KYUNGPOOK Math. J. 64(2024), 287-301
https://doi.org/10.5666/KMJ.2024.64.2.287
pISSN 1225-6951 eISSN 0454-8124


# Distribution of Runs and Patterns in Four State Trials

Jungtaek Oh
*Department of Biomedical Science, School of Medicine, Kyungpook National University, Daegue, 41566, Republic of Korea*
*Clinical Omics Center, School of Medicine, Kyungpook National University, Daegue, 41566, Republic of Korea*
*The Institute of Industrial Technology, Changwon National University, Changwon, 51140, Republic of Korea*
*e-mail* : jungtaekoh0191@gmail.com *and* mathguide@nate.com

Abstract. From the mathematical and statistical point of view, a segment of a DNA strand can be viewed as a sequence of four-state (A, C, G, T) trials. Herein, we consider the distributions of runs and patterns related to the run lengths of multi-state sequences, especially for four states (A, B, C, D). Let $X_1, X_2, \ldots$ be a sequence of four state independent and identically distributed trials taking values in the set $\mathscr{S} = \{A,\ B,\ C,\ D\}$. In this study, we obtain exact formulas for the probability distribution function for the discrete distribution of runs of B's of order $k$. We obtain longest run statistics, shortest run statistics, and determine the distributions of waiting times and run lengths.

## 1. Introduction

Runs and run related statistics have been extensively studied in literature due to their wide range of applications in various areas including statistics (e.g., hypothesis testing), engineering (e.g., system reliability, health services monitoring, and quality control), molecular biology and bioinformatics (e.g., population genetics, and DNA sequence homology), physics, psychology, radar astronomy, computer science

Received January 17, 2024; accepted March 11, 2024.
2020 Mathematics Subject Classification: 62E15, 60C05, 62E15.
Key words and phrases: runs and patterns, multi-state trials, discrete distribution of order $k$, waiting time distribution, distribution of run length, longest run, shortest run, DNA sequence.
This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2019R1A6A3A01090663, and NRF-2020R1I1A1A01057479) and partially supported under the framework of international cooperation program managed by the National Research Foundation of Korea(NRF-2022K2A9A1A01098016).

(e.g., encoding/decoding and transmission of digital information), and finance (e.g., financial engineering, risk analysis, and prediction). The significant progress made in runs and related statistics during the previous few decades was been nicely surveyed in [6] as well as in [14] and the references therein. More recent contributions are exemplified by such papers as [5], [12], [17], [20], and [2].

There are two main types of problems concerning runs and related statistics.

(1) The number of trials until the first occurrence of a run or pattern (or until the $r$-th occurrence of a run or pattern), which is called a waiting time problem.

(2) The number of occurrences of a run or pattern until the $n$-th trial.

Waiting time distributions have attracted a lot of interest in applied probability. Consequently, the properties of waiting time distributions have been extensively studied [10], [1], [7], [18], [4], [3], [16], [15].

The geometric distribution of order $k$ is one of the best known waiting time distributions. It is defined as the distribution of the number of trials until obtaining the first consecutive $k$ successes. This definition is due to [24]. It is clear that the geometric distribution of order $k$ reduces to the classical geometric distribution for the case $k = 1$ with probability mass function (PMF) $f(x) = q^{x-1}p$ $(x \geq 1)$ (the number of Bernoulli trials needed to get one success).

Let $X_1, X_2, \ldots$ be a sequence of four state independent and identically distributed (IID) trials taking values from a set $\mathscr{S} = \{A,\ B,\ C,\ D\}$ of four symbols. For the number of trials $W_1^k$ until the first consecutive $k$ successes we have the following equivalent definitions of $W_1^k$.

$$\begin{aligned} W_1^k &= \min\{n : X_{n-k+1} = \cdots = X_n = 1\} \\ &= \min\left\{n : \prod_{n-k+1}^{n} X_j = 1\right\} \\ &= \min\left\{n : \sum_{n-k+1}^{n} X_j = k\right\}. \end{aligned}$$

Various statistical properties of the geometric distribution of order $k$ have been applied to multiple areas, ranging from quality control to reliability. In particular, the consecutive-$k$-out-of-$n$:$F$ system (refer to [9], [8], [19], [11], [27]) is very closely related to this distribution. Another closely related random variable is the length $L_n$ of the longest success run in $n$ four state trials, as shown above. Because $W_1^k \leq n$ if and only if $L_n \geq k$, there is a relationship between the probability distribution functions of $W_1^k$ and $L_n$ given by the identity $\mathbf{P}(W_1^k \leq n) = \mathbf{P}(L_n \geq k)$.

Let $X_1, X_2, \ldots, X_n$ be independent random variables distributed identically as $W_1^k$. The distribution of the number $W_r^k$ of trials until the $r$-th appearance of a success run of length $k$ is a negative binomial distribution of order $k$. This follows from the fact that it is the distribution of the $r$-fold convolution of the geometric

distribution of order $k$, namely

$$W_r^k = \sum_{i=1}^{r} X_i.$$

Some main aspects of occurrences of runs and patterns are: where, when and how many times they occur. To study these, one defines statistics that count runs and patterns according to various enumerating schemes. A classical counting scheme for enumerating runs of fixed length was presented in[13]. Once $k$ consecutive successes are observed, the number of occurrences of $k$ consecutive successes increases and counting procedure starts anew. This is referred to as a *non-overlapping* counting scheme. It follows a binomial distribution of order $k$. In another counting scheme, we count only the number $E_n^k$ of the success runs of a length exactly $k$, preceded and succeeded either by failures or by nothing ([21]).

In a sequence of $n$ four state trials, we can define other statistics via runs and patterns, such as the longest (maximal) run length and shortest (minimal) run length (refer to [26]). It is obvious that the longest run and shortest run are, respectively, upper and lower bounds of the number of consecutive successes that appears in a sequence of $n$ four state trials. These distributions can be applied in DNA type sequence comparison by observing the frequencies of the longest runs of matches or mismatches.

Sequences of categorical outcomes arise frequently in biomedical research, representing, for example, deoxyribonucleic acid (DNA) strand segments or findings of health care evaluations. They are typically analyzed by defining problem-specific statistics involving runs and patterns. A molecule of deoxyribonucleic acid is a chain or sequence of pairs of nucleotides with the four base structures adenine, cytosine, guanine, and thymine, or A, C, G, and T. The occurrence of a specified sequence of nucleotides in some portion of the chain is the event that the specified run of A's, C's, G's, and T's occurs. Mathematically, a random DNA strand segment can be viewed as a sequence of four-state (A, C, G, T) trials. We consider some distributions of runs and patterns related to run lengths for multi-state, especially four state trials.

For a sequence $X_1, X_2, \ldots, X_n$ of IID trials with values taken from a set of symbols $\mathscr{S} = \{A,\ B,\ C,\ D\}$ with probabilities $P(A) = P_a$, $P(B) = P_b$, $P(C) = P_c$ and $P(D) = P_d$, such that $P_a + P_b + P_c + P_d = 1$, we consider the following stochastic variables.

- Let $N_n^k$ denotes the number of the non-overlapping runs of B's of order $k$ in $n$ independent trials,
- Let $E_n^k$ denotes the number of the runs of B's of exact length $k$, preceded and succeeded by A or C or D or nothing,
- Let $W_1^k$ denotes the waiting time for the first occurrence of a run of B's of length $k$,

- Let $W_r^k$ denotes the waiting time for the $r$-th occurrence of run of B's of length $k$,
- Let $L_n$ denotes the maximum length of a run of B's (longest run statistics),
- Let $M_n$ denotes the minimum length of a run of B's (shortest run statistics),
- Let $NL_n$ denotes the number of times a non-overlapping run of length $L_n$ appears,
- Let $NM_n$ denotes the number of times a non-overlapping run of length $M_n$ appears.

To illustrate the above mentioned quantities, we consider the runs of B's in the following example for $n = 40$.

$$DAAABBBBBCAABBBCCADBBBBCCDABBBBBBCACCDBB,$$

for which one can check $N_{40}^2 = 9$, $N_{40}^3 = 5$, $L_{40} = 6$, $M_{40} = 2$, $E_{40}^2 = 1$, $E_{40}^3 = 1$, $E_{40}^4 = 1$, $E_{40}^5 = 1$, $E_{40}^6 = 1$, $NL_{40} = 1$ and $NM_{40} = 1$.

In this study, we are obtain probability distribution functions for runs and patterns in four state IID trials in terms of multinomial coefficients. The exact PMFs are derived via combinatorial analysis. In Section 2, we obtain the PMF of a binomial distribution of order $k$ and distribution of runs of a length exactly $k$. In Section 3, we obtain PMFs of a geometric distribution of order $k$ and a negative binomial distribution of order $k$. In Section 4, we obtain the PMFs for longest and shortest runs.

## 2. Discrete Dstribution of Order $k$

We consider a sequence $X_1, X_2, \ldots, X_n$ of multistate trials defined on the state space $\mathscr{S} = \{A,\ B,\ C,\ D\}$ with probabilities $P(A) = P_a$, $P(B) = P_b$, $P(C) = P_c$ and $P(D) = P_d$, such that $P_a + P_b + P_c + P_d = 1$. In this section, we obtain the PMFs of a binomial distribution of order $k$, and the distribution of runs of length exactly $k$ using combinatorial analysis.

### 2.1. Binomial Distribution of Order $k$

First, we consider the number $N_n^k$ of runs of B's of length $k$ in $n$ IID four state trials. We establish the PMF of the random variable $N_n^k$ using combinatorial analysis.

**Theorem 2.1.** *The* PMF *of* $N_n^k$, *for* $0 \le x \le \left[\frac{n}{k}\right]$, *is given by*

$$P(N_n^k = x) = P_b^n \sum_{i=0}^{k-1} \sum_{\star i} \binom{\sum_{t=1}^k (x_t + y_t + z_t) + x}{x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k, x} \left(\frac{P_a}{P_b}\right)^{\sum_{t=1}^k x_t} \times \left(\frac{P_c}{P_b}\right)^{\sum_{t=1}^k y_t} \left(\frac{P_d}{P_b}\right)^{\sum_{t=1}^k z_t},$$

*where the inner summation* $\sum_{\star}$ *is over all nonnegative integers* $x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k$ *for which* $\sum_{t=1}^{k} t(x_t + y_t + z_t) + kx + i = n$, *for* $i = 0, 1, \ldots, k-1$.

*Proof.* Let $\underbrace{B \cdots B}_{t-1} A = O_t$, $\underbrace{B \cdots B}_{t-1} C = J_t$, and $\underbrace{B \cdots B}_{t-1} D = T_t$, where $1 \le t \le k$. A typical element of the event $\left\{N_n^k = x\right\}$ is a sequence

$$\cdots \boxed{R_t} \cdots \overbrace{\boxed{\underbrace{B \ldots B}_{k}}}^{1} \cdots \boxed{R_t} \cdots \overbrace{\boxed{\underbrace{B \ldots B}_{k}}}^{2} \cdots \boxed{R_t} \cdots \overbrace{\boxed{\underbrace{B \ldots B}_{k}}}^{x} \cdots \boxed{R_t} \cdots \underbrace{B \ldots B}_{i},$$

where $i \in \{0, \ldots, k-1\}$, and $\boxed{R_t}$ represents any combination of the strings $O_t, J_t$, and $T_t$ appearing altogether $x_t$, $y_t$, and $z_t$ times, respectively, in the sequence, satisfying

$$\sum_{t=1}^{k} t(x_t + y_t + z_t) + kx + i = n.$$

The number of different ways of arranging this sequence equals

$$\binom{\sum_{t=1}^{k} x_t + \sum_{t=1}^{k} y_t + \sum_{t=1}^{k} z_t + x}{x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k, x}.$$

Because of the independence of the trials, the probability of the above sequence is

$$P_a^{\sum_{t=1}^{k} x_t} P_b^{\sum_{t=1}^{k} (t-1)(x_t+y_t+z_t)+kx} P_c^{\sum_{t=1}^{k} y_t} P_d^{\sum_{t=1}^{k} z_t}.$$

The probability of the run of B's of length $i$ at the end of each of these sequences is $P_b^i$ $(0 \le i < k)$, which leads to the overall probability

$$P_a^{\sum_{t=1}^{k} x_t} P_b^{\sum_{t=1}^{k} (t-1)(x_t+y_t+z_t)+kx+i} P_c^{\sum_{t=1}^{k} y_t} P_d^{\sum_{t=1}^{k} z_t}.$$

Summing over $i = 0, 1, \ldots, k-1$ the result follows. □

**Remark 1.** For $P_a = q$, $P_b = p$ such that $p + q = 1$, $P_c = P_d = 0$ and $y_1, \ldots, y_k, z_1, \ldots, z_k = 0$, Theorem 2.1 reduces to Theorem 2.1 of [25].

### 2.2. Distributions of Runs of Length Exactly $k$

In this subsection we consider the number $E_n^k$ of runs of B's of length $k$ in $n$ IID four state trials. We establish the PMF of the random variable $E_n^k$ using combinatorial analysis.

**Theorem 2.2.** *The PMF of $E_n^k$, for $0 \le x \le \left[\frac{n+1}{k+1}\right]$, is given by*

$$P(E_n^k = x) =$$
$$P_b^n \sum_{i=0}^{n-x(k+1)} \sum_{\star(k,i)} \left[ \binom{\sum_{t=1}^{n-x(k+1)} x_t + \sum_{t=1}^{n-x(k+1)} y_t + \sum_{t=1}^{n-x(k+1)} z_t}{x_1, \ldots, x_{n-x(k+1)}, y_1, \ldots, y_{n-x(k+1)}, z_1, \ldots, z_{n-x(k+1)}} \right.$$
$$\left. \times \left(\frac{P_a}{P_b}\right)^{\sum_{t=1}^{n-x(k+1)} x_t} \left(\frac{P_c}{P_b}\right)^{\sum_{t=1}^{n-x(k+1)} y_t} \left(\frac{P_b}{P_d}\right)^{\sum_{t=1}^{n-x(k+1)} z_t} \right],$$

*where the inner summation $\sum_{\star(k,i)}$ is over all nonnegative integers $x_1$, …, $x_{n-x(k+1)}$, $y_1$, …, $y_{n-x(k+1)}$, $z_1$, …, $z_{n-x(k+1)}$ for which $\sum_{t=1}^{n-x(k+1)} t(x_t+y_t+z_t) = n-i$ and*

$$x_{k+1} + y_{k+1} + z_{k+1} = \begin{cases} x & \text{if } i \ne k, \\ x-1 & \text{if } i = k. \end{cases}$$

*Proof.* Let $\underbrace{B \cdots B}_{t-1} A = O_t$, $\underbrace{B \cdots B}_{t-1} C = J_t$, and $\underbrace{B \cdots B}_{t-1} D = T_t$, for $t = 1, 2, \ldots, n - x(k+1)$. A typical element of the event $\left\{E_n^k = x\right\}$ is a sequence

$$\cdots \boxed{R_t} \cdots \overbrace{\boxed{O_{k+1} \text{ or } J_{k+1} \text{ or } T_{k+1}}}^{1} \cdots \boxed{R_t} \cdots \overbrace{\boxed{O_{k+1} \text{ or } J_{k+1} \text{ or } T_{k+1}}}^{x} \cdots \boxed{R_t} \cdots \underbrace{B \ldots B}_{i},$$

where $0 \le i \le n - x(k+1)$, and $\boxed{R_t}$ represents any string $O_t$, $J_t$, and $T_t$ appearing $x_t$, $y_t$, and $z_t$ times, respectively, in the sequence, satisfying

$$\sum_{t=1}^{n-x(k+1)} t(x_t + y_t + z_t) = n - i,$$

subject to the condition $x_{k+1} + y_{k+1} + z_{k+1} = \begin{cases} x & \text{if } i \ne k, \\ x-1 & \text{if } i = k. \end{cases}$ The number of different ways of arranging the sequence equals

$$\binom{x_1 + \cdots + x_{n-x(k+1)} + y_1 + \cdots + y_{n-x(k+1)} + z_1 + \cdots + z_{n-x(k+1)}}{x_1, \ldots, x_{n-x(k+1)}, y_1, \ldots, y_{n-x(k+1)}, z_1, \ldots, z_{n-x(k+1)}}.$$

Because of the independence of the trials, the sequence has probability

$$P_a^{x_1+\cdots+x_{n-x(k+1)}} P_c^{y_1+\cdots+y_{n-x(k+1)}} P_d^{z_1+\cdots+z_{n-x(k+1)}}$$
$$\times P_b^{(x_2+y_2+z_2)+2(x_3+y_3+z_3)+\cdots+\{n-x(k+1)-1\}(x_{n-x(k+1)}+y_{n-x(k+1)}+z_{n-x(k+1)})}.$$

The probability of the run of B's of length $i \in \{0, 1, \ldots, n - x(k+1)\}$ at the end of each of these sequences is $P_b^i$, which leads to the overall probability

$$P_a^{x_1+\cdots+x_{n-x(k+1)}} P_c^{y_1+\cdots+y_{n-x(k+1)}} P_d^{z_1+\cdots+z_{n-x(k+1)}}$$
$$\times P_b^{(x_2+y_2+z_2)+2(x_3+y_3+z_3)+\cdots+\{n-x(k+1)-1\}(x_{n-x(k+1)}+y_{n-x(k+1)}+z_{n-x(k+1)})+i}.$$

Summing over $i = 0, 1, \ldots, n - x(k+1)$, the result follows. □

## 3. Waiting Time Distributions

We consider a sequence $X_1, X_2, \ldots$ of multistate trials defined on the state space $\mathscr{S} = \{A,\ B,\ C,\ D\}$ with probabilities $P(A) = P_a$, $P(B) = P_b$, $P(C) = P_c$, and $P(D) = P_d$, such that $P_a + P_b + P_c + P_d = 1$. In this section, we obtain PMFs for the geometric distribution of order $k$ and the negative binomial distribution of order $k$, by employing combinatorial analysis.

### 3.1. Geometric Distribution of Order $k$

First, we consider the waiting time $W_1^k$ for the first occurrence of a run of B's of length $k$. We establish the PMF of the random variable $W_1^k$ using combinatorial analysis.

**Theorem 3.1.** *The PMF of $W_1^k$ is given by*

$$P(W_1^k = n) =$$
$$P_b^n \sum_{\star} \sum_{t=1}^{k} \binom{(x_t + y_t + z_t)}{x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k} \left(\frac{P_a}{P_b}\right)^{\sum_{t=1}^k x_t} \left(\frac{P_c}{P_b}\right)^{\sum_{t=1}^k y_t} \left(\frac{P_d}{P_b}\right)^{\sum_{t=1}^k z_t},$$

*where the outer summation $\sum_\star$ is over all nonnegative integers $x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k$ for which $\sum_{t=1}^k t(x_t + y_t + z_t) = n - k$.*

*Proof.* Let $\underbrace{B \cdots B}_{t-1} A = O_t$, $\underbrace{B \cdots B}_{t-1} C = J_t$, and $\underbrace{B \cdots B}_{t-1} D = T_t$, $t = 1, \ldots, k$. A typical element of the event $\left\{W_1^k = x\right\}$ is a sequence

$$\cdots\cdots\cdots \boxed{R_t} \cdots\cdots\cdots \boxed{\underbrace{B \ldots B}_{k}},$$

where $\boxed{R_t}$ represents any string $O_t$, $J_t$, and $T_t$ appearing $x_t$, $y_t$, and $z_t$ times, respectively, in the sequence, satisfying $\sum_{t=1}^k t(x_t + y_t + z_t) = n - k$. The number of different ways of arranging the sequence equals

$$\binom{\sum_{t=1}^k (x_t + y_t + z_t)}{x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k}.$$

Because of the independence of the trials, the probability of the sequence is

$$P_a^{\sum_{t=1}^k x_t} P_b^{\sum_{t=1}^k (t-1)(x_t+y_t+z_t)} P_c^{\sum_{t=1}^k y_t} P_d^{\sum_{t=1}^k z_t}.$$

The probability of the run of B's of length $k$ at the end of each these sequences is $P_b^k$, which leads to the overall probability

$$P_a^{\sum_{t=1}^k x_t} P_b^{\sum_{t=1}^k (t-1)(x_t+y_t+z_t)+k} P_c^{\sum_{t=1}^k y_t} P_d^{\sum_{t=1}^k z_t}.$$

□

**Remark 2.** For $P_a = q$, $P_b = p$ such that $p + q = 1$, $P_c = P_d = 0$ and $y_1, \ldots, y_k, z_1, \ldots, z_k = 0$, Theorem 3.1 reduces to Theorem 3.1 of [23].

### 3.2. Negative Binomial Distribution of Order $k$

Let $W_r^k$ be the random variable denoting the waiting time for the $r$-th occurrence of a run of B's of length $k$. We establish the PMF of random variables $W_r^k$ using combinatorial analysis.

**Theorem 3.2.** *The PMF of $W_r^k$ in $n$ four state IID trials is given by*

$$P(W_r^k = n) = P_b^n \sum_{\star} \sum_{t=1}^{k} \binom{(x_t + y_t + z_t) + r - 1}{x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k, r-1} \times \left(\frac{P_a}{P_b}\right)^{\sum_{t=1}^k x_t} \left(\frac{P_c}{P_b}\right)^{\sum_{t=1}^k y_t} \left(\frac{P_d}{P_b}\right)^{\sum_{t=1}^k z_t},$$

*where the outer summation $\sum_\star$ is over all nonnegative integers $x_1, \ldots, x_k$, $y_1, \ldots, y_k$, $z_1, \ldots, z_k$ for which $\sum_{t=1}^k t(x_t + y_t + z_t) + kr = n$.*

*Proof.* Let $\underbrace{B \cdots B}_{t-1} A = O_t$, $\underbrace{B \cdots B}_{t-1} C = J_t$, and $\underbrace{B \cdots B}_{t-1} D = T_t$, $t = 1, \ldots, k$. A typical element of the event $\left\{W_r^k = x\right\}$ is a sequence

$$\cdots \boxed{R_t} \cdots \overbrace{\boxed{\underbrace{B \ldots B}_{k}}}^{1} \cdots \boxed{R_t} \cdots \overbrace{\boxed{\underbrace{B \ldots B}_{k}}}^{2} \cdots \boxed{R_t} \cdots \overbrace{\boxed{\underbrace{B \ldots B}_{k}}}^{r-1} \cdots \boxed{R_t} \cdots \overbrace{\boxed{\underbrace{B \ldots B}_{k}}}^{r},$$

where $\boxed{R_t}$ represents any string $O_t$, $J_t$, and $T_t$ appearing $x_t$, $y_t$, and $z_t$ times, respectively, in the sequence, satisfying $\sum_{t=1}^k t(x_t + y_t + z_t) + kr = n$. The number of different ways of arranging the sequence equals

$$\binom{\sum_{t=1}^k (x_t + y_t + z_t) + r - 1}{x_1, \ldots, x_k, y_1, \ldots, y_k, z_1, \ldots, z_k, r-1}.$$

Because of the independence of the trials, the probability of the sequence is

$$P_a^{\sum_{t=1}^{k} x_t} P_b^{\sum_{t=1}^{k}(t-1)(x_t+y_t+z_t)+k(r-1)} P_c^{\sum_{t=1}^{k} y_t} P_d^{\sum_{t=1}^{k} z_t}.$$

The probability of the run of B's of length $k$ at the end of each of these sequences is $P_b^k$, which leads to the overall probability

$$P_a^{\sum_{t=1}^{k} x_t} P_b^{\sum_{t=1}^{k}(t-1)(x_t+y_t+z_t)+kr} P_c^{\sum_{t=1}^{k} y_t} P_d^{\sum_{t=1}^{k} z_t}.$$

□

**Remark 3.** For $P_a = q$, $P_b = p$ such that $p + q = 1$, $P_c = P_d = 0$ and $y_1, \ldots, y_k, z_1, \ldots, z_k = 0$, Theorem 3.2 reduces to Theorem 3.1 (a) of [22].

## 4. Distributions of Run Lengths

We consider a sequence $X_1, X_2, \ldots, X_n$ of multistate trials defined on the state space $\mathscr{S} = \{A,\ B,\ C,\ D\}$ with probabilities $P(A) = P_a$, $P(B) = P_b$, $P(C) = P_c$ and $P(D) = P_d$, such that $P_a + P_b + P_c + P_d = 1$. In this section, we obtain the PMFs of the distribuions of the longest and shortest runs and establish the PMFs and their joint distributions using combinatorial analysis.

### 4.1. Distribution of Tehe Longest Run Length

Let $L_n$ be the maximum length of a run of B's in $n$ four state IID trials, which is called longest run statistics. We establish the PMF of the random variable $L_n$ using combinatorial analysis.

**Theorem 4.1.** *The PMF of $L_n$ is given by*

$$P(L_n = \ell) = P_b^n \sum_{i=0}^{\ell} \sum_{\star} \binom{\sum_{t=1}^{\ell+1}(x_t + y_t + z_t)}{x_1, \ldots, x_{\ell+1}, y_1, \ldots, y_{\ell+1}, z_1, \ldots, z_{\ell+1}} \left(\frac{P_a}{P_b}\right)^{\sum_{t=1}^{\ell+1} x_t} \times \left(\frac{P_c}{P_b}\right)^{\sum_{t=1}^{\ell+1} y_t} \left(\frac{P_d}{P_b}\right)^{\sum_{t=1}^{\ell+1} z_t},$$

*where the inner summation $\sum_{\star}$ is over all nonnegative integers $x_1, \ldots, x_{\ell+1}$, $y_1, \ldots, y_{\ell+1}$, $z_1, \ldots, z_{\ell+1}$ such that $\sum_{t=1}^{\ell+1} t(x_t + y_t + z_t) = n - i$, for $0 \le i \le \ell$, and satisfying at least one of the conditions $x_{\ell+1} \ge 1$, $y_{\ell+1} \ge 1$, $z_{\ell+1} \ge 1$, and $i = \ell$.*

*Proof.* Let $\underbrace{B \cdots B}_{t-1} A = O_t$, $\underbrace{B \cdots B}_{t-1} C = J_t$, and $\underbrace{B \cdots B}_{t-1} D = T_t$, where $t = 1, \ldots, \min(\ell + 1, n)$. A typical element of the event $\{L_n = \ell\}$ is a sequence

$$\cdots \boxed{R_t} \cdots \underbrace{B \cdots B}_{i},$$

where $\boxed{R_t}$ represents any of the strings $O_t$, $J_t$, and $T_t$ appearing $x_t$, $y_t$, and $z_t$ times, respectively, in the sequence, satisfying $\sum_{t=1}^{\ell+1} t(x_t + y_t + z_t) = n - i$, for $0 \le i \le \ell$, and satisfying at least one of the conditions: $x_{\ell+1} \ge 1$, $y_{\ell+1} \ge 1$, $z_{\ell+1} \ge 1$, and $i = \ell$. The number of different ways of arranging the sequence equals

$$\binom{\sum_{t=1}^{\ell+1}(x_t + y_t + z_t)}{x_1, \ldots, x_{\ell+1}, y_1, \ldots, y_{\ell+1}, z_1, \ldots, z_{\ell+1}}.$$

Because of the independence of the trials, the probability of the above sequence is

$$P_a^{\sum_{t=1}^{\ell+1} x_t} P_b^{\sum_{t=1}^{\ell+1}(t-1)(x_t+y_t+z_t)} P_c^{\sum_{t=1}^{\ell+1} y_t} P_d^{\sum_{t=1}^{\ell+1} z_t}.$$

The probability of the run of B's of length $i$ $(0 \le i \le \ell)$ at the end of each of these sequences is $P_b^i$, which leads to the overall probability

$$P_a^{\sum_{t=1}^{\ell+1} x_t} P_b^{\sum_{t=1}^{\ell+1}(t-1)(x_t+y_t+z_t)+i} P_c^{\sum_{t=1}^{\ell+1} y_t} P_d^{\sum_{t=1}^{\ell+1} z_t}.$$

Summing over $i = 0, 1, \ldots, \ell$ the result follows. □

### 4.2. Joint Distributions of Maximum Length and Number of Times

Let $L_n$ denote the maximum length of runs of B's and $NL_n$ the number of times a run of length $L_n$ appears in a sequence of size $n$. We establish the joint PMF of the random variables $L_n$ and $NL_n$ using combinatorial analysis.

**Theorem 4.2.** *The joint PMF of $L_n$ and $NL_n$ is given by*

$$P(L_n = \ell \wedge NL_n = x) = P_b^n \sum_{i=0}^{\ell} \sum_{\star} \left[ \binom{x_1 + \cdots + x_{\ell+1} + y_1 + \cdots + y_{\ell+1} + z_1 + \cdots + z_{\ell+1}}{x_1, \ldots, x_{\ell+1}, y_1, \ldots, y_{\ell+1}, z_1, \ldots, z_{\ell+1}} \times \left(\frac{P_a}{P_b}\right)^{x_1+\cdots+x_{\ell+1}} \left(\frac{P_c}{P_b}\right)^{y_1+\cdots+y_{\ell+1}} \left(\frac{P_d}{P_b}\right)^{z_1+\cdots+z_{\ell+1}} \right],$$

*where the inner summation $\sum_{\star}$ is over all nonnegative integers $x_1, \ldots, x_{\ell+1}, y_1, \ldots, y_{\ell+1}, z_1, \ldots, z_{\ell+1}$ for which $\sum_{t=1}^{\ell+1} t(x_t + y_t + z_t) = n - i$, and satisfying at least a conditions*

$$\begin{cases} x_{\ell+1} \ge 1 \text{ or} \\ y_{\ell+1} \ge 1 \text{ or} \\ z_{\ell+1} \ge 1 \text{ or} \\ i = \ell \text{ and } 0 \le i \le \min(\ell, n - x(\ell+1)), \end{cases}$$

*subject to*

$$x_{\ell+1} + y_{\ell+1} + z_{\ell+1} = \begin{cases} x & \text{if } i \ne \ell, \\ x - 1 & \text{if } i = \ell. \end{cases}$$

*Proof.* Let $\underbrace{B\cdots B}_{t-1}A = O_t$, $\underbrace{B\cdots B}_{t-1}C = J_t$, and $\underbrace{B\cdots B}_{t-1}D = T_t$, where $1 \le t \le \ell+1$. A typical element of the event $\{L_n = \ell \wedge NL_n = x\}$ is a sequence

$$\cdots \boxed{R_t} \cdots \overset{1}{\boxed{O_{\ell+1} \text{ or } J_{\ell+1} \text{ or } T_{\ell+1}}} \cdots \boxed{R_t} \cdots \overset{x}{\boxed{O_{\ell+1} \text{ or } J_{\ell+1} \text{ or } T_{\ell+1}}} \cdots \boxed{R_t} \cdots \underbrace{B\cdots B}_{i},$$

where $0 \le i \le \min(\ell, n - x(\ell+1))$, and $\boxed{R_t}$ represents any string $O_t$, $J_t$, and $T_t$ appearing $x_t$, $y_t$, and $z_t$ times, respectively, in the sequence, satisfying

$$\sum_{t=1}^{\ell+1} t(x_t + y_t + z_t) = n - i,$$

subject to the condition $x_{\ell+1} + y_{\ell+1} + z_{\ell+1} = \begin{cases} x & \text{if } i \ne \ell, \\ x-1 & \text{if } i = \ell. \end{cases}$ The number of different ways of arranging the sequence equals

$$\binom{x_1 + \cdots + x_{\ell+1} + y_1 + \cdots + y_{\ell+1} + z_1 + \cdots + z_{\ell+1}}{x_1, \ldots, x_{\ell+1}, y_1, \ldots, y_{\ell+1}, z_1, \ldots, z_{\ell+1}}.$$

Because of the independence of the trials, the probability of the above sequence is

$$P_a^{x_1+\cdots+x_{\ell+1}} P_c^{y_1+\cdots+y_{\ell+1}} P_d^{z_1+\cdots+z_{\ell+1}} P_b^{(x_2+y_2+z_2)+2(x_3+y_3+z_3)+\cdots+l(x_{\ell+1}+y_{\ell+1}+z_{\ell+1})}.$$

The probability of the run of B's of length $i$ for $0 \le i \le n - x(\ell+1))$ at the end of each of these sequences is $P_b^i$, which leads to the overall probability.

$$P_a^{x_1+\cdots+x_{\ell+1}} P_c^{y_1+\cdots+y_{\ell+1}} P_d^{z_1+\cdots+z_{\ell+1}} P_b^{(x_2+y_2+z_2)+2(x_3+y_3+z_3)+\cdots+\ell(x_{\ell+1}+y_{\ell+1}+z_{\ell+1})+i}.$$

Summing over $i = 0, 1, \ldots, n - x(\ell+1)$ the results follows. □

### 4.3. Distribution of The Smallest Run Length

Let $M_n$ denote the minimum length of a run of B's in $n$ IID four state trials. We establish the PMF of the random variable $M_n$ using combinatorial analysis.

**Theorem 4.3.** *The PMF of $M_n$, for $0 \le s \le n$, is given by*

$$P(M_n = s) =$$
$$\sum_i \sum_{\star} \left[ \binom{x_1 + x_{s+1} + \cdots + x_n + y_1 + y_{s+1} + \cdots + y_n + z_1 + z_{s+1} + \cdots + z_n}{x_1, x_{s+1}, \ldots, x_n, y_1, y_{s+1}, \ldots, y_n, z_1, z_{s+1}, \ldots, z_n} \right.$$
$$\left. \times \left(\frac{P_a}{P_b}\right)^{x_1+x_{s+1}+\cdots+x_{n-s}} \left(\frac{P_c}{P_b}\right)^{y_1+y_{s+1}+\cdots+y_{n-s}} \left(\frac{P_d}{P_b}\right)^{z_1+z_{s+1}+\cdots+z_{n-s}} \right],$$

*where the summation* $\sum_\star$ *is over all nonnegative integers* $x_1, x_{s+1}, \ldots, x_n, y_1, y_{s+1}, \ldots, y_n, z_1, z_{s+1}, \ldots, z_n$ *such that* $(x_1+y_1+z_1)+(s+1)(x_{s+1}+y_{s+1}+z_{s+1})+\cdots+n(x_n+y_n+z_n)=n-i$, *where* $i \in \{0, s, s+1, \ldots, n\}$, *satisfying at least one of the conditions:* $x_{s+1} \geq 1$, $y_{s+1} \geq 1$, $z_{s+1} \geq 1$, *and* $i = s$.

*Proof.* Let $\underbrace{B\cdots B}_{t-1}A = O_t$, $\underbrace{B\cdots B}_{t-1}C = J_t$, and $\underbrace{B\cdots B}_{t-1}D = T_t$, where $t = 1, s+1, \ldots, n$. A typical element of the event $\{M_n = s\}$ is a sequence

$$\cdots \boxed{R_t} \cdots \underbrace{B\cdots B}_{i},$$

where $\boxed{R_t}$ represents any of the strings $O_t$, $J_t$, and $T_t$ appearing $x_t$, $y_t$, and $z_t$ times, respectively, in the sequence, satisfying $(x_1+y_1+z_1)+(s+1)(x_{s+1}+y_{s+1}+z_{s+1})+\cdots+n(x_n+y_n+z_n)=n-i$, for $i \in \{0, s, s+1, \ldots, n\}$, and satisfying at least one of the conditions: $x_{s+1} \geq 1$, $y_{s+1} \geq 1$, $z_{s+1} \geq 1$, and $i = s$. The number of different ways of arranging the sequence equals

$$\binom{x_1+x_{s+1}+\cdots+x_n+y_1+y_{s+1}+\cdots+y_n+z_1+z_{s+1}+\cdots+z_n}{x_1, x_{s+1}, \ldots, x_n, y_1, y_{s+1}, \ldots, y_n, z_1, z_{s+1}, \ldots, z_n}.$$

By the independence of trials, the probability of the above sequence is given by

$$P_a^{x_1+x_{s+1}+\cdots+x_{n-s}} P_b^{s(x_{s+1}+y_{s+1}+z_{s+1})+(s+1)(x_{s+2}+y_{s+2}+z_{s+2})+\cdots+(n-1)(x_n+y_n+z_n)} \times P_c^{y_1+y_{s+1}+\cdots+y_{n-s}} P_d^{z_1+z_{s+1}+\cdots+z_{n-s}}.$$

The probability of a run of B's of length $i \in \{0, s, s+1, \ldots, n\}$ at the end of each of these sequences is $P_b^i$, which leads to the overall probability

$$P_a^{x_1+x_{s+1}+\cdots+x_{n-s}} P_b^{s(x_{s+1}+y_{s+1}+z_{s+1})+(s+1)(x_{s+2}+y_{s+2}+z_{s+2})+\cdots+(n-1)(x_n+y_n+z_n)+i} \times P_c^{y_1+y_{s+1}+\cdots+y_{n-s}} P_d^{z_1+z_{s+1}+\cdots+z_{n-s}}.$$

Summing over $i = 0, s, s+1, \ldots, n$ the result follows. □

### 4.4. Joint Distributions of Minimum Length and Number of Times

Let $M_n$ denote the minimum length of a run of B's and $NM_n$ be the number of times a run of length $M_n$ appears in a sequence of size $n$. We establish the joint PMF of the random variables $M_n$ and $NM_n$ using combinatorial analysis.

**Theorem 4.4.** *The joint PMF of* $M_n$ *and* $NM_n$, *for* $0 \leq s \leq n$, $0 \leq x \leq \left[\frac{n+1}{s+1}\right]$, *is given by*

$$P(M_n = s \wedge NM_n = x) = P_b^n \sum_{i=0}^{l} \sum_{\star} \binom{x_1+x_{s+1}+\cdots+x_n+y_1+y_{s+1}+\cdots+y_n+z_1+z_{s+1}+\cdots+z_n}{x_1, x_{s+1}, \ldots, x_n, y_1, y_{s+1}, \ldots, y_n, z_1, z_{s+1}, \ldots, z_n} \times \left(\frac{P_a}{P_b}\right)^{x_1+x_{s+1}+\cdots+x_n} \left(\frac{P_c}{P_b}\right)^{y_1+y_{s+1}+\cdots+y_n} \left(\frac{P_b}{P_d}\right)^{z_1+z_{s+1}+\cdots+z_n},$$

*where the inner summation* $\sum_{\star}$ *is over all nonnegative integers* $x_1, x_{s+1}, \ldots, x_n, y_1, y_{s+1}, \ldots, y_n, z_1, z_{s+1}, \ldots, z_n$ *for which* $(x_1+y_1+z_1)+(s+1)(x_{s+1}+y_{s+1}+z_{s+1})+\cdots+n(x_n+y_n+z_n)=n-i$, *for* $i \in \{0, s, s+1, \ldots, n-x(s+1)\}$, *while satisfying at least one of the conditions:* $x_{s+1} \geq 1$, $y_{s+1} \geq 1$, $z_{s+1} \geq 1$, *and* $i = s$, *and subject to* $x_{s+1}+y_{s+1}+z_{s+1} = \begin{cases} x & \text{if } i \neq s, \\ x-1 & \text{if } i = s. \end{cases}$

*Proof.* Let $\underbrace{B\cdots B}_{t-1}A = O_t$, $\underbrace{B\cdots B}_{t-1}C = J_t$, and $\underbrace{B\cdots B}_{t-1}D = T_t$, where $s+1 \leq t \leq n$. A typical element of the event $\{M_n = s \wedge NM_n = x\}$ is a sequence

$$\cdots \boxed{R_t} \cdots \overbrace{\boxed{O_{s+1} \text{ or } J_{s+1} \text{ or } T_{s+1}}}^{1} \cdots \boxed{R_t} \cdots \overbrace{\boxed{O_{s+1} \text{ or } J_{s+1} \text{ or } T_{s+1}}}^{x} \cdots \boxed{R_t} \cdots \underbrace{B\cdots B}_{i},$$

where $i \in \{0, s, s+1, \ldots, n-x(s+1)\}$, and $\boxed{R_t}$ represents any string $O_t$, $J_t$, and $T_t$, appearing $x_t$, $y_t$ and $z_t$ times, respectively, in the sequence, satisfying

$$(x_1+y_1+z_1)+(s+1)(x_{s+1}+y_{s+1}+z_{s+1})+\cdots+n(x_n+y_n+z_n)=n-i,$$

subject to the condition $x_{s+1}+y_{s+1}+z_{s+1} = \begin{cases} x & \text{if } i \neq s, \\ x-1 & \text{if } i = s. \end{cases}$ The number of different ways of arranging the sequence equals

$$\binom{x_1+x_{s+1}+\cdots+x_n+y_1+y_{s+1}+\cdots+y_n+z_1+z_{s+1}+\cdots+z_n}{x_1, x_{s+1}, \ldots, x_n, y_1, y_{s+1}, \ldots, y_n, z_1, z_{s+1}, \ldots, z_n}.$$

By the independence of trials, the probability of the above sequence is given by

$$P_a^{x_1+x_{s+1}+\cdots+x_n} P_c^{y_1+y_{s+1}+\cdots+y_n} P_d^{z_1+z_{s+1}+\cdots+z_n} \times P_b^{s(x_{s+1}+y_{s+1}+z_{s+1})+\cdots+(n-1)(x_n+y_n+z_n)}.$$

The probability of the run of B's of length $i \in \{0, s, s+1, \ldots, n\}$ at the end of each of these sequences is $P_b^i$, which leads to the overall probability

$$P_a^{x_1+x_{s+1}+\cdots+x_n} P_c^{y_1+y_{s+1}+\cdots+y_n} P_d^{z_1+z_{s+1}+\cdots+z_n} \times P_b^{s(x_{s+1}+y_{s+1}+z_{s+1})+\cdots+(n-1)(x_n+y_n+z_n)+i}.$$

Summing over $i = 0, s, s+1, \ldots, n$ the result follows. □

**Acknowledgements.** The author would like to thank Dr. habil. Tommy Rene Jensen whose comments led to significant improvements in this manuscript. This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2019R1A6A3A01090663, NRF-2020R1I1A1A01057479) and partially supported under the framework of international cooperation program managed by the National Research Foundation of Korea(NRF-2022K2A9A1A01098016).

# References


[1] S. Aki, *Waiting time problems for a sequence of discrete random variables*, Ann. Inst. Statist. Math., **44(2)**(1992), 363–378.

[2] S. Aki, *Waiting time for consecutive repetitions of a pattern and related distributions*, Ann. Inst. Statist. Math., **71(2)**(2019), 307–325.

[3] D. L. Antzoulakos, *On waiting time problems associated with runs in Markov dependent trials*, Ann. Inst. Statist. Math., **51(2)**(1999), 323–330.

[4] D. L. Antzoulakos and A. N. Philippou, *Probability distribution functions of succession quotas in the case of Markov dependent trials*, Ann. Inst. Statist. Math., **49(3)**(1997), 531–539.

[5] A. N. Arapis, F. S. Makri and Z. M. Psillakis, *Distributions of statistics describing concentration of runs in non homogeneous Markov-dependent trials*, Comm. Statist. Theory Methods, **47(9)**(2018), 2238–2250.

[6] N. Balakrishnan and M. V. Koutras, *Runs and scans with applications* John Wiley & Sons, New York, 2003.

[7] K. Balasubramanian, R. Viveros and N. Balakrishnan, *Sooner and later waiting time problems for Markovian Bernoulli trials*, Stat. Probab. Lett., **18(2)**(1993), 153–161.

[8] G. J. Chang, L. Cui and F. K. Hwang, *Reliabilities of consecutive-k systems*, Springer Science & Business Media, 2000.

[9] M. T. Chao, J. C. Fu, and M. V. Koutras, *Survey of reliability studies of consecutive-k-out-of-n: F and related systems*, IEEE Trans. Reliab., **44(1)**(1995), 120–127.

[10] M. Ebneshahrashoob and M. Sobel, *Sooner and later waiting time problems for Bernoulli trials: frequency and run quotas*, Stat. Probab. Lett., **9(1)**(1990), 5–11.

[11] S. Eryilmaz, *Review of recent advances in reliability of consecutive k-out-of-n and related systems*, Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, **224(3)**(2010), 225–237.

[12] S. Eryilmaz, *On success runs in a sequence of dependent trials with a change point*, Stat. Probab. Lett., **132**(2018), 91–98.

[13] W. Feller, *An introduction to probability theory and its applications Vol.* 1, John Wiley & Sons, New York, 1968.

[14] J. C. Fu and W. W. Lou, *Distribution theory of runs and patterns and its applications: a finite Markov chain imbedding approach*, World Scientific, 2003.

[15] K. Inoue and S. Aki, *On sooner and later waiting time distributions associated with simple patterns in a sequence of bivariate trials*, Metrika, **77(7)**(2014), 895–920.

[16] S. Kim, C. Park and J. Oh, *On waiting time distribution of runs of ones or zeros in a Bernoulli sequence*, Stat. Probab. Lett., **83(1)**(2013), 339–344.

[17] Y. Kong, *Joint distribution of rises, falls, and number of runs in random sequences*, Comm. Statist. Theory Methods, **48(3)**(2019), 493–499.

[18] M. V. Koutras, *On a waiting time distribution in a sequence of Bernoulli trials*, Ann. Inst. Statist. Math., **48(4)**(1996), 789–806.

[19] W. Kuo and M. J. Zuo, *Optimal reliability modeling: principles and applications*, John Wiley & Sons, New York, 2003.

[20] F. S. Makri, Z. M. Psillakis and A. N. Arapis, *we*, J. Appl. Stat., **46(1)**(2019), 85–100.

[21] A. M. Mood, *The distribution theory of runs*, Ann. Math. Statist., **11(4)**(1940), 367–392.

[22] A. N. Philippou, *The negative binomial distribution of order k and some of its properties*, Biom. J., **26(7)**(1984), 789-794.

[23] A. N. Philippou and A. A. Muwafi, *Waiting for the kth consecutive success and the fibonacci sequence of order k*, Fibonacci Quart, **20(1)**(1982), 28–32.

[24] A. N. Philippou, C. Georghiou and G. N. Philippou, *A generalized geometric distribution and some of its properties*, Stat. Probab. Lett., **1(4)**(1983), 171–175.

[25] A. N. Philippou and F. S. Makri, *Successes, runs and longest runs*, Stat. Probab. Lett., **4(4)**(1986), 101–105.

[26] K. Sen, M. L. Agarwal and S. Chakraborty, *Lengths of runs and waiting time distributions by using Polya-Eggenberger sampling scheme*, Studia Sci. Math. Hungar., **39(3-4)**(2002), 309–332.

[27] I. S. Triantafyllou, *Consecutive-type reliability systems: an overview and some applications*, J. Qual. Reliab. Eng., **2015(1)**(2015), 1–20.