J. Appl. Prob. **43,** 1137–1154 (2006)
Printed in Israel
© *Applied Probability Trust* 2006

ON THE ASYMPTOTIC DISTRIBUTION OF THE DISCRETE SCAN STATISTIC

MICHAEL V. BOUTSIKAS [*] [**] AND
MARKOS V. KOUTRAS,[*] [***] *University of Piraeus*

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 and k, $n \geq k$, being 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 and subsequently use it to establish asymptotic results for the distribution of the discrete scan statistic as $n, k \to \infty$ 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

2000 Mathematics Subject Classification: 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 indicating 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 (the 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 RV is usually referred to in the literature as the *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 [1, pp. 377–387], [9], [12, Part I], and [8, pp. 140–151].

To fix our notation, let X_i, $i \in \mathbb{Z}$, be a sequence of independent, identically distributed (i.i.d.) binary random variables (RVs) with

$$\mathrm{P}(X_i = 1) = p, \quad \mathrm{P}(X_i = 0) = q = 1 - p, \qquad i = 1, 2, \ldots, n,$$

and denote by

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

Received 25 May 2005; revision received 12 July 2006.

[*] Postal address: Department of Statistics and Insurance Science, University of Piraeus, 80 Karaoli and Dimitriou Street, 185 34 Piraeus, Greece.
[**] Email address: mbouts@unipi.gr
[***] Email address: mkoutras@unipi.gr

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

the k-scan process (a 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 1s due to local positive dependence between X_i, $i = 1, 2, \dots, n$, or due to the existence of subsequences of consecutive X_i with $\mathrm{P}(X_i = 1) > p$. As Glaz and Naus [10] 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 [11] 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 such that a prespecified significance level is achieved 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 $n, k \to \infty$) 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 in molecular biology. In the study of amino acid sequences, various classification schemes are in common use, including a chemical alphabet of eight letters, a functional alphabet of four letters, a charge alphabet of three 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} and 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 $\mu_j = \mathrm{P}(Z_{i1} = a_j) = \mathrm{P}(Z_{i2} = a_j)$, $j = 1, 2, \dots, l$. 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 (we let X_i be 0 otherwise). Then X_i, $i = 1, 2, \dots, n$, form a sequence of binary i.i.d. RVs with success probabilities

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

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 sufficiently close 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 evidence of the lack of a local match between the two sequences under inspection. It should be stressed that, in this application, we are also interested in large values of both n (long amino acid sequences) and k (long matching regions).

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

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 the Z_i are i.i.d. RVs with cumulative distribution function F, and denote by

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

the corresponding RVs, which indicate whether or not the ith claim exceeds the threshold u. (Here $\mathbf{1}_A(\cdot)$ denotes the indicator function of the (generic) set A.) Then

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

and $S_{n,k}$ will describe the maximum number of ‘large claims’ (i.e. claims exceeding the threshold u) in a period of k consecutive days. Since the primary interest in this situation is also focused on extremely long periods ($n \to \infty$, $k \to \infty$), 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 RV $S_{n,k}$. Exact results for the distribution of the scan statistic were discussed in [7], [1, pp. 291–301], and [8, pp. 88–96]. 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 *et al.* [12] and Balakrishnan and Koutras [1] for up-to-date reviews of this topic.

In a recent article by Boutsikas and Koutras [4], a compound Poisson approximation was established for the distribution of the enumerating RV

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

As a by-product, an approximation for

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

was established, along with an upper bound for the error incurred in its use. However, the asymptotic result given there holds under the conditions $n \to \infty$, $p \to 0$ with k and r fixed, which are of no interest in the examples mentioned above. One might suspect that, even in the case of interest (p fixed and $n, k \to \infty$) a compound Poisson law underlies the behavior, yet the tests for this provided by the results of [4] are inconclusive. This is due to the fact that, for $r < k$, the upper bound appearing there is of order $O(p)$ and, therefore, does not converge to 0 as $n, k \to \infty$ while p is fixed.

In the present article, motivated by the abovementioned remarks, we establish a new compound Poisson 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, we develop a compound Poisson approximation for the distribution of W_n, 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 in Section 5 we present

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

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 RVs X and Y is defined as

$$d(X, Y) = \sup_{w} |\mathrm{P}(X \le w) - \mathrm{P}(Y \le w)|$$

and offers a very efficient tool for establishing convergence in distribution; a sequence of RVs converges weakly to Y if the corresponding sequence of distances converges to 0. By the term 'compound Poisson distribution $\mathrm{CP}(\lambda, H)$ with parameter λ and compounding distribution H', we shall refer to the distribution of a random sum of the form $\sum_{i=1}^{N} Z_i$ where N is a Poisson RV with $\lambda = \mathrm{E}(N)$ and the Z_i are i.i.d. RVs (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 [3]. 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 nonnegative RVs Z_a, $a = 1, 2, \dots$. For each $a = 2, 3, \dots,$ introduce a subset, B_a, of $\{1, 2, \dots, a-1\}$ (the left neighborhood of dependence of Z_a) such 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 $\sum_{a=1}^{\nu} Z_a$ (with ν a fixed, positive integer) and a compound Poisson distribution $\mathrm{CP}(\lambda, H)$ with suitably chosen λ and H.

Theorem 1. (Boutsikas and Koutras [3].) *If Z_a, $a = 1, 2, \dots, \nu$, is a sequence of nonnegative RVs, then*

$$d\bigg(\sum_{a=1}^{\nu} Z_a, \mathrm{CP}(\lambda, H)\bigg) \le \sum_{a=2}^{\nu}\bigg(\mathrm{P}\bigg(Z_a > 0, \sum_{b\in B_a} Z_b > 0\bigg) + \mathrm{P}(Z_a > 0)\,\mathrm{P}\bigg(\sum_{b\in B_a} Z_b > 0\bigg)\bigg) + \frac{1}{2}\sum_{i=1}^{\nu} \mathrm{P}(Z_i > 0)^2, \qquad (2)$$

where $\lambda = \sum_{a=1}^{\nu} \lambda_a$ and

$$H(x) = \frac{1}{\lambda}\sum_{a=1}^{\nu} \lambda_a \mathrm{P}(Z_a \le x \mid Z_a > 0), \qquad x \in \mathbb{R},$$

with $\lambda_a = \mathrm{P}(Z_a > 0)$, $a = 1, 2, \dots, \nu$.

Theorem 1 states that if the RVs 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.

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

If X and Y are nonnegative RVs, it is evident that $|\mathrm{P}(X=0)-\mathrm{P}(Y=0)| \le d(X, Y)$ and, therefore, that

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

is also bounded from above by the right-hand side of (2). It is worth stressing that, should one wish to establish bounds for $\mathrm{P}(\sum_{a=1}^{\nu} Z_a = 0)$ 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.

Now let $b(x; n, p)$ and $B(x; n, p)$ respectively denote the probability mass function and cumulative distribution function of a binomial RV X, i.e.

$$b(x; n, p) = \mathrm{P}(X = x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, \dots, n,$$
$$B(x; n, p) = \mathrm{P}(X \le x) = \sum_{r=0}^{x} b(r; n, p), \qquad x \in \mathbb{R}.$$

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

$$\begin{aligned} f(s; k, p) &= \mathrm{P}(S_1 < s,\ S_2 < s,\ \dots,\ S_k < s,\ S_{k+1} \ge s), \\ G(s; k, p) &= \mathrm{P}(S_1 < s,\ S_2 < s,\ \dots,\ S_{k+1} < s), \end{aligned} \tag{3}$$

which can be expressed using $b(x; n, p)$ and $B(x; n, p)$ as follows (cf. [10]), for $1 \le s \le k$ (if $s > k$ or $s < 0$ then we set $f(s, k; p) = 0$):

$$\begin{aligned} f(s; k, p) &= \frac{p}{s} b(s-1; k-1, p)[sqb(s-1; k-1, p) + (s - kp)B(s-2; k-1, p)], \\ G(s; k, p) &= B(s-1; k, p)^2 - b(s; k, p)[(s-1)B(s-2; k, p) - kpB(s-3; k-1, p)]. \end{aligned} \tag{4}$$

The standard symbols '$\sim$', $o(\cdot)$, and $O(\cdot)$ will assume their usual meanings, i.e.

$$\begin{aligned} &f(t) \sim g(t) && \text{as } t \to t_0 && \text{if } \lim_{t \to t_0} \frac{f(t)}{g(t)} = 1, \\ &f(t) = o(g(t)) && \text{as } t \to t_0 && \text{if } \lim_{t \to t_0} \frac{f(t)}{g(t)} = 0, \\ &f(t) = O(g(t)) && \text{if } \frac{f(t)}{g(t)} \text{ is bounded.} \end{aligned}$$

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

$$\begin{aligned} |\mathrm{P}(S_{n,k} < r) - \mathrm{e}^{-\lambda}| &= |\mathrm{P}(W_n = 0) - \mathrm{e}^{-\lambda}| \\ &= |\mathrm{P}(\mathbf{1}_{[r,\infty)}(S_a) = 0 \text{ for all } a = 1, \dots, n-k+1) - \mathrm{e}^{-\lambda}|, \end{aligned}$$

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

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

$$d(W_n, \mathrm{CP}(\lambda, H)) \le d\Big(W_n, \sum Z_a\Big) + d\Big(\sum Z_a, \mathrm{CP}(\lambda, H)\Big). \tag{5}$$

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 RVs possessing the aforementioned properties is provided by

$$C_a = \Bigg[\prod_{j=a-k}^{a-1} (1 - \mathbf{1}_{[r,\infty)}(S_j))\Bigg]\Bigg[\mathbf{1}_{[r,\infty)}(S_a) \sum_{m=a}^{a+k} \mathbf{1}_{[r,\infty)}(S_m)\Bigg], \qquad a = 1, 2, \dots.$$

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 RV, which counts the total number of clumps located in 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 fewer than r successes. As a matter of fact, it is the inclusion of this extra term that makes the construction of sharp bounds feasible; were we to represent the declumping procedure exclusively by the second bracket and the last term of the first bracket, then the resulting bounds would exhibit a slow convergence rate (of order $O(p)$) for $r < k$, and only the case $r = k$ could exhibit a better rate of order $O(p^k)$. For more details on this approach we refer the reader to [4]. We are now ready to prove the next theorem.

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

$$\begin{aligned} &d(W_n, \mathrm{CP}(\lambda, H)) \\ &\qquad \le (2k-1)\lambda pq b(r-1; k-1, p) + 3\lambda k f(r; k, p) + (\lambda+2)(1 - G(r; k, p)), \end{aligned}$$

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

$$\begin{aligned} H(x) &= \mathrm{P}(C_1 \le x \mid C_1 > 0) \\ &= \mathrm{P}\Bigg(\sum_{m=k+1}^{2k+1} \mathbf{1}_{[r,\infty)}(S_m) \le x \;\Bigg|\; \mathbf{1}_{[r,\infty)}(S_j) = 0,\ j = 1, 2, \dots, k,\ \mathbf{1}_{[r,\infty)}(S_{k+1}) = 1\Bigg). \end{aligned}$$

Proof. By applying inequality (5) with $Z_a = C_a$, $a = 1, 2, \dots$, we may write

$$d(W_n, \mathrm{CP}(\lambda, H)) \le d\Bigg(W_n, \sum_{a=1}^{n-k+1} C_a\Bigg) + d\Bigg(\sum_{a=1}^{n-k+1} C_a, \mathrm{CP}(\lambda, H)\Bigg), \tag{6}$$

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

where (see also (3) and (4))

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

The second term on the right-hand side of (6) can be bounded from above with the aid of Theorem 1. More specifically, if we introduce

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

the left neighborhoods of dependence, we deduce that

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

On the other hand, for the first term of (6) we have (using the well-known coupling inequality for the total variation distance, d_{TV})

$$d\biggl(W_n, \sum_{a=1}^{n-k+1} C_a\biggr) \le d_{\mathrm{TV}}\biggl(W_n, \sum_{a=1}^{n-k+1} C_a\biggr) \le \mathrm{P}\biggl(W_n \ne \sum_{a=1}^{n-k+1} C_a\biggr).$$

The RVs 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 until trial $i+2k-1$, $i = 1, 2, \dots, n-2k$.

(ii) One of the scanning windows starting at $n-k+2, \dots, n+1$ contains at least r 1s.

(iii) One of the scanning windows starting at $-k+1, \dots, 0$ contains at least r 1s.

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

Cases (ii) and (iii) occur because of so-called 'edge effects', while case (i) occurs because, for computational convenience, we used truncated clumps. Hence, the following inequality will hold:

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

This concludes the proof of the theorem.

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

Corollary 1. *Let $F_{n,k}(r) = \mathrm{P}(S_{n,k} < r)$, $r = 1, 2, \dots, k$, denote the cumulative distribution function of the discrete scan statistic $S_{n,k}$. Then*

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

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

Roos [15] has developed several results that can be used to establish compound Poisson approximations for sums of dependent RVs (see also [2] for additional references on this topic). For the problem at hand, it is unclear whether these results can be profitably exploited to produce as manageable an 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 θ with respect to the Bernoulli distribution with parameter p) or Kullback–Leibler distance, which is given by

$$H(\theta, p) = \theta \ln \frac{\theta}{p} + (1 - \theta) \ln \frac{1 - \theta}{1 - p} = \ln \frac{\theta^\theta (1 - \theta)^{1-\theta}}{p^\theta (1 - p)^{1-\theta}}. \tag{7}$$

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

$$h(\theta, p) = \frac{\mathrm{d}}{\mathrm{d}\theta} H(\theta, p) = \ln\bigg(\frac{\theta/(1-\theta)}{p/(1-p)}\bigg) > 0,$$

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

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

We shall now present a simple auxiliary lemma that will prove useful in the investigation of the asymptotic distribution of $S_{n,k}$. Henceforth, we shall assume that $r \equiv r_n$ and $k \equiv k_n$, with both sequences, $\{r_n\}$ and $\{k_n\}$, tending to ∞ as $n \to \infty$. Where not stated explicitly, all convergences and limits apply as $n \to \infty$.

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

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

then

$$\binom{k}{r}\theta^r(1-\theta)^{k-r} = \frac{1 + O((\rho^2+1)/k)}{\sqrt{2\pi\theta(1-\theta)k}}, \tag{8}$$

$$\sum_{i=r}^{k}\binom{k}{i}p^i(1-p)^{k-i} \sim \frac{\theta(1-p)}{\theta - p}\,\frac{\mathrm{e}^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}}, \tag{9}$$

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

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

$$\frac{(1 - a_n/k_n)^{k_n}}{\mathrm{e}^{-a_n}} = 1 + O\biggl(\frac{a_n^2}{k_n}\biggr) \to 1. \tag{10}$$

This is readily ascertainable if we apply the elementary inequality $x \le -\ln(1-x) \le x/(1-x)$ with $x = a_n/k_n < 1$ (note that $\lim_{n\to\infty}(a_n/k_n) = 0$ and assume that k_n is large enough that $a_n/k_n < 1$), to obtain

$$\exp\biggl(-\frac{1}{1 - a_n/k_n}\frac{a_n^2}{k_n}\biggr) \le \frac{(1 - a_n/k_n)^{k_n}}{\mathrm{e}^{-a_n}} \le 1.$$

In view of the last inequality we may write

$$\biggl|1 - \frac{(1 - a_n/k_n)^{k_n}}{\mathrm{e}^{-a_n}}\biggr| \le 1 - \exp\biggl(-\frac{1}{1 - a_n/k_n}\frac{a_n^2}{k_n}\biggr) = O\biggl(\frac{a_n^2}{k_n}\biggr),$$

which proves the asymptotic expression (10).

Next, a straightforward application of Stirling's formula yields

$$\begin{aligned}\binom{k}{r} &= \frac{k^k\sqrt{2\pi k}\,\mathrm{e}^{c_k/12k}}{\mathrm{e}^k}\,\frac{\mathrm{e}^r}{r^r\sqrt{2\pi r}\,\mathrm{e}^{c_r/12r}}\,\frac{\mathrm{e}^{k-r}}{(k-r)^{k-r}\sqrt{2\pi(k-r)}\exp\,(c_{k-r}/[12(k-r)])} \\ &= \frac{1}{\sqrt{2\pi r(1-r/k)}}\,\frac{k^k}{r^r(k-r)^{k-r}}\exp\biggl(\frac{1}{12k}\biggl(c_k - \frac{c_r}{r/k} - \frac{c_{k-r}}{1-r/k}\biggr)\biggr),\end{aligned}$$

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

$$\frac{k^k}{r^r(k-r)^{k-r}}\theta^r(1-\theta)^{k-r} = \biggl(\frac{\theta k}{r}\biggr)^r\biggl(\frac{(1-\theta)k}{k-r}\biggr)^{k-r}$$

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

and taking into account the asymptotic expansions

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

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

(resulting from (10) with $a_n = \rho_n k_n / r_n$ and $a_n = \rho_n k_n/(k_n - r_n)$, respectively), we conclude that

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

The proof of (8) is now easily completed by observing that

$$\sqrt{\frac{\theta(1-\theta)}{(r/k)(1-r/k)}} - 1 = O\left(\frac{\rho}{k}\right),$$

$$\exp\left(\frac{1}{12k}\left(c_k - \frac{c_r}{r/k} - \frac{c_{k-r}}{1-r/k}\right)\right) - 1 = O\left(\frac{1}{k}\right).$$

To prove (9), first note that

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

$(q = 1-p)$. Since $r/k \to \theta > p$, we may choose r and k to be large enough that

$$\frac{k-r}{r}\frac{p}{q} = \frac{1-r/k}{r/k}\frac{p}{1-p} < 1,$$

whence

$$\frac{\sum_{i=r}^{k}\binom{k}{i}p^i q^{k-i}}{\binom{k}{r}p^r q^{k-r}} \le \frac{1 - ([(k-r)/r]p/q)^{k-r+1}}{1 - [(k-r)/r]p/q} \to \frac{1}{1 - [(1-\theta)/\theta]p/q} = \frac{\theta - \theta p}{\theta - p}.$$

Observe next that, for k and r large enough that

$$\frac{k - r - \lfloor\sqrt{k-r}\rfloor}{r + \lfloor\sqrt{k-r}\rfloor}\frac{p}{1-p} < 1,$$

we may write

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

Hence,

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

and the proof is easily completed if we use (8) and take into account the fact that

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

It is worth mentioning that (9) can be viewed as a special case of Petrov's [14] well-known large deviation theorem (see also [13] for an extension of Petrov's result).

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

Theorem 3. *Let p be fixed, let $\theta \in (p,1)$, and let $\{r_n\}$ and $\{k_n\}$ be two sequences satisfying the condition*

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

If the sequence

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

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

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

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

Proof. Recalling the notation used in Corollary 1, we may write

$$
|\mathrm{P}(S_{n,k}<r)-\mathrm{e}^{-l_n}| \leq |F_{n,k}(r)-\mathrm{e}^{-\lambda_{r,k,n}}|+|\mathrm{e}^{-\lambda_{r,k,n}}-\mathrm{e}^{-l_n}|. \tag{11}
$$

By (8), we deduce that

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

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

while (9) yields

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

It is not difficult to check that

$$\lambda_{r,k,n} = (n-k+1)f(r;k,p) = l_n\biggl(1 + O\biggl(\frac{\rho^2+1}{k}\biggr)\biggr),$$

$$|\mathrm{e}^{-\lambda_{r,k,n}} - \mathrm{e}^{-l_n}| = \mathrm{e}^{-l_n}|1 - \mathrm{e}^{l_n-\lambda_{r,k,n}}| = O(l_n - \lambda_{r,k,n}) = O\biggl(\frac{\rho^2+1}{k}\biggr).$$

On the other hand, the upper bound provided for $|F_{n,k}(r) - \mathrm{e}^{-\lambda_{r,k,n}}|$ by Corollary 1 takes the asymptotic form

$$\begin{aligned}&l_n(2k-1)\frac{q\theta\mathrm{e}^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} + 3l_nk\frac{(\theta-p)\mathrm{e}^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} \\ &\quad + (l_n+2)\frac{(\theta-p)k\mathrm{e}^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} \\ &\quad \sim \frac{k\mathrm{e}^{-kH(\theta,p)-\rho h(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}}(l_n(6\theta - 2\theta p - 4p) + 2(\theta-p)) \\ &\quad = O(\sqrt{k}\mathrm{e}^{-kH(\theta,p)-\rho h(\theta,p)}),\end{aligned}$$

and the proof is easily completed by incorporating the above results into (11).

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 [5] and the references cited therein.

Let $Y_1, Y_2, \dots$ be a sequence of i.i.d. RVs with $\mathrm{E}(Y_i) = 0$, $i = 1, 2, \dots$, and define the statistic

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

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 Erdős–Rényi theorem [6] states that if $k \equiv k_n = \lfloor c \ln n \rfloor$ for $c > 0$, then $U_n/ak_n \to 1$ almost surely for a large class of distributions of Y_i (here $a > 0$ is a number

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

depending on the distribution of Y_i and the constant c). Deheuvels and Devroye [5] derived an extreme value result for the same statistic. More specifically, they proved that if the Y_i obey any nonlattice distribution and $k \equiv k_n = \lfloor c \ln n \rfloor$, $c > 0$, then

$$\lim_{n\to\infty} \mathrm{P}\biggl(\frac{U_n - b_n}{a_n} \le x\biggr) = \Lambda(x), \qquad x \in \mathbb{R},$$

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

We shall now exploit Theorem 3 to establish a similar extreme value result when the sequence of i.i.d. RVs are binary Bernoulli variables (lattice distribution with span 1).

Theorem 4. *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. binary RVs with constant success probabilities $p = \mathrm{P}(X_1 = 1) = 1 - \mathrm{P}(X_1 = 0)$, let $\theta \in (p, 1)$, and let $k \equiv k_n = \lfloor \ln n / H(\theta, p) \rfloor$. If $\Lambda(x) = \exp(-\mathrm{e}^{-x})$ denotes the cumulative distribution function of the Gumbel distribution and*

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

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

$$\lim_{n\to\infty}\biggl[\mathrm{P}\biggl(\frac{S_{n,k} - b_n}{1/h(\theta, p)} < y\biggr) - \Lambda(y - \varepsilon_n(y)h(\theta, p))\biggr] = 0, \tag{12}$$

where

$$\varepsilon_n(y) = \biggl(b_n + \frac{y}{h(\theta, p)}\biggr) - \biggl\lfloor b_n + \frac{y}{h(\theta, p)}\biggr\rfloor.$$

Moreover, the rate of convergence in (12) 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 (12) as

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

In order to make use of Theorem 3, observe that $\varepsilon_n(y) = r_n(y) - \lfloor r_n(y) \rfloor$ while

$$\begin{aligned} r_n - \theta k_n &= r_n(y) - \varepsilon_n(y) - k_n\theta \\ &= \frac{1}{h(\theta, p)}\biggl[y + \ln\frac{\theta - p}{\sqrt{2\pi\theta(1-\theta)}} - \frac{1}{2}\ln k_n + \biggl(\frac{\ln n}{H(\theta, p)} - k_n\biggr)H(\theta, p)\biggr] - \varepsilon_n(y) \\ &= O(\ln k_n) \\ &= o(\sqrt{n}). \end{aligned}$$

Moreover, note that both r_n and k_n tend to ∞ as $n \to \infty$, while the quantity l_n used in Theorem 3 takes the form

$$\begin{aligned} l_n &= n\frac{(\theta - p)\mathrm{e}^{-k_n H(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k_n}} \exp\biggl(-y - \ln\frac{n(\theta - p)\mathrm{e}^{-k_n H(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k_n}} + \varepsilon_n(y)h(\theta, p)\biggr) \\ &= \mathrm{e}^{-y+\varepsilon_n(y)h(\theta,p)}. \end{aligned}$$

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

Since l_n is bounded (note that $\varepsilon_n(y) \in [0, 1)$), a direct application of Theorem 3 yields the limiting expression (12). The rate of convergence is given by

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

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

$$\begin{aligned} \mathrm{P}\bigg((U_n - k_n(\theta - p))h(\theta, p) + \frac{1}{2}\ln k_n - \ln \frac{np(1-\theta)(\theta - p)\mathrm{e}^{-k_n H(\theta,p)}}{(1-p)\theta\sqrt{2\pi\theta(1-\theta)}} \le y\bigg) \\ = \exp\bigg(-\exp\bigg(-y + \varepsilon_n(y)\ln\frac{(1-p)\theta}{p(1-\theta)}\bigg)\bigg) + O\bigg(\frac{(\ln k)^2}{k}\bigg), \end{aligned} \tag{13}$$

where $U_n = \max_{1\le i\le n-k+1} \sum_{j=i}^{i+k-1}(X_j - p)$. The last expression is almost the same as that of Theorem 6 of [5] (when applied to Bernoulli variables), the only difference being in the additional oscillating term $\varepsilon_n(y)\ln((1-p)\theta/p(1-\theta))$ appearing on the left-hand side of (13). This is because the result of [5] holds only for nonlattice distributions, whereas (our) Theorem 4 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 RVs, and the same will hold for all lattice distributions. Nevertheless, if we can determine appropriate sequences $\{n_i \in \mathbb{N}\}$, such that $\varepsilon_{n_i}(y) \to \varepsilon(y)$ as $i \to \infty$ for every y, we may obtain an extreme value distribution for the (normalized) U_{n_i}, $i = 1, 2, \dots$, of the form

$$\exp\bigg(-\exp\bigg(-y + \varepsilon(y)\ln\frac{(1-p)\theta}{p(1-\theta)}\bigg)\bigg).$$

6. Numerical results

In the previous sections, three different approximations were developed for the cumulative distribution function, $F_{n,k}(r) = \mathrm{P}(S_{n,k} < r)$, 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, that $F_{n,k}(r) = \mathrm{P}(S_{n,k} < r) \approx 0$ when $r \le pk$. For this reason, in the sequel we shall assume that $r > pk$. According to Corollary 1, $F_{n,k}(r)$ can be approximated by the quantity

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

with $f(r; k, p)$ as given in (4).

Theorem 3 states that the asymptotic behavior of $F_{n,k}(r)$ can be investigated with the use of the expression $\exp(-l_n)$. With $\theta = r/k$, the quantity $\exp(-l_n)$ reduces to

$$F_2(n, k, r; p) = \exp\bigg(-n\frac{(r/k - p)\mathrm{e}^{-kH(r/k,p)}}{\sqrt{2\pi(r/k)(1-r/k)k}}\bigg), \qquad r > kp,$$

with $H(\theta, p)$ as given in (7).

Finally, Theorem 4 offers a third asymptotic approximation for $F_{n,k}(r)$ in terms of the cumulative distribution function of the Gumbel distribution. This third approximation converges quite slowly (especially when r is not very close to θk), a fact that holds for the majority of Erdős–Rényi-type laws as well. Therefore, this result is primarily of theoretical interest.

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

The other two expressions, $F_1(n, k, r; p)$ and $F_2(n, k, r; p)$, can be used to obtain quite reasonable approximations for the cumulative distribution $F_{n,k}(r)$.

Were we interested in the expected value of $S_{n,k}$, we could make use of the well-known formula

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

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

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

Next, by replacing $F_{n,k}(r)$ by $F_1(n, k, r; p)$ and then $F_2(n, k, r; p)$, we may write

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

and, respectively,

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

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

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

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

In Tables 1–6 we provide Monte Carlo estimations (denoted *Sim*) of the exact values of $F_{n,k}(r) = \mathrm{P}(S_{n,k} < r)$ and $\mathrm{E}(S_{n,k})$ along with the respective approximations, for a variety

TABLE 1: $n = 1000$, $k = 30$, $p = 0.5$.

r	*Sim*	F_1	*UB*	F_2	Q'_L
20	0.0030	0.010 481	5.814 920	0.009 012	0.003 081
21	0.0543	0.074 823	1.838 730	0.067 685	0.053 611
22	0.2703	0.290 652	0.445 214	0.276 236	0.270 714
23	0.6039	0.612 020	0.089 468	0.599 248	0.604 788
24	0.8507	0.851 214	0.017 160	0.845 072	0.849 814
25	0.9580	0.957 953	0.003 415	0.955 991	0.957 749
26	0.9904	0.990 954	0.000 646	0.990 486	0.990 927
27	0.9985	0.998 532	0.000 100	0.998 446	0.998 529
	$\mathrm{E}(S_{n,k}) = 22.272$	$\mathrm{E}(S_{n,k}) = 22.212$	—	$\mathrm{E}(S_{n,k}) = 22.256$	—

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

TABLE 2: $n = 10\,000$, $k = 100$, $p = 0.5$.

r	*Sim*	F_1	*UB*	F_2	Q'_L
61	0.0002	0.000 426	6.211 840	0.000 393	0.000 166
63	0.0259	0.030 981	1.196 450	0.029 690	0.026 104
65	0.2697	0.277 177	0.172 825	0.272 680	0.270 811
67	0.6742	0.676 179	0.021 687	0.672 762	0.674 519
69	0.9056	0.906 263	0.003 008	0.905 096	0.906 049
71	0.9795	0.979 849	0.000 487	0.979 585	0.979 825
73	0.9966	0.996 562	0.000 077	0.996 515	0.996 559
75	0.9995	0.999 527	0.000 010	0.999 520	0.999 526
	$\mathrm{E}(S_{n,k}) = 65.80$	$\mathrm{E}(S_{n,k}) = 65.766$	—	$\mathrm{E}(S_{n,k}) = 65.788$	—

TABLE 3: $n = 100\,000$, $k = 1000$, $p = 0.5$.

r	*Sim*	F_1	*UB*	F_2	Q'_L
539	0.0062	0.009 528	4.078 320	0.009 118	0.006 942
543	0.0577	0.070 025	1.270 740	0.068 191	0.063 361
547	0.2288	0.243 519	0.353 627	0.240 034	0.236 715
551	0.4878	0.497 304	0.090 433	0.493 741	0.493 687
555	0.7169	0.724 794	0.022 406	0.722 389	0.723 520
559	0.8669	0.870 856	0.005 799	0.869 612	0.870 493
563	0.9444	0.946 087	0.001 658	0.945 544	0.945 989
567	0.9796	0.979 482	0.000 518	0.979 272	0.979 455
571	0.9928	0.992 786	0.000 166	0.992 711	0.992 778
	$\mathrm{E}(S_{n,k}) = 551.55$	$\mathrm{E}(S_{n,k}) = 551.17$	—	$\mathrm{E}(S_{n,k}) = 551.23$	—

TABLE 4: $n = 100\,000$, $k = 100$, $p = 0.5$.

r	*Sim*	F_1	*UB*	F_2	Q'_L
66	0.0004	0.000 660	0.462 0750	0.000 642	0.000 617
67	0.0188	0.019 289	0.135 3200	0.018 994	0.018 919
68	0.1303	0.131 504	0.037 2180	0.130 446	0.130 825
69	0.3728	0.370 420	0.009 8961	0.368 931	0.369 945
70	0.6282	0.629 420	0.002 6684	0.628 218	0.629 231
71	0.8157	0.814 325	0.000 7715	0.813 621	0.814 269
72	0.9199	0.916 981	0.000 2460	0.916 640	0.916 966
73	0.9673	0.965 843	0.000 0847	0.965 696	0.965 839
74	0.9867	0.986 855	0.000 0300	0.986 797	0.986 854
75	0.9947	0.995 233	0.000 0105	0.995 211	0.995 233
76	0.9979	0.998 367	0.000 0035	0.998 359	0.998 367
	$\mathrm{E}(S_{n,k}) = 69.21$	$\mathrm{E}(S_{n,k}) = 69.172$	—	$\mathrm{E}(S_{n,k}) = 69.177$	—

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press

TABLE 5: $n = 100\,000$, $k = 1000$, $p = 0.7$.

r	*Sim*	F_1	*UB*	F_2	Q'_L
735	0.0028	0.006 738	4.391 180	0.006 420	0.004 679
739	0.0566	0.067 919	1.206 990	0.066 100	0.061 369
743	0.2612	0.267 633	0.287 480	0.264 010	0.261 151
747	0.5516	0.554 803	0.061 936	0.551 410	0.551 898
751	0.7796	0.786 442	0.013 116	0.784 471	0.785 624
755	0.9093	0.914 489	0.003 028	0.913 635	0.914 301
759	0.9661	0.970 108	0.000 789	0.969 800	0.970 065
763	0.9877	0.990 646	0.000 218	0.990 548	0.990 636
767	0.9951	0.997 350	0.000 059	0.997 322	0.997 347
	$\mathrm{E}(S_{n,k}) = 746.46$	$\mathrm{E}(S_{n,k}) = 746.28$	—	$\mathrm{E}(S_{n,k}) = 746.33$	—

TABLE 6: $n = 10\,000$, $k = 100$, $p = 0.7$.

r	*Sim*	F_1	*UB*	F_2	Q'_L
81	0.0062	0.008 590	2.104 690	0.008 042	0.006 245
82	0.0546	0.060 537	0.730 446	0.058 116	0.054 386
83	0.2081	0.215 001	0.228 520	0.210 098	0.208 207
84	0.4521	0.457 658	0.066 394	0.452 187	0.453 727
85	0.6895	0.692 245	0.018 849	0.688 244	0.690 780
86	0.8502	0.852 425	0.005 543	0.850 228	0.852 004
87	0.9371	0.938 287	0.001 731	0.937 294	0.938 179
88	0.9774	0.977 020	0.000 557	0.976 630	0.976 992
89	0.9920	0.992 309	0.000 174	0.992 173	0.992 302
90	0.9976	0.997 685	0.000 051	0.997 642	0.997 683
	$\mathrm{E}(S_{n,k}) = 86.56$	$\mathrm{E}(S_{n,k}) = 86.565$	—	$\mathrm{E}(S_{n,k}) = 86.573$	—

of the parameters n, k, and r, and for $p = 0.5, 0.7$. The quantity *UB* is a bound for the discrepancy between $F_{n,k}(r)$ and F_1 (see Corollary 1). 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 (denoted Q'_L) for the same quantity which was suggested by Glaz *et al.* [12, p. 45, Equation (4.3)]. Note that Q'_L also provides very accurate approximations, especially for large values of n, k, and r. However, the computational difficulty of 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 with the use of Q'_L.

It should be stressed that the arguments used to derive Q'_L do not offer any clue as to how the approximation error can be bounded. Corollary 1, on the contrary, offers an explicit, computationally tractable bound for the discrepancy between $F_{n,k}(r)$ and $F_1(n, k, r; p)$, namely

$$UB = (2k - 1)\lambda pqb(r - 1; k - 1, p) + 3\lambda kf(r; k, p) + (\lambda + 2)(1 - G(r; k, p)).$$

As 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. Note 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 to establish interval estimates this good.

In closing, we mention that we could formally write down an exact formula for the distribution of $S_{n,k}$ by embedding the RV of interest in an appropriate Markov chain (see [7] and [1, p. 297]). However, the dimension of the transition probability matrix of the chain becomes extremely large even for moderate values of r and k (it is nearly one billion for the smallest tabulated values, $k = 30$ and $r = 20$), a fact that makes the evaluation unfeasible. In cases where the parameter values lead to intractable computations, the approach taken in this article is of special interest.

References

[1] Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications.* John Wiley, Chichester.

[2] Barbour, A. D. and Chryssaphinou, O. (2001). Compound Poisson approximation: a user's guide. *Ann. Appl. Prob.* **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*, eds C. A. Charalambides, M. V. Koutras and N. Balakrishnan, Chapman and Hall/CRC Press, Boca Raton, FL, pp. 63–86.

[4] Boutsikas, M. V. and Koutras, M. V. (2002). Modeling claim exceedances over thresholds. *Insurance Math. Econom.* **30,** 67–83.

[5] Deheuvels, P. and Devroye, L. (1987). Limit laws of Erdős–Rényi–Shepp type. *Ann. Prob.* **15,** 1363–1386.

[6] Erdős, P. and Rényi, A. (1970). On a new law of large numbers. *J. Anal. Math.* **23,** 103–111.

[7] Fu, J. C. (2001). Distribution of the scan statistic for a sequence of bistate trials. *J. Appl. Prob.* **38,** 908–916.

[8] Fu, J. C. and Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns and Its Applications. A Finite Markov Chain Imbedding Approach.* World Scientific, Singapore.

[9] Glaz, J. and Balakrishnan, N. (eds) (1999). *Scan Statistics and Applications.* Birkhäuser, Boston, MA.

[10] Glaz, J. and Naus, J. I. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *Ann. Appl. Prob.* **1,** 306–318.

[11] Glaz, J. and Zhang, Z. (2004). Multiple window discrete scan statistics. *J. Appl. Stat.* **31,** 967–980.

[12] Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*. Springer, New York.

[13] Höglund, T. (1979). A unified formulation of the central limit theorem for small and large deviations from the mean. *Z. Wahrscheinlichkeitsth.* **49,** 105–117.

[14] Petrov, V. V. (1965). On the probabilities of large deviations for sums of random variables. *Theory Prob. Appl.* **10,** 287–298.

[15] Roos, M. (1994). Stein's method for compound Poisson approximation: the local approach. *Ann. Appl. Prob.* **4,** 1177–1187.

https://doi.org/10.1239/jap/1165505213 Published online by Cambridge University Press