Journal of Statistical Planning and Inference 142 (2012) 908–924

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

Counting certain binary strings

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

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

ARTICLE INFO

Article history:
Received 16 October 2010
Received in revised form
8 June 2011
Accepted 27 October 2011
Available online 10 November 2011

Keywords:
Binary trials
Exchangeable trials
Independent trials
Runs
Strings
Waiting time
Urn models
Records
Non-parametric tests of randomness

ABSTRACT

Consider a sequence of exchangeable or independent binary (i.e. zero-one) random variables. Numbers of strings with a fixed number of ones between two subsequent zeros are studied under an overlapping enumeration scheme. The respective waiting times are examined as well. Exact probability functions are obtained by means of combinatorial analysis and via recursive schemes in the case of an exchangeable and of an independent sequence, respectively. Explicit formulae for the mean values and variances of the number of strings are provided for both types of the sequences. For a Bernoulli sequence the asymptotic normality of the numbers of strings is established too. Indicative exchangeable and independent sequences, combined with numerical examples, clarify further the theoretical results.

1. Introduction and preliminaries

Recently, some research associated with the number of patterns which are composed of runs of ones between subsequent zeros in binary sequences have appeared in the literature. The patterns are counted according to at most/exactly/at least, overlapping/non-overlapping enumerating scheme and they are defined on binary sequences of several internal structures; e.g. sequences of independently (identically/non-identically) distributed elements or dependently distributed elements in a Markovian fashion of some order. The methods used to derive exact/limiting, marginal/joint probability distributions include combinatorial analysis, generating functions, Markov chain imbedding technique (MCIT) and recursion schemes. See, for instance Koutras (1996), Antzoulakos (2001), Sarkar et al. (2004), Holst (2007), Dafnis et al. (2010b) and the references therein. Although some of these numbers have been studied in the past in connection to particular cases of run and scan statistics (see, e.g. Glaz and Balakrishnan, 1999; Balakrishnan and Koutras, 2002; Fu and Lou, 2003) the prementioned papers provide alternative formulae and suggest additional meaning and possible applications. For a recent literature on runs/scans and related statistics in a sequence of exchangeable binary trials see, e.g. Eryilmaz (2010, 2011) and Inoue et al. (2011).

Let a sequence of binary random variables (RVs) $X_1, X_2, \ldots$ taking the values $x_i = 1$ or 0 for $i = 1, 2, \ldots$. In such a sequence the one (1) may denote either a success (S) or a failure (F) depending on the pattern of interest. The 0–1 sequence considered in the article may be a Poisson sequence (that is, a sequence of independently distributed binary RVs) with

* Corresponding author. Tel.: +30 2610 996738.
E-mail addresses: makri@math.upatras.gr (F.S. Makri), psillaki@physics.upatras.gr (Z.M. Psillakis).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.10.015

$p_i = P(X_i = 1) = 1 - q_i = 1 - P(X_i = 0)$, $i = 1, 2, \ldots$ or an exchangeable (or symmetrically dependent) sequence (that is, a sequence of binary RVs the joint distribution of which is invariant under permutation of its arguments) with $p_{n,y} = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ for $n > 0$ and $y = n - \sum_{i=1}^{n} x_i$. A Bernoulli sequence (that is, a sequence of independent and identically distributed (iid) RVs with a common probability of 1's, p) is a particular Poisson sequence with $p_i = p = 1 - q$ or an exchangeable sequence with $p_{n,y} = p^{n-y} q^y$, $n = 1, 2, \ldots$.

A d-string is a string $0\underbrace{11\ldots1}_{d \geq 0}0$ of exactly d consecutive ones (i.e. a run of ones of length d) between two subsequent zeros. The RV counting-in the overlapping sense (that is, a zero which is not at either end of the sequence can contribute at most to two strings) the number of such strings in the first n ($n \geq 2$) trials is denoted by $M_{d,n}$. Including the string $0\underbrace{11\ldots1}_{d \geq 0}$ with exactly d ones after the last zero in the count we obtain the RV $N_{d,n}$.

As an illustration, we consider the first $n = 20$ outcomes of a binary sequence 10010110011101010111. Then we have $M_{0,20} = N_{0,20} = 2$, $M_{1,20} = N_{1,20} = 3$, $M_{2,20} = N_{2,20} = 1$, $M_{3,20} = 1$, $N_{3,20} = 2$, $M_{d,20} = N_{d,20} = 0$, $d > 3$.

The waiting time RVs, respective to $M_{d,n}$ and $N_{d,n}$, $W_{d,r}^{(M)}$ and $W_{d,r}^{(N)}$, denote the number of trials required for the r-th, $r \geq 1$, occurrence of the corresponding strings. The RVs $M_{d,n}$ and $W_{d,r}^{(M)}$ are related via the dual relationship

$$W_{d,r}^{(M)} > n \quad \text{iff } M_{d,n} < r, \ r \geq 1, \ n \geq 2, \tag{1}$$

which offers an alternative way of obtaining results for $W_{d,r}^{(M)}$ through formulae established for $M_{d,n}$ and vice versa.

The RVs $M_{d,n}$ and $W_{d,r}^{(M)}$ were formally introduced by Sarkar et al. (2004) and Sen and Goyal (2004) whereas $N_{d,n}$ was introduced by Holst (2007). For $d = 0$, $M_{d,n}$ is the Ling's (1988) RV $M_n^{(k)}$, for $k = 2$, counting the number of overlapping runs of zeros of length 2. For $d \geq 1$, let $E_{n,d}$ denote the number of runs of ones of length exactly equal to d in n binary trials (see, e.g. Mood, 1940). Then it holds: $M_{d,n} \leq N_{d,n} \leq E_{n,d}$, $d = 1, 2, \ldots, n-2$. Actually, for $d \geq 1$, $E_{n,d}$ counts the number of runs of 1's of length d whereas $M_{d,n}$ considers the runs of 1's as distances of size d between two subsequent 0's, counting them as the number of 1's, ignoring possible occurrences of runs of 1's of length d at the beginning and at the end of the sequence. Of course, there are some cases for which the two RVs take the same value, e.g. when the sequence starts and ends with a zero. An analogous interpretation also holds for $N_{d,n}$ which for $d \geq 1$ admits an additional meaning under the general framework of (k_1, k_2)-events; see, e.g. Dafnis et al. (2010a). If we consider the (k_1, k_2)-event: at least k_1 consecutive 0's are followed by exactly k_2 consecutive 1's, with k_1, k_2 positive integers, then $N_{d,n}$, $d \geq 1$, enumerates $(1,d)$-events. Although the values of $M_{d,n}$ and $N_{d,n}$ do not differ more than 1 in a binary sequence, $M_{d,n}$ does not admit such an interpretation.

The previous discussion clarifies that the RVs $M_{d,n}$ and $N_{d,n}$ enumerate 0–1 patterns (strings) and not just runs of either 1's or 0's as their related RVs $M_n^{(2)}$ and $E_{n,d}$ do. Besides, they are flexible since they combine, in a simple manner, the main characteristics of the latter RVs; specifically, the overlapping counting in the sense of $M_n^{(k)}$ and the definition of a run as an uninterrupted sequence of same symbols in the sense of $E_{n,k}$. These features motivate their use in several areas of applied research associated with the study of cycle lengths in random permutations and record models in financial engineering as well as with urn models used in population genetics, health sciences and evolution of species. In these studies, of both theoretical and practical interest, certain sequences of independent and exchangeable binary RVs are used as models. For example, let us define as records in a given sequence of numbers the elements in the sequence with values strictly larger than all previous ones (see, e.g. Nevzorov, 2001). Then, $M_{d,n}$, $d \geq 1$, counts how many isolated ($d = 1$), double ($d = 2$), triple ($d = 3$) etc. records occur between two non-records in a sequence representing for instance financial data (e.g. exchange rates or stock market prices). The number of two consecutive non-records, possibly overlapping, in the sequence is enumerated by $M_{0,n}$. Another example is the connection of the joint distribution of $N_{d,n}$, $d = 0, 1, \ldots, n-1$ to Ewens sampling formula in population genetics which can be obtained conceptually from a Polya-like urn scheme (see Mahmoud, 2009).

Finally, $M_{d,n}$ and $N_{d,n}$, like other RVs related to runs and scans (see, e.g. Koutras and Alexandrou, 1997; Boutsikas and Koutras, 2002; Antzoulakos et al., 2003) might be useful in the context of testing (a) of randomness of a series of observations, reduced to a binary sequence according to some rule, where the null hypothesis of iid is tested against an alternative hypothesis of some kind of dependence, and (b) of similarity between two sequences of observations where a success is interpreted as a match of the sequences at a given position. To accomplish such postulates, a systematic study of the critical values of several significant levels and of the empirical powers of the tests using $M_{d,n}$ and $N_{d,n}$ compared to the respective ones of other popular statistics is required. Such a study, which also depends on the tested binary sequences, would reveal advantages/disadvantages of the employed statistics.

Sen and Goyal (2004) studied, via combinatorial methods, the exact distribution of $M_{d,n}$ and $W_{d,r}^{(M)}$ defined on Bernoulli sequences whereas Sarkar et al. (2004) studied the exact and asymptotic distributions of them defined on homogeneous Markov chains of some order. In a series of papers Holst (2007, 2008a,b, 2009, 2010) deriving recursions for the binomial moments and also using an embedding in a marked Poisson process, systematically examined exact and asymptotic probability distributions of $M_{d,n}$ and $N_{d,n}$, $d = 0, 1, \ldots$ and related RVs, defined on certain Poisson sequences including as a particular case sequences of records. He also unified previous results due to Chern et al. (2000), Mori (2001) and Chern and Hwang (2005) among others. Employing an enriched version of MCIT (see, Koutras and Alexandrou, 1995), Dafnis et al. (2010b) studied the exact distributions of $M_{d,n}$ and $W_{d,r}^{(M)}$ defined on Bernoulli sequences and Dafnis and Philippou (2010) examined the exact distribution of $W_{d,r}^{(M)}$ defined on a homogeneous Markov chain of first order via the same method.

In this paper we study the RVs $M_{d,n}$, $N_{d,n}$, $W_{d,r}^{(M)}$ and $W_{d,r}^{(N)}$ defined on binary sequences of independent and exchangeable RVs. In Section 2, exact probability mass functions (PMFs) of these RVs are derived by means of combinatorial analysis in