Open Journal of Applied Sciences
Supplement：2012 world Congress on Engineering and Technology

Counting Runs of Ones and Ones in Runs of Ones in Binary Strings

Counting runs in binary strings

Frosso S. Makri, Zaharias M. Psillakis, Nikolaos Kollas
Departments of Mathematics and Physics
University of Patras
Patras, Greece
makri@math.upatras.gr; psillaki@physics.upatras.gr

Abstract—Consider a binary string (a symmetric Bernoulli sequence) of length n. For a positive integer $k, 1 \le k \le n$, we exactly enumerate, in all 2^n possible binary strings of length n, the number of all runs of 1s of length (equal, at least) k and the number of 1s in all runs of 1s of length at least k. To solve these counting problems, we use probability theory and we obtain simple and easy to compute explicit formulae as well as recursive schemes, for these potential useful in engineering numbers.

Keywords-runs; symmetric Bernoulli trials; probability theory; combinatorial problems

1. Introduction and Preliminaries

Nowadays, the increasing use of the computer science in diverse applications including encoding, compression and transmission of digital information calls for understanding the distribution of runs of 1s or 0s. For instance, such knowledge would help in analyzing, and comparing also, several techniques used in communication networks (wired or wireless). In such networks binary data, ranging from a few bytes (e.g. e-mails) to many gigabytes of greedy multimedia applications (e.g. video on demand), are highly processed. For details, see [1-2] and the references therein.

Another area where the study of the distribution of runs of 1s and 0s has become increasingly useful is the field of bioinformatics or computational biology. In particular, molecular biologists examine tandem repeats among DNA (Deoxyribonucleic acid) segments trying to specify how probable are runs of matches, denoted as 1s, in adjacent segments of a DNA sequence. See, e.g. [3-5].

In such applications, as the indicative ones mentioned above, a key point is the understanding how 1s and 0s are distributed and combined among the elements of a binary sequence (finite or infinite, memoryless or not) and eventually forming runs of 1s and 0s according to certain enumeration rules (counting schemes). Each enumeration rule defines how runs of same symbols (i.e. 1s or 0s) are formed and consequently counted. A rule may depend on, among other considerations, whether overlapping counting is allowed or not as well as if the counting starts or not from scratch when a run of a certain size has been so far enumerated. For extensive reviews of the runs literature we refer to [6-8]. The topic is still active and attractive too, because of the wide range of its application in many areas of applied probability and engineering including hypothesis testing, quality control, system reliability and financial engineering. Some recent contributions on the subject, among others, are the works of [9] – [22].

Let $\{X_i\}_{i\ge1}$ be a sequence of binary (two-state) random variables (RVs) taking on the values zero (0) or one (1) ordered on a line. According to Mood's [23] enumeration scheme a run of 1s (1-run) is defined to be a sequence of consecutive 1s preceded and succeeded by 0s or by nothing. The number of 1s in a 1-run is referred to as its length (or size). For a positive integer k, let $E_{n,k}$ denote the number of 1-runs of length exactly k in the first n, $n \ge k \ge 1$, binary trials. Following Makri and Psillakis [19], we use the indicator functions $U_j = (1 - X_{j-k})(1 - X_{j+1}) \prod_{i=j-k+1}^{j} X_i$, ending at $j, k \le j \le n$, with the convention $X_0 \equiv X_{n+1} = 0$. Consequently, the statistic $E_{n,k}$ can be expressed as $E_{n,k} = \sum_{j=k}^{n} U_j$. The RV $E_{n,k}$, which is a fundamental one in the run literature, besides its independent merit, may be used for the representation of other interesting statistics, too. Among them, the following two have been frequently discussed in the literature and in particular in financial engineering and bioinformatics [4, 17-18]. They refer to 1-runs of length exceeding a positive integer k, $1 \le k \le n$, in n binary trials, and they are the number $G_{n,k}$ of 1-runs of length at least k; $G_{n,k} = \sum_{i=k}^{n} E_{n,i}$, and the number $S_{n,k}$ of 1s in all 1-runs of length at least k; $S_{n,k} = \sum_{i=k}^{n} iE_{n,i}$. An alternative interpretation of $S_{n,k}$ is that it denotes the sum of the lengths of the 1-runs of length greater than or equal to k. The statistics $E_{n,k}, G_{n,k}, S_{n,k}$ have been studied on binary sequences of several internal structures by many researches who used various methods. See, e.g. [1, 4, 9, 11-14, 17-19, 23-30].

In this brief note we show how someone can easily enumerate explicitly the (total) number of occurrences of all 1-runs associated with the first two mentioned statistics, as well as, the (total) number of 1s according to the third one, in all possible 2^n binary strings of length n. Our approach is relied on simple and efficient probabilistic arguments. It provides an

alternative way to recapture explicit formulae for numbers associated with 1-runs and it also establishes a new explicit expression for the number of 1s in certain 1-runs. A unified recursive scheme for these numbers is provided, too.

2. Main Results

Let $X_{n,k}^{(a)}$ stand for the RVs $E_{n,k}, G_{n,k}, S_{n,k}$ for $a = e, g, s$, respectively. The support (range set) of $X_{n,k}^{(a)}$ is

$$S(X_{n,k}^{(a)}) = \begin{cases} \{0,1,\dots,\lfloor (n+1)/(k+1) \rfloor\} \; for \; a = e, g \\ \{0, k, k+1, \dots, n\} \; for \; a = s \end{cases}. \quad (1)$$

Next, we consider a sequence $\{X_i\}_{i=1}^{n}$ of length n of independent (i.e. derived by a memoryless source) and identically distributed 0-1 RVs with a common probability of 1s p; i.e. $p = P(X_i = 1) = 1 - P(X_i = 0) = 1 - q$, $i = 1,2,\dots,n$. Such a sequence, called a finite Bernoulli sequence, is of particular importance in studies of applied probability because of its simplicity, and also since it may be considered as a special case of a sequence with dependent elements; e.g. a Markovian or an exchangeable one.

For a Bernoulli sequence of length n, let $f^{(a)}(x; n, k; p)$ denote the probability mass function (PMF) of the RV $X_{n,k}^{(a)}$; i.e.

$$f^{(a)}(x; n, k; p) = P(X_{n,k}^{(a)} = x), x \in S(X_{n,k}^{(a)}), a = e, g, s. \quad (2)$$

Then, the expected value of $X_{n,k}^{(a)}$,

$$E\left(X_{n,k}^{(a)}; p\right) = \sum_{x \in S(X_{n,k}^{(a)})} x f^{(a)}(x; n, k; p), \quad (3)$$

is given by (see Makri et al. [11])

$$E\left(X_{n,k}^{(a)}; p\right) = \begin{cases} 0, a = e, g, s, for \; k > n \\ p^n, a = e, for \; k = n \\ qp^k[2 + (n-k-1)q], a = e, for \; 1 \le k \le n-1 \\ p^k[1 + (n-k)q], a = g, for \; 1 \le k \le n \\ p^k[k + (n-k)(kq+p)], a = s, for \; 1 \le k \le n. \end{cases} \quad (4)$$

In the sequel, we consider a symmetric $(p = 1/2)$ finite Bernoulli sequence (i.e. a finite binary string) of length n for which we obtain our main results. Since the cardinality of a proper sample space is 2^n (i.e. there are 2^n binary strings that are equally likely to occur) the classical definition of probability implies that

$$f^{(a)}(x; n, k; 1/2) = N_{n,k;x}^{(a)}/2^n. \quad (5)$$

The numbers $N_{n,k;x}^{(a)}, x \in S(X_{n,k}^{(a)})$, for $a = e, g, s$ admit the following interpretation: (i) $N_{n,k;x}^{(a)}, a = e, g$ is the number of all binary strings of length n with exactly x, $x = 0,1,\dots,\lfloor (n+1)/(k+1) \rfloor$, 1-runs of length [exactly $(a = e)$, at least $(a = g)$] k, among all the 2^n possible binary strings of length n, $n \ge k \ge 1$. (ii) $N_{n,k;x}^{(s)}$ is the number of all binary strings of length n with exactly x, $x = 0, k, k+1, \dots, n$, 1s contained in all 1-runs of length at least k, among all the 2^n possible binary strings of length n, $n \ge k \ge 1$.

Simple explicit expressions of $N_{n,k;x}^{(a)}$ (not repeated here) are given in Makri and Psillakis [19]. The authors provided an explicit formula of $N_{n,k;x}^{(a)}$, for $a = e, g, s$ in terms of binomial coefficients. Their method is based on the solution of a combinatorial problem; specifically, the allocation of balls into cells under certain constrains (see Lemma 2.2 of [11]). An explicit expression of $N_{n,k;x}^{(e)}$, in terms of binomial coefficients too, is given by Sinha and Sinha [1] who used a generating function approach. The latter expression contains an additional sum; therefore it may be evaluated slower computationally than that provided in [19].

The numbers $N_{n,k;x}^{(a)}$ allow us to establish (we do not actually need their specific expressions according to the new proposed approach) respective numbers referring to all possible 2^n binary strings of length n. They are defined as

$$R_{n,k}^{(a)} = \sum_{x \in S(X_{n,k}^{(a)})} x N_{n,k;x}^{(a)}, \quad (6)$$

i.e. $R_{n,k}^{(a)}$ is the total number of occurrences of all 1-runs of length [exactly $(a = e)$, at least $(a = g)$] k, and the total number of 1s in all 1-runs of length at least k $[a = s]$, in all possible 2^n binary strings of length n, $n \ge k \ge 1$.

Since $E\left(X_{n,k}^{(a)}; 1/2\right) = \sum_{x \in S(X_{n,k}^{(a)})} x f^{(a)}(x; n, k; 1/2)$, hence by (5) and (6), $R_{n,k}^{(a)} = 2^n E(X_{n,k}^{(a)}; 1/2)$. Therefore (4) implies

$$R_{n,k}^{(a)} = \begin{cases} 1, a = e, g; n, a = s, for \; k = n \\ 2, a = e; 3, a = g; 3n-2, a = s, for \; k = n-1 \\ (n-k+3)2^{n-k-2}, a = e, for \; 1 \le k \le n-2 \\ (n-k+2)2^{n-k-1}, a = g, for \; 1 \le k \le n-2 \\ [n(k+1) - k(k-1)]2^{n-k-1}, a = s, for \; 1 \le k \le n-2. \end{cases} \quad (7)$$

Readily, by symmetry, $R_{n,k}^{(a)}$ provides the respective numbers associated with 0-runs and 0s in 0-runs.

By (7) it is noted that for a fixed n, $R_{n,k}^{(a)}$, $a = e, g, s$ decreases exponentially as k increases, and for a fixed $k, 1 \le k \le n-2$, as $n \to \infty$, it holds

$$R_{n,k}^{(g)}/R_{n,k}^{(e)} \to 2, R_{n,k}^{(s)}/R_{n,k}^{(g)} \to k+1, R_{n,k}^{(s)}/R_{n,k}^{(e)} \to 2(k+1). \quad (8)$$

Furthermore, $R_{n,k}^{(e)} \le R_{n,k}^{(g)} \le R_{n,k}^{(s)}$, since $E_{n,k} \le G_{n,k} \le S_{n,k}$, $1 \le k \le n$.

Table I presents the three numbers $R_{n,k}^{(e)}, R_{n,k}^{(g)}, R_{n,k}^{(s)}$ in binary strings of length $n = 2^l$ bits, $l = 2,4$, and for $1 \le k \le n$. The entries of the table confirm the previously noted behavior of the depicted numbers.

Sinha [31] was the first who addressed the usefulness of the number $R_{n,k}^{(e)}$ and also provided its formula. Then, Sinha and Sinha [1] obtained an explicit expression of $R_{n,k}^{(e)}$ whereas

Makri and Psillakis [19] derived the same formula for it and they also established an explicit expression of $R_{n,k}^{(g)}$. In both papers ([1] and [19]) the approach was relied on the definition of $R_{n,k}^{(a)}$ via $N_{n,k;x}^{(a)}$, $a = e, g$. Recently, Sinha and Sinha [2] reestablished $R_{n,k}^{(e)}$ solving explicitly a recursive generation scheme for it.

The proposed, in the present note, approach is a new one and treats under the same frame all the numbers $R_{n,k}^{(a)}$ in a simple, unified and systematic way. Accordingly, by (7) we effortless get a recursive scheme for $R_{n,k}^{(a)}$, $a = e, g, s$. It gives a way to generate $R_{n,k}^{(a)}$ from $R_{n,k+1}^{(a)}$ and it offers further insight in understanding the interdependencies among the studied numbers. Specifically, it holds

$$R_{n,k}^{(a)} = \begin{cases} 1, a = e, g; n, a = s, for\ k = n \\ 2, a = e; 3, a = g; 3n - 2, a = s, for\ k = n - 1 \\ 2R_{n,k+1}^{(a)} + 2^{n-k-1}A_{n,k}^{(a)}, a = e, g, s, for\ k = 1,2, \dots, n - 2, \end{cases} \quad (9)$$

with $A_{n,k}^{(a)} = 2^{-1}, a = e; 1, a = g; 2k - n, a = s$. We note that for the particular case $a = e$ we capture by the relevant entries of (9) and (7), Theorems 2 and 3 of [2], respectively.

TABLE I. NUMBERS OF OCCURRENCES OF 1-RUNS, $R_{n,k}^{(e)}, R_{n,k}^{(g)}$ AND NUMBER OF 1S, $R_{n,k}^{(s)}$, IN BINARY STRINGS OF LENGTH n

n	k	$R_{n,k}^{(e)}$	$R_{n,k}^{(g)}$	$R_{n,k}^{(s)}$
4	1	12	20	32
	2	5	8	20
	3	2	3	10
	4	1	1	4
16	1	147456	278528	524288
	2	69632	131072	376832
	3	32768	61440	237568
	4	15360	28672	139264
	5	7168	13312	77824
	6	3328	6144	41984
	7	1536	2816	22016
	8	704	1280	11264
	9	320	576	5632
	10	144	256	2752
	11	64	112	1312
	12	28	48	608
	13	12	20	272
	14	5	8	116
	15	2	3	46
	16	1	1	16

3. Conlusions

In this note we stated three run statistics which are important in many areas of applied probability. We defined them on a binary (0-1) sequence, and we then provided explicitly their mean values for a Bernoulli sequence. After that, we considered binary strings (symmetric Bernoulli sequences) and we showed how the analytic expressions of the means of these RVs provide eventually the respective explicit expressions of three numbers studied recently by different methods. Finally, as a byproduct of our approach, we proposed a unified recursive scheme which clarifies further the interdependencies among these numbers. The examined numbers are potential useful in many engineering applications like the ones mentioned briefly in the Introduction. Early results are encouraging in this direction.

REFERENCES

[1] K. Sinha and B. P. Sinha, "On the distribution of runs of ones in binary strings," Comput. Math. Appl., vol. 58, pp. 1816-1829, 2009.

[2] K. Sinha and B. P. Sinha, "Energy-efficient communication: understanding the distribution of runs in binary strings," 1st International Conference on Recent Advances in Information Technology (RAIT-2012), pp. 177-181, 2012.

[3] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences," Nucleic Acids Res., vol. 27, pp. 573-580, 1999.

[4] W.Y. W. Lou, "The exact distribution of the "k-tuple statistic for sequence homoloy," Statist. Probab. Lett., vol. 61, pp. 51-59, 2003.

[5] G. Nuel, L. Regad, J. Martin and A.C. Camproux, "Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data," Alg. Mol. Biol., vol. 5, pp. 1-18, 2010.

[6] N. Balakrishnan and M. V. Koutras, Runs and Scans with Applications, New York: Wiley, 2002.

[7] J. C. Fu and W. Y. W. Lou, Distribution Theory of Runs and Patterns and its Applications: A Finite Markov Imbedding Approach, New Jersey: World Scientific, 2003.

[8] M. V. Koutras, "Applications of Markov chains to the distribution theory of runs and patterns," in Handbook of Statistics, vol. 21, D. N. Shanbhag, C. R. Rao, Eds. North Holland: Elsevier, 2003, pp. 431-472.

[9] S. Eryilmaz, "Success runs in a sequence of exchangeable binary trials," J. Statist. Plann. Inference, vol. 137, pp. 2954-2963, 2007.

[10] 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. Statist. Theory Methods, vol. 36, pp. 657-668, 2007.

[11] 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.

[12] S. Eryilmaz, "Run statistics defined on the multicolor urn model," J. Appl. Probab., vol. 45, pp. 1007-1023, 2008.

[13] S. Demir and S. Eryilmaz, "Run statistics in a sequence of arbitrarily dependent binary trials," Stat. Papers, vol. 51, pp. 959-973, 2010.

[14] K. Inoue and S. Aki, "On the conditional and unconditional distributions of the number of success runs on a circle with applications," Statist. Probab. Lett., vol. 80, pp. 874-885, 2010.

[15] F. S. Makri, "On occurences of $F-S$ strings in linearly and circularly ordered binary sequences," J. Appl. Probab., vol. 47, pp. 157-178, 2010.

[16] F. S. Makri, "Minimum and maximum distances between failures in binary sequences," Statist. Probab. Lett., vol. 81, pp. 402-410, 2011.

[17] F. S. Makri and Z. M. Psillakis, "On runs of length exceeding a threshold: normal approximation," Stat. Papers, vol. 52, pp. 531-551, 2011.

[18] F. S. Makri and Z. M. Psillakis, "On success runs of length exceeded a threshold," Methodol. Comput. Appl. Probab., vol. 13, pp. 269-305, 2011.

[19] F. S. Makri and Z. M. Psillakis, "On success runs of a fixed length in Bernoull sequences: exact and aymptotic results," Comput. Math. Appl., vol. 61, pp. 761-772, 2011.

[20] S. D. Dafnis, A. N. Philippou and D. L. Anztoulakos, "Distributions of patterns of two successes separeted by a string of $k-2$ failures," Stat. Papers, vol. 53, pp. 323-344, 2012.

[21] M. V. Koutras and F. S. Milienos, "Exact and asymptotic results for pattern waiting times," J. Statist. Plann. Inference, vol. 142, pp. 1464-1479, 2012.

[22] F. S. Makri and Z. M. Psillakis, "Counting certain binary strings," J. Statist. Plann. Inference, vol. 142, pp. 908-924, 2012.

[23] A. M. Mood, "The distribution theory of runs," Ann. Math. Stat., vol. 11, pp. 367-392, 1940.

[24] J. C. Fu and M. V. Koutras, "Distribution theory of runs: a Markov chain approach," J. Amer. Statist. Assoc., vol. 89, pp. 1050-1058, 1994.

[25] D. L. Antzoulakos, "On waiting time problems associated with runs in Markov dependent trials," Ann. Inst. Statist. Math., vol. 51, pp. 323-330, 1999.

[26] Q. Han and S. Aki, "Joint distributions of runs in a sequence of multi-trials," Ann. Inst. Statist. Math., vol. 51, pp. 419-447, 1999.

[27] J. C. Fu, W. Y. W. Lou, Z. Bai and G. Li, "The exact and limiting distributions for the number of successes in success runs within a sequence of Markov-dependent two-state trials," Ann. Inst. Statist. Math., vol. 54, pp. 719-730, 2002.

[28] K. Sen, M. L. Agarwal and S. Chakraborty, "Lenths of runs and waiting time distributions by using Polya-Eggenberger sampling scheme," Studia Sci. Math. Hungar., vol. 2, pp. 309-332, 2002.

[29] D. L. Antzoulakos, S. Bersimis and M. V. Koutras, "On the distribution of the total number of run lengths," Ann. Inst. Statist. Math., vol. 55, pp. 865-884, 2003.

[30] D. E. Martin, "Distribution of the number of successes in success runs of length at least k in higher-order Markovian sequences," Methodol. Comput. Appl. Probab., vol. 7, pp. 543-554, 2005.

[31] K. Sinha, Location and communication issues in mobile networks. Ph. D. Dissertaion, Department of Computer Science and Engineering, Jadavpur University, Calcutta, India , 2007.

