

ASYMPTOTIC RESULTS FOR THE MULTIPLE SCAN STATISTIC

M.V. Boutsikas, M.V. Koutras and F.S. Milienos*

Dept. of Statistics and Insurance Science, University of Piraeus

Abstract

The contribution of the theory of scan statistics to the study of many real life applications has been rapidly expanding during the last decades. The multiple scan statistic, defined on a sequence of n Bernoulli trials, enumerates the number of occurrences of k consecutive trials which contain at least r successes among them ($r \leq k \leq n$). In the present article, we establish some asymptotic results for the distribution of the multiple scan statistic, as $n, k, r \rightarrow \infty$ and illustrate their accuracy through a simulation study. Our approach is based on an appropriate combination of compound Poisson approximation and random walk theory.

Keywords: Scan statistics; compound Poisson approximations; Kolmogorov distance; total variation distance; convergence in distribution; random walks.

MSC 2010: Primary 62E17, Secondary 60F05

*The research of M.V. Koutras and F.S. Milienos was supported by the European Union (European Social Fund-ESF) and Greek national funds through the Operational Program Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)-Research Funding Program: Aristeia II. Investing in knowledge society through the ESF.

1 Introduction

Quite often a statistician faces the problem of studying a sequence of discrete or continuous random variables (outcomes) in order to investigate the occurrence of events beyond her/his expectations. If the sequence under study consists of Bernoulli random variables then plausible criteria, for testing the uniformity and independence of outcomes, are usually based on the theory of runs; a well known example is the runs test (Wald and Wolfowitz, 1943). Increased flexibility in the test criteria to be practiced can be achieved by resorting to the scan statistic and its variations. In its most popular version, the scan statistic $S_{n,k}$ is defined as the maximum number of successes within k consecutive Bernoulli trials, among a total number of n trials ($k \leq n$). The scan statistic has been fruitfully exploited in a diverse spectrum of research areas such as quality control, reliability theory, biosurveillance, molecular biology, start-up demonstration testing, actuarial science etc (e.g. Glaz and Naus, 1991; Koutras, 1996; Boutsikas and Koutras, 2002, 2003; Balakrishnan and Koutras, 2002; Glaz et al. 2009; Balakrishnan et al. 2014).

Additional descriptive power, for studying the underlying problems, is provided by making use of the multiple scan statistic $W_{n,k,r}$. The multiple scan statistic denotes the total number of occurrences of overlapping moving windows of length k (among n Bernoulli trials) which contain at least r successes ($r \leq k$); note that $P(W_{n,k,r} = 0) = P(S_{n,k} < r)$. Since in most applications the interest focuses on extremely long periods ($n, k \rightarrow \infty$) and, as stated in the related literature, the evaluation of the exact distribution of scan related statistics becomes computationally intractable even for n, k and r of moderate size (e.g. Koutras and Alexandrou, 1995; Fu, 2001; Fu

et al. 2012; Wu et al. 2013), the development of asymptotic results and approximations has received special research interest. The approximations that have been suggested so far could be divided into three main categories: Poisson, compound Poisson and product-type approximations (e.g. Glaz et al. 1994; Chen and Glaz, 1999). Since windows with high concentration of successes tend to occur in clumps (local groups), an approximation of $W_{n,k,r}$'s distribution by an appropriate compound Poisson distribution is expected to offer the most efficient approach; Boutsikas et al. (2009) reviewed a number of asymptotic results of this kind.

Boutsikas and Koutras (2006) considered the problem of establishing compound Poisson approximations for $W_{n,k,r}$ that yield upper bounds which are manageable under the conditions $n, k, r \rightarrow \infty$, with the success probability remaining fixed. In this paper, motivated by the fact that the compounding distribution exploited there does not admit a neat form, we practiced a new approach which makes profitable use of an appropriate random walk scheme. After introducing the necessary definitions and notations in Section 2, the limiting form of the compounding distribution is investigated in Section 3 by studying its relation to the total number of non-negative partial sums, of a properly defined random walk. Finally, in Section 4, a simulation study is carried out for examining the accuracy of the approximation.

2 Definitions and notations

Let $\{X_i, i \in \mathbb{N}\}$ be a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables (r.v.'s) with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p = q$, $i \in \mathbb{N}$. The k -scan process $\{S_i, i \in \mathbb{N}\}$ and the multiple scan

statistic $W_{n,r,k}$ are defined as

$$S_i := \sum_{j=i}^{i+k-1} X_j, i \in \mathbb{N}, \quad \text{and} \quad W_n = W_{n,r,k} := \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_i),$$

for $n, k, r \in \mathbb{N}$ ($r \leq k \leq n$); the symbol $I_A(x)$ denotes the indicator function of A , i.e. $I_A(x) := 1$ if $x \in A$, and 0 otherwise. The (simple) scan statistic $S_{n,k}$ is then given by $S_{n,k} := \max_{1 \leq i \leq n-k+1} S_i$. Let

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

denote the Kolmogorov distance between the distributions of the r.v's (or vectors) X and Y . Moreover, the total variation distance between the distributions of X, Y , taking values in a countable space D , is defined by

$$d_{TV}(X, Y) := \frac{1}{2} \sum_{d \in D} |P(X = d) - P(Y = d)|.$$

Boutsikas and Koutras (2006) proved that $d(W_n, CP(\lambda_{n,r,k}, H_{r,k})) \leq b_{n,r,k}$, where $CP(\lambda_{n,r,k}, H_{r,k})$ denotes the compound Poisson distribution with parameter $\lambda_{n,r,k}$ and compounding distribution $H_{r,k}$. The bound $b_{n,r,k}$, and the parameter $\lambda_{n,r,k}$ possess an explicit closed form (cf. Th. 2 in Boutsikas and Koutras, 2006) while

$$H_{r,k}(x) = P\left(\sum_{i=k+1}^{2k+1} I_{[r,\infty)}(S_i) \leq x | S_1 < r, \dots, S_k < r, S_{k+1} \geq r\right). \quad (2.1)$$

Throughout this paper, the parameters $r = r_n, k = k_n$ are considered as functions of n , with $r_n, k_n \rightarrow \infty$ as $n \rightarrow \infty$ ($r_n \leq k_n$), although n will be sometimes suppressed in the notation, for typographical convenience. It can be verified that, if p is fixed and $\theta \in (p, 1)$ such that $(r_n - \theta k_n)/\sqrt{k_n} \rightarrow 0$ as $n \rightarrow \infty$ then the upper bound $b_{n,r,k}$ vanishes, and thus W_n converges weakly to a compound Poisson distribution. Unfortunately, $H_{r,k}$ does not admit an explicit closed form, a fact making the task of calculating the asymptotic

distribution of W_n almost infeasible. The aim of the next section is to study the limiting form of $H_{r,k}$, so that a numerically tractable scheme is produced.

3 The asymptotic distribution of the multiple scan statistics

Taking into account that the differences $S_i - S_{i-1}$ can take on only the values $0, 1$ or -1 , it seems plausible to examine the resemblance of the sequence $\{S_i, i \in \mathbb{N}\}$ to a random walk process with similar steps. In Theorem 3, which is one of the key results of this article, we prove that there does exist an appropriate random walk process with feasible steps $\{-1, 0, 1\}$ which describe the asymptotic behavior of $H_{r,k}$. Before establishing this Theorem, we need to develop some auxiliary results. To start with, let us consider two independent collections of i.i.d. Bernoulli r.v.'s $X_1, \dots, X_k$ and $Z_1, \dots, Z_k$ such that $P(X_i = 1) = p$ and $P(Z_i = 1) = \pi$, for $i = 1, \dots, k$ and $p, \pi \in (0, 1)$. If $\mathbf{H}_m = (H_1, \dots, H_m)$ denotes a random vector distributed according to the conditional distribution of $\mathbf{X}_m = (X_1, \dots, X_m)$ given $\sum_{i=1}^k X_i = r$, one may easily verify that

$$P(\mathbf{H}_m = \mathbf{x}_m) = \frac{(r)_{\sum_{i=1}^m x_i} (k-r)_{\sum_{i=1}^m (1-x_i)}}{(k)_m}.$$

The total variation distance between $\mathbf{H}_m$ and $\mathbf{Z}_m = (Z_1, \dots, Z_m)$ can be expressed as

$$d_{TV}(\mathbf{Z}_m, \mathbf{H}_m) = \frac{1}{2} \sum_{i=0}^m \sum_{\mathbf{x} \in \{0,1\}^m: \sum x_j = i} \left| \pi^i (1-\pi)^{m-i} - \frac{(r)_i (k-r)_{m-i}}{(k)_m} \right|,$$

and denoting by $b(m, \pi)$ and $HG(r, k, m)$ the binomial and the hypergeometric distribution with parameters $m \in \{1, 2, \dots\}$, $\pi \in (0, 1)$ and $r, k, m \in$

$\{1, 2, \dots\}$, respectively, we may write that

$$d_{TV} \left(\mathbf{Z}_m, \mathbf{X}_m \mid \sum_{i=1}^k X_i = r \right) = d_{TV}(b(m, \pi), HG(r, k, m)). \quad (3.1)$$

Let us next introduce the events $C_0 := [S_{k+1} \geq r, S_k < r]$, $C_1 := [S_{k-1} < r, \dots, S_{k-h+1} < r]$ and $C_2 := [S_{k-h} < r, \dots, S_1 < r]$, with h being positive integer with $h \leq k/2$. The next two Lemmas provide some useful upper bounds for the conditional probabilities associated with the events C_0, C_1, C_2 and the random variables

$$Q_1 := \sum_{i=k+1}^{k+h} I_{[r, \infty)}(S_i), \quad Q_2 := \sum_{i=k+h+1}^{2k} I_{[r, \infty)}(S_i).$$

The proofs of both Lemmas are given in the Appendix.

Lemma 1. *Let $B := p(1-p) + \frac{r-1}{k-1} \left(1 - \frac{r-1}{k-1}\right)$. Then*

$$P(C_2^c | C_0) \leq \frac{k-1}{(h\mu)^2} B, \quad P(Q_2 > 0 | C_0) \leq \frac{k-1}{(h\mu+1)^2} B.$$

Assume now that $(H_{k+1}, H_{k+2}, \dots, H_{2k-1})$ is a random vector distributed according to the conditional distribution of $(X_{k+1}, X_{k+2}, \dots, X_{2k-1})$ given that $\sum_{j=k+1}^{2k-1} X_j = r-1$, and introduce a set of $k-1$ independent r.v.'s $H'_{k+1}, \dots, H'_{2k-1}$ such that $H'_{k+i} \stackrel{D}{=} H_{k+i}, i = 1, \dots, k-1$. Denote also by $P_{x,h}, P_h$ the probabilities

$$P_{x,h} := P \left(\sum_{i=1}^{h-1} I_{[0, \infty)} \left(\sum_{j=1}^i (X_{2k+j} - H'_{k+j}) \right) \leq x-1 \right), x = 1, 2, \dots, \quad (3.2)$$

$$P_h := P \left(\sum_{i=1}^{h-1} I_{(0, \infty)} \left(\sum_{j=1}^i (X_{k-j} - H'_{2k-j}) \right) = 0 \right). \quad (3.3)$$

We are now ready to state the second Lemma.

Lemma 2. *We have that $|P(C_1 | C_0) - P_h| < \frac{h-2}{k-2}$, and*

$$|P(Q_1 \leq x, C_1 | C_0) - P_{x,h} P_h| \leq \frac{2h-3}{k-2}. \quad (3.4)$$

Making use of the aforementioned results we shall now proceed to the proof of the next Theorem which associates the compounding distribution (2.1) to a three step random walk distribution.

Theorem 3. *Let $\{\xi_i, i \in \mathbb{N}_0\}$ be a sequence of i.i.d. r.v.'s taking on the values 1, -1 and 0 with probabilities $\alpha = p(1 - \theta)$, $\gamma = \theta(1 - p)$ and $\beta = 1 - p(1 - \theta) - \theta(1 - p)$ respectively, where $0 < p < \theta < 1$. Define also $\Xi_n := \sum_{i=1}^n \xi_i, n \geq 0$ and*

$$\Xi := \sum_{n=0}^{\infty} I_{[0, \infty)}(\Xi_n), \quad (3.5)$$

($\Xi_0 := 0$). If $r/k \xrightarrow{n \rightarrow \infty} \theta \in (p, 1)$, then $H_{r,k}(x) \rightarrow P(\Xi \leq x)$ for $x = 0, 1, \dots$

Proof. Note first that on introducing the auxiliary function $H^*(x) = P(Q_1 + Q_2 \leq x | C_0 C_1 C_2)$ we may prove that

$$0 \leq H^*(x) - H_{r,k}(x) \leq P(S_{2k+1} \geq r) \rightarrow 0, \quad (3.6)$$

and the RHS converges exponentially fast to zero when $(r_n - \theta k_n)/\sqrt{k_n} \rightarrow 0$.

Moreover, one can easily verify that

$$\begin{aligned} H^*(x) &\leq \frac{P(Q_1 \leq x, C_1 C_0)}{P(C_0 C_1) - P(C_2^c C_0 C_1)} \leq \frac{P(Q_1 \leq x, C_1 | C_0)}{P(C_1 | C_0) - P(C_2^c | C_0)}, \\ H^*(x) &\geq P(AC_2 | C_0 C_1) \geq P(A | C_0 C_1) - P(C_2^c | C_0 C_1) \\ &\geq P(Q_1 \leq x | C_0 C_1) - P(Q_2 > 0 | C_0 C_1) - P(C_2^c | C_0 C_1) \\ &\geq \frac{P(Q_1 \leq x, C_1 | C_0) - P(Q_2 > 0 | C_0) - P(C_2^c | C_0)}{P(C_1 | C_0)}. \end{aligned} \quad (3.7)$$

Combining (3.7) with Lemmas 1, 2 we arrive at the next inequality

$$\frac{P_{x,h} P_h - \frac{2h-3}{k-2} - \frac{(k-1)B}{(h\mu+1)^2} - \frac{(k-1)B}{h^2 \mu^2}}{P_h + \frac{h-2}{k-2}} \leq H^*(x) \leq \frac{P_{x,h} P_h + \frac{2h-3}{k-2}}{P_h - \frac{h-2}{k-2} - \frac{(k-1)}{h^2 \mu^2} B}. \quad (3.8)$$

Making use of the fact that the sum $\sum_{j=1}^i (X_{k-j} - H'_{2k-j}), i = 1, 2, \dots$ describes a random walk with steps 1, -1, 0 and respective limiting probabilities $p(1 - \theta)$, $\theta(1 - p)$ and $1 - p(1 - \theta) - \theta(1 - p)$, we may verify that under

Theorem's conditions, the choice $h = \lfloor k^{2/3} \rfloor$ yields

$$H^*(x) \rightarrow \lim_{n \rightarrow \infty} P_{x,h} = P\left(\sum_{i=0}^{\infty} I_{[0,\infty)}(\Xi_i) \leq x\right) = P(\Xi \leq x). \quad (3.9)$$

For the proof of (3.9) one should keep in mind that

$$P_h \rightarrow P\left(\sum_{i=1}^{\infty} I_{(0,\infty)}(\Xi_i) = 0\right) = P(\Xi \leq 1)/P(\xi_1 = -1) > 0.$$

□

From the proof of the above theorem (cf. (3.8),(3.6)), one may confer that the rate of convergence of $H_{r,k}(x)$ to $P(\Xi \leq x)$ is at least $O(k^{-1/3})$. The usefulness of the above asymptotic result can be substantiated only if a simple form for the distribution of Ξ can be developed. The next Lemma is the first step towards obtaining a very simple form for the probability generating function of Ξ . Let $\mathcal{E}$ stand as abbreviation for the attribute “*the random walk process $\{\Xi_n, n \in \mathbb{N}\}$ returns to zero*”.

Lemma 4. *The generating function of the sequence $u_0 = 1$, $u_n = P(\mathcal{E} \text{ occurs at the } n\text{-th trial}) = P(\Xi_n = 0), n = 1, \dots$ is given by*

$$U(s) = \sum_{n=0}^{\infty} u_n s^n = \frac{1}{\sqrt{(1 - \beta s)^2 - 4\alpha\gamma s^2}}.$$

Proof. Apparently, $u_n = P(\sum_{j=1}^n \xi_j = 0)$ equals

$$\sum_{i=0}^n P\left(\sum_{j=1}^n \xi_j = 0, \sum_{j=1}^n I_{\{0\}}(\xi_j) = i\right) = \sum_{i=0}^n \binom{n}{\frac{n-i}{2}, i} (\alpha\gamma)^{\frac{n-i}{2}} \beta^i$$

(if $\frac{n-i}{2} \notin \mathbb{N}$ the binomial operator is set to 0). Therefore,

$$U(s) = \sum_{n=0}^{\infty} u_n s^n = \sum_{n=0}^{\infty} \sum_{i=0}^n \binom{n}{\frac{n-i}{2}, i} (\alpha\gamma)^{\frac{n-i}{2}} \beta^i s^n = \sum_{i=0}^{\infty} G_i(s) \frac{(\beta s)^i}{i!}, \quad (3.10)$$

where

$$\begin{aligned} G_i(s) &:= \sum_{n=i}^{\infty} i! \binom{n}{\frac{n-i}{2}, i} (\alpha\gamma s^2)^{\frac{n-i}{2}} = \sum_{j=0}^{\infty} i! \binom{2j+i}{j, i} (\alpha\gamma s^2)^j \\ &= \sum_{j=0}^{\infty} \frac{(2j+i)!}{(2j)!} \binom{-\frac{1}{2}}{j} (-1)^j (\sqrt{4\alpha\gamma s^2})^{2j}. \end{aligned}$$

In order to simplify the above sum we consider the function

$$g_i(x) := \sum_{j=0}^{\infty} \binom{-\frac{1}{2}}{j} (-1)^j x^{2j+i} = x^i h(x),$$

where $h(x) = (1 - x^2)^{-\frac{1}{2}}$. By taking the i -th order derivatives of the above we get

$$g_i^{(i)}(x) = \sum_{j=0}^{\infty} \frac{(2j+i)!}{(2j)!} \binom{-\frac{1}{2}}{j} (-1)^j x^{2j} = \sum_{j=0}^{\infty} \binom{i}{j} h^{(j)}(x) \frac{i!}{j!} x^j.$$

Since $G_i(s) = g_i^{(i)}(\sqrt{4\alpha\gamma s^2})$, from (3.10), we have

$$\begin{aligned} U(s) &= \sum_{i=0}^{\infty} g_i^{(i)}(\sqrt{4\alpha\gamma s^2}) \frac{\beta^i s^i}{i!} = \frac{1}{1 - \beta s} \sum_{j=0}^{\infty} \frac{h^{(j)}(\sqrt{4\alpha\gamma s^2})}{j!} \left(\frac{\beta s \sqrt{4\alpha\gamma s^2}}{1 - \beta s} \right)^j \\ &= \frac{1}{1 - \beta s} h \left(\frac{\beta s \sqrt{4\alpha\gamma s^2}}{1 - \beta s} + \sqrt{4\alpha\gamma s^2} \right) = \frac{1}{\sqrt{(1 - \beta s)^2 - 4\alpha\gamma s^2}}. \end{aligned}$$

□

We are ready to establish the next Theorem pertaining to the exact distribution of Ξ .

Theorem 5. (i) The probability generating function of Ξ is given by

$$M(s) = E(s^{\Xi}) = \frac{2s(\theta - p)}{1 - s + s(\theta - p) + \sqrt{(1 - (1 - \theta - p)s)^2 - 4\theta ps}}. \quad (3.11)$$

(ii) Ξ follows a compound geometric distribution; more specifically, $\Xi \stackrel{D}{=} 1 + \sum_{i=1}^M V_i$ where M follows a geometric distribution with success probability $f = 1 - (\gamma - \alpha)$ and V_i 's are i.i.d. r.v's (independent of M) with

$$P(V_i = x) = \frac{1}{x(x-1)f} \sum_{i=1}^{\lfloor x/2 \rfloor} i \binom{x}{i, i} (\alpha\gamma)^i \beta^{x-2i}, \quad x = 2, 3, \dots$$

and $P(V_i = 1) = \frac{\beta}{2f} + \frac{1}{2}$ (where α, β, γ are as given in Theorem 5).

Proof. (i) Let $F(s) = \sum_{n=1}^{\infty} f_n s^n$ and $f_n = P(\mathcal{E} \text{ occurs for the first time at the } n\text{-th trial})$ ($f_0 = 0$). It is known (e.g. Feller, 1968) that $F(s) = 1 - U(s)^{-1}$. Moreover, since $u = U(1) = (\theta - p)^{-1} < \infty$, $\mathcal{E}$ is transient and the probability that $\mathcal{E}$ ever occurs equals $f = F(1) = 1 - (\theta - p)$. Observe now that the number of times the process $\{\Xi_j, j \in \mathbb{N}\}$ returns to zero equals $M = \sum_{j=1}^{\infty} I_{\{0\}}(\Xi_j)$ and is a geometrically distributed r.v. with $P(M = m) = f^m(1 - f), m = 0, 1, \dots$. Denote next by V the number of steps of the random walk on a non-negative path until the first occurrence of $\mathcal{E}$. Exploiting the reflection principle we conclude that $P(V = \eta, M > 0) = f_\eta/2$, for $\eta \geq 2$. It is obvious that,

$$\Xi = \sum_{i=0}^{\infty} I_{[0, \infty)}(\Xi_i) \stackrel{D}{=} 1 + V_1 + \dots + V_M, \quad (3.12)$$

where $V_1, V_2, \dots$ is a sequence of i.i.d. r.v.'s, independent of M , such that $V_i \stackrel{D}{=} V | M > 0$ (recall that $\Xi_0 = 0$). This is valid because $p < \theta$ and thus, after the last occurrence of $\mathcal{E}$, the process will remain below 0 almost surely. The generating function of V_i equals

$$E(s^{V_i}) = sP(V_i = 1) + \sum_{\eta=2}^{\infty} s^\eta \frac{f_\eta}{2f},$$

and since $E(s^{V_i}) = 1$ for $s = 1$, we get $P(V_i = 1) = \frac{1}{2} + \frac{f_1}{2f}$. Therefore, $E(s^{V_i}) = \frac{1}{2}(s + \frac{1}{f}F(s))$ and the generating function of Ξ takes on the form

$$\begin{aligned} E(s^\Xi) &= E(s^{1+\sum_{i=1}^M V_i}) = sE(E(s^{\sum_{i=1}^M V_i} | M)) = sE(E(s^{V_i})^M) = \frac{s(1-f)}{1-fE(s^{V_i})} \\ &= \frac{2s(1-f)}{2-F(s)-sf} = \frac{2sU(s)}{U(s)(u-us+s)+u}, \end{aligned}$$

The proof of part (i) is easily completed by applying Lemma 4 while part (ii) follows from (3.12) by inverting the expression of $E(s^{V_i})$ given above. $\square$

We are now ready to state the next Theorem which provides a very simple form of the asymptotic distribution of W_n . The quantity $H(\theta, p)$ used below, stands for the Kullback-Leibler distance, i.e. $H(\theta, p) = \ln \frac{\theta^\theta (1-\theta)^{1-\theta}}{p^\theta (1-p)^{1-\theta}}$.

Theorem 6. *Let $n, r, k \rightarrow \infty$ so that $r/k = \theta + o(k^{-1})$ for some $\theta \in (p, 1)$, and denote by $l \in (0, \infty)$ the limit of the sequence*

$$l_{n,r,k} := n \frac{(\theta - p) e^{-kH(\theta,p)}}{\sqrt{2\pi\theta(1-\theta)k}} \rightarrow l \in (0, \infty),$$

as $n \rightarrow \infty$. Then, W_n converges to the compound Poisson distribution with parameter l and compounding distribution H , where H is the c.d.f. of the compound geometric distribution described in Theorem 5.

The proof of the above Theorem is an immediate consequence of Theorems 3 and 5, and Th. 3 in Boutsikas and Koutras (2006).

4 Numerical results

In this section, a simulation study is carried out for assessing the accuracy of the new asymptotic results. In view of Theorem 6, for large enough n and k , which satisfy the conditions mentioned there, the distribution of the number of k consecutive trials with at least $100 \cdot \theta\%$ successes, could be approximated by a properly defined compound Poisson distribution. Special attention should be paid to the condition $\theta > p$ and the fact that, in practice, the success probability is usually unknown; so, an application of Theorem 6 requires an accurate estimation $\hat{p}$ of the success probability, in order to determine the range of θ 's where the approximation can be used (i.e. for $\theta > \hat{p}$). This can be achieved by the aid of an initial random sample that will be exploited for producing an estimate of p .

Figure 1 depicts the asymptotic and simulated values of $P(W_n \leq x)$, $x \geq 0$ for $p = 0.4$ and several values of n, k, r . The quantity d provided below each figure is the square root of the mean (squared) difference between the approximation and the simulated distribution of W_n (over the support of W_n). It is clear that, as n increases, the difference between the simulated and the asymptotic value becomes smaller. However, when θ comes close to p then the approximation becomes less accurate; this is mainly due to the fact that the limit of $l_{n,r,k}$ is very close to 0.

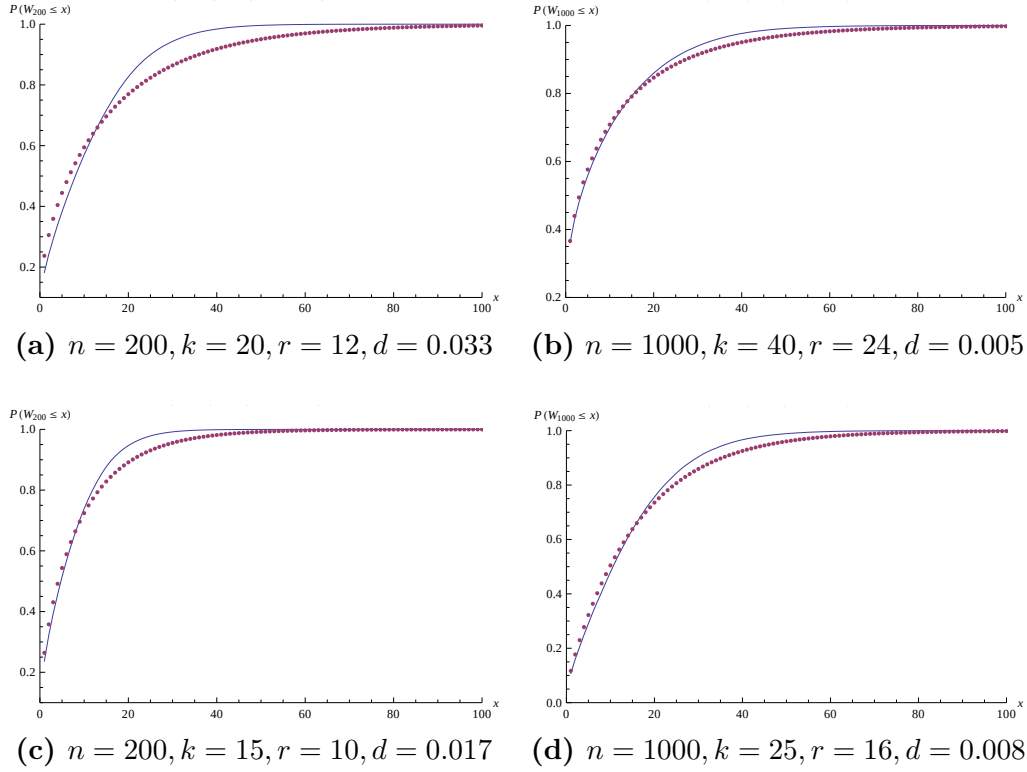


Figure 1. Asymptotic (dotted line) and simulated distribution (solid line; $3 \cdot 10^5$ iterations), with $p = 0.4$ and $\theta = r/k$.

Appendix

Proof of Lemma 1 Note first that the r.v.'s $H_{k+1}, \dots, H_{2k-1}$ are negatively associated (NA, cf. Joag-Dev and Proschan, 1983) and hence $-H_{k+1}, \dots, -H_{2k-1}$, are also NA r.v.'s. The independent r.v.'s $X_1, \dots, X_{k-1}$ are trivially NA and thus $X_1, \dots, X_{k-1}, -H_{k+1}, \dots, -H_{2k-1}$ are also NA r.v.'s, as a union of two independent families of NA r.v.'s.

In order to prove the first part of Lemma we observe that the probability $P(C_2^c|C_0)$ can be written as follows,

$$\begin{aligned} 1 - P \left(S_1 < r, \dots, S_{k-h} < r \mid X_k = 0, \sum_{j=k+1}^{2k-1} X_j = r-1, X_{2k} = 1 \right) \\ = 1 - P \left(\sum_{j=1}^s L_j \leq 0, s = h, \dots, k-1 \right) = P \left(\max_{s \in \{h, \dots, k-1\}} \sum_{j=1}^s L_j > 0 \right), \end{aligned}$$

where $L_j := X_{k-j} - H_{2k-j}$ with $\mu := E(L_j) = p - \frac{r-1}{k-1}$.

Note next that the r.v.'s $Y_1 := \sum_{j=1}^h (L_j - \mu)$, $Y_i := L_{h+i-1} - \mu$, $i = 2, \dots, k-h$ are also NA as nondecreasing functions over disjoint subsets of $X_1, \dots, X_{k-1}, -H_{k+1}, \dots, -H_{2k-1}$ and recall Th. 2 in Christofides and Vaggelatou (2004), to establish the next inequalities

$$P(C_2^c|C_0) \leq P \left(\max_{s \in \{1, \dots, k-h\}} \sum_{i=1}^s Y_i > -h\mu \right) \leq \frac{\sum_{i=1}^{k-h} V(Y_i)}{h^2 \mu^2} = \frac{(k-1)}{h^2 \mu^2} B$$

(since $r/k \rightarrow \theta > p$ we may assume that $-h\mu > 0$).

The proof of the second part is established by similar arguments, on introducing $L'_j := X_{2k+j} - H_{k+j}$, verifying the identity we can write

$$P(Q_2 > 0|C_0) = P \left(\max_{s \in \{h, \dots, k-1\}} \sum_{j=1}^s L'_j > -1 \right),$$

and making use of the fact that the r.v.'s $Y_1 := \sum_{j=1}^h (L'_j - \mu)$, $Y_2 := L'_{h+1} - \mu, \dots, Y_{k-h} = L'_{k-1} - \mu$, are NA. A direct application of Th. 2 in Christofides

and Vaggelatou (2004), yields

$$P(Q_2 > 0 | C_0) \leq P\left(\max_{s \in \{1, \dots, k-h\}} \sum_{i=1}^s Y_i > -h\mu - 1\right) \leq \frac{(k-1)}{(h\mu+1)^2} B.$$

Proof of Lemma 2. The probability $P(Q_1 \leq x, C_1 | C_0)$ equals

$$\begin{aligned} & P\left(\sum_{i=k+2}^{k+h} I_{[r, \infty)}(S_i) \leq x-1, S_{k-h+1} < r, \dots, S_{k-1} < r, | C_0\right) \\ &= P\left(\sum_{i=1}^{h-1} I_{[0, \infty)}\left(\sum_{j=1}^i L'_j\right) \leq x-1, \sum_{i=1}^{h-1} I_{(0, \infty)}\left(\sum_{j=1}^i L_j\right) = 0\right) \\ &= P((\mathbf{X}, \mathbf{H}) \in B_x), \end{aligned}$$

where

$$\begin{aligned} \mathbf{H} &:= (H_{k+1}, \dots, H_{k+h-1}, H_{2k-h+1}, \dots, H_{2k-1}), \\ \mathbf{X} &:= (X_{k-h+1}, \dots, X_{k-1}, X_{2k+1}, \dots, X_{2k+h-1}). \end{aligned}$$

and B_x is a subset of $\{0, 1\}^{4(h-1)}$ that makes the above equality true. Define $H'_{k+1}, H'_{k+2}, \dots, H'_{2k-1}$ to be independent r.v.'s, also independent of X_i 's, such that $H'_{k+i} \stackrel{D}{=} H_{k+i}, i = 1, \dots, k-1$, and set $\mathbf{H}' := (H'_{k+1}, \dots, H'_{k+h-1}, H'_{2k-h+1}, \dots, H'_{2k-1})$. Exploiting the independence of $\mathbf{X}$ from $\mathbf{H}$ and $\mathbf{H}'$ applying (3.1) we may write

$$\begin{aligned} d_{TV}((\mathbf{X}, \mathbf{H}'), (\mathbf{X}, \mathbf{H})) &= d_{TV}\left(\mathbf{H}', \mathbf{X} \mid \sum_{i=k+1}^{2k-1} X_i = r-1\right) \\ &= d_{TV}\left(b(2h-2, \frac{r-1}{k-1}), HG(r-1, k-1, 2h-2)\right) < \frac{2h-3}{k-2}. \end{aligned}$$

with the last inequality following from Example 1 in Soon (1996). The second inequality of Lemma 2 can be easily deduced by noting that

$$\begin{aligned} |P(Q_1 \leq x, C_1 | C_0) - P_{x,h} P_h| &= |P((\mathbf{X}, \mathbf{H}) \in B_x) - P((\mathbf{X}, \mathbf{H}') \in B_x)| \\ &\leq d_{TV}((\mathbf{X}, \mathbf{H}'), (\mathbf{X}, \mathbf{H})) \leq \frac{2h-3}{k-2}. \end{aligned}$$

Similarly, $P(C_1|C_0)$ can be expressed as

$$P(C_1|C_0) = P\left(\sum_{i=1}^{h-1} I_{(0,\infty)}\left(\sum_{j=1}^i L_j\right) = 0\right) = P((\widetilde{\mathbf{X}}, \widetilde{\mathbf{H}}) \in D),$$

where $\widetilde{\mathbf{H}} := (H_{2k-1}, \dots, H_{2k-h+1})$, $\widetilde{\mathbf{X}} := (X_{k-1}, \dots, X_{k-h+1})$. Introducing again a random vector $\widetilde{\mathbf{H}}' := (H'_{2k-1}, \dots, H'_{2k-h+1})$, we may write

$$|P(C_1|C_0) - P_h| = \left|P((\widetilde{\mathbf{X}}, \widetilde{\mathbf{H}}) \in D) - P((\widetilde{\mathbf{X}}, \widetilde{\mathbf{H}}') \in D)\right| \leq d_{TV}(\widetilde{\mathbf{H}}', \widetilde{\mathbf{H}}),$$

and the first inequality of Lemma 2 results by noting that the RHS is bounded above by $(h-2)/(k-2)$.

References

- [1] Balakrishnan, N. and Koutras, M.V. (2002). *Runs, Scans and Applications*. J. Wiley & Sons, NY.
- [2] Balakrishnan, N., Koutras, M.V. and Milienos, F.S. (2014). Start-up demonstration tests: models, methods and applications, with some unifications. *Appl. Stoch. Models Bus. Ind.* 30, 373-413 (discussion paper).
- [3] Boutsikas, M.V. and Koutras, M.V. (2002). Modelling claim exceedances over thresholds. *Insurance Math. Econom.* 30, 67-83.
- [4] Boutsikas, M.V. and Koutras, M.V. (2003). Bounds for the distribution of two-dimensional binary scan statistics. *Probab. Eng. Inform. Sci.* 17, 509-525.
- [5] Boutsikas, M.V. and Koutras, M.V. (2006). On the asymptotic distribution of the discrete scan statistic. *J. Appl. Prob.* 43, 1137-1154.
- [6] Boutsikas, M.V., Koutras, M.V. and Milienos, F.S. (2009). Extreme value results for scan statistics. In *Scan Statistics Methods and Applications* (Eds., Glaz, Pozdnyakov and Wallenstein), 57-82, Birkhauser.
- [7] Chen, J. and Glaz, J. (1999). Approximations for the distribution and the moments of discrete scan statistics. In *Scan Statistics and Applications* (Eds. Glaz and Balakrishnan), 27-66, Birkhauser.

- [8] Christofides, T. and Vaggelatou, E. (2004). A connection between supermodular ordering and positive/negative association. *J. Multivar. Anal.* 88,138-151.
- [9] Fu, J.C. (2001). Distribution of the scan statistic for a sequence of bivariate trials. *J. Appl. Prob.* 38, 908-916.
- [10] Fu, J. C., Wu, T. L. and Lou, W. W. (2012). Continuous, discrete, and conditional scan statistics. *J. Appl. Prob.* 49, 199-209.
- [11] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I. J.Wiley & Sons, NY.
- [12] Glaz, J. and Naus, J. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *Ann.Appl.Prob.* 1, 306-318.
- [13] Glaz, J., Naus, J., Roos, M. and Wallenstein, S. (1994). Poisson approximation for the distribution and moments of ordered m-spacings. *J. Appl. Prob.* 31, 271-281.
- [14] Glaz, J., Pozdnyakov, V. and Wallenstein, S. (2009). *Scan Statistics: Methods and Applications*. Birkhauser, Boston.
- [15] Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *Ann. Statist.* 11, 286-295.
- [16] Koutras, M.V. (1996). On a waiting time distribution in a sequence of Bernoulli trials. *Ann.In.Stat.Math.* 48, 789-806.
- [17] Koutras, M.V. and Alexandrou, V. (1995). Runs, scans and urn model distributions: A unified Markov chain approach. *Ann.In.Stat.Math.* 47, 743-766.
- [18] Soon, S. (1996). Binomial approximation for dependent indicators. *Stat.Sin.* 6, 703-714.
- [19] Wald, A. and Wolfowitz, J. (1943). An exact test for randomness in the nonparametric case based on serial correlation. *An.Math.Stat.* 14, 378-388.
- [20] Wu, T.L., Glaz, J. and Fu, J.C. (2013). Discrete, continuous and conditional multiple window scan statistics. *J. Appl. Prob.* 50, 1089-1101.