



Binomial Distribution of Order k in a Modified Binary Sequence

Spiros D. Dafnis¹ · Markos V. Koutras² · Frosso S. Makri³

Accepted: 1 May 2022 / Published online: 8 June 2022
© Grace Scientific Publishing 2022

Abstract

Let us consider a sequence of n binary trials (signals). A counter registers successes, but once a success is registered the mechanism is locked for a number of trials following each registration. Under this framework the observed sequence of outcomes turns to a dependent sequence with non-identical success probabilities even if the original trials were independent and identically distributed. In the present paper, we study the distribution of the number of success runs registered by the counter after the completion of the n signals. Our study covers the general case where the original trials are independent but not necessarily identically distributed. The special case of identically distributed trials gives birth to the modified binomial distribution of order k , which generalizes binomial distributions extensively studied in the literature. In this case, we derive neat recursive relations for the probability mass function, the probability generating function and the moments. The applicability of the modified binomial distribution of order k in several research areas is highlighted and after devel-

Spiros D. Dafnis and Frosso S. Makri have contributed equally to this work.

This article is part of the topical collection “Advances in Probability and Statistics: an Issue in Memory of Theophilos Cacoullas” guest edited by Narayanaswamy Balakrishnan, Charalambos A. Charalambides, Tasos Christofides, Markos Koutras, and Simos Meintanis.

✉ Markos V. Koutras
mkoutras@unipi.gr

Spiros D. Dafnis
sdafnis@aia.gr; dafnisspyros@gmail.com

Frosso S. Makri
makri@math.upatras.gr

¹ Department of Crop Science, Agricultural University of Athens, Iera Odos 75, 11855 Athens, Greece

² Department of Statistics and Insurance Science, University of Piraeus, Karaoli and Dimitriou 80, 18534 Piraeus, Greece

³ Department of Mathematics, University of Patras, Rio Campus, 26504 Patras, Greece

oping theoretical results we discuss how they can be exploited to study a biomedical engineering problem.

Keywords Binary sequence · Run · Counter · Markov-chain · Biomedical engineering

1 Introduction

The concept of success run has been extensively used in many applications of several research areas, where the interest focuses on the study of experimental trials with two outcomes. For example,

a. For a mechanical engineer performing a start-up test for a unit, it is reasonable to couch his/her decision (accepting the machine or rejecting it) on the number of consecutive successful or unsuccessful attempted start-ups (see [4, 5]).

b. Many quality control plans base the acceptance/rejection of the sample lot on the occurrence of prolonged sequences of successive working/failed components (see e.g. [6, 22]).

c. In biosurveillance, the occurrence of many consecutive days with the number of new Covid infections exceeding a warning threshold, may trigger restriction measures in specific locations or countries.

Similar set-ups may also be encountered in molecular biology, finance, actuarial science, ecology, reliability engineering, etc., see [3].

Run-related problems have attracted the attention of probabilists and statisticians as far back as the 18-th century. As mentioned in [19] “The Probability of throwing a Chance assigned a given number of times without intermission, in any given number of Trials” (De Moivre [11], p. 243) was interpreted by Todhunter [29] to mean the probability that a run of r successes is completed at the n -th trial in a sequence of Bernoulli trials each with probability of success p . Feller [13] exploited the theory of recurrent events to establish a formula for the probability generating function (pgf) of the distribution of the trial number X at which the first run of length r occurs.

The paper by Philippou et al. [25] inspired a remarkable upsurge of interest in these distributions under the name *success runs distributions of order k* . From 1984 onward there has been a vast research literature on run-related distributions. The classical framework for a fixed length run-related problem is mentioned in [13]. A sequence of n Bernoulli trials is observed, with the possible outcomes labelled as success (S) or failure (F), and the number $N_{n,k}$ of non-overlapping and recurrent success runs of length k is counted (k is a fixed positive integer). The nomenclature used for the distribution of $N_{n,k}$ is “binomial distribution of order k ”. The classical geometric, negative binomial, logarithmic and Poisson distributions have been generalized as well in a runs context.

The distributions of order k have many practical applications in areas as diverse as statistical quality control, nonparametric statistical inference, molecular biology, ecology, meteorology, psychology, reliability, start-up demonstration testing, etc., see Balakrishnan and Koutras [3] where a good overview of several applications has been included. In the same book, properties, asymptotics, and estimation of the param-

eters of the distributions of order k are covered. The problems of deriving explicit expressions for the probability mass functions and cumulative distribution functions are discussed and appropriate references are provided. Several diagrams illustrating the shapes of the distributions are also provided. For more recent results in the area of run-related distributions the interested reader may refer to [2, 7, 23] where good overviews up to date at the time the articles were written are also included.

There are, however, many interesting applications in the aforementioned areas, where the Bernoulli model is not appropriate. For this reason, several modifications of the traditional framework have been suggested in the statistics bibliography, the most popular ones being the two-state Markov model, the exchangeable binary trials and the independent but not identical binary model.

In the present article, we shall study the distribution of $N_{n,k}$ in a sequence of binary trials where the probability of success p does not remain constant; whenever at any given trial success results with probability p , then in the next $r - 1$ trials we assume the probability of success vanishes.

The aforementioned set-up was first used by Dandekar [10] to introduce a “modified” binomial distribution. He also discussed an application of it to a fertility enquiry problem.

Another interesting application of Dandekar’s model arises in the study of the Geiger counters used for cosmic rays and α -particles. As indicated in [13] (see page 306) counters of this type may be described by the following simplified model. Bernoulli trials are performed at a uniform rate. A counter is used to register successes, but once a success is registered the mechanism is locked for the next $r - 1$ trials. In other words, a success at the n -th trial is registered if, and only if, no registration has occurred in the preceding $r - 1$ trials. The counter is then locked at trials numbered $n + 1, \dots, n + r - 1$ and is freed at the $(n + r)$ -th trial, until another registration occurs and the system locks again for the subsequent $r - 1$ trials. Manifestly, the output of the counter consists of dependent trials. It should be stressed that the original sequence of signals arriving at the counter could be either a sequence of identical trials or non-identical ones.

In the present article, we shall study a generalization of Dandekar’s [10] modified binomial distribution by looking at the number of non-overlapping success runs of length k in a non-iid binary sequence. More specifically, we assume that we have a sequence of n independent binary trials (signals) with success (failure) probabilities p_t (q_t), $t = 1, 2, \dots, n$. A counter registers only the S outcomes, and each time an S is registered the counter keeps locked for the next $r - 1$ incoming signals (trial outcomes). The random variable (r.v.) of interest, to be denoted by $N_{n,k,r}$, is the number of success runs registered by the counter after the completion of the n S/F signals ($k \geq 1$, $r \geq 1$ and $n \geq (k - 1)r + 1$).

As an illustration, let us consider the case $n = 6$, $k = 2$, $r = 3$ and $n = 6$, $k = 2$, $r = 2$. Then the event $N_{6,2,3} = 1$ contains the next 3 realizations

$$S * S * S * *, \quad FS * S * S *, \quad FFS * S * S *$$

while the event $N_{6,2,2} = 1$ contains the next 8 realizations

$$\begin{aligned} S * S * S *, \quad S * S * FS, \quad S * S * FF, \quad FS * S * F, \\ FS * S * S, \quad FFS * S *, \quad FFFS * S, \quad S * FS * S, \end{aligned}$$

where $*$ stands for one discrete instance (trial) where the counter is locked (so we are not interested in the specific outcome).

In order to exemplify further the usefulness and applicability of the framework presented before and make clear the motivation for studying $N_{n,k,r}$ we provide below a number of pertinent applications:

a. When investigating a Geiger-counter record, it is natural to look at the probability $P(N_{n,k,r} \geq c)$ to assess the hazard generated by the registered signals. Large values of $P(N_{n,k,r} \geq c)$ indicate increased high levels of temporarily concentrated radiation.

b. In the insurance section and premium pricing, it is reasonable to assume that after an incidence (e.g. a burglary in an insured house), the probability of having a second incidence for the next, say $r - 1$, time periods becomes negligible. Thus the sequence of success (burglary occurrence) - failure (no burglary occurrence) trials resembles the modified binary framework suggested above. Then one may get interested to evaluate $P(N_{n,k,r} \geq c)$ in order to proceed to an appropriate premium pricing.

c. In start-up demonstration testing, let us assume that the tests are performed by an automated mechanism on the same unit, and involve multilevel inspection that increases the mechanical stress on the unit. Then, arriving at an S (failure of the unit) will result at a high stress level for the inspected unit; therefore it is reasonable not to take into account the outcomes obtained for, say $r - 1$, start-up tests following the failure. If the unit rejection is associated with the number of consecutive failures in n tests (i.e. success runs of length k) then $N_{n,k,r}$ can be exploited to study the stochastic behaviour of the whole testing plan.

d. In Covid surveillance, when modelling a characteristic (i.e. number of new cases, number of casualties, etc., in a specific time period) as a binary variable (e.g. low-high value) it may be reasonable to ignore a number of time periods, say $r - 1$, following a high value (success, S); doing so, short-time fluctuations that will result in alternating S - F 's after the observed high value will not be taken into account. Then, large values of $N_{n,k,r}$ will indicate a persisting "bad" situation calling for state decisions.

At this point it should be noted that the special case $p_t = p$ is of great importance. In this case, we name the distribution of $N_{n,k,r}$, $(r - 1)$ -modified binomial distribution of order k . For $r = 1$, the 0-modified binomial distribution of order k is the classical binomial distribution of order k (see [12, 14, 16, 18, 26, 30]), while for $r = 1$ and $k = 1$ it is the usual binomial distribution. Moreover, for $k = 1$, the $(r - 1)$ -modified binomial distribution of order 1 is the modified binomial distribution, studied in [10]. For recent generalizations of distributions of order k we refer to Dafnis and Makri [8], Dafnis et al. [9] and Kumar [24].

Sen et al. [27] studied the distribution of $N_{n,k,r}$ considering a Pólya–Eggenberger sampling scheme and employing interesting combinatorial arguments. In the present paper, we consider two cases which are completely different in nature. In the more general one, the original binary trials are considered to be independent but not identi-

cally distributed. In the second case, the success probability of a trial is considered to be constant and equal to p , i.e. the original binary trials are independent and identically distributed. However, as already been stated, in both cases trials where the enumeration is carried out to determine $N_{n,k,r}$ are neither independent nor identically distributed. We employ the Markov chain imbedding (MCI) technique to study the distribution of $N_{n,k,r}$. We, finally, present an interesting application of the new results, along with numerical results and figures that provide a better illustration of our theoretical study.

Throughout the paper we denote by $[x]$ the greater integer which is less than or equal to x and by $\delta_{i,j}$ the Kronecker's Delta function, i.e. $\delta_{i,j}$ equals 1, if $i = j$ and 0, otherwise.

2 Exact Distribution of $N_{n,k,r}$

In the present section, we shall study the distribution of $N_{n,k,r}$, defined in the Introduction. We shall employ the MCI technique, taking into consideration that trials where the enumeration is carried out to determine $N_{n,k,r}$ may be divided in two different types of subperiods: Subperiods when the counter is not locked and the probability of success of the t -th trial equals p_t and subperiods when the counter is locked and the probability of success of the t -th trial vanishes.

The MCI technique projects the enumerating r.v. of interest to appropriate subspaces of the state space of a properly defined Markov chain. This approach was introduced in the novel paper of Fu and Koutras [14] and was further popularized by the monograph of Fu and Lou [15]. Koutras and Alexandrou [21] refined the method by providing a general recursive scheme for the probability distribution of a Markov chain imbeddable r.v. of binomial type (MVB). The MCI technique was further developed in a series of papers since then (see, among others, [2, 17]). We, now, recall the definitions of a Markov chain imbeddable variable and an MVB.

Definition 1 The integer valued random variable X_n with support $\{0, 1, \dots, \ell_n\}$ (n a nonnegative integer) will be called Markov chain imbeddable variable if

- (i) there exists a Markov chain $\{Y_t; t \geq 0\}$ defined on a state space Ω ,
- (ii) there exists a partition $\{C_x, x = 0, 1, \dots\}$ on Ω ,
- (iii) for every $x = 0, 1, \dots, \ell_n$ the probabilities $P(X_n = x)$ can be deduced by considering the projection of the probability space of Y_n onto C_x i.e.

$$P(X_n = x) = P(Y_n \in C_x), \quad x = 0, 1, \dots, \ell_n.$$

Before we proceed to Definition 2, let us assume that the sets C_x of the partition $\{C_x, x = 0, 1, \dots\}$ have the same cardinality $s = |C_x|$, $x = 0, 1, \dots$, more specifically

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

Definition 2 A nonnegative integer random variable X_n will be called MVB if

- (a) X_n can be embedded into a Markov chain as in Definition 1,

(b) $P(Y_t \in C_{yj} | Y_{t-1} \in C_{xi}) = 0$, for all $y \neq x, x + 1$.

Definition 2 gives birth to the next two $s \times s$ transition probability matrices

$$A_t(x) = (P(Y_t \in c_{xj} | Y_{t-1} \in c_{xi})), \quad B_t(x) = (P(Y_t \in c_{x+1,j} | Y_{t-1} \in c_{xi})).$$

Let $\mathbf{f}_t(x)$ be the probability vector associated with time t and sub-state set C_x , i.e.

$$\mathbf{f}_t(x) = (P(Y_t = c_{x,0}), P(Y_t = c_{x,1}), \dots, P(Y_t = c_{x,s-1})), \quad 0 \leq t \leq n.$$

Then, it is straightforward that the probability mass function of X_n can be expressed as follows

$$P(X_n = x) = \mathbf{f}_n(x)\mathbf{1}', \quad x = 0, 1, \dots, \ell_n \tag{1}$$

with $\mathbf{1} = (1, 1, \dots, 1) \in R^s$. In the sequel we shall adopt the convention $P(X_0 = 0) = 1$ and denote by $\boldsymbol{\pi}_x$ the (row) vector of initial probabilities of the Markov chain. The following lemma (see [21]) provides a recursive scheme for the probability vectors $\mathbf{f}_t(x)$.

Lemma 2.1 *For an MVB X_n the sequence $\mathbf{f}_t(x), t = 1, 2, \dots, n$ satisfies the recurrence relations*

$$\begin{aligned} \mathbf{f}_t(0) &= \mathbf{f}_{t-1}(0)A_t(0), \\ \mathbf{f}_t(x) &= \mathbf{f}_{t-1}(x)A_t(x) + \mathbf{f}_{t-1}(x-1)B_t(x-1), \quad 1 \leq x \leq \ell_n, \end{aligned}$$

with initial conditions $\mathbf{f}_0(x) = \boldsymbol{\pi}_x, 0 \leq x \leq \ell_n$.

We may, now, proceed to derive our new results regarding the distribution of $N_{n,k,r}$. We shall first consider the general case where the trials are not identically distributed while the system is not locked. Under this set-up we have

Theorem 2.1 *The probability mass function (pmf) of the r.v. $N_{n,k,r} (k \geq 1, r \geq 1)$ is given by*

$$P(N_{n,k,r} = x) = \mathbf{f}_n(x)\mathbf{1}', \quad x = 0, 1, \dots, \left\lfloor \frac{n+r-1}{kr} \right\rfloor, \tag{2}$$

where $\mathbf{f}_n(x)$ are probability vectors satisfying the recursive relations of Lemma 2.1, with $A_t, B_t, t = 1, 2, \dots, n$ being defined as follows:

a. A_t is a $kr \times kr$ matrix which has all its entries 0 except from the entries:

- $(1 + ir, 1), i = 0, \dots, k - 1$, which are all equal to q_t ,
- $(1 + ir, 2 + ir)$ for $k \geq 2, i = 0, \dots, k - 2$, which are all equal to p_t ,
- $(2 + ir + j, 3 + ir + j)$ for $k \geq 2$ and $r \geq 2, i = 0, \dots, k - 2, j = 0, \dots, r - 2$, which are all equal to 1,
- $(2 + (k - 1)r + j, 3 + (k - 1)r + j)$ for $r \geq 3, j = 0, \dots, r - 3$, which are all equal to 1,
- $(rk, 1)$ for $r \geq 2$, which equals 1.

b. B_t is a $kr \times kr$ matrix with all its elements vanishing except from the column $(k - 1)r + 1$, if $r \geq 2$, or the first column, if $r = 1$, which equals $\mathbf{1}' - A_t\mathbf{1}'$.

Proof We shall first prove that $N_{n,k,r}$ is an MVB.

We set $\ell_n = \lfloor \frac{n+r-1}{kr} \rfloor$ and introduce the state space $\Omega = \bigcup_{x=0}^{\ell_n} C_x$ where $C_x, x = 0, 1, \dots, \ell_n$ are disjoint subspaces with $|C_x| = kr$ elements labelled as $C_x = \{c_{xi}, i = 0, \dots, kr - 1\}$.

We introduce next a Markov chain $\{Y_t, t \geq 0\}$ on Ω as follows: $Y_t \in c_{x,i} = \{(x, i)\}$, or equivalently $Y_t = (x, i)$, if at the first t outcomes the number of non-overlapping occurrences of k consecutive successes registered by the counter is x , and

(a) $i = 0$, if

(1) at the t -th outcome the counter is not locked and the outcome is an F

or

(2) at the t -th outcome the counter is locked, $x \geq 1$ and the last registration occurred at trial $t - r + 1$.

(b) $i = 1 + rj$, for $j = 0, 1, \dots, k - 2$, with $k \geq 2$, if the counter is not locked and the t -th outcome was the $(1 + j)$ -th consecutive S registered by the counter.

(c) $i = 1 + rj + s$, for $j = 0, \dots, k - 2, s = 1, \dots, r - 1$, with $k \geq 2$ and $r \geq 2$ (if $r \geq 3, i$ gets the additional values $i = 1 + (k - 1)r + s$, for $s = 1, \dots, r - 2$), if at the t -th outcome the counter is locked for exactly s consecutive trials.

(d) $i = 1 + (k - 1)r$, if the counter is not locked, $x \geq 1$ and the t -th outcome is the k -th consecutive S registered by the counter after the $(x - 1)$ -th success run of length k was completed ($Y_t = (x, i)$ and $Y_{t-1} = (x - 1, (k - 1)r)$).

Under this set-up, one may easily verify that the r.v. $N_{n,k,r}$ becomes an *MVB* with initial probability vector

$$\pi_0 = (1, 0, 0, \dots, 0)_{1 \times kr},$$

and respective matrices $A_t = A_t(x)$ and $B_t = B_t(x), x = 0, 1, \dots, \ell_n$, the ones described in the theorem.

The result follows by Lemma 2.1 and Eq. (2). □

Matrices A_t and B_t resulting from Theorem 2.1 in the special case $r = 1$, reduce to the ones given in [21].

As an illustration let us treat the special cases $k = 2, r = 2$ and $k = 2, r = 3$. In the first case the matrices A_t and B_t of the Markov chain are given by

$$A_t = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 3) \\ q_t & p_t & 0 & 0 \\ 0 & 0 & 1 & 0 \\ q_t & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad B_t = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 3) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_t \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

while in the second we get

$$A_t = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 3) & (\cdot, 4) & (\cdot, 5) \\ q_t & p_t & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ q_t & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$B_t = \begin{pmatrix} (\cdot, 0) & (\cdot, 1) & (\cdot, 2) & (\cdot, 3) & (\cdot, 4) & (\cdot, 5) \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_t & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We shall next proceed to the computation of the distribution of $N_{6,2,3}$ and $N_{6,2,2}$. For typographical convenience, let us consider the case of trials having a common success probability p when the system is not locked. In this case the structure of A_t, B_t is exactly the same as given before, with the p_t, q_t 's being replaced by p, q , respectively. Apparently, now we have

$$A_1 = A_2 = \dots = A_6 = A, \quad B_1 = B_2 = \dots = B_6 = B.$$

Furthermore, a repeated application of the recursive scheme of Lemma 2.1 yields, for $k = 2, r = 3$

$$\begin{aligned} \mathbf{f}_0(0) &= (1, 0, 0, 0, 0, 0), \mathbf{f}_1(0) = \mathbf{f}_0(0) \cdot A = (q, p, 0, 0, 0, 0), \\ \mathbf{f}_2(0) &= \mathbf{f}_1(0) \cdot A = (q^2, pq, p, 0, 0, 0), \mathbf{f}_3(0) = \mathbf{f}_2(0) \cdot A = (q^3, pq^2, pq, p, 0, 0), \\ \mathbf{f}_4(0) &= \mathbf{f}_3(0) \cdot A = (pq + q^4, pq^3, pq^2, pq, 0, 0), \\ \mathbf{f}_5(0) &= \mathbf{f}_4(0) \cdot A = (pq^2 + q(pq + q^4), p(pq + q^4), pq^3, pq^2, 0, 0), \\ \mathbf{f}_6(0) &= \mathbf{f}_5(0) \cdot A = (q^3(2p + q^2), q^2(2p + q^3), pq(p + q^3), pq^3, 0, 0) \text{ and} \\ \mathbf{f}_0(1) &= (0, 0, 0, 0, 0, 0), \mathbf{f}_1(1) = \mathbf{f}_0(1) \cdot A + \mathbf{f}_0(0) \cdot B = (0, 0, 0, 0, 0, 0), \\ \mathbf{f}_2(1) &= \mathbf{f}_1(1) \cdot A + \mathbf{f}_1(0) \cdot B = (0, 0, 0, 0, 0, 0), \\ \mathbf{f}_3(1) &= \mathbf{f}_2(1) \cdot A + \mathbf{f}_2(0) \cdot B = (0, 0, 0, 0, 0, 0), \\ \mathbf{f}_4(1) &= \mathbf{f}_3(1) \cdot A + \mathbf{f}_3(0) \cdot B = (0, 0, 0, 0, p^2, 0), \\ \mathbf{f}_5(1) &= \mathbf{f}_4(1) \cdot A + \mathbf{f}_4(0) \cdot B = (0, 0, 0, 0, p^2q, p^2), \\ \mathbf{f}_6(1) &= \mathbf{f}_5(1) \cdot A + \mathbf{f}_5(0) \cdot B = (p^2, 0, 0, 0, p^2q^2, p^2q). \end{aligned}$$

Applying now Theorem 2.1 for $x = 0$ and $x = 1$ we may obtain the exact distribution of $N_{6,2,3}$ as follows

$$\begin{aligned} P(N_{6,2,3} = 0) &= \mathbf{f}_6(0) \cdot \mathbf{1}' = q^3(2p + q^2) + q^2(2p + q^3) + pq(p + q^3) + pq^3, \\ P(N_{6,2,3} = 1) &= \mathbf{f}_6(1) \cdot \mathbf{1}' = p^2(1 + q + q^2). \end{aligned}$$

Following exactly the same procedure for $k = 2, r = 2$ we may easily derive the exact distribution of $N_{6,2,2}$ as follows

$$\begin{aligned} P(N_{6,2,2} = 0) &= 1 - (p^3 + 3p^3q + 3p^2q^2 + p^2q^3), \\ P(N_{6,2,2} = 1) &= p^3 + 3p^3q + 3p^2q^2 + p^2q^3. \end{aligned}$$

It should be noted that, the two cases worked out before serve only illustration purposes; the formulas established by the suggested methodology could be easily established by taking into account the realizations of the events $N_{6,2,3} = 1$ and $N_{6,2,2} = 1$ provided in the Introduction. Thus,

$$\begin{aligned} P(N_{6,2,3} = 1) &= P(S**S**, FS**S**, FFS**S) = p^2(1 + q + q^2), \\ P(N_{6,2,3} = 0) &= 1 - P(N_{6,2,3} = 1) = 1 - p^2(1 + q + q^2) \end{aligned}$$

and

$$\begin{aligned}
 P(N_{6,2,2} = 1) &= P(S * S * S*, S * S * FS, S * S * FF, FS * S * F, FS * S * S, \\
 &\quad FFS * S*, FFFS * S, S * FS * S) \\
 &= p^3 + 3p^3q + 3p^2q^2 + p^2q^3, \\
 P(N_{6,2,2} = 0) &= 1 - P(N_{6,2,3} = 1) = 1 - (p^3 + 3p^3q + 3p^2q^2 + p^2q^3).
 \end{aligned}$$

It goes without saying that, for larger values of n it is infeasible in practice to register all realizations of the events $N_{n,k,r}$; in these cases the use of Theorem 2.1 is unavoidable.

Let us, now, denote by $\varphi_n(z)$ and $\Phi(z, w)$ the single and double generating functions of the r.v. $N_{n,k,r}$, i.e.

$$\varphi_n(z) = \sum_{x=0}^{\infty} P(N_{n,k,r} = x)z^x = \sum_{x=0}^{\infty} f_n(x)z^x, \quad \Phi(z, w) = \sum_{n=0}^{\infty} \varphi_n(z)w^n.$$

The next Proposition provides a closed expression for the double generating function and neat recursive relations for the pmf, pgf and moments of $N_{n,k,r}$.

Proposition 2.1 *If the trials have a common success probability p when the system is not locked, then the following results hold true.*

(a) *The double generating function $\Phi(z, w)$ of the r.v. $N_{n,k,r}$ ($k \geq 1, r \geq 1$) equals*

$$\Phi(z, w) = \frac{1 + pw(\sum_{i=0}^{k-2} p^i w^{ir})(\sum_{i=0}^{r-1} w^i) + p^k w^{(k-1)r+1} z \sum_{i=0}^{r-2} w^i}{1 - qw \sum_{i=0}^{k-1} p^i w^{ir} - p^k w^{kr} z}. \tag{3}$$

(b) *The pgf $\varphi_n(z)$ of the r.v. $N_{n,k,r}$ ($k \geq 1, r \geq 1$) satisfies the recursive scheme*

$$\varphi_n(z) = p^k z \varphi_{n-kr}(z) + q \sum_{i=0}^{k-1} p^i \varphi_{n-ir-1}(z), \quad n \geq kr + 1, \tag{4}$$

with initial conditions $\varphi_{(k-1)r+1}(z) = p^k z + 1 - p^k$, $\varphi_n(z) = p^{k-1}(1 - q^{i+1})z + 1 - p^{k-1}(1 - q^{i+1})$, $n = (k - 1)r + 1 + i$, $i = 1, \dots, r - 1$ and $\varphi_n(z) = 1$, $0 \leq n \leq (k - 1)r$.

(c) *The pmf $f_n(x)$ of the r.v. $N_{n,k,r}$ ($k \geq 1, r \geq 1$) satisfies the recursive scheme*

$$f_n(x) = p^k f_{n-kr}(x - 1) + q \sum_{i=0}^{k-1} p^i f_{n-ir-1}(x), \quad n \geq kr + 1, x \geq 0, \tag{5}$$

with initial conditions

$$\begin{aligned}
 f_n(x) &= 0, \quad \text{if } x < 0, \text{ or } x > [(n + r - 1)/(kr)], \text{ or } n < 0, \\
 f_{(k-1)r+1}(1) &= p^k, \quad f_{(k-1)r+1}(0) = 1 - p^k,
 \end{aligned}$$

$$f_n(1) = p^{k-1}(1 - q^{i+1}), \quad f_n(0) = 1 - f_n(1), \quad n = (k - 1)r + 1 + i, \quad i = 1, \dots, r - 1 \text{ and } f_n(x) = \delta_{x,0}, \quad 0 \leq n \leq (k - 1)r.$$

(d) The m -th moments $\mu_{n,m} = E[(N_{n,k,r})^m]$, $m \geq 1$, of the r.v. $N_{n,k,r}$ ($k \geq 1, r \geq 1$) satisfy the recursive scheme

$$\mu_{n,m} = p^k \sum_{i=0}^m \binom{m}{i} \mu_{n-kr,i} + q \sum_{i=0}^{k-1} p^i \mu_{n-ir-1,m}, \quad n \geq (k - 1)r + 2, \quad (6)$$

with $\mu_{n,0} = 1, \mu_{n,m} = 0$ for $n \leq (k - 1)r$ and $m \geq 1, \mu_{(k-1)r+1,m} = p^k$, for $m \geq 1$ and $\mu_{n,m} = p^{k-1}(1 - q^{i+1}), n = (k - 1)r + 1 + i, i = 1, \dots, r - 1, m \geq 1$.

Proof (a) Under the assumption that the trials have a common success probability p when the system is not locked, it is apparent that $N_{n,k,r}$ turns into a homogeneous MVB . Therefore, its double generating function can be expressed as (see [21])

$$\Phi(z, w) = \pi_0 [I - w(A + zB)]^{-1} \mathbf{1}', \quad (7)$$

where I is the identity $s \times s$ matrix and $A_t = A, B_t = B$ are the matrices from Theorem 2.1. The result follows using some algebra.

(b) Exploiting Eq. (3) we get

$$\begin{aligned} & \left(1 - qw \sum_{i=0}^{k-1} p^i w^{ir} - p^k w^{kr} z \right) \sum_{n=0}^{\infty} \varphi_n(z) w^n \\ &= 1 + pw \left(\sum_{i=0}^{k-2} p^i w^{ir} \right) \left(\sum_{i=0}^{r-1} w^i \right) + p^k w^{(k-1)r+1} z \sum_{i=0}^{r-2} w^i, \end{aligned} \quad (8)$$

and the result follows by comparing the coefficients of w^n in both sides of (8).

(c) It suffices to replace $\varphi_n(z), n \geq 0$, in (4), by the power series $\varphi_n(z) = \sum_{x=0}^{\infty} f_n(x) z^x$, and then consider the coefficients of z^x in both sides of the resulting identity.

(d) The moment generating function $M(z)$ of $N_{n,k,r}$ can be expressed as $E(\exp(zN_{n,k,r})) = \varphi_n(e^z)$. Accordingly, replacing z by e^z in (4), we may easily derive a recursive scheme for the moment generating function of $N_{n,k,r}$. The desired result follows by taking the m -th order derivative with respect to z on both sides of this recursive scheme and using the well known identity

$$\frac{d^m}{dz^m} \left(e^{wz} M(z) \right) \Big|_{z=0} = \sum_{i=0}^m \binom{m}{i} w^{m-i} \mu_{n,i}.$$

In Proposition 2.1, for $k = 1$ or $r = 1$, the convention $\sum_{i=0}^{-1} = 1$ was used. □

It is worth mentioning that one may establish an alternative proof of (c) by conditioning on the number of S 's appearing before the first occurrence of an F in the sequence of n trials.

As far as part (d) of Proposition 2.1 is concerned, it provides an effective recurrence scheme for computing the moments of $N_{n,k,r}$ up to a desired order for all $n = 1, 2, \dots$. If one is interested in the evaluation only of the means $\mu_{n,1} = E(N_{n,k,r})$, he/she might use the next matrix-based expression (see e.g. [21])

$$\mu_{n,1} = \pi_0 \left[\sum_{i=1}^n (A+B)^{i-1} \right] B \mathbf{1}' = \pi_0 \frac{I - (A+B)^n}{I - (A+B)} B \mathbf{1}'$$

or the respective expression for the generating function of $\mu_{n,1}$, $n = 1, 2, \dots$, namely

$$\sum_{n=1}^{\infty} \mu_{n,1} w^n = \frac{w}{1-w} \pi_0 \left[I - w(A+B) \right]^{-1} B \mathbf{1}'.$$

As it was mentioned in the Introduction Sen et al. [27] examined the distribution of the r.v. $N_{n,k,r}$ under the Pólya–Eggenberger sampling scheme with parameters a , b and s . Setting $\frac{a}{a+b} = p$ and $s = 0$ (sampling with replacement) an expression for the pmf of $N_{n,k,r}$ in the special case that the trials have a common success probability p when the counter is not locked, containing multiple sums involving binomial coefficients, can be deduced.

To our knowledge, the recursive formulae of Proposition 2.1 have not appeared in the literature before. In addition, several published results can be derived as special cases of it. For $r = 1$, (5) reduces to (5.4) of Balakrishnan and Koutras [3] while, for $k = 1$, we may obtain a formula that relates to the closed formula for the cumulative distribution function of $N_{n,1,r}$ derived by Dandekar [10]. For $k = 1$ and $r = 1$, (5) reduces to a well-known recurrence satisfied by the pmf of the binomial distribution.

In the next section we present an application of the distribution proposed in the current work.

3 An Application

The analysis of the long-term fluctuation of Peak Expiratory Flow (PEF) and Forced Expiratory Volume at 1 second (FEV1) has been successfully used at research level to identify asthmatic patients at high risk and for the prognosis of imminent seizures (see e.g. [28]). In practice, however, the daily measurement of these parameters and recording of their prices in special diaries, has proved to be a very complicated and time-consuming process [20]. Recent developments in biosensor technology have made feasible the development of small-scale spirometries that interconnect with telematic systems and provide distant measurements of PEF and FEV1 and real-time assessment of results [1]. This capability makes it possible, appropriate statistical tools to be utilized towards the development of a system for forecasting of asthma exacerbations in children and adolescents, through the analysis of real-time variability of

PEF and FEV1. The system may, subsequently, detect changes in the daily fluctuation pattern of the above spirometric parameters and: (a) automatically notify, both the treating physician and the patient himself, about the apparent loss of control and the probability of an imminent seizure of the disease (b) empower optimization decisions on the type and dosage of the appropriate medication to be used.

In the present paper, we suggest that the $(r - 1)$ -modified binomial distribution of order k can be used to calibrate the aforementioned system. We shall focus on one of the spirometric parameters (or a weighted average of all of them) and study the empirical values collected regularly, say every 1 min. We will denote by 0 and 1 the occurrence of a value in and out of a prespecified comfort zone (CZ), respectively. The occurrence of a 0 is a sign of a stabilized medical condition and the next value of the spirometric parameter will be generated in 1 minute. On the other hand, the occurrence of an 1 reveals a non-stabilized medical condition and the patient should be given some time, say $r - 1$ minutes, before an additional value is taken into account. The occurrence of k consecutive 1's is a sign of a stabilized bad medical condition. Thus, the distribution of the r.v. $N_{n,k,r}$ may provide significant information regarding patient's progress and facilitate the establishment of valuable decision criteria for selecting the type and dosage of medication. What the values of k and r should be depended on the clinical evidence and the level of risk one is willing to accept.

To substantiate the last declaration let us assume that one of the spirometric parameters, e.g. PEF, is monitored by recording it at 1 minute intervals for 1 hour. Apparently, the collected data for a specific subject (monitored patient) can be transformed to a sequence of $n = 60$ binary trials 0 – 1 by labelling as 1 (success) an observed PEF lying outside the CZ and 0 (failure) otherwise. Then, extremely large values of $N_{n,k,r}$ provide evidence of a stabilized critical medical condition, so it seems plausible to assign that condition to the subject under study if $N_{n,k,r} > c$ where k , r and n are design parameters of our decision process. Making use of the distribution of PEF for patients that according to past knowledge are not in critical medical condition, we may calculate the probability p_0 that such a patient produces a PEF within the CZ. Then the choice of the design parameters could be based on the condition

$$P(N_{n,k,r} > c) \leq a, \text{ for all } p \leq g(p_0), \quad (9)$$

where a is the (maximum) acceptance risk of assigning a critical medical condition to a patient that is not in such a condition and $g(\cdot)$ denotes a non-decreasing function. The last quantity is determined by the practitioner accordingly to past experience; for simplicity, we assume that $g(x) = x$, however the approach taken in the sequel can be easily adapted to the general case.

Condition (9) can be expressed as

$$F_{n,k,r}(c; p) \geq 1 - a, \text{ for all } p \leq p_0, \quad (10)$$

where $F_{n,k,r}(x; p) = P(N_{n,k,r} \leq x)$ denotes the cumulative distribution function of $N_{n,k,r}$ when the success probability of the binary sequence equals p . Since the last

Table 1 Cumulative distribution function of $N_{60,k,4}$

x	$F_{60,k,4}(x; 0.6) = P(N_{60,k,4} \leq x)$			
	$k = 2$	$k = 3$	$k = 4$	$k = 5$
0	0.0001	0.0356	0.2286	0.4896
1	0.0025	0.2337	0.6886	0.9146
2	0.0251	0.6150	0.9575	0.9982
3	0.1312	0.9160	1	1
4	0.4005	0.9964		
5	0.7535	1		
6	0.9630			
7	1			
$E(N_{60,k,4})$	4.7237	2.2033	1.1251	0.5975
$\text{Var}(N_{60,k,4})$	1.2198	0.9535	0.6515	0.4184

quantity is a decreasing function of p , (10) is guaranteed if

$$F_{n,k,r}(c; p_0) \geq 1 - a, \tag{11}$$

a condition that could be exploited for selecting the design parameters k , r and c .

In practice, the parameter r will be provided by the practitioner, since it indicates the time period for using an additional PEF measurement after an out of CZ recording.

Since we are trying to determine two parameters (k and c), couching on a single condition (i.e. (11)), it is evident that several combinations of them could be used. An additional criterion could therefore be exploited to select one of the available alternatives; a reasonable approach along these lines might be to keep the combination of (k , c) values that minimizes the variance of the statistic $N_{n,k,r}$.

As an illustration of the aforementioned procedure let us assume that the p_0 value deduced by analysing past PEF data equals $p_0 = 0.6$ and that the value of r provided by the practitioner is $r = 4$, while the maximum risk we are willing to take equals $a = 0.05$. Then, (11) reads

$$F_{60,k,4}(c; 0.6) \geq 0.95,$$

and from Table 1 we may obtain the following acceptable pairs of (k , c):

$$(k, c) = (2, 6), (3, 4), (4, 2), (5, 2).$$

Among these choices, the optimal, in terms of the minimum variance of $N_{60,k,4}$ is $k = 5$, $c = 2$. Therefore, our decision rule takes on the form: the subject under study is considered to be in a stabilized critical condition if $N_{60,5,4} \geq 2$, i.e. if $N_{60,5,4} = 2$ or $N_{60,5,4} = 3$.

It should be noted that the variance of $N_{n,k,r}$ is a decreasing function of k ; therefore, the minimum variance criterion leads to the choice of (k , c) pair with the maximum k value, among the feasible pairs spotted out at the first step.

Closing this section, we provide, for illustration purposes, some numerical results for the distribution of the r.v. $N_{n,k,r}$. Table 2 depicts the distribution, the mean and the

Table 2 Distribution of $N_{60,3,2}$

x	$P(N_{60,3,2} = x)$			
	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$
0	0.7317	0.1455	0.0035	*
1	0.2346	0.3255	0.0301	*
2	0.0313	0.3080	0.1102	0.0006
3	0.0022	0.1602	0.2246	0.0059
4	*	0.0498	0.2789	0.0320
5	*	0.0094	0.2164	0.1107
6	*	0.0010	0.1032	0.2414
7	*	*	0.0285	0.3165
8	*	*	0.0039	0.2230
9	*	*	0.0002	0.0657
10	*	*	*	0.0037
$E(N_{60,3,2})$	0.3043	1.6764	3.9757	6.7791
$Var(N_{60,3,2})$	0.2890	1.2669	1.9486	1.55731

*Value less than 10^{-4}

Table 3 Distribution of $N_{30,k,r}$ for $p = 0.6$ and different choices of k, r

x	$P(N_{30,k,r} = x)$			
	$k = 2, r = 2$	$k = 2, r = 3$	$k = 3, r = 2$	$k = 3, r = 3$
0	0.0010	0.0046	0.0682	0.1414
1	0.0139	0.0557	0.2737	0.4353
2	0.0751	0.2350	0.3865	0.3600
3	0.2100	0.4151	0.2243	0.0631
4	0.3227	0.2567	0.0456	
5	0.2642	0.0326	0.0014	
6	0.1007			
7	0.0120			
$E(N_{30,k,r})$	4.0956	2.9615	1.9095	1.3450
$Var(N_{30,k,r})$	1.4194	0.8856	0.9582	0.6352

*Value less than 10^{-4}

variance of $N_{60,3,2}$ for $p = 0.2, 0.4, 0.6$ and 0.8 . Table 3 shows the distribution of the r.v. $N_{30,k,r}$ for $p = 0.6$ and a variety of choices of the parameters k and r depending on the desirable level of acceptable risk. Numerics in all tables have been rounded down to 4 decimal points.

Figure 1 depicts the pmf of $N_{60,3,2}$, for $p = 0.2, 0.4, 0.6$ and 0.8 .

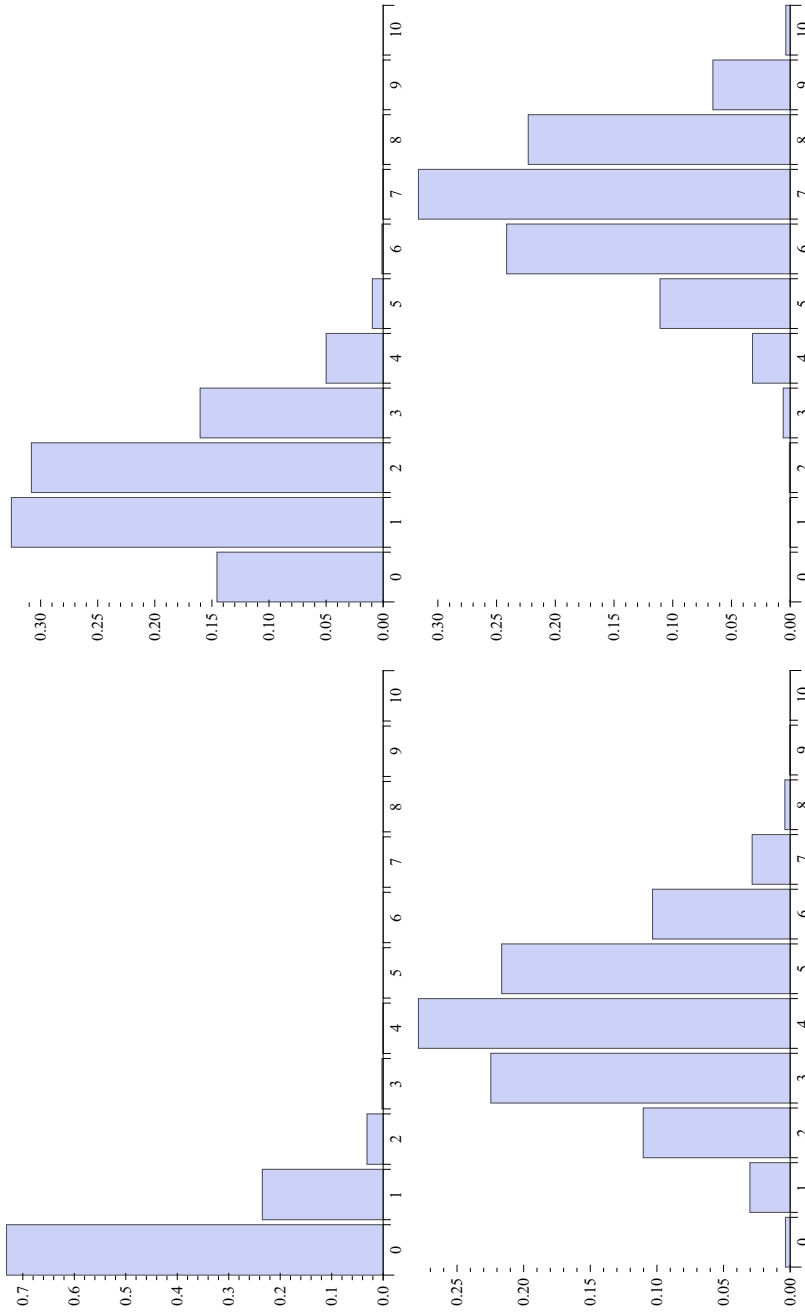


Fig. 1 The pmf of $N_{60,3,2}$ for $p = 0.2$, $p = 0.6$ (left) and $p = 0.4$, $p = 0.8$ (right)

4 Conclusion and Future Work

In the present paper, we studied the $(r - 1)$ -modified binomial distribution of order k and derived neat recursive relations for the probability mass function, probability generating function and moments. We also studied the general case where the original trials are independent but not necessarily identically distributed. We illustrated how our new results can be applied in biomedical engineering.

Regarding the development of the system for forecasting of asthma exacerbations, it is apparent that the existence of the two design parameters k and r offers flexibility for setting up different decision criteria, which can be adapted to the desirable level of risk and additional medical characteristics of the clinical case under study. The suitable choice of the parameters in any case will be explored using real data and parametric tests. The fitting of the distributions to real-life applications may point out that a classification of the empirical data in more than two categories is beneficial, and can provide a more robust stochastic model. If this is the case, our theoretical results will have to be enriched so that the case of multi-state trials can be treated.

Acknowledgements The authors wish to thank Dr. Aris Dermizakis, researcher at Department of Medicine, University of Patras, for his valuable help regarding the proposed application of the current work in Biomedical Engineering. The authors wish to thank the referees for their comments and suggestions which helped to improve the article.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aliverti A (2017) Wearable technology: role in respiratory health and disease. *Breathe* 13:e27–e36
2. Antzoulakos DL, Bersimis S, Koutras MV (2003) On the distribution of the total number of run lengths. *Ann Inst Stat Math* 55:865–884
3. Balakrishnan N, Koutras MV (2002) *Runs and scans with applications*. Wiley, New York
4. Balakrishnan N, Koutras MV, Milienos FS (2014) Start-up demonstration tests: models, methods and applications, with some unifications. *Appl Stoch Model Bus Ind* 30:373–413
5. Balakrishnan N, Koutras M, Milienos F (2021) Reliability analysis and plans for successive testing: start-up demonstration tests and applications. Academic Press, Cambridge
6. Bersimis S, Koutras MV, Papadopoulos GK (2014) Waiting time for an almost perfect run and applications in statistical process control. *Methodol Comput Appl Probab* 16:207–222
7. Chadjiconstantinidis S, Antzoulakos DL, Koutras MV (2000) Joint distributions of successes, failures and patterns in enumeration problems. *Adv Appl Probab* 32:866–884
8. Dafnis SD, Makri FS (2021) Weak runs in sequences of binary trials. *Metrika*. <https://doi.org/10.1007/s00184-021-00842-1>
9. Dafnis SD, Makri FS, Koutras MV (2021) Generalizations of runs and patterns distributions for sequences of binary trials. *Methodol Comput Appl Probab* 23:165–185
10. Dandekar, V. M. (1955). Certain modified forms of binomial and Poisson distributions. *Sankhyā Indian J Stat* 15:237–250
11. De Moivre A (1738) *The doctrine of chance*, Third. Chelsea Publishing Co, New York
12. Eryilmaz S, Demir S (2007) Run statistics in a sequence of arbitrarily dependent binary trials. *Stat Pap* 45:1007–1023
13. Feller W (1968) *An introduction to probability theory and its applications*, vol 1. Wiley, New York

14. Fu JC, Koutras MV (1994) Distribution theory of runs: a Markov chain approach. *J Am Stat Assoc* 89:1050–1058
15. Fu JC, Lou WYW (2003) Distribution theory of runs and patterns and its applications: a finite Markov imbedding approach. World Scientific Publishing, New Jersey
16. Godbole AP (1990) Specific formulae for some success run distributions. *Stat Probab Lett* 10:119–124
17. Han Q, Aki S (1999) Joint distributions of runs in a sequence of multi-state trials. *Ann Inst Stat Math* 51:419–447
18. Hirano K (1986) Some properties of the distributions of order k . In: Philippou AN et al (eds) *Fibonacci numbers and their applications*. Reidel, Dordrecht, pp 43–53
19. Johnson NL, Kotz S, Kemp AW (1992) *Univariate discrete distributions*, 2nd edn. Wiley, New York
20. Kamps AW, Roorda RJ, Brand PL (2001) Peak flow diaries in childhood asthma are unreliable. *Thorax* 56:180–182
21. Koutras MV, Alexandrou VA (1995) Runs, scans and urn model distributions: a unified Markov chain approach. *Ann Inst Stat Math* 47:743–766
22. Koutras MV, Bersimis S, Antzoulakos DL (2006) Improving the performance of the chi-square control chart via runs rules. *Methodol Comput Appl Probab* 8:409–426
23. Koutras MV, Eryilmaz S (2017) Compound geometric distribution of order k . *Methodol Comput Appl Probab* 19:377–393
24. Kumar CS (2009) A class of discrete distributions of order k . *J Stat Theory Pract* 3(4):795–803. <https://doi.org/10.1080/15598608.2009.10411960>
25. Philippou AN, Georghiou C, Philippou GN (1983) A generalized geometric distribution and some of its properties. *Statist Probab Lett* 1:171–175
26. Philippou AN, Makri FS (1986) Successes, runs and longest runs. *Statist Probab Lett* 4:101–105
27. Sen K, Agarwal ML, Bhattacharya S (2015) Geiger counter-type Pólya–Eggenberger distributions. *Commun Stat Theory Methods* 44:4912–4926
28. Thamrin C, Stern G, Strippoli MP, Kuehni CE, Suki B, Taylor DR, Frey U (2009) Fluctuation analysis of lung function as a predictor of long-term response to beta2-agonists. *Eur Respir J* 33:486–493
29. Todhunter I (1865) *A history of the mathematical theory of probability*, Cambridge, Macmillan, reprinted 1965. Chelsea, New York
30. Yalcin F, Eryilmaz S (2014) q -geometric and q -binomial distributions of order k . *J Comput Appl Math* 71:31–38

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