |
| | 0.01 | 2 | 1* | T | 0.266138 | | 0.01 | 2 | 1* | T | 0.896528 |
| | | | 2** | F | 0.004867 | | | | 2** | F | 0.013350 |

accuracy of the results, (e.g. $d_{\text{TV}} \leq 0.5 \times 10^{-\ell}$, $\ell = 1, 2, \ldots$) then we use a Poisson approximation of $E_{n,k}$. Otherwise, we record all the previous steps and we give a warning about the risk of using such approximations of the distribution of $E_{n,k}$ with parameter vector $(n, k, p)$. Numerical investigation suggested that although conditions (17) and $d_{\text{TV}} \leq 0.5 \times 10^{-\ell}$, $\ell = 1, 2, \ldots$ are very conservative; when they are satisfied we obtain satisfactory enough normal or Poisson approximations of $E_{n,k}$.

As an example let us consider $n = 2048, 8192$; $p = 0.90, 0.50, 0.10, 0.05, 0.01$. Since $n$ is large we computed (17) and $\delta(n, k, p) = \min\{1, d_{\text{TV}}(n, k, p)\}$ for several indicative values of $k$. The $k$'s with one (two) stars are the maximum (minimum) ones such that condition (17) ($d_{\text{TV}} \leq 0.5 \times 10^{-1}$) is satisfied. In Table 7, $\lceil E(L_n) \rceil$ which is a characteristic length of every binary sequence is included in the table, too. $E(L_n)$ is the expected length of the longest success run in a Bernoulli sequence of length $n$ with success probability $p$. Numerical investigation suggested that the normal approximation of $E_{n,k}$ behaves well for small $k$'s upper bounded by $\lceil E(L_n) \rceil$, whereas the Poisson approximation offers acceptable results for large $k$'s lower bounded by $\lceil E(L_n) \rceil$. Accordingly, the bell shaped experimental distribution of $E_{2048,k}$, $k = 1, 2$ presented in Fig. 9 of [13] might be approximated by smoothing it by a normal distribution with proper $p$'s determined by the method presented in Section 3.2.

(B) Next, we clarify Propositions 2.1 and 2.2 by numerical examples. Specifically, Proposition 2.1 provides easily computed expressions of $E(E_{n,k})$ and $V(E_{n,k})$ which in turn, via Proposition 2.2, can be used to obtain bounds and approximations of the distribution of $E_{n,k}$ (see Remark 2.4). This information is complementary to the one derived using normal or Poisson approximations (see Remark 2.3).

Table 8 demonstrates exact means and variances of $E_{n,k}$ along with the values of $n\mu$, $n\sigma^2$ and $\lambda$ for the indicative values of $p = 0.95, 0.5, 0.05$ and for $n = 10^\ell$, $\ell = 3, 4, 6$. The chosen values of $k$ are equal to $\lceil E(L_n) \rceil$. The entries of the table suggest that as $n$ increases $n\mu$ and $n\sigma^2$ become good approximates of $E(E_{n,k})$ and $V(E_{n,k})$, respectively. The same is true for $\lambda$ when $p$ is small enough.

In Table 9 we tabulate exact values, approximate values, bounds and upper bounds of the relative error of the probability $P(E_{n,k} = 0)$ of having no success runs of length exactly equal to $k$, for a variety of parameters $n$, $k$ and $p$. In the table, in addition to $k = k^* = \lceil E(L_n) \rceil$, we selected for comparison $k$'s so that to have the same percentages $k/n$ in different values of $n$, $p$. The subscripts $N$ and Po stand for the normal and Poisson approximation, respectively, of $G(1)$ whereas the subscripts $MC$ and $C$ refer to values calculated via Proposition 2.2.

**Table 8**
Exact and approximate means and variances of $E_{n,k}$ for $k = \lceil E(L_n) \rceil$.

| $p$ | $n$ | $k$ | $E(E_{n,k})$ | $V(E_{n,k})$ | $n\mu$ | $n\sigma^2$ | $\lambda$ |
|---|---|---|---|---|---|---|---|
| 0.95 | 1 000 | 88 | 0.02604940 | 0.02395299 | 0.02739159 | 0.02728730 | 10.95663680 |
| | 10 000 | 132 | 0.02840521 | 0.02838682 | 0.02867186 | 0.02865320 | 11.46874504 |
| | 1000 000 | 222 | 0.02834660 | 0.02834627 | 0.02835178 | 0.02835146 | 11.34071337 |
| 0.50 | 1 000 | 10 | 0.24243164 | 0.24131209 | 0.24414063 | 0.24300814 | 0.97656250 |
| | 10 000 | 13 | 0.30487061 | 0.30463815 | 0.30463815 | 0.30494295 | 1.22070313 |
| | 1000 000 | 20 | 0.23841453 | 0.23841231 | 0.23841858 | 0.23841636 | 0.95367432 |
| 0.05 | 1 000 | 2 | 2.25423125 | 2.22936446 | 2.25625000 | 2.23133254 | 2.50000000 |
| | 10 000 | 3 | 1.12791125 | 1.12703409 | 1.12812500 | 1.12724753 | 1.25000000 |
| | 1000 000 | 5 | 0.28203015 | 0.28202928 | 0.28203125 | 0.28203038 | 0.31250000 |

**Table 9**
Exact values, approximate values and bounds of $G(1) = P(E_{n,k} < 1) = P(W_{1,k} > n)$.

| $p$ | $n$ | $k$ | $k/n(\%)$ | $G(1)$ | $\hat{G}_N(1)$ | $\hat{G}_{\mathrm{Po}}(1)$ | $\hat{B}_{\mathrm{Po}}(1)$ | $L_{MC}(1)$ | $U_C(1)$ | $\hat{G}_{MC}(1)$ | $\hat{B}_{MC}(1)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.95 | 100 | 1 | 1 | 0.728448 | 0.697616 | – | – | 0.672250 | 0.765310 | 0.718780 | 0.069215 |
| | | 10 | 10 | 0.825480 | 0.814382 | – | – | 0.806907 | 0.839909 | 0.823408 | 0.020450 |
| | | 43* | 43 | 0.973707 | 0.997922 | – | – | 0.973556 | 0.973854 | 0.973705 | 0.000153 |
| 0.5 | 100 | 1 | 1 | 0.000002 | 0.000143 | – | – | – | 0.063715 | – | – |
| | | 6* | 6 | 0.679010 | 0.570986 | – | – | 0.621094 | 0.716579 | 0.668837 | 0.076869 |
| | | 10 | 10 | 0.977502 | 0.998859 | 0.994119 | – | 0.977295 | 0.977703 | 0.977499 | 0.000209 |
| | 1000 | 10* | 1 | – | 0.698129 | 0.922720 | – | 0.757568 | 0.804145 | 0.780857 | 0.030741 |
| | | 100 | 10 | – | 1.000000 | 1.000000 | 0.000000 | 1.000000 | – | – | – |
| 0.05 | 100 | 1 | 1 | 0.007890 | 0.021391 | 0.502021 | – | – | 0.161513 | – | – |
| | | 2* | 2 | 0.798647 | 0.719328 | 0.822070 | 0.058826 | 0.776394 | 0.815611 | 0.796003 | 0.025256 |
| | | 10 | 10 | 1.000000 | 1.000000 | 1.000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000 | 0.000000 |
| | 1000 | 2* | 0.2 | – | 0.119853 | 0.256556 | – | – | 0.304936 | – | – |
| | | 10 | 1 | – | 1.000000 | 1.000000 | 0.000000 | 1.000000 | 1.000000 | 1.000000 | 0.000000 |

## 4. Discussion and further results

In the present paper we studied the exact and limiting distribution of $E_{n,k}$ and we obtained its mean value and variance. The presented results are either new or they reconsider under a new aspect, results of other researchers by providing alternative formulae or by giving to them an additional meaning and possible applicabilities.

Besides its independent merit $E_{n,k}$ may be used in the representation of other interesting statistics. For instance, we refer to the following ones that have been frequently discussed in the literature: (a) The number $G_{n,k}$ of success runs of length at least $k$, i.e. $G_{n,k} = \sum_{i=k}^{n} E_{n,i}$ (see, e.g. [5,7–12,15,17–20]). (b) The (total) number $S_{n,k}$ of successes in all success runs of length greater than or equal to $k$, i.e. $S_{n,k} = \sum_{i=k}^{n} iE_{n,i}$ (see, e.g. [3,4,6,8,12,33]). An alternative interpretation of $S_{n,k}$ is that it denotes the sum of the lengths of the success runs of length at least equal to $k$.

The approach used to derive the number $N_{n;e_k}$ can also be employed to establish analogous numbers associated with the statistics $G_{n,k}$ and $S_{n,k}$ for $1 \le k \le n$. Accordingly, we have

(a) Let $N_{n;g_k}$ denote the number of binary strings which contain for a given $g_k$, $g_k = 0, 1, \ldots, \left\lfloor \frac{n+1}{k+1} \right\rfloor$, exactly $g_k$ runs of 1's of length at least $k$ in all possible binary strings of length $n$.

(b) Let $N_{n;g_k,s_k}$ denote the number of binary strings which contain for given $g_k$ and $s_k$, $g_k = 0, 1, \ldots, \left\lfloor \frac{s_k}{k} \right\rfloor$, $s_k = 0, k, k+1, \ldots, n$, exactly $g_k$ runs of 1's of length at least $k$ with total number of 1's (with sum of lengths of runs of 1's) exactly equal to $s_k$ in all possible binary strings of length $n$.

The previous numbers can be proved to be useful in engineering applications similar to the ones discussed in [13]. Employing analogous arguments to those used in Theorems 3.3 and 4.1 of Makri et al. [8] we obtain

$$N_{n;g_k} = \sum_{y=0}^{n-kg_k} \binom{y+1}{g_k} \sum_{j=0}^{\left\lfloor \frac{n-y-kg_k}{k} \right\rfloor} (-1)^j \binom{y+1-g_k}{j} \binom{n-k(g_k+j)}{n-y-k(g_k+j)}, \tag{23}$$

$$N_{n;g_k,s_k} = \sum_{y=0}^{n-s_k} \binom{y+1}{g_k} \binom{s_k-(k-1)g_k-1}{g_k-1} \sum_{j=0}^{\left\lfloor \frac{n-y-s_k}{k} \right\rfloor} (-1)^j \binom{y+1-g_k}{j} \binom{n-s_k-kj-g_k}{n-s_k-kj-y} \tag{24}$$

for $g_k = 1, \ldots, \left\lfloor \frac{s_k}{k} \right\rfloor$, $s_k = k, k+1, \ldots, n$; $N_{n;g_k,s_k} = \sum_{y=0}^{n} \sum_{j=0}^{\left\lfloor \frac{n-y}{k} \right\rfloor} (-1)^j \binom{y+1}{j} \binom{n-kj}{n-y-kj}$, for $s_k = g_k = 0$.

**Table 10**
Total number of 1's in binary strings of length $n = 2, 4, 8$ and 16.

| $n$ | $k$ | $R_{n,k}^{(e)}$ | $R_{n,k}^{(g)}$ | $R_{n,k}^{(s)}$ | $n$ | $k$ | $R_{n,k}^{(e)}$ | $R_{n,k}^{(g)}$ | $R_{n,k}^{(s)}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 2 | 3 | 4 | 16 | 1 | 147 456 | 278 528 | 544 288 |
| | 2 | 1 | 1 | 2 | | 2 | 69 632 | 131 072 | 376 832 |
| 4 | 1 | 12 | 20 | 32 | | 3 | 32 768 | 61 440 | 237 568 |
| | 2 | 5 | 8 | 20 | | 4 | 15 360 | 28 672 | 139 264 |
| | 3 | 2 | 3 | 10 | | 5 | 7 168 | 13 312 | 77 824 |
| | 4 | 1 | 1 | 4 | | 6 | 3 328 | 6 144 | 41 984 |
| 8 | 1 | 320 | 576 | 1024 | | 7 | 1 536 | 2 816 | 22 016 |
| | 2 | 144 | 256 | 704 | | 8 | 704 | 1 280 | 11 264 |
| | 3 | 64 | 112 | 416 | | 9 | 320 | 576 | 5 632 |
| | 4 | 28 | 48 | 224 | | 10 | 144 | 256 | 2 752 |
| | 5 | 12 | 20 | 112 | | 11 | 64 | 112 | 1 312 |
| | 6 | 5 | 8 | 52 | | 12 | 28 | 48 | 608 |
| | 7 | 2 | 3 | 22 | | 13 | 12 | 20 | 272 |
| | 8 | 1 | 1 | 8 | | 14 | 5 | 8 | 116 |
| | | | | | | 15 | 2 | 3 | 46 |
| | | | | | | 16 | 1 | 1 | 16 |

Using $N_{n;g_k}$ and $N_{n;g_k,s_k}$, $1 \le k \le n$, the total number $R_{n,k}^{(g)}$ of occurrences of all runs of 1's of length at least $k$ and the total number $R_{n,k}^{(s)}$ of 1's in all runs of 1's of length at least $k$, in all possible strings of length $n$, are

$$R_{n,k}^{(g)} = \sum_{g_k=1}^{\left\lfloor \frac{n+1}{k+1} \right\rfloor} g_k N_{n;g_k}, \qquad R_{n,k}^{(s)} = \sum_{s_k=k}^{n} s_k \sum_{g_k=1}^{\lfloor s_k/k \rfloor} N_{n;g_k,s_k}. \tag{25}$$

We note that $R_{n,k}^{(e)} \le R_{n,k}^{(g)} \le R_{n,k}^{(s)}$ since $E_{n,k} \le G_{n,k} \le S_{n,k}$, $1 \le k \le n$. Table 10 presents the three numbers $R_{n,k}^{(e)}$, $R_{n,k}^{(g)}$ and $R_{n,k}^{(s)}$ in binary strings of $n = 2^\ell$, $\ell = 1, 2, 3, 4$ bits for $1 \le k \le n$. From the entries of the table we observe that for a fixed $n$, $R_{n,k}^{(\alpha)}$, $\alpha = e, g$ decreases exponentially as $k$ increases. In fact, we have

$$\begin{aligned} R_{n,k}^{(\alpha)} &= 1, \text{ for } \alpha = e, g, \ k = n; \quad 2, \text{ for } \alpha = e, \ k = n-1; \quad 3, \text{ for } \alpha = g, \ k = n-1 \\ &= (n-k+3)2^{n-k-2}, \quad \text{for } \alpha = e, \ k = 1, 2, \ldots, n-2 \\ &= (n-k+2)2^{n-k-1}, \quad \text{for } \alpha = g, \ k = 1, 2, \ldots, n-2 \end{aligned} \tag{26}$$

and for fixed $k$, $1 \le k \le n-2$, $\frac{R_{n,k}^{(g)}}{R_{n,k}^{(e)}} \to 2$, as $n \to \infty$. Note that in (26) for $\alpha = e$ we rephrased the solution of (4) of Sinha and Sinha [13]. A simple explicit form of $R_{n,k}^{(s)}$ remains an open issue.

## Acknowledgements

The authors would like to thank the referees for the thorough reading, helpful comments and suggestions which helped to improve this paper.

## References

[1] N. Balakrishnan, M.V. Koutras, Runs and Scans with Applications, Wiley, New York, 2002.
[2] J.C. Fu, W.Y.W. Lou, Distribution Theory of Runs and Patterns and its Applications: A Finite Markov Imbedding Approach, World Scientific, New Jersey, 2003.
[3] J.C. Fu, W.Y.W. Lou, Z. Bai, G. Li, The exact and limiting distributions for the number of successes in success runs within a sequence of Markov-dependent two-state trials, Ann. Inst. Statist. Math. 54 (2002) 719–730.
[4] D.L. Antzoulakos, S. Bersimis, M.V. Koutras, On the distribution of the total number of run lengths, Ann. Inst. Statist. Math. 55 (2003) 865–884.
[5] M.V. Koutras, Applications of Markov chains to the distribution theory of runs and patterns, in: D.N. Shanbhag, S.C. Rao (Eds.), Stochastic Processes: Modelling and Simulation, in: Handbook Statist., vol. 21, North-Holland, Amsterdam, 2003, pp. 431–472.
[6] W.Y.W. Lou, The exact distribution of the $k$-tuple statistic for sequence homology, Statist. Probab. Lett. 61 (2003) 51–59.
[7] S. Eryilmaz, S. Demir, Success runs in an sequence of exchangeable binary trials, J. Statist. Plann. Inference 137 (2007) 2954–2963.
[8] F.S. Makri, A.N. Philippou, Z.M. Psillakis, Success run statistics defined on an urn model, Adv. in Appl. Probab. 39 (2007) 991–1019.
[9] S. Eryilmaz, Run statistics defined on the multicolor urn model, J. Appl. Probab. 45 (2008) 1007–1023.
[10] S. Demir, S. Eryilmaz, Run statistics in a sequence of arbitrarily dependent binary trials, Statist. Papers (2008) 1–15. doi:10.1007/s00362-008-0191-7.
[11] F.S. Makri, Z.M. Psillakis, On success runs of length exceeded a threshold, Methodol. Comput. Appl. Probab. (2009) 1–37. doi:10.1007/s11009-009-9147-1.
[12] F.S. Makri, Z.M. Psillakis, On runs of length exceeding a threshold: normal approximation, Statist. Papers (2009) 1–21. doi:10.1007/s00362-009-0268-y.
[13] K. Sinha, B. Sinha, On the distributions of runs of ones in binary strings, Comput. Math. Appl. 58 (2009) 1816–1829.

[14] J.C. Fu, W.Y.W. Lou, On the normal approximation for the distribution of the number of simple or compound patterns in a random sequence of multi-state trials, Methodol. Comput. Appl. Probab. 9 (2007) 195–205.
[15] A.M. Mood, The distribution theory of runs, Ann. Math. Stat. 11 (1940) 367–392.
[16] G. Blom, L. Holst, D. Sandell, Problems and Snapshots from the World of Probability, Springer-Verlag, New York, 1994.
[17] J.C. Fu, M. Koutras, Distribution theory of runs: a Markov chain approach, J. Amer. Statist. Assoc. 89 (1994) 1050–1058.
[18] D.L. Antzoulakos, On waiting time problems associated with runs in Markov dependent trials, Ann. Inst. Statist. Math. 51 (1999) 323–330.
[19] Q. Han, S. Aki, Joint distributions of runs in a sequence of multi-state trials, Ann. Inst. Statist. Math. 51 (1999) 419–447.
[20] K. Sen, M.L. Agarwal, S. Chakraboty, Lengths of runs and waiting time distributions by using Polya–Eggenberger sampling scheme, Studia Sci. Math. Hungar. 2 (2002) 309–332.
[21] W. Mendenhall, T. Sincich, A Second Course in Statistics: Regression Analysis, fifth ed., Pentice Hall, New Jersey, 1996.
[22] F.S. Makri, Z.M. Psillakis, Bounds for reliability of $k$-within connected-$(r, s)$-out-of-$(m, n)$ failure systems, Microelectron. Reliab. 37 (1997) 1217–1224.
[23] W. Feller, An Introduction to Probability Theory and its Applications, 3rd ed., vol. 1, Wiley, New York, 1968.
[24] C.A. Charalambides, Enumerative Combinatorics, Chapman and Hall, CRC, Boca Raton, 2002.
[25] N. Johnson, S. Kotz, Urn Models and their Applications, John Wiley, New York, 1977.
[26] R. Arratia, L. Goldstein, L. Gordon, Two moments suffice for Poisson approximations: the Chen–Stein method, Ann. Probab. 17 (1989) 9–25.
[27] W. Hoeffding, H. Robbins, The central limit theorem for dependent random variables, Duke Math. J. 15 (1948) 773–780.
[28] N. Yannaros, Poisson approximation for random sums of Bernoulli random variables, Statist. Probab. Lett. 11 (1991) 161–165.
[29] E.O. George, D. Bowman, A full likelihood procedure for analyzing exchangeable binary data, Biometrics 51 (1995) 512–523.
[30] F.S. Makri, On occurrences of $F$–$S$ strings in linearly and circularly ordered binary sequences, J. Appl. Probab. 47 (2010) 157–178.
[31] V.B. Nevzorov, Records: Mathematical Theory, American Mathematical Society, Providence, RI, 2001.
[32] B. Gottfried, Programming with Pascal, Schaum McGraw-Hill, 1985.
[33] D.E. Martin, Distribution of the number of successes in success runs of length at least $k$ in higher-order Markovian sequences, Methodol. Comput. Appl. Probab. 7 (2005) 543–554.