

# On the asymptotic distribution of the discrete scan statistic

Michael V. Boutsikas and Markos V. Koutras

Department of Statistics and Insurance Science, University of Piraeus, Greece

July 11, 2006

## Abstract

The discrete scan statistic in a binary (0-1) sequence of  $n$  trials is defined as the maximum number of successes within any  $k$  consecutive trials ( $n \geq k$  are two positive integers). It has been used in many areas of science (quality control, molecular biology, psychology, etc.) to test the null hypothesis of uniformity against a clustering alternative. In this article we provide a compound Poisson approximation which is subsequently used to establish asymptotic results for the distribution of the discrete scan statistic when  $n, k \rightarrow \infty$  and the success probability of the trials is kept fixed. An extreme value theorem is also provided for the celebrated Erdős-Rényi Statistic.

**Keywords:** Discrete scan statistic, compound Poisson approximation, randomness test, Erdős-Rényi statistic, Kolmogorov distance, extreme value theorem.

**MSC 2000:** Primary 62E17, 60F05, Secondary: 62E20, 60F10.

## 1 Introduction

Scientists dealing with experimental data modeled by independent Bernoulli trials frequently seek reasonable criteria providing clustering evidence (lack of randomness) or detecting changes in the underlying process. The length of the longest success run is definitely a very powerful statistic for studying problems of this nature, a fact that explains the continuing interest in its probabilistic characteristics since de Moivre's era (17th century). A natural and intuitively appealing generalization of the success run principle arises if instead of looking at pure success runs we consider the maximum number of successes within any  $k$  contiguous (consecutive) trials. The resulting random variable is usually referred in the literature as *binary discrete scan statistic* and has widespread applicability in a significant number of scientific areas such as quality control, molecular biology,

psychology, epidemiological studies, reliability theory etc; see Balakrishnan and Koutras (2002), Glaz and Balakrishnan (1999), Glaz, Naus and Wallenstein (2001) and Fu and Lou (2003).

To fix our notation, let  $X_i, i \in \mathbb{Z}$  be a sequence of iid binary r.v.'s with

$$P(X_i = 1) = p, \quad P(X_i = 0) = q = 1 - p, \quad i = 1, 2, \dots, n$$

and denote by

$$S_i = \sum_{j=i}^{i+k-1} X_j, \quad i \in \mathbb{Z}$$

the  $k$ -scan process (moving window of length  $k \geq 1$ ) generated by the sequence  $X_i, i \in \mathbb{Z}$ . Then the discrete scan statistic is defined as

$$S_{n,k} = \max_{1 \leq i \leq n-k+1} S_i = \max_{1 \leq i \leq n-k+1} \sum_{j=i}^{i+k-1} X_j$$

where  $n \geq k$  is a fixed positive integer.

An instance where  $S_{n,k}$  arises in quite a natural way is in randomness tests when the null hypothesis of uniformity and independence of  $X_i, i = 1, 2, \dots, n$  is to be tested against the alternative hypothesis of clustering of 1's due to local positive dependence between  $X_i, i = 1, 2, \dots, n$ , or due to the existence of subsequences of consecutive  $X_i$ 's with  $P(X_i = 1) > p$ . As Glaz and Naus (1991) indicated, the generalized likelihood ratio test for checking the hypothesis of uniformity, rejects the null hypothesis of uniformity whenever  $S_{n,k} \geq c$ , with the value of  $c$  being determined by the significance level of the test. Recently, Glaz and Zhang (2004) introduced an alternative more sensitive procedure exploiting a multiple scan statistic of variable size instead of the single (fixed window length) scan statistic  $S_{n,k}$ .

Apparently, the evaluation of  $c$  so that a prespecified significance level is acquired calls for the distribution of the test statistic  $S_{n,k}$ . Since randomness tests are frequently applied to large data sets, theoretical developments related to the asymptotic distribution of  $S_{n,k}$  (as  $n, k \rightarrow \infty$ ) will play a primary role in the analysis of the test.

Another instance where  $S_{n,k}$  could be used is offered by the following model which originates from molecular biology. In the study of amino acid sequences, various classification schemes are in common use, including a chemical alphabet of 8 letters, a functional alphabet of 4 letters, a charge alphabet of 3 letters etc. In order to introduce quantitative means for assessing and interpreting genomic inhomogeneities between sequences of different species or sequences subject to different

chemical infections and/or several levels of corruption, molecular biologists look for long aligned subsequences that match in most of their positions and try to specify what is an unusually long match. In order to construct an appropriate mathematical model, let  $Z_{i1}, Z_{i2}, i = 1, 2, \dots, n$  be two amino acid sequences from a finite alphabet  $A = \{a_1, a_2, \dots, a_l\}$  with  $\mu_j = P(Z_{i1} = a_j) = P(Z_{i2} = a_j)$  for  $j = 1, 2, \dots, l$ . The two sequences will be said to match in position  $i \in \{1, 2, \dots, n\}$  if  $Z_{i1} = Z_{i2}$ , in which case we let  $X_i$  be 1 (and 0 otherwise). Then  $X_i, i = 1, 2, \dots, n$  form a sequence of binary i.i.d. r.v.'s with success probabilities

$$p = P(X_i = 1) = P(Z_{i1} = Z_{i2}) = \sum_{j=1}^l \mu_j^2$$

and the number of matches over a window of length  $k$  will be described by the corresponding  $k$ -scan process  $S_i, i = 1, 2, \dots, n$ . Moreover, a "near perfect" match at position  $i$  can be described by the event  $S_i \geq c$ , with  $c$  being an integer, close enough to  $k$ . It is clear that the condition

$$S_{n,k} = \max_{1 \leq i \leq n-k+1} S_i < c$$

can then be used as an evidence of lack of local match between the two sequences under inspection. It should be stressed that, in this application we are also interested in large values of  $n$  (long amino acid sequences) and  $k$  (long matching regions).

As a final example we provide the following actuarial model. Let  $Z_i, i = 1, 2, \dots, n$  be the daily claim sizes over an  $n$ -day period and  $u \geq 0$  a given threshold. Assume that  $Z_i$  are iid r.v.'s with cdf  $F$  and denote by

$$X_i = I_{(u, \infty)}(Z_i) = \begin{cases} 1 & \text{if } Z_i > u \\ 0 & \text{if } Z_i \leq u \end{cases}, \quad i = 1, 2, \dots, n$$

the corresponding random variables which indicate whether the  $i$ -th claim exceeds threshold  $u$  or not. Then

$$P(X_i = 1) = E(X_i) = P(Z_i > u) = 1 - F(u) = p, \quad i = 1, 2, \dots, n$$

and  $S_{n,k}$  will describe the maximum number of "large claims" (i.e. claims exceeding threshold  $u$ ) in a period of  $k$  consecutive days. Since the primary interest in this situation is also focused in extremely long periods ( $n \rightarrow \infty, k \rightarrow \infty$ ) one should look at the asymptotic distribution of  $S_{n,k}$ .

In all the aforementioned examples, it is clear that, the study of the underlying model calls for the investigation of the distribution of the random variable  $S_{n,k}$ . Exact results for the distribution

of the scan statistic are discussed in Fu (2001), Balakrishnan and Koutras (2002) and Fu and Lou (2003). Since the evaluation of the exact distribution is computationally intractable, especially for large values of the parameters, several approximations and bounds have been developed during the last decade. The interested reader may refer to the recent monographs by Glaz, Naus and Wallenstein (2001) and Balakrishnan and Koutras (2002) for up to date review on this topic.

In a recent article by Boutsikas and Koutras (2002), a compound Poisson approximation was established for the distribution of the enumerating random variable

$$W_n = \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_i).$$

As a by-product, an approximation for

$$P(S_{n,k} < r) = P(W_n = 0) \tag{1}$$

was established, along with an upper bound for the error incurred by it. However, the asymptotic result given there, holds true under the conditions  $n \rightarrow \infty, p \rightarrow 0$  and  $k, r$  fixed, which are of no interest for the examples mentioned before. One might suspect that, even in the case of interest ( $p$  fixed and  $n, k \rightarrow \infty$ ) a compound Poisson law is hidden below the surface, yet the hooks provided by the results of Boutsikas and Koutras (2002) leave us without a catch. This is due to the fact that, for  $r < k$ , the upper bound appearing there is of order  $O(p)$  and therefore it does not converge to 0 as  $n, k \rightarrow \infty$  while  $p$  is fixed.

In the present article, motivated by the abovementioned remarks, we establish a new CP approximation for  $W_n$  that offers an upper bound manageable under the conditions of interest.

In Section 2, we introduce all necessary notation and preliminary material. In Section 3, exploiting an appropriate declumping technique, a compound Poisson approximation for the distribution of  $W_n$  is developed, along with tight upper bounds for the Kolmogorov distance between the distribution of  $W_n$  and the approximating distribution. In Section 4 an asymptotic result for the distribution of the scan statistic  $S_{n,k}$  is established while Section 5 presents an extreme value theorem for the same statistic, that is comparable to the well known Erdős-Rényi results (when applied to binary sequences). Finally, in Section 6 an extensive numerical experimentation is carried out in order to investigate the quality of the approximations and bounds.

## 2 Preliminaries

The Kolmogorov distance between the distributions of two random variables  $X$  and  $Y$  is defined as

$$d(X, Y) = \sup_w |P(X \leq w) - P(Y \leq w)|,$$

and offers a very efficient tool for establishing convergence in distribution; a sequence of random variables converges weakly to  $Y$  if the corresponding sequence of distances converges to 0. By the term compound Poisson distribution  $CP(\lambda, H)$  with parameter  $\lambda$  and compounding distribution  $H$ , we shall refer to the distribution of a random sum of the form  $\sum_{i=1}^N Z_i$  with  $N$  being a Poisson random variable with  $\lambda = E(N)$  and  $Z_i$  being i.i.d random variables (also independent of  $N$ ) whose distribution function is  $H$ .

The main result of the next section is an application of a general theorem on compound Poisson approximation published by Boutsikas and Koutras (2001). For the purposes of the present exposition, we shall retain a simplified version of their result which is more than adequate to meet our needs.

Consider first a sequence of non-negative random variables  $Z_a, a = 1, 2, \dots$ . For each  $a = 2, 3, \dots$  introduce a subset  $B_a$  of  $\{1, 2, \dots, a - 1\}$  (left neighborhood of dependence of  $Z_a$ ) so that  $Z_a$  is independent of all  $Z_b, b \in \{1, 2, \dots, a - 1\} \setminus B_a$ . The next theorem provides an upper bound for the Kolmogorov distance between the distribution of the sum  $\sum_{a=1}^\nu Z_a$  ( $\nu$  a fixed positive integer) and a compound Poisson distribution  $CP(\lambda, H)$  with suitably chosen  $\lambda, H$ .

**Theorem 1** (*Boutsikas and Koutras (2001)*). *If  $Z_a, a = 1, 2, \dots, \nu$  is a sequence of non-negative random variables, then*

$$d\left(\sum_{a=1}^\nu Z_a, CP(\lambda, H)\right) \leq \sum_{a=2}^\nu \left( P(Z_a > 0, \sum_{b \in B_a} Z_b > 0) + P(Z_a > 0)P\left(\sum_{b \in B_a} Z_b > 0\right) \right) \quad (2) \\ + \frac{1}{2} \sum_{i=1}^\nu P(Z_i > 0)^2,$$

where  $\lambda = \sum_{a=1}^\nu \lambda_a$ , and  $H(x) = \frac{1}{\lambda} \sum_{a=1}^\nu \lambda_a P(Z_a \leq x | Z_a > 0), x \in \mathbf{R}$  with  $\lambda_a = P(Z_a > 0), a = 1, 2, \dots, \nu$ .

Theorem 1 states that, if the random variables  $Z_a, a = 1, 2, \dots$  are "locally" dependent and the masses of their distributions are concentrated on 0, then  $\sum_{a=1}^\nu Z_a$  can be satisfactorily approximated by an appropriate compound Poisson distribution.

If  $X, Y$  are non-negative random variables, it is evident that  $|P(X = 0) - P(Y = 0)| \leq d(X, Y)$  and therefore,

$$\left| P\left(\sum_{a=1}^{\nu} Z_a = 0\right) - e^{-\lambda} \right|$$

is also upper bounded by the RHS in (2). It is worth stressing that, should one wish to establish bounds for  $P(\sum_{a=1}^{\nu} Z_a = 0)$  only (and not for the whole distribution of  $\sum_{a=1}^{\nu} Z_a$ ), there is no need to proceed to the calculation of the compounding distribution  $H$ .

Let now  $b(x; n, p), B(x; n, p)$  denote the probability mass function and cumulative distribution function respectively of a binomial random variable  $X$ , i.e.

$$b(x; n, p) = P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \dots, n$$

$$B(x; n, p) = P(X \leq x) = \sum_{r=0}^x b(r; n, p), \quad x \in R.$$

In the following sections we shall make frequent use of the quantities

$$f(s; k, p) = P(S_1 < s, S_2 < s, \dots, S_k < s, S_{k+1} \geq s) \quad (3)$$

$$G(s; k, p) = P(S_1 < s, S_2 < s, \dots, S_{k+1} < s) \quad (4)$$

which can be expressed through  $b(x; n, p), B(x; n, p)$  as follows (cf. Glaz and Naus (1991))

$$f(s; k, p) = \frac{p}{s} b(s-1; k-1, p) [sq \cdot b(s-1; k-1, p) + (s-kp)B(s-2; k-1, p)], \quad (5)$$

$$G(s; k, p) = (B(s-1; k, p))^2 - b(s; k, p) [(s-1)B(s-2; k, p) - kpB(s-3; k-1, p)], \quad (6)$$

for  $1 \leq s \leq k$  (if  $s > k$  or  $s < 0$  we set  $f(s, k; p) = 0$ ).

The standard notations  $\sim, o(\cdot), O(\cdot)$  will assume their usual meaning and  $I_A(\cdot)$  will denote the indicator function of the set  $A$ , i.e.

$$f(t) \sim g(t) \text{ as } t \rightarrow t_0 \text{ if } \lim_{t \rightarrow t_0} \frac{f(t)}{g(t)} = 1, \quad f(t) = o(g(t)) \text{ as } t \rightarrow t_0 \text{ if } \lim_{t \rightarrow t_0} \frac{f(t)}{g(t)} = 0,$$

$$f(t) = O(g(t)) \text{ if } \frac{f(t)}{g(t)} \text{ is bounded, } I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

In addition, summations of the form  $\sum_{i=a}^b x_i$  with  $a > b$ , will be assumed to vanish. Finally, we shall write  $\lfloor x \rfloor$  for the integer part of  $x$ .

### 3 An Approximation for the cumulative distribution function of $S_{n,k}$

As already stated after Theorem 1, should we wish to exploit (2) for establishing fine upper bounds (i.e. bounds converging to 0) for  $d(W_n, CP(\lambda, H))$  or simply for

$$\left| P(S_{n,k} < r) - e^{-\lambda} \right| = \left| P(W_n = 0) - e^{-\lambda} \right| = \left| P(I_{[r,\infty)}(S_a) = 0 \text{ for all } a = 1, \dots, n - k + 1) - e^{-\lambda} \right|,$$

care should be taken so that  $Z_a, a = 1, 2, \dots$  be locally dependent and their probability mass functions be concentrated on 0. Since scans exhibit a strong tendency to clustering (especially if  $p$  does not converge to 0), a direct application of Theorem 1 for  $Z_a = I_{[r,\infty)}(S_a), a = 1, 2, \dots$  does not yield powerful estimates for the approximation error. One convenient way to improve the performance of the upper bound is by introducing first a set of weakly dependent random variables  $Z_a = C_a, a = 1, 2, \dots$  so that a small upper bound is gained for  $d(\sum Z_a, CP(\lambda, H))$  through Theorem 1 and then making use of the triangle inequality

$$d(W_n, CP(\lambda, H)) \leq d(W_n, \sum Z_a) + d(\sum Z_a, CP(\lambda, H)). \quad (7)$$

It goes without saying that an efficient upper bound for the quantity  $d(W_n, \sum Z_a)$  will also be needed.

A set of random variables possessing the aforementioned properties is provided by

$$C_a = \left[ \prod_{j=a-k}^{a-1} (1 - I_{[r,\infty)}(S_j)) \right] \left[ I_{[r,\infty)}(S_a) \sum_{m=a}^{a+k} I_{[r,\infty)}(S_m) \right], \quad a = 1, 2, \dots$$

The second bracket enumerates the number of scanning windows of length  $k$  that begin at positions  $a, a+1, \dots, a+k$  and contain at least  $r$  successes each (such a random variable, that counts the total number of clumps located at a specific area, is usually called a declumping variable). On the other hand, the first bracket guarantees that, at the previous  $k$  positions  $a-k, a-k+1, \dots, a-1$ , all scanning windows of length  $k$  contain less than  $r$  successes. As a matter of fact, it is the inclusion of this extra term that makes the construction of sharp bounds feasible; should one couch the declumping procedure exclusively on the second bracket, and the last term of the first one, the resulting bounds will exhibit a slow convergence rate (of order  $O(p)$ ) for  $r < k$  and only the case  $r = k$  could be covered by a better rate of order  $O(p^k)$ . For more details on this approach we refer to Boutsikas and Koutras (2002). We are now ready to prove the next theorem.

**Theorem 2** Let  $W_n = \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_i)$  be the number of moving sums that contain at least  $r$  1's. Then

$$d(W_n, CP(\lambda, H)) \leq (2k-1)(\lambda pq)b(r-1; k-1, p) + 3\lambda k f(r; k, p) + (\lambda+2)(1-G(r; k, p))$$

where  $\lambda = \lambda_{r,k,n} = (n-k+1)f(r; k, p)$  and

$$\begin{aligned} H(x) &= P(C_1 \leq x | C_1 > 0) \\ &= P\left(\sum_{m=k+1}^{2k+1} I_{[r,\infty)}(S_m) \leq x \mid I_{[r,\infty)}(S_j) = 0, j = 1, 2, \dots, k, \quad I_{[r,\infty)}(S_{k+1}) = 1\right) \end{aligned}$$

**Proof.** Applying inequality (7) for  $Z_a = C_a, a = 1, 2, \dots$  we may write

$$d(W_n, CP(\lambda, H)) \leq d(W_n, \sum_{a=1}^{n-k+1} C_a) + d\left(\sum_{a=1}^{n-k+1} C_a, CP(\lambda, H)\right) \quad (8)$$

where (see also (3),(5))

$$\lambda = \sum_{a=1}^{n-k+1} P(C_a > 0) = (n-k+1)f(r; k, p).$$

The second term in the RHS of inequality (8) can be bounded from above by the aid of Theorem 1. More specifically, if we introduce the left neighborhoods of dependence as

$$B_a = \{\max\{1, a-3k+1\}, \dots, a-1\}, \quad a = 2, 3, \dots$$

we deduce

$$\begin{aligned} d\left(\sum_{a=1}^{n-k+1} C_a, CP(\lambda, H)\right) &\leq \sum_{i=2}^{n-k+1} \sum_{b=\max\{1, i-3k+1\}}^{i-1} (P(C_b > 0, C_i > 0) + P(C_b > 0)P(C_i > 0)) \\ &\quad + (n-k+1)P(C_1 > 0)^2 \\ &\leq \sum_{i=2}^{n-k+1} \sum_{b=\max\{1, i-3k+1\}}^{i-k-1} P(C_b > 0, C_i > 0) + 3k(n-k+1)P(C_1 > 0)^2 \\ &\leq (n-k) \sum_{b=1}^{2k-1} P(C_b > 0, C_{3k} > 0) + 3(n-k+1)k f^2(r; k, p) \\ &\leq (n-k) \sum_{b=1}^{2k-1} P(S_{b-k} < r, \dots, S_{b-1} < r, S_b \geq r) P(S_{3k-1} < r, S_{3k} \geq r) \\ &\quad + 3(n-k+1)k f^2(r; k, p) \\ &\leq \lambda(2k-1) \binom{k-1}{r-1} p^r q^{k-r+1} + 3\lambda k f(r; k, p). \quad (9) \end{aligned}$$

On the other hand, for the first term of (8) we have (using the well-known coupling inequality for the total variation distance  $d_{TV}$ )

$$d(W_n, \sum_{a=1}^{n-k+1} C_a) \leq d_{TV}(W_n, \sum_{a=1}^{n-k+1} C_a) \leq P(W_n \neq \sum_{a=1}^{n-k+1} C_a).$$

The random variables  $W_n$  and  $\sum_{i=1}^{n-k+1} C_i$  are unequal only in the following three cases:

- i) A loose clump that starts at trial  $i$  does not end up till trial  $i + 2k - 1$  for  $i = 1, 2, \dots, n - 2k$ .
- ii) One of the scanning windows starting at  $n - k + 2, \dots, n + 1$  contains at least  $r$  1's.
- iii) One of the scanning windows starting at  $-k + 1, \dots, 0$  contains at least  $r$  1's.

Cases (ii),(iii) are due to the so called "edge effects", while (i) results from the fact that, for computational convenience, we used truncated clumps. Hence, the following inequality will hold true

$$\begin{aligned} d(W_n, \sum_{i=1}^{n-k+1} C_i) &\leq \sum_{i=1}^{n-2k} P(C_i > 0, (S_{i+k+1} \geq r \text{ or } \dots \text{ or } S_{i+2k} \geq r)) + 2(1 - P(S_1 < r, \dots, S_k < r)) \\ &\leq \sum_{i=1}^{n-2k} P(S_{i-k} < r, \dots, S_{i-1} < r, S_i \geq r)(1 - P(S_{i+k+1} < r, \dots, S_{i+2k+1} < r)) \\ &\quad + 2(1 - P(S_1 < r, \dots, S_k < r, S_{k+1} < r)) \\ &= (n - 2k)f(r; k, p)(1 - G(r; k, p)) + 2(1 - G(r; k, p)) \\ &\leq (\lambda + 2)(1 - G(r; k, p)). \end{aligned} \tag{10}$$

This concludes the proof of the Theorem. ■

The following Corollary is an immediate consequence of the above theorem and formula (1).

**Corollary 3** *If  $F_{n,k}(r) = P(S_{n,k} < r)$ ,  $r = 1, 2, \dots, k$ , denotes the cumulative distribution function of the discrete scan statistic  $S_{n,k}$ , then*

$$\left| F_{n,k}(r) - e^{-\lambda} \right| \leq (2k - 1)(\lambda pq)b(r - 1; k - 1, p) + 3\lambda k f(r; k, p) + (\lambda + 2)(1 - G(r; k, p)).$$

where  $\lambda = \lambda_{r,k,n} = (n - k + 1)f(r; k, p)$ .

Roos (1994) has developed several results that can be used for establishing CP approximations for sums of dependent r.v.'s (see also Barbour and Chryssaphinou (2001) for additional references on this topic). For the problem at hand, it is unclear whether these results can be profitably

exploited to produce a manageable upper bound as the one given in Theorem 2. Moreover, even if such a bound was established, it is not expected to improve on the order of convergence offered by our result.

## 4 The asymptotic distribution of $S_{n,k}$

In the present section we are going to present a large deviation result for  $S_{n,k}$ . Let us first introduce some additional notation that will be used in the sequel.

For  $0 < p < \theta < 1$  we shall denote by  $H(\theta, p)$  the relative entropy (of the bernoulli distribution with parameter  $\theta$  with respect to the bernoulli distribution with parameter  $p$ ) or Kullback-Liebler distance which is given by the formula

$$H(\theta, p) = \theta \ln \frac{\theta}{p} + (1 - \theta) \ln \frac{1 - \theta}{1 - p} = \ln \frac{\theta^\theta (1 - \theta)^{1 - \theta}}{p^\theta (1 - p)^{1 - \theta}}. \quad (11)$$

The derivative of  $H(\theta, p)$  with respect to  $\theta$

$$h(\theta, p) = \frac{d}{d\theta} H(\theta, p) = \ln \left( \frac{\theta}{1 - \theta} / \frac{p}{1 - p} \right) > 0 \quad (12)$$

measures the log odds ratio between two biased coins. It is clear that  $H(\theta, p)$  increases from 0 to  $\ln(1/p)$  as  $\theta$  increases from  $p$  to 1.

We shall now present a simple auxiliary lemma, that will be proved useful in the investigation of  $S_{n,k}$ 's asymptotic distribution. From now on we shall assume that  $r = r_n$  and  $k = k_n$  with both sequences  $r_n$  and  $k_n$  tending to  $\infty$  as  $n \rightarrow \infty$ .

**Lemma 4** *If  $p$  is fixed,  $\theta \in (p, 1)$  and  $r_n, k_n$  satisfy the condition*

$$\lim \frac{r_n - \theta k_n}{\sqrt{k_n}} = 0$$

*then*

$$a) \quad \binom{k}{r} \theta^r (1 - \theta)^{k-r} = \frac{1 + O(\frac{\rho^2+1}{k})}{\sqrt{2\pi\theta(1-\theta)k}}, \quad (13)$$

$$b) \quad \sum_{i=r}^k \binom{k}{i} p^i (1-p)^{k-i} \sim \frac{\theta(1-p)}{\theta-p} \frac{e^{-kH(\theta,p) - \rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} \quad (14)$$

where  $\rho = \rho_n = r_n - \theta k_n = o(\sqrt{k_n})$ .

**Proof.** Note first that, for any sequence  $\{a_n\}$  of real numbers with  $a_n = o(\sqrt{k_n})$  we have

$$\frac{\left(1 - \frac{a_n}{k_n}\right)^{k_n}}{e^{-a_n}} = 1 + O\left(\frac{a_n^2}{k_n}\right) \rightarrow 1. \quad (15)$$

This is readily ascertainable if we apply first the elementary inequality  $x \leq -\ln(1-x) \leq \frac{x}{1-x}$ ,  $x < 1$  for  $x = a_n/k_n < 1$  (note that  $\lim_{n \rightarrow \infty} (a_n/k_n) = 0$  and assume that  $k_n$  is large enough so that  $a_n/k_n < 1$ ) to get

$$e^{-\frac{1}{1-\frac{a_n}{k_n}} \frac{a_n^2}{k_n}} \leq \frac{\left(1 - \frac{a_n}{k_n}\right)^{k_n}}{e^{-a_n}} \leq 1.$$

In view of the last inequality we may write

$$\left|1 - \frac{\left(1 - \frac{a_n}{k_n}\right)^{k_n}}{e^{-a_n}}\right| \leq 1 - e^{-\frac{1}{1-\frac{a_n}{k_n}} \frac{a_n^2}{k_n}} = O\left(\frac{a_n^2}{k_n}\right)$$

which proves the asymptotic expression (15).

Next, a straightforward application of Stirling's formula yields

$$\begin{aligned} \binom{k}{r} &= \frac{k^k \sqrt{2\pi k} e^{\frac{c_k}{12k}}}{e^k} \frac{e^r}{r^r \sqrt{2\pi r} e^{\frac{c_r}{12r}}} \frac{e^{k-r}}{(k-r)^{k-r} \sqrt{2\pi(k-r)} e^{\frac{c_{k-r}}{12(k-r)}}} \\ &= \frac{1}{\sqrt{2\pi r(1-\frac{r}{k})}} \frac{k^k}{r^r (k-r)^{k-r}} \exp\left(\frac{1}{12k} \left(c_k - \frac{c_r}{r/k} - \frac{c_{k-r}}{1-r/k}\right)\right), \end{aligned} \quad (16)$$

where  $c_i \in (0, 1)$ ,  $i = 1, 2, \dots$ . Making use of the obvious equality

$$\frac{k^k}{r^r (k-r)^{k-r}} \theta^r (1-\theta)^{k-r} = \left(\frac{\theta k}{r}\right)^r \left(\frac{(1-\theta)k}{k-r}\right)^{k-r}, \quad (17)$$

and taking into account the asymptotic expansions (resulting from (15) for  $a_n = \rho_n k_n / r_n$  and  $a_n = \rho_n k_n / (k_n - r_n)$  respectively)

$$\left(\frac{\theta k}{r}\right)^r = \left(1 + \frac{\theta k - r}{r}\right)^r = \left(1 - \frac{\rho}{r}\right)^r = e^{-\rho} \left(1 + O\left(\frac{\rho^2}{k}\right)\right), \quad (18)$$

$$\left(\frac{(1-\theta)k}{k-r}\right)^{k-r} = \left(1 + \frac{r - k\theta}{k-r}\right)^{k-r} = \left(1 + \frac{\rho}{k-r}\right)^{k-r} = e^{\rho} \left(1 + O\left(\frac{\rho^2}{k}\right)\right) \quad (19)$$

we conclude that

$$\begin{aligned} \binom{k}{r} \theta^r (1-\theta)^{k-r} &= \frac{1}{\sqrt{2\pi r(1-\frac{r}{k})}} \left(\frac{\theta k}{r}\right)^r \left(\frac{(1-\theta)k}{k-r}\right)^{k-r} e^{\frac{1}{12k} \left(c_k - \frac{c_r}{r/k} - \frac{c_{k-r}}{1-r/k}\right)} \\ &= \frac{\left(1 + O\left(\frac{\rho^2}{k}\right)\right)}{\sqrt{2\pi r(1-\frac{r}{k})}} e^{\frac{1}{12k} \left(c_k - \frac{c_r}{r/k} - \frac{c_{k-r}}{1-r/k}\right)}. \end{aligned}$$

The proof of part (a) is now easily completed by observing that

$$\sqrt{\frac{\theta(1-\theta)}{\frac{r}{k}(1-\frac{r}{k})}} - 1 = O\left(\frac{\rho}{k}\right), \quad e^{\frac{1}{12k}\left(c_k - \frac{c_r}{r/k} - \frac{c_{k-r}}{1-r/k}\right)} - 1 = O\left(\frac{1}{k}\right).$$

For the proof of part (b) note first that

$$\frac{\sum_{i=r}^k \binom{k}{i} p^i q^{k-i}}{\binom{k}{r} p^r q^{k-r}} = 1 + \sum_{i=1}^{k-r} \frac{(k-r)(k-r-1)\dots(k-r-i+1)}{(r+1)(r+2)\dots(r+i)} \left(\frac{p}{q}\right)^i \leq \sum_{i=0}^{k-r} \left(\frac{k-r}{r} \frac{p}{q}\right)^i.$$

( $q = 1 - p$ ). Since  $r/k \rightarrow \theta > p$ , we may choose  $r, k$  large enough so that  $\frac{k-r}{r} \frac{p}{q} = \frac{1-r/k}{r/k} \frac{p}{1-p} < 1$  which results to

$$\frac{\sum_{i=r}^k \binom{k}{i} p^i q^{k-i}}{\binom{k}{r} p^r q^{k-r}} \leq \frac{1 - \left(\frac{k-r}{r} \frac{p}{q}\right)^{k-r+1}}{1 - \frac{k-r}{r} \frac{p}{q}} \rightarrow \frac{1}{1 - \frac{1-\theta}{\theta} \frac{p}{q}} = \frac{\theta - \theta p}{\theta - p}.$$

Observe next that, for  $k, r$  large enough so that  $\frac{k-r - \lfloor \sqrt{k-r} \rfloor}{r + \lfloor \sqrt{k-r} \rfloor} \frac{p}{q} < 1$ , we may write

$$\begin{aligned} \frac{\sum_{i=r}^k \binom{k}{i} p^i q^{k-i}}{\binom{k}{r} p^r q^{k-r}} &\geq 1 + \sum_{i=1}^{\lfloor \sqrt{k-r} \rfloor} \frac{(k-r)(k-r-1)\dots(k-r-i+1)}{(r+1)(r+2)\dots(r+i)} \left(\frac{p}{q}\right)^i \\ &\geq \sum_{i=0}^{\lfloor \sqrt{k-r} \rfloor} \left(\frac{k-r - \lfloor \sqrt{k-r} \rfloor}{r + \lfloor \sqrt{k-r} \rfloor}\right)^i \left(\frac{p}{q}\right)^i \\ &= \frac{1 - \left(\frac{k-r - \lfloor \sqrt{k-r} \rfloor}{r + \lfloor \sqrt{k-r} \rfloor} \frac{p}{q}\right)^{\lfloor \sqrt{k-r} \rfloor + 1}}{1 - \frac{k-r - \lfloor \sqrt{k-r} \rfloor}{r + \lfloor \sqrt{k-r} \rfloor} \frac{p}{q}} \rightarrow \frac{1}{1 - \frac{1-\theta}{\theta} \frac{p}{q}} = \frac{\theta - \theta p}{\theta - p}. \end{aligned}$$

Hence,

$$\frac{\sum_{i=r}^k \binom{k}{i} p^i q^{k-i}}{\binom{k}{r} p^r q^{k-r}} \rightarrow \frac{\theta - \theta p}{\theta - p},$$

and the proof is easily completed if we use part (a) and take into account that

$$e^{-kH(\theta,p) - \rho h(\theta,p)} = \frac{p^r (1-p)^{k-r}}{\theta^r (1-\theta)^{k-r}}.$$

■

It is worth mentioning that part (b) of Lemma 4 can be viewed as a special case of the well known Petrov's (1965) large deviation theorem (see also Höglund (1979) for an extension of Petrov's result).

We are now ready to elucidate the asymptotic behaviour of  $S_{n,k}$ .

**Theorem 5** Let  $p$  be fixed,  $\theta \in (p, 1)$  and  $r = r_n, k = k_n$  two sequences satisfying the condition

$$\lim_{n \rightarrow \infty} \frac{r_n - \theta k_n}{\sqrt{k_n}} = 0.$$

If the sequence

$$l_n = n \frac{(\theta - p) e^{-kH(\theta,p) - \rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}}, \quad n = 1, 2, \dots$$

( $\rho = \rho_n = r_n - \theta k_n$ ) is bounded from above, then

$$P(S_{n,k} < r) \sim e^{-l_n}.$$

Moreover, the rate of convergence in the above approximation is of order  $O((\rho^2 + 1)/k)$ .

**Proof.** Recalling the notations used in Corollary 3 we may write

$$\left| P(S_{n,k} < r) - e^{-l_n} \right| \leq \left| F_{n,k}(r) - e^{-\lambda_{r,k,n}} \right| + \left| e^{-\lambda_{r,k,n}} - e^{-l_n} \right|. \quad (20)$$

By virtue of part (a) of Lemma 4 we deduce

$$\begin{aligned} f(r; k, p) &= \frac{r}{k} \binom{k}{r} p^r q^{k-r} \left[ \frac{qr}{pk} \binom{k}{r} p^r q^{k-r} + \left(1 - \frac{kp}{r}\right) \left(1 - \sum_{i=r}^k \frac{i}{kp} \binom{k}{i} p^i q^{k-i}\right) \right] \\ &= \frac{(\theta - p) e^{-kH(\theta,p) - \rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} \left(1 + O\left(\frac{\rho^2 + 1}{k}\right)\right), \end{aligned}$$

while part (b) of Lemma 4 yields

$$\begin{aligned} 1 - G(r; k, p) &= 1 - \left(1 - \sum_{i=r}^k \binom{k}{i} p^i q^{k-i}\right)^2 + kp \binom{k}{r} p^r q^{k-r} \\ &\quad \times \left( \frac{r-1-pk}{pk} - \frac{r-1}{pk} \sum_{i=r-1}^k \binom{k}{i} p^i q^{k-i} + \sum_{i=r-2}^{k-1} \binom{k-1}{i} p^i q^{k-1-i} \right) \\ &\sim 2 \frac{\theta(1-p)}{\theta-p} \frac{e^{-kH(\theta,p) - \rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} + kp \frac{e^{-kH(\theta,p) - \rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} \frac{\theta-p}{p} \\ &\sim \frac{(\theta-p) k e^{-kH(\theta,p) - \rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}}. \end{aligned}$$

It is not difficult to check that

$$\begin{aligned} \lambda_{r,k,n} &= (n - k + 1) f(r; k, p) = l_n \left(1 + O\left(\frac{\rho^2 + 1}{k}\right)\right), \\ \left| e^{-\lambda_{r,k,n}} - e^{-l_n} \right| &= e^{-l_n} \left| 1 - e^{l_n - \lambda_{r,k,n}} \right| = O(l_n - \lambda_{r,k,n}) = O\left(\frac{\rho^2 + 1}{k}\right). \end{aligned}$$

On the other hand, the upper bound provided by Corollary 3 for  $|F_{n,k}(r) - e^{-\lambda_{r,k,n}}|$  takes on the asymptotic form

$$l_n(2k-1) \frac{q\theta e^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} + 3l_n k \frac{(\theta-p)e^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} + (l_n+2) \frac{(\theta-p)ke^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} \\ \sim \frac{ke^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} (l_n(6\theta-2\theta p-4p)+2(\theta-p)) = O(\sqrt{k}e^{-kH(\theta,p)-\rho h(\theta,p)})$$

and the proof is easily completed by incorporating the above results in (20). ■

## 5 An extreme value theorem for the Erdős-Rényi Statistic

A substantial literature on asymptotic results has been published under the heading of Erdős-Rényi laws. A nice collection of results of this type may be found in the publication of Deheuvels and Devroye (1987) and the references cited therein.

Let  $Y_1, Y_2, \dots$  be a sequence of iid random variables with  $E(Y_i) = 0, i = 1, 2, \dots$  and define the statistic

$$U_n = \max_{1 \leq i \leq n-k+1} \sum_{j=i}^{i+k-1} Y_j$$

which measures the maximum of the moving sums  $\sum_{j=i}^{i+k-1} Y_j, i = 1, 2, \dots, n-k+1$ . The classical Erdos-Renyi theorem (cf. Erdos and Renyi (1970)) declares that if  $k = k_n = \lfloor c \ln n \rfloor, c > 0$  then  $U_n/(ak_n) \rightarrow 1$  almost surely for a large class of distributions for  $Y_i$  ( $a > 0$  is a number depending on the distribution of  $Y_i$  and the constant  $c$ ). Deheuvels and Devroye (1987) derived an extreme value result for the same statistic. More specifically, they proved that, if  $Y_i$  obey any nonlattice distribution and  $k = k_n = \lfloor c \ln n \rfloor, c > 0$ , then

$$\lim_{n \rightarrow \infty} P\left(\frac{U_n - b_n}{a_n} \leq x\right) = \Lambda(x), \quad x \in R$$

where  $\Lambda(x) = \exp(-e^{-x})$  is the cumulative distribution function of the Gumbel distribution and  $a_n > 0, b_n \in R$  are appropriate normalizing constants.

We shall now exploit Theorem 5 to establish a similar extreme value result when the sequence of iid random variables are binary bernoulli variables (lattice distribution with span 1).

**Theorem 6** *Let  $X_1, X_2, \dots$  be a sequence of iid binary random variables with constant success probabilities  $p = P(X_1 = 1) = 1 - P(X_1 = 0)$ ,  $\theta \in (p, 1)$  and  $k = k_n = \lfloor \ln n / H(\theta, p) \rfloor$ . If*

$\Lambda(x) = \exp(-e^{-x})$  denotes the cdf of the Gumbel distribution and

$$b_n = k_n\theta + \frac{1}{h(\theta, p)} \ln \frac{n(\theta - p)e^{-k_n H(\theta, p)}}{\sqrt{2\pi\theta(1-\theta)k_n}}$$

then for the discrete scan statistic  $S_{n,k} = \max_{1 \leq i \leq n-k+1} \sum_{j=i}^{i+k-1} X_j$  we have

$$\lim_{n \rightarrow \infty} \left[ P \left( \frac{S_{n,k} - b_n}{1/h(\theta, p)} < y \right) - \Lambda(y - \epsilon_n(y)h(\theta, p)) \right] = 0 \quad (21)$$

where

$$\epsilon_n(y) = \left( b_n + \frac{y}{h(\theta, p)} \right) - \left\lfloor b_n + \frac{y}{h(\theta, p)} \right\rfloor.$$

Moreover, the rate of convergence in (21) is of order  $O((\ln k)^2/k)$ .

**Proof.** On introducing the notation

$$r_n(y) = b_n + \frac{y}{h(\theta, p)}$$

we may express the probability appearing in (21) as

$$P(S_{n,k} < \lfloor r_n(y) \rfloor) = P(S_{n,k} < r_n), \quad r_n = \lfloor r_n(y) \rfloor.$$

In order to make use of Theorem 5, observe that  $\epsilon_n(y) = r_n(y) - \lfloor r_n(y) \rfloor$  while

$$\begin{aligned} r_n - \theta k_n &= r_n(y) - \epsilon_n(y) - k_n\theta \\ &= \frac{y + \ln \frac{\theta - p}{\sqrt{2\pi\theta(1-\theta)}} - \frac{1}{2} \ln k_n + \left( \frac{\ln n}{H(\theta, p)} - k_n \right) H(\theta, p)}{h(\theta, p)} - \epsilon_n(y) \\ &= O(\ln k_n) = o(\sqrt{n}). \end{aligned}$$

Moreover, note that both  $r_n$  and  $k_n$  tend to  $\infty$  as  $n \rightarrow \infty$  while the quantity  $l_n$  used in Theorem 5 takes on the form

$$l_n = n \frac{(\theta - p)e^{-k_n H(\theta, p)}}{\sqrt{2\pi\theta(1-\theta)k_n}} e^{-y - \ln \frac{n(\theta - p)e^{-k_n H(\theta, p)}}{\sqrt{2\pi\theta(1-\theta)k_n}} + \epsilon_n(y)h(\theta, p)} = e^{-y + \epsilon_n(y)h(\theta, p)}.$$

Since  $l_n$  is bounded (note that  $\epsilon_n(y) \in [0, 1)$ ) a direct application of Theorem 5 yields the limiting expression (21). The rate of convergence is given by

$$O\left(\frac{(r_n - \theta k_n)^2 + 1}{k}\right) = O\left(\frac{(\ln k)^2}{k}\right).$$

■

It is worth mentioning that the above asymptotic result can be written in the equivalent form

$$\begin{aligned} & P \left( (U_n - k_n(\theta - p)) h(\theta, p) + \frac{1}{2} \ln k_n - \ln \frac{np(1-\theta)(\theta-p)e^{-k_n H(\theta, p)}}{(1-p)\theta\sqrt{2\pi\theta(1-\theta)}} \leq y \right) \\ &= \exp \left( -e^{-y + \epsilon_n(y) \ln \frac{(1-p)\theta}{p(1-\theta)}} \right) + O\left(\frac{(\ln k)^2}{k}\right), \end{aligned} \quad (22)$$

where  $U_n = \max_{1 \leq i \leq n-k+1} \sum_{j=i}^{i+k-1} (X_j - p)$ . The last expression is almost the same as Theorem 6 in Deheuvels and Devroye (1987) (when applied to bernoulli variables) the only difference being in the additional oscillating term  $\epsilon_n(y) \ln \frac{(1-p)\theta}{p(1-\theta)}$  appearing in the LHS of (22). This is due to the fact that their result holds true only for nonlattice distributions whereas Theorem 6 refers to the Bernoulli distribution. Apparently,  $U_n$  does not belong to the domain of attraction of an extreme value distribution in the case of Bernoulli r.v.'s, and the same will hold true for all lattice distributions. Nevertheless, if we can spot out appropriate sequences  $n_i \in N$  such that  $\epsilon_{n_i}(y) \rightarrow \epsilon(y)$  as  $i \rightarrow \infty$  for every  $y$ , we may obtain an extreme value distribution for the (normalized)  $U_{n_i}$ ,  $i = 1, 2, \dots$  of the form  $\exp \left( -e^{-y + \epsilon(y) \ln \frac{(1-p)\theta}{p(1-\theta)}} \right)$ .

## 6 Numerical results

In the previous sections, three different approximations were developed for the cumulative distribution function  $F_{n,k}(r) = P(S_{n,k} < r)$  of the discrete scan statistic  $S_{n,k}$ . It is worth mentioning that the expected number of successes within a scanning window of length  $k$  is  $kp$  and therefore  $F_{n,k}(r) = P(S_{n,k} < r) \approx 0$  when  $r \leq kp$ . For this reason, in the sequel we shall assume that  $r > kp$ . According to Corollary 3,  $F_{n,k}(r)$  can be approximated by the quantity

$$F_1(n, k, r; p) = \exp(-\lambda) = \exp(-(n - k + 1)f(r; k, p)), \quad r > kp,$$

with  $f(r; k, p)$  being evaluated through formula (5).

Theorem 5 states that the asymptotic behaviour of  $F_{n,k}(r)$  can be investigated by the aid of the expression  $\exp(-l_n)$ . On choosing  $\theta = r/k$ , the quantity  $\exp(-l_n)$  reduces to

$$F_2(n, k, r; p) = \exp \left( -n \frac{\left(\frac{r}{k} - p\right) e^{-kH\left(\frac{r}{k}, p\right)}}{\sqrt{2\pi\frac{r}{k}\left(1 - \frac{r}{k}\right)k}} \right), \quad r > kp,$$

with  $H(\theta, p)$  being given in (11).

Finally, Theorem 6 offers a third asymptotic approximation for  $F_{n,k}(r)$  in terms of the cumulative distribution function of the Gumbel distribution. The third approximation converges quite

slowly (especially when  $r$  is not very close to  $\theta k$ ) a fact that holds true for the majority of Erdős-Rényi type laws as well. Therefore, this result is primarily of theoretical interest. The other two expressions  $F_1(n, k, r; p), F_2(n, k, r; p)$  can be used to obtain quite reasonable approximations for the cumulative distribution  $F_{n,k}(r)$ .

Should one be interested on the expected value of  $S_{n,k}$ , he could make use of the well known formula

$$E(S_{n,k}) = \sum_{r=1}^{\infty} P(S_{n,k} \geq r) = \sum_{r=1}^{\infty} (1 - P(S_{n,k} < r)) = \sum_{r=1}^{\infty} (1 - F_{n,k}(r)),$$

which, on taking into account that  $F_{n,k}(r) \approx 0$  for  $r \leq kp$  and  $F_{n,k}(r) = 1$  for  $r > k$  yields

$$E(S_{n,k}) \approx r_0 + \sum_{r=r_0+1}^k (1 - F_{n,k}(r)), \quad r_0 = \lfloor kp \rfloor.$$

Replacing next  $F_{n,k}(r)$  by  $F_1(n, k, r; p), F_2(n, k, r; p)$  we may write

$$E(S_{n,k}) \approx r_0 + \sum_{r=r_0+1}^{r_1} (1 - e^{-(n-k+1)f(r;k,p)}),$$

and

$$E(S_{n,k}) \approx r_0 + \sum_{r=r_0+1}^{r_1} \left( 1 - \exp \left( -n \frac{\left(\frac{r}{k} - p\right) e^{-kH\left(\frac{r}{k}, p\right)}}{\sqrt{2\pi \frac{r}{k} \left(1 - \frac{r}{k}\right) k}} \right) \right),$$

with the summations terminated whenever the approximate value for  $F_{n,k}(r)$  is almost 1 (i.e.  $F_{r,k}(r_1) \approx 1$ ). In the same vain, one could also use the expression

$$E(S_{n,k}^m) = \sum_{r=1}^k (r^m - (r-1)^m) P(S_{n,k} \geq r), \quad m = 1, 2, \dots$$

to obtain reasonable and computationally tractable approximations for the higher moments of  $S_{n,k}$ .

In Table 1 we provide monte carlo estimations of the exact values of  $F_{n,k}(r) = P(S_{n,k} < r)$  and  $E(S_{n,k})$  along with the respective approximations for a variety of the parameters  $n, k, r$  and  $p = 0.5, 0.7$ . It is clear that, as  $n, k$  and  $r$  increase, the quality of the approximation of  $F_{n,k}(r)$  improves substantially. For comparison reasons we have also included in the table, a third approximation (labeled as  $Q'_L$ ) for the same quantity, which was suggested by Glaz, Naus and Wallenstein (2001) (see formula (4.3) in page 45). Note that  $Q'_L$  provides also very accurate approximations, especially for large  $n, k, r$  values. However, the computational difficulty in evaluating  $Q'_L$  is much higher (as compared to  $F_1(n, k, r; p)$  and especially  $F_2(n, k, r; p)$ ); in addition no estimate is available for the convergence rate of the approximation established by the aid of  $Q'_L$ .

It should be stressed that, the arguments used to derive  $Q'_L$ , are not offering any clue on how the error of approximation could be bounded. On the contrary, Corollary 3 offers an explicit computationally tractable bound for the discrepancy between  $F_{n,k}(r)$  and  $F_1(n, k, r; p)$ , namely

$$UB = (2k - 1)(\lambda pq)b(r - 1; k - 1, p) + 3\lambda k f(r; k, p) + (\lambda + 2)(1 - G(r; k, p))$$

When  $r$  increases, the quantity  $UB$  becomes extremely small (less than  $10^{-4}$ ) and, as a consequence, a very tight interval estimate for  $F_{n,k}(r)$  may be developed. In closing, we mention that, the rate of convergence for the approximation provided by  $F_2(n, k, r; p)$  is also available; however this cannot be used efficiently for establishing as good interval estimates.

In closing we mention that, one could formally write down an exact formula for the distribution of  $S_{n,k}$  by embedding the r.v. of interest in an appropriate Markov Chain (cf. Fu (2001), Balakrishnan and Koutras (2002), p. 297). However, the dimension of the transition probability matrix of the chain becomes extremely large even for moderate values of  $r, k$  (it is nearly a billion for the smallest tabulated values  $k = 30, r = 20$ ), a fact that makes the evaluation practically infeasible. In cases where the parameter values lead to intractable computations, the approach taken at this article is of special interest.

## References

- [1] Balakrishnan, N. and Koutras, M. V. (2002). *Runs, Scans and Applications.*, John Wiley & Sons.
- [2] Barbour A.D. and Chryssaphinou O. (2001). Compound Poisson approximation: A user's guide. *The Annals of Applied Probability* **11**, 964-1002.
- [3] Boutsikas, M.V. and Koutras, M.V. (2001) Compound Poisson approximation for sums of dependent random variables. In *Probability and Statistical Models with Applications: A volume in honor of Prof. T. Cacoullos* (Eds. Ch.A. Charalambides, M.V. Koutras, N. Balakrishnan), 63-86, Chapman and Hall/CRC press.
- [4] Boutsikas, M.V. and Koutras, M.V. (2002) Modelling claim exceedances over thresholds. *Insurance: Mathematics and Economics* **30**, 67-83.

- [5] Deheuvels, P. and Devroye, L. (1987). Limit laws of Erdős-Rényi-Shepp type. *The Annals of Probability* **15**, 1363-1386.
- [6] Erdős, P. and Rényi, A. (1970) On a new law of large numbers. *J. Analyse Math.* **23**, 103-111.
- [7] Fu, J.C. (2001). Distribution of the scan statistic for a sequence of bistate trials. *Journal of Applied Probability* **38**, 908-916.
- [8] Fu, J.C. and Lou, W.Y.W. (2003) *Distribution Theory of runs and Patterns and its Applications* (Singapore: World Scientific Publishing).
- [9] Glaz, J. and Balakrishnan, N. (eds.) (1999) *Scan Statistics and Applications*, Boston, MA: Birkhauser.
- [10] Glaz, J. and Naus, J.I. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *The Annals of Applied Probability* **1**, 306-318.
- [11] Glaz, J. , Naus, J. and Wallenstein, S.(2001). *Scan Statistics*. Springer.
- [12] Glaz, J. and Zhang, Z. (2004). Multiple Window Scan Statistics. *Journal of Applied Statistics* **31**, 967-980.
- [13] Höglund, T (1979) A unified formulation of the central limit theorem for small and large deviations from the mean. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **49**, 105-117.
- [14] Petrov V.V. (1965) On the probabilities of large deviations of sums of random variables. *Theory Probab. Appl.* **10**, 287-298.
- [15] Roos, M. (1994) Stein's method for compound poisson approximation: the local approach. *The Annals of Applied Probability* **4**, 1177-1187.

**Table 1.** Approximation of  $F_{n,k}(r) = P(S_{n,k} < r)$  and  $E(S_{n,k})$  by the aid of formulae  $F_1, F_2$  (cf. Section 6). *Sim* is the respective monte carlo estimates while **UB** is a bound for the discrepancy between  $F_{n,k}(r)$  and  $F_1$  (cf. Corollary 3).  $Q_L'$  refers to the formula (4.3), page 45 in Glaz, Naus and Wallenstein (2001).

$n = 1000, k = 30, p = 0.5$					
$r$	<i>sim</i>	$F_1$	<b>UB</b>	$F_2$	$Q_L'$
20	0.0030	0.010481	5.814920	0.009012	0.003081
21	0.0543	0.074823	1.838730	0.067685	0.053611
22	0.2703	0.290652	0.445214	0.276236	0.270714
23	0.6039	0.612020	0.089468	0.599248	0.604788
24	0.8507	0.851214	0.017160	0.845072	0.849814
25	0.9580	0.957953	0.003415	0.955991	0.957749
26	0.9904	0.990954	0.000646	0.990486	0.990927
27	0.9985	0.998532	0.000100	0.998446	0.998529
$E(S_{n,k})$	22.272	22.212		22.256	
$n = 10000, k = 100, p = 0.5$					
$r$	<i>sim</i>	$F_1$	<b>UB</b>	$F_2$	$Q_L'$
61	0.0002	0.000426	6.211840	0.000393	0.000166
63	0.0259	0.030981	1.196450	0.029690	0.026104
65	0.2697	0.277177	0.172825	0.272680	0.270811
67	0.6742	0.676179	0.021687	0.672762	0.674519
69	0.9056	0.906263	0.003008	0.905096	0.906049
71	0.9795	0.979849	0.000487	0.979585	0.979825
73	0.9966	0.996562	0.000077	0.996515	0.996559
75	0.9995	0.999527	0.000010	0.999520	0.999526
$E(S_{n,k})$	65.80	65.766		65.788	
$n = 100000, k = 1000, p = 0.5$					
$r$	<i>sim</i>	$F_1$	<b>UB</b>	$F_2$	$Q_L'$
539	0.0062	0.009528	4.078320	0.009118	0.006942
543	0.0577	0.070025	1.270740	0.068191	0.063361
547	0.2288	0.243519	0.353627	0.240034	0.236715
551	0.4878	0.497304	0.090433	0.493741	0.493687
555	0.7169	0.724794	0.022406	0.722389	0.723520
559	0.8669	0.870856	0.005799	0.869612	0.870493
563	0.9444	0.946087	0.001658	0.945544	0.945989
567	0.9796	0.979482	0.000518	0.979272	0.979455
571	0.9928	0.992786	0.000166	0.992711	0.992778
$E(S_{n,k})$	551.55	551.17		551.23	
$n=100000, k=100, p=0.5$					
$r$	<i>sim</i>	$F_1$	<b>UB</b>	$F_2$	$Q_L'$
66	0.0004	0.000660	0.4620750	0.000642	0.000617
67	0.0188	0.019289	0.1353200	0.018994	0.018919
68	0.1303	0.131504	0.0372180	0.130446	0.130825
69	0.3728	0.370420	0.0098961	0.368931	0.369945
70	0.6282	0.629420	0.0026684	0.628218	0.629231
71	0.8157	0.814325	0.0007715	0.813621	0.814269
72	0.9199	0.916981	0.0002460	0.916640	0.916966
73	0.9673	0.965843	0.0000847	0.965696	0.965839
74	0.9867	0.986855	0.0000300	0.986797	0.986854
75	0.9947	0.995233	0.0000105	0.995211	0.995233
76	0.9979	0.998367	0.0000035	0.998359	0.998367
$E(S_{n,k})$	69.21	69.172		69.177	
$n=100000, k=1000, p=0.7$					
$r$	<i>sim</i>	$F_1$	<b>UB</b>	$F_2$	$Q_L'$
735	0.0028	0.006738	4.391180	0.006420	0.004679
739	0.0566	0.067919	1.206990	0.066100	0.061369
743	0.2612	0.267633	0.287480	0.264010	0.261151
747	0.5516	0.554803	0.061936	0.551410	0.551898
751	0.7796	0.786442	0.013116	0.784471	0.785624
755	0.9093	0.914489	0.003028	0.913635	0.914301
759	0.9661	0.970108	0.000789	0.969800	0.970065
763	0.9877	0.990646	0.000218	0.990548	0.990636
767	0.9951	0.997350	0.000059	0.997322	0.997347
$E(S_{n,k})$	746.46	746.28		746.33	
$n=10000, k=100, p=0.7$					
$r$	<i>sim</i>	$F_1$	<b>UB</b>	$F_2$	$Q_L'$
81	0.0062	0.008590	2.104690	0.008042	0.006245
82	0.0546	0.060537	0.730446	0.058116	0.054386
83	0.2081	0.215001	0.228520	0.210098	0.208207
84	0.4521	0.457658	0.066394	0.452187	0.453727
85	0.6895	0.692245	0.018849	0.688244	0.690780
86	0.8502	0.852425	0.005543	0.850228	0.852004
87	0.9371	0.938287	0.001731	0.937294	0.938179
88	0.9774	0.977020	0.000557	0.976630	0.976992
89	0.9920	0.992309	0.000174	0.992173	0.992302
90	0.9976	0.997685	0.000051	0.997642	0.997683
$E(S_{n,k})$	86.56	86.565		86.573	