BENTHAM OPEN

CrossMark

The Open Statistics & Probability Journal

Content list available at: www.benthamopen.com/TOSPJ/

DOI: 10.2174/1876527001708010001

RESEARCH ARTICLE

On Limited Length Binary Strings with an Application in Statistical Control

F.S. Makri[1,*] and Z.M. Psillakis[2]

[1]*Department of Mathematics, University of Patras, 26500 Patras, Greece*
[2]*Department of Physics, University of Patras, 26500 Patras, Greece*

Received: August 04, 2016 Revised: September 01, 2016 Accepted: September 12, 2016

Abstract: In a 0 - 1 sequence of Markov dependent trials we consider a statistic which counts strings of a limited length run of 0s between subsequent 1s. Its probability mass function is used to determine the chance that a stochastic process remains or not in statistical control. Illustrative numerics are presented.

Keywords: Binary strings, runs, overlapping counting, Markov chain, Statistical control.

1. INTRODUCTION AND PRELIMINARIES

Let $\{X_i\}_{i\geq 1}$ be a sequence of binary random variables (RVs) with values $x_i \in \{0,1\}$ordered on a line. For $n \geq 2$ and two non-negative integer numbers k and l, $0 \leq k \leq l \leq n - 2$, we consider the statistic (random variable) $M_{n;k,l}$ which enumerates strings (binary patterns) $1\underbrace{00\ldots0}_{d}1$ of a limited (constrained) length equal to d+2, $k \leq d \leq l$; *i.e.* strings which consist of a zero run (consecutive 0s) of length at least equal to k and at most equal to l between two subsequent ones. The counting of such constrained (k, l) strings is considered in the overlapping sense; that is, a 1 which is not at either end of the sequence may contribute towards counting two possible strings, the one which ends with the occurrence of it and the next one which starts with it.

As an illustration let 0100010000111011001001100 be the first n = 25 outcomes of a 0 - 1 sequence. Then $M_{25;1,2}$ = 3, $M_{25;1,3}$ = 4, $M_{25;0,0}$ = 4 and $M_{25;0,4}$ = 9.

Let $A = \{k + 2, k + 3,..., n\}$ be an index set and $\{B, B^c\}$ a partition of A, where $B = \{k + 2, k + 3,..., l + 1\}$ and $B^c = \{l + 2, l + 3,..., n\}$. Then formally, $M_{n;k,l}$, $0 \leq k \leq l \leq n$ - 2, can be written, on a 0 - 1 sequence $\{X_i\}_{i=1}^{n}$, as (see Makri and Psillakis [1]):

$$M_{n;k,\ell} = \sum_{i\in A} I_i, \quad I_i = \Big[\sum_{r=1}^{\beta_i-k} X_{i-k-r} \prod_{j=i-k-r+1}^{i-1} (1 - X_j)\Big] X_i, \tag{1}$$

with

$$\beta_i = \begin{cases} i-1, & \text{if } i \in B \\ \ell+1, & \text{otherwise.} \end{cases} \tag{2}$$

[*] Address correspondence to this author at the Department of Mathematics, University of Patras, 26500 Patras, Greece; Tel: 0030-2610-996738; E-mail: makri@math.upatras.gr

1876-5270/17 2017 Bentham Open

Readily, for $n < k + 2$, $M_{n;k,l} = 0$. We notice that throughout the article we apply the conventions $\sum_{i=a}^{b} = 0, \prod_{i=a}^{b} = 1$, for $a > b$; that is, an empty sum (product) is to be interpreted as a zero (unity). The support (range set) of $M_{n;k,l}$ is:

$$\mathcal{R}(M_{n;k,\ell}) = \Big\{0, 1, \ldots, \lfloor \frac{n-1}{k+1} \rfloor \Big\}, \tag{3}$$

where by $\lfloor x \rfloor$ we denote the greatest integer less than or equal to a real number x.

A statistic related to $M_{n;k,l}$ is the waiting time $W_{m;k,l}$ until the m-th, $m \geq 1$, occurrence of a constrained (k, l) string. It is defined and connected to $M_{n;k,l}$ as follows:

$$W_{m;k,\ell} = \min\{n \geq m(k+1)+1 : M_{n;k,\ell} = m\}; \quad W_{m;k,\ell} > n \quad \text{iff} \quad M_{n;k,\ell} < m. \tag{4}$$

Hence, Eq. (4) offers an alternative way of obtaining results for the waiting time RV $W_{m;k,l}$ through formulae established for the string enumerative RV $M_{n;k,l}$ and *vise versa*.

The study, *via* $M_{n;k,l}$, of constrained (k, l) strings where a 1 is followed by at least k and at most l 0s before the next 1 in a 0 - 1 sequence covers as particular cases strings with zero run length at most equal ($M_{n;0,l}$, $d \leq l$), exactly equal ($M_{n;k,k}$, $d = k$) and at least equal ($M_{n;k,n-2}$, $d \geq k$) to a non-negative integer number. Some relevant contributions on the subject are the works of Sarkar *et al.* [2], Sen and Goyal [3], Holst [4, 5], Huffer *et al.* [6], Eryilmaz and Zuo [7], Eryilmaz and Yalcin [8], Dafnis *et al.* [9], and Makri and Psillakis [10]. Moreover, $M_{n;0,0}$ is the Ling's [11] RV $M_n^{(2)}$ which counts overlapping runs of 1s of length 2 (with overlapping part of length at most 1), see *e.g.* Balakrishnan and Koutras [12].

In Section 2 of the present paper we first introduce the required notation and second we recall a recent result of Makri and Psillakis [1] which refers to the probability mass function (PMF) of $M_{n;k,l}$ defined on a 0 - 1 Markov chain. The presented expressions for the PMF are then used in an application case study which is discussed in Section 3 and it refers to statistical process control.

Throughout the article, $\delta_{i,j}$ denotes the Kronecker delta function of the integer arguments i and j.

2. PMF OF $M_{n;k,l}$ FOR MARKOV DEPENDENT TRIALS

In this Section we provide the PMF $f_{n;k,l}(x) = P(M_{n;k,l} = x)$, $x \in \mathcal{R}(M_{n;k,\ell})$, $0 \leq k \leq l \leq n - 2$. The 0 - 1 sequence $\{X_i\}_{i=1}^{n}$, $n \geq 2$ on which $M_{n;k,l}$ is defined, is generated by a time-homogeneous first order Markov chain (MRKV) with one step transition probability matrix $\mathbf{P} = (p_{i,j})$ and initial probability vector $\mathbf{p}^{(1)} = (p_0^{(1)}, p_1^{(1)})$ with $p_{ij} = P(X_t = j \mid X_{t-1} = i)$, $t \geq 2$, $p_0^{(1)} = P(X_1 = 0) = 1 - P(X_1 = 1) = 1 - p_1^{(1)}$ and $i, j \in \{0, 1\}$. Readily, a 0 - 1 sequence $\{X_i\}_{i=1}^{n}$, $n \geq 2$ of independent and identically distributed (IID) RVs with a common probability of 1s, $p = P(X_i = 1) = 1 - P(X_i = 0) = 1 - q$, $i = 1,2,...,n$ is a particular MRKV sequence with $p_{00} = p_{10} = q$, $p_{01} = p_{11} = p$ and $(p_0^{(1)}, p_1^{(1)}) = (q, p)$.

We next present two combinatorial results (Lemma 1 and Corollary 1) which then are used for the calculation of $f_{n;k,l}(x)$ given by Proposition 1.

Lemma 1 (Makri *et al.* [13]) Let $C_{i,r-i}(\alpha, m, k_1 - 1, k_2 - 1)$ be the number of allocations of α indistinguishable balls into m distinguishable cells, i specified of which have capacity $k_1 - 1$ and each of r-i specified cells has capacity $k_2 - 1$, $0 \leq r \leq m$, $0 \leq i \leq r$, $k_1 \geq 1$, $k_2 \geq 1$. Then,

$$C_{i,r-i}(\alpha, m, k_1-1, k_2-1) = \sum_{j_1=0}^{\lfloor \frac{\alpha}{k_1} \rfloor} \binom{i}{j_1} \sum_{j_2=0}^{\lfloor \frac{\alpha - k_1 j_1}{k_2} \rfloor} (-1)^{j_1+j_2} \binom{r-i}{j_2} \binom{\alpha + m - k_1 j_1 - k_2 j_2 - 1}{\alpha - k_1 j_1 - k_2 j_2}. \tag{5}$$

We note that the number $C_{i,r-i}(\alpha, m, k_1 - 1, k_2 - 1)$ equivalently gives the number of integer solutions of the equation $x_1 + x_2 + ... + x_m = \alpha$ such that $0 \leq x_j < k_1$, $1 \leq j \leq i$ and $0 \leq x_{i+j} < k_2$, $1 \leq j \leq r$-i. Setting $k_1 = k_2 = k$ in Lemma 1 we obtain, as a corollary, the following result.

Corollary 1 (Makri *et al.* [14]). Let $H_r(\alpha, m, k - 1)$ be the number of allocations of α indistinguishable balls into m distinguishable cells where each of the r, $0 \le r \le m$, specified cells is occupied by at most k - 1 balls. Then, $H_r(\alpha, m, k - 1) = C_{i,r-i}(\alpha, m, k - 1, k - 1)$. Accordingly,

$$\begin{aligned} H_r(\alpha, m, k-1) &= \sum_{j=0}^{\lfloor \frac{\alpha}{k} \rfloor} (-1)^j \binom{r}{j} \binom{\alpha + m - kj - 1}{\alpha - kj}, \text{ if } k > 1 \\ &= \binom{\alpha + m - r - 1}{\alpha}, \text{ if } k = 1. \end{aligned} \quad \textbf{(6)}$$

Proposition 1 (Makri and Psillakis [1]). For $\mathbf{p}^{(1)} = (p_0^{(1)}, p_1^{(1)}) = (p_0, p_1)$ and $\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$ the PMF $f_{n;k,l}(x)$, $x \in \mathcal{R}(M_{n;k,\ell})$, $n \ge k + 2$ is given by:

$$\begin{aligned} f_{n;k,\ell}(x) &= \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{s=x+1}^{n-kx-i-j} \sum_{r=r_{k,\ell}}^{\min\{n-s,s-1+i+j\}} (1-p_i) p_{01}^{r-j} p_{10}^{r-i} p_{00}^{n-s-r} p_{11}^{s-r-1+i+j} \Delta_{x,s,r,i,j}(k,\ell) \\ &\quad + p_0 p_{00}^{n-1} \delta_{x,0} \end{aligned} \quad \textbf{(7)}$$

where $r_{k,l} = x + i + j$ if $0 < k \le l$; $i + j$ if $0 = k < l$,

$$\begin{aligned} \Delta_{x,s,r,i,j}(k,\ell) &= \binom{s-1}{r-i-j} \binom{r-i-j}{x} \sum_{z=0}^{r-x-i-j} \binom{r-x-i-j}{z} C_{x,z}(a_{k,\ell}, r, \ell-k, k-2), \\ &\qquad \text{if} \quad k \ge 2, \ell \ge k \\ &= \binom{s-1}{r-i-j} \binom{r-i-j}{x} H_x(a_{1,\ell}, r, \ell-1), \qquad \text{if} \quad k = 1, \ell \ge k \\ &= \binom{s-1}{r-i-j} \binom{r-i-j}{x-s+r+1-i-j} H_{x-s+r+1-i-j}(a_{0,\ell}, r, \ell-1), \\ &\qquad \text{if} \quad k = 0, \ell \ge 1 \end{aligned}$$

with

$$\begin{aligned} a_{k,\ell} &= n-s-r-(k-1)x-\ell(r-x-z-i-j), \quad \text{if} \quad k \ge 2, \ell \ge k \\ &= n-s-r-\ell(r-x-i-j), \quad \text{if} \quad k = 1, \ell \ge k \\ &= n-s-r-\ell(s-x-1), \quad \text{if} \quad k = 0, \ell \ge 1 \end{aligned}$$

and

$$\begin{aligned} f_{n;0,0}(x) &= \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{s=x+1}^{\lfloor \frac{n+x+1-i-j}{2} \rfloor} \binom{n-s-1}{s-x-2+i+j} \binom{s-1}{s-x-1} \\ &\quad \times (1-p_i) p_{01}^{s-x-1+i} p_{10}^{s-x-1+j} p_{00}^{n-2s+x+1-i-j} p_{11}^{x} + p_0 p_{00}^{n-1} \delta_{x,0}, \\ &\qquad \text{for} \quad x = 0, 1, \ldots, n-2 \\ &= p_1 p_{11}^{n-1}, \quad \text{for} \quad x = n-1. \end{aligned} \quad \textbf{(8)}$$

3. A NOTE FOR APPLICATION: STATISTICAL PROCESS CONTROL

Suppose we observe values of a stochastic process $\{Y_i\}_{i \ge 1}$ which operates under chance causes in or out an acceptable zone of interest (*ZI*); *i.e.* the process is in statistical control. We are interested in remaining the process in *ZI*, otherwise a risky or an awkward situation of the process might appear. For instance, a *ZI* might be a control zone in a statistical control chart, a trading zone in a stock or exchange market, a comfort zone in a patient mechanical support system, a zone of acceptable items with specific characteristics in an assembly line, *etc*.

We denote by 0 and 1 the occurrence of the value of the process in and out of *ZI*, respectively. We assume that each value of the process is out of *ZI* with probability p_{01} or p_{11} if the previous value was in or out of *ZI*, respectively. Therefore, the indicator RVs $X_i = 0$, if $Y_i \in ZI$; 1, otherwise constitute a 0 - 1 Markov chain $\{X_i\}_{i\geq 1}$ with transition probability matrix $\mathbf{P} = \begin{pmatrix} 1-p_{01} & p_{01} \\ 1-p_{11} & p_{11} \end{pmatrix}$ and initial probability vector $\mathbf{p}^{(1)} = (1 - p_1, p_1)$. In practise, *ZI* and p_{01}, p_{11} might be evaluated from a reference sample in Phase I (retrospective) analysis of the process whereas the monitoring of the process, we are interested in, is considered as Phase II (prospective) analysis of the process (see, *e.g*. Chakraborti *et al*. [15]).

Following Dafnis and Philippou [16] we interpret the occurrence of two subsequent 1s separated by at least *k* and at most *l* 0s as a sign that the process leaves its *ZI* and it enters in a risky situation for which additional care has to be paid. In such a case the number of values in *ZI* is not enough to compensate for the others out of it. What values of *k* and *l* should depend on the under study process.

Accordingly, the event of having *m* (≥1) or more constrained (*k*, *l*) strings in a stream of *n* future values of the process could serve as an index that the process will not any more be bounded in its *ZI*. The risk of experiencing of such an event would be measured by $P(M_{n;k,l}\geq m) = P(W_{m;k,l}\leq n)$. Therefore, the distributions and the characteristics of the related RVs $M_{n;k,l}$ and $W_{m;k,l}$ would be of importance in understanding the future process behavior. Usually in practice, we consider a detector that counts such events and consequently sends an alarm signal to a process monitoring system. The detector's tolerance (the false alarm probability) is taken to be at most equal to γ ($0 < \gamma < 1$) so that an expected number, γ100%, of false alarms happen. For a given γ a critical value of *m* (if there is such a value for the selected γ and the process parameters *n*, *k*, *l* and p_1, p_{01}, p_{11}) m_γ is: $m_\gamma = \min\{m \geq 1: P(M_{n;k,l}\geq m) = \gamma^* \leq \gamma\}$. The probability γ^* is the largest real number which does not exceed γ. It may not be equal (in fact it is rare to be equal) to the assigned probability γ, as it refers to the discrete RV $M_{n;k,l}$.

As a numerical example we consider Markov dependent trials with $p_1 = 0.5$, $p_{01} = 0.45$ and $p_{11} = 0.90$. That is, we assume that the process has 50-50 chance to be initially in or out its *ZI* and the occurrence of a value out of *ZI* implies that the process continues to remain out of *ZI* with large probability, in fact twice the probability that the process leaves its *ZI*. The latter one is assumed to be a little smaller than 0.5.

For a reasonable value of the detector's tolerance $\gamma = 0.05$ and for large enough values of *n*, we present in Table **1** critical values m_γ, and exact probabilities $\gamma^* = P(M_{n;k,l}\geq m_\gamma)$ for several constrained (*k*, *l*) strings of type at most (AM), at least (AL), at least - at most (AL-AM) and equal (EQ). For comparison we have included in the Table the respective m_γ and γ^* for IID trials with $p = 0.5$.

Table 1. Critical values m_γ and exact probabilities $\gamma^* = P(M_{n;k,l} \geq m_\gamma)$ for $\gamma = 0.05$.

				MRKV		**IID**	
Type	***k***	***l***	***n***	m_γ	γ^*	m_γ	γ^*
AM	0	2	50	47	0.032167	29	0.048855
			100	90	0.044433	54	0.049354
			200	175	0.040410	103	0.040483
AL-AM	1	2	50	6	0.038992	14	0.032755
			100	10	0.039564	25	0.040928
			200	17	0.048307	46	0.047713
	1	3	50	7	0.023274	15	0.034682
			100	11	0.046771	28	0.030000
			200	20	0.032868	52	0.039427
AL	2	48	50	5	0.026426	9	0.032183
		98	100	8	0.037465	17	0.017336
		198	200	14	0.032873	31	0.024284
EQ	2	2	50	4	0.010796	7	0.016457
			100	5	0.040582	11	0.028616

(Table 1) contd.....

				MRKV		IID	
Type	***k***	***l***	***n***	m_γ	γ^*	m_γ	γ^*
			200	8	0.041990	19	0.029692

The entries of Table **1** suggest that in order to have at most 5 false alarms in a total of 100 alarms in a stream of n future values of a process, defined on MRKV with the prementioned values p_{01}, p_{11} and p_1, we have to take larger m_γ when we use a (0, 2) string than m_γ of a (2, 2) string. This is so, because a (2, 2) string is one of the strings counted as a (0, 2) string and therefore as a more strict pattern is a less frequently appearing string, as a warning pattern, than a (0, 2) string. Analogous arguments/interpretations are also suggested for the other presented types of strings as well as when someone compares MRKV and IID trials.

CONCLUSION

In this study the RV $M_{n;k;l}$ counting a flexible class of binary strings of a limited length is considered. A potential application of its PMF in describing the probabilistic behavior of a binary stochastic process is discussed. The stochastic process is assumed to operate under chance in or out an acceptable zone of interest. Such a zone can be a control chart zone, a trading zone of a stock market, *etc*. An index connected to the number of occurrences of limited length binary strings is properly defined via the PMF of $M_{n;k;l}$. Accordingly, it is then used to determine whether or not the process is or not under statistical control. Numerical results clarify the implementation of several types of binary strings used in the control process.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous referee for the through reading and suggestions which helped to improve the paper.

REFERENCES

[1] F.S. Makri, and Z.M. Psillakis, "Exact distributions of constrained (k, l) strings of failures between subsequent successes", *Stat. Papers*, vol. 54, pp. 783-806, 2013.
[http://dx.doi.org/10.1007/s00362-012-0462-1]

[2] A. Sarkar, K. Sen, and Anuradha, "Waiting time distributions of runs in higher order Markov chains", *Ann. Inst. Stat. Math.,* vol. 56, pp. 317-349, 2004.
[http://dx.doi.org/10.1007/BF02530548]

[3] K. Sen, and B. Goyal, "Distributions of patterns of two failures separated by success runs of length k", *J. Korean Stat. Soc.*, vol. 33, pp. 35-58, 2004.

[4] L. Holst, "Counts of failure strings in certain Bernoulli sequences", *J. Appl. Probab.,* vol. 44, pp. 824-830, 2007.
[http://dx.doi.org/10.1017/S0021900200003454]

[5] L. Holst, "The number of two consecutive successes in a Hope-Polya urn", *J. Appl. Probab.,* vol. 45, pp. 901-906, 2008.
[http://dx.doi.org/10.1017/S0021900200004770]

[6] F.W. Huffer, J. Sethuraman, and S. Sethuraman, "A study of counts of Bernoulli strings via conditional Poisson processes", *Proc. Am. Math. Soc.,* vol. 137, pp. 2125-2134, 2009.
[http://dx.doi.org/10.1090/S0002-9939-08-09793-1]

[7] S. Eryilmaz, and M. Zuo, "Constrained (k, d)-out-of-n systems", *Int. J. Syst. Sci.,* vol. 41, pp. 679-685, 2010.
[http://dx.doi.org/10.1080/00207720903144537]

[8] S. Eryilmaz, and F. Yalcin, "On the mean and extreme distances between failures in Markovian binary sequences", *J. Comput. Appl. Math.,* vol. 236, pp. 1502-1510, 2011.
[http://dx.doi.org/10.1016/j.cam.2011.09.013]

[9] S.D. Dafnis, A.N. Philippou, and D.L. Antzoulakos, "Distributions of patterns of two successes separated by a string of k-2 failures", *Stat. Papers,* vol. 53, pp. 323-344, 2012.
[http://dx.doi.org/10.1007/s00362-010-0340-7]

[10] F.S. Makri, and Z.M. Psillakis, "Counting certain binary strings", *J. Stat. Plan. Inference,* vol. 142, pp. 908-924, 2012.
[http://dx.doi.org/10.1016/j.jspi.2011.10.015]

[11] K.D. Ling, "On binomial distributions of order k", *Stat. Probab. Lett.,* vol. 6, pp. 247-250, 1988.
[http://dx.doi.org/10.1016/0167-7152(88)90069-7]

[12] N. Balakrishnan, and M.V. Koutras, *Runs and scans with applications.,* Wiley: New York, 2002.

[13] F.S. Makri, A.N. Philippou, and Z.M. Psillakis, "Polya, inverse Polya, and circular Polya distributions of order *k* for *l*-overlapping success runs", *Commun. Stat. Theory Methods*, vol. 36, pp. 657-668, 2007.
[http://dx.doi.org/10.1080/03610920601033942]

[14] F.S. Makri, A.N. Philippou, and Z.M. Psillakis, "Success run statistics defined on an urn model", *Adv. Appl. Probab.,* vol. 39, pp. 991-1019, 2007.
[http://dx.doi.org/10.1017/S0001867800002202]

[15] S. Chakraborti, S. Eryilmaz, and S.W. Human, "A phase II nonparametric control chart based on precedence statistics with runs-type signaling rules", *Comput. Stat. Data Anal.,* vol. 53, pp. 1054-1065, 2009.
[http://dx.doi.org/10.1016/j.csda.2008.09.025]

[16] S.D. Dafnis, and A.N. Philippou, "Distributions of patterns with applications in engineering", *IAENG Int. J. Appl. Math.,* vol. 41, pp. 68-75, 2011.