J. Appl. Prob. **49,** 199–209 (2012)
Printed in England

CONTINUOUS, DISCRETE, AND CONDITIONAL SCAN STATISTICS

JAMES C. FU * ** AND
TUNG-LUNG WU,* *University of Manitoba*
W.Y. WENDY LOU,*** *University of Toronto*

Abstract

The distributions for continuous, discrete, and conditional discrete scan statistics are studied. The approach of finite Markov chain imbedding, which has been applied to random permutations as well as to runs and patterns, is extended to compute the distribution of the conditional discrete scan statistic, defined from a sequence of Bernoulli trials. It is shown that the distribution of the continuous scan statistic induced by a Poisson process defined on (0, 1] is a limiting distribution of weighted distributions of conditional discrete scan statistics. Comparisons of rates of convergence as well as numerical comparisons of various bounds and approximations are provided to illustrate the theoretical results.

Keywords: Scan statistics; Markov chain imbedding; random permutation; Poisson process

2010 Mathematics Subject Classification: Primary 60E05
Secondary 60J10

1. Introduction

Let $N(t)$ be a Poisson process defined on (0, 1] with rate parameter λ. Given $0 < \omega \leq 1$, define

$$S(\omega, t) = N(t + \omega) - N(t)$$

as the number of events that occurred in the interval $(t, t + \omega]$. The unconditional continuous scan statistic of window size ω is defined as

$$S(\omega) = \sup_{0<t\leq 1-\omega} S(\omega, t). \tag{1.1}$$

Two recent books on scan statistics by Glaz and Balakrishnan (1999) and Glaz *et al.* (2001) provide a very good overview of the history, applications, and recent developments in this field. Under a very restricted range of parameters, the exact distribution of $S(\omega)$ has been derived; finding the exact distribution for $S(\omega)$ in general remains a hot topic. Approximations and upper and lower bounds have also been studied by many authors. For example, the upper and lower bounds of a continuous scan statistic for a Poisson process have been studied in Janson (1984). More recently, Alm (1999) and Haiman (2000) derived several approximations for scan statistics.

Received 22 March 2011; revision received 13 October 2011.
* Postal address: Department of Statistics, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada.
** Email address: fu@cc.umanitoba.ca
*** Postal address: Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, M5T 3M7, Canada.

Let $X_1, \ldots, X_n$ be a sequence of n Bernoulli trials with success probability p and failure probability $q = 1 - p$. Given an integer $1 \leq r \leq n$, the statistic

$$S_n(r) = \sup_{1 \leq i \leq n-r+1} S_n(r, i), \tag{1.2}$$

where

$$S_n(r, i) = \sum_{j=i}^{i+r-1} X_j,$$

is often referred to as the discrete scan statistic of window size r. The unconditional probability $\mathrm{P}(S_n(r) > a)$ can be obtained by averaging the conditional probabilities $\mathrm{P}(S_n(r) > a \mid \sum_{i=1}^n X_i = N)$ for $0 \leq a \leq \min(r-1, N-1)$, $N = 0, 1, \ldots, n$, over the binomial distribution, i.e.

$$\mathrm{P}(S_n(r) > a) = \sum_{N=0}^{n} \binom{n}{N} p^N q^{n-N} \mathrm{P}\left(S_n(r) > a \,\middle|\, \sum_{i=1}^{n} X_i = N \right). \tag{1.3}$$

The statistic $S_n(r)$ given $\sum_{i=1}^n X_i = N$ is often referred to as the conditional discrete scan statistic. The exact formula for computing the distribution of the discrete scan statistic, given in Karlin and McGregor (1959) and Naus (1974), is rather restrictive on the size of the window and can require excessive computational time and space. The numerical complexity and the parameter restrictions make computing the probability $\mathrm{P}(S_n(r) > a)$ very difficult via combinatorics. There are a considerable number of approximations and bounds that have been developed, such as in Naus (1982), Glaz (1989), (1992), and Chen and Glaz (1997), and many of them perform very well numerically. Fu (2001) showed that the scan statistic $S_n(r)$ is finite Markov chain imbeddable, and, hence, the probability $\mathrm{P}(S_n(r) > a)$ can be cast in terms of transition probability matrices of an imbedded Markov chain.

In this paper, the concept of finite Markov chain imbedding is extended to study the distributions of runs and patterns on random permutations, which paves a direct way to computing the conditional probability $\mathrm{P}(S_n(r) > a \mid N)$. The main purpose of this paper is to provide a rigorous proof that, for $0 < \omega \leq 1$ and considering the Poisson process with $t = 1$ and $N(1) = N$,

$$\lim_{n \to \infty} \mathrm{P}(S_n([n\omega] + k) > a \mid N) = \mathrm{P}(S(\omega) > a \mid N) \tag{1.4}$$

for any fixed integer k, where $S_n([n\omega])$ stands for $S_n(r)$ with $r = [n\omega]$, the integer part of $n\omega$. Hence, the distribution of the continuous scan statistic $S(\omega)$ can then be approximated, for large n, via

$$\mathrm{P}(S(\omega) > a) \sim \sum_{N=0}^{\infty} \frac{\lambda^N}{N!} \mathrm{e}^{-\lambda} \mathrm{P}(S_n([n\omega]) > a \mid N). \tag{1.5}$$

The connection to the conditional scan statistic for a Poisson process on $(0, 1]$ has been pointed out in Naus (1974): when n and r are large compared to a and N, (1.4) holds for n and r approaching ∞ at the rate of $r/n \to \omega$.

2. Discrete and conditional scan statistics

For the sequence $X_1, \ldots, X_n$ of Bernoulli trials with $p = \mathrm{P}(X_1 = 1) = 1 - \mathrm{P}(X_1 = 0)$, the distribution of the unconditional scan statistic $S_n(r)$ of window size r defined by (1.2)

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

has been studied directly without using (1.3) and the conditional probability of $S_n(r)$ given $\sum_{i=1}^n X_i = N$; see Koutras and Alexandrou (1995) and Fu (2001). The main tool for finding the probability $\mathrm{P}(S_n(r) < s)$, $1 \le s \le r$, is the finite Markov chain imbedding technique, which shows that the event $S_n(r) < s$ occurs if and only if a corresponding compound pattern $\Lambda_{r,s}$, generated by the event $S_n(r) < s$, does not appear in the sequence $X_1, \ldots, X_n$. It then follows that

$$\mathrm{P}(S_n(r) < s) = \mathrm{P}(W(\Lambda_{r,s}) > n) = \boldsymbol{\xi}_0 \boldsymbol{N}_{r,s}^n(p) \mathbf{1}^\top \quad \text{for all } p \in (0,1), \tag{2.1}$$

where $W(\Lambda_{r,s})$ is the waiting time random variable of seeing the compound pattern $\Lambda_{r,s}$, $\boldsymbol{N}_{r,s}^n(p)$ is the essential transition probability matrix of the imbedded Markov chain associated with the waiting time random variable $W(\Lambda_{r,s})$, $\boldsymbol{\xi}_0$ is a suitable initial state vector, and $\mathbf{1}^\top$ is the transpose of the vector $(1, \ldots, 1)$.

Following Fu (2001), given r and s, where $1 \le s \le r$, we define a collection of simple patterns

$$\mathcal{F}_{r,s} = \{\Lambda_i : \Lambda_1 = \underbrace{1 \cdots 1}_{s},\ \Lambda_2 = 101 \cdots 1, \ldots,\ \Lambda_l = \underbrace{1 \cdots 10 \cdots 01}_{r}\},$$

representing all possible simple patterns containing s successes ('1's) that begin and end with a success and that have a length no longer than r. It is easy to check that there are

$$l = \sum_{v=0}^{r-s} \binom{s-2+v}{v}$$

such simple patterns. The compound pattern $\Lambda_{r,s}$ generated by the event $S_n(r) < s$ is defined as

$$\Lambda_{r,s} = \bigcup_{\Lambda_i \in \mathcal{F}_{r,s}} \Lambda_i.$$

Furthermore, let

$$\mathcal{P}_{n,N} = \left\{ \boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) : \pi_i = 0, 1 \text{ and } \sum_{i=1}^n \pi_i = N \right\}$$

be the family of random permutations with $n - N$ '0's and N '1's. The conditional probability of the event $S_n(r) < s$ given $\sum_{i=1}^n X_i = N$ can be viewed as the probability of no pattern $\Lambda_{r,s}$ occurring in an $[n - N, N]$-specified random permutation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$. We adopt the finite Markov chain imbedding technique to obtain this conditional probability.

Before proceeding, we provide the following simple example to illustrate the foregoing formalism. Given $\sum_{i=1}^n \pi_i = 10$, $r = 4$, and $s = 3$, then the event $S_n(4) < 3$ occurs in an $[n - 10, 10]$-specified random permutation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ if and only if none of the three patterns $\Lambda_1 = 111$, $\Lambda_2 = 1011$, and $\Lambda_3 = 1101$ occur in the random permutation $\boldsymbol{\pi}$. The set of simple patterns $\mathcal{F}_{4,3} = \{\Lambda_1, \Lambda_2, \text{and } \Lambda_3\}$ and the compound pattern $\Lambda_{4,3} = \bigcup_{\Lambda_i \in \mathcal{F}_{4,3}} \Lambda_i$ are induced by the event $S_n(4) < 3$.

Continuing with the general case, let us define $W(\Lambda_{r,s}, \boldsymbol{\pi})$ as the waiting time of the compound pattern on the random permutation $\boldsymbol{\pi} \in \mathcal{P}_{n,N}$. It follows that

$$\mathrm{P}\left(S_n(r) < s \,\middle|\, \sum_{i=1}^n X_i = N \right) = \mathrm{P}(W(\Lambda_{r,s}, \boldsymbol{\pi}) > n \mid \boldsymbol{\pi} \in \mathcal{P}_{n,N}). \tag{2.2}$$

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

The right-hand probability of the above equation can be obtained by using the following nonhomogeneous Markov chain imbedding procedure. Let us consider taking balls out one by one from an urn containing $n - N$ '0' balls and N '1' balls, sampling without replacement until all the balls are taken out. Each new realization generated by the procedure is in one-to-one correspondence with a permutation $\boldsymbol{\pi} \in \mathcal{P}_{n,N}$. Define a set of ending blocks (or subpatterns) generated by the compound pattern $\Lambda_{r,s}$ as

$$E_{r,s} = \{\phi, \omega_1 = 1,\ \omega_2 = 10, \ldots,\ \omega_k = 1 \cdots 10 \cdots 01, \alpha\},$$

where ϕ stands for the empty set and α for the absorbing state corresponding to the compound pattern $\Lambda_{r,s}$, and then define a state space as

$$\Omega = \{(l, \omega) \colon l = 0, 1, \ldots, N \text{ and } \omega \in E_{r,s}\} \cup \{\phi, \alpha\},$$

with the understanding that $\phi = (0, \phi)$ represents the initial state and $\alpha \equiv (\cdot, \alpha)$ the absorbing state. We define a nonhomogeneous Markov chain $\{Y_t\}_0^n$ on Ω having an initial probability $\mathrm{P}(Y_0 = \phi) = 1$ and an absorbing state α. Given $Y_{t-1} = (l_{t-1}, \omega_{t-1})$, the first coordinate keeps track of the number of successes that have occurred up to time $t - 1$, $\sum_{i=1}^{t-1} \pi_i = l_{t-1}$, and the second coordinate, $\omega_{t-1} \in E_{r,s}$, is the ending block of the sequence $\pi_1, \ldots, \pi_{t-1}$ (the first $t - 1$ draws). The transition probabilities, for given $u = (l, \omega)$ and $v = (l', \omega')$, are given by

$$\begin{aligned} p_{u,v}^{(t)} &= \mathrm{P}(Y_t = (l', \omega') \mid Y_{t-1} = (l, \omega)) \\ &= \begin{cases} \dfrac{N - l}{n - t + 1} & \text{if } \pi_t = 1, l' = l + 1, \text{ and } \omega' = \langle \omega, 1 \rangle_{E_{r,s}}, \\ \dfrac{n - N - t + l + 1}{n - t + 1} & \text{if } \pi_t = 0, l' = l, \text{ and } \omega' = \langle \omega, 0 \rangle_{E_{r,s}}, \\ 1 & \text{if } \omega = \omega' = \alpha, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{2.3}$$

where $\langle \omega, \pi_t \rangle_{E_{r,s}}$ denotes the longest ending block in $E_{r,s}$ of the sequence $\pi_1, \ldots, \pi_{t-1}, \pi_t$.

Clearly, $\{Y_t\}$ is a nonhomogeneous Markov chain defined on the state space Ω with transition probability matrices of the form

$$\boldsymbol{M}_t(r, s) = \begin{matrix} \Omega \setminus \alpha \\ \alpha \end{matrix} \left[\begin{array}{c|c} \boldsymbol{N}_t(r, s, N) & \boldsymbol{C}_t(r, s) \\ \hline \boldsymbol{O} & \boldsymbol{I} \end{array} \right] \tag{2.4}$$

for $t = 1, \ldots, n$, where $\boldsymbol{N}_t(r, s, N) = (p_{u,v}^{(t)})$, $t = 1, \ldots, n$, are the essential transition probability matrices and its elements $p_{u,v}^{(t)}$ are given by (2.3). Our construction, together with (2.3) and (2.4), shows that $W(\Lambda_{r,s}, \boldsymbol{\pi})$ is finite Markov chain imbeddable. Hence, (2.2) yields the first part of the following theorem.

Theorem 2.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of Bernoulli trials.*

(i) *The conditional probability of $S_n(r) < s$ given $\sum_{i=1}^n X_i = N$ is equal to*

$$\mathrm{P}\left(S_n(r) < s \;\middle|\; \sum_{i=1}^n X_i = N \right) = \boldsymbol{\xi}_0 \prod_{t=1}^n \boldsymbol{N}_t(r, s, N) \mathbf{1}^\top \tag{2.5}$$

for $1 \le s \le \min(r, N)$, and equal to 1 if $N < s$.

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

(ii) *The unconditional probability is given by*

$$\mathrm{P}(S_n(r) < s) = \sum_{N=0}^{n} \binom{n}{N} p^N q^{n-N} \left(\boldsymbol{\xi}_0 \prod_{t=1}^{n} \boldsymbol{N}_t(r, s, N) \mathbf{1}^\top \right).$$

The result in part (ii) follows directly from part (i) and the fact that N has a binomial distribution. Note that the conditional probability $\mathrm{P}(S_n(r) < s \mid \sum_{i=1}^{n} X_i = N)$ does not depend on the parameter value p. The main reason for developing the distribution of the conditional discrete scan statistic is to approximate the distribution of the continuous scan statistic $S(\omega)$, with Poisson process arrival function $N(t)$ as defined by (1.1). As a byproduct, (2.5) gives the p-value for the conditional test based on the scan statistic $S_n(r)$ given $\sum_{i=1}^{n} X_i = N$.

3. Continuous scan statistic

It is well known that, for a Poisson process, given $N(1) = N$, the N points are uniformly distributed on $(0, 1]$. Let us divide the interval $(0, 1]$ into n subintervals $(t_0, t_1], \ldots, (t_{n-1}, t_n]$, with $t_0 = 0, t_n = 1$, and each of equal length $t_i - t_{i-1} = \Delta t = 1/n$ for all $i = 1, \ldots, n$. This implies that the $[n - N, N]$-specified random permutations in $\mathcal{P}_{n,N}$ have the same probability.

Let us consider the continuous scan statistics $S(\omega) = \sup_{0<t\le 1-\omega}(N(t + \omega) - N(t))$ of window size ω, $0 < \omega \le 1$.

Theorem 3.1. *For $a < N$,*

$$\mathrm{P}(S(\omega) > a \mid N) = \lim_{n\to\infty} \mathrm{P}(S_n([n\omega]) > a \mid N),$$

where $[n\omega]$ is the integer part of $n\omega$.

To prove the above theorem, we introduce the following two lemmas.

Lemma 3.1. *It holds that*

$$\max_{1\le i\le n-[n\omega]} S_n([n\omega], i) \le \sup_{0<t\le 1-\omega} S(\omega, t) \le \max_{1\le i\le n-[n\omega]-1} S_n([n\omega] + 2, i).$$

Proof. It is easy to see that

$$\sup_{0<t\le 1-\omega} S(\omega, t) = \max_{1\le i\le n-[n\omega]} \sup_{t_{i-1}<t\le t_i} S(\omega, t).$$

For given $i = 1, \ldots, n - [n\omega]$, it follows from the definition of $\sup_{t_{i-1}<t\le t_i} S(\omega, t)$ that

$$\max(S_n([n\omega], i), S_n([n\omega], i + 1)) \le \sup_{t_{i-1}<t\le t_i} S(\omega, t). \tag{3.1}$$

Note also that

$$\max_{1\le i\le n-[n\omega]} S_n([n\omega], i) = \max_{1\le i\le n-[n\omega]-1} \max(S_n([n\omega], i), S_n([n\omega], i + 1)).$$

By the same token, we have, for $i = 1, \ldots, n - [n\omega] - 1$,

$$\sup_{t_{i-1}<t\le t_i} S(\omega, t) \le S_n([n\omega] + 2, i). \tag{3.2}$$

The result follows from (3.1) and (3.2) by taking the maximum of both sides.

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

Lemma 3.2. *Given* $0 < \omega \le 1$ *and* $a < N$, *we have, for all* $i = 1, \ldots, n - [n\omega] - 1$,

$$\text{(i)} \quad |\mathrm{P}(S_n([n\omega]+2, i) > a \mid N) - \mathrm{P}(S_n([n\omega], i) > a \mid N)| = O\left(\frac{1}{n}\right),$$

$$\text{(ii)} \quad |\mathrm{P}(S_n([n\omega]+2, i) > a \mid N) - \mathrm{P}(S_n([n\omega], i+1) > a \mid N)| = O\left(\frac{1}{n}\right).$$

Proof. For given i and n, it follows that

$$S_n([n\omega]+2, i) = \sum_{j=i}^{i+[n\omega]+1} \pi_j \ge S_n([n\omega], i) = \sum_{j=i}^{i+[n\omega]-1} \pi_j.$$

For given a, N, and large n, the above equation yields

$$\begin{aligned}
&\mathrm{P}(S_n([n\omega]+2, i) > a \mid N) - \mathrm{P}(S_n([n\omega], i) > a \mid N) \\
&\quad = \mathrm{P}(S_n([n\omega]+2, i) > a,\ S_n([n\omega], i) \le a \mid N) \\
&\quad = \mathrm{P}(S_n([n\omega]+2, i) = a+1,\ S_n([n\omega], i) = a \text{ or } a-1 \mid N) \\
&\qquad + \mathrm{P}(S_n([n\omega]+2, i) = a+2,\ S_n([n\omega], i) = a \mid N) \\
&\quad = \mathrm{P}(S_n([n\omega]+2, i) = a+1,\ \pi_{i+[n\omega]} + \pi_{i+[n\omega]+1} = 1 \text{ or } 2 \mid N) \\
&\qquad + \mathrm{P}(S_n([n\omega]+2, i) = a+2,\ \pi_{i+[n\omega]} + \pi_{i+[\omega]+1} = 2 \mid N) \\
&\quad \le \frac{2N}{n},
\end{aligned}$$

which is independent of i. This proves part (i). Part (ii) can be proved in the same fashion, and is thus omitted.

Proof of Theorem 3.1. For $N \ge 1$,

$$\mathrm{P}(S_n([n\omega]) > 0 \mid N) \equiv 1 \quad \text{and} \quad \mathrm{P}(S_n([n\omega]) > N \mid N) \equiv 0$$

for all n; hence, Theorem 3.1 holds. It follows from Lemma 3.1 that we have the inequality

$$\begin{aligned}
\mathrm{P}\Big(\max_{1 \le i \le n-[n\omega]} S_n([n\omega], i) > a \;\Big|\; N\Big) &\le \mathrm{P}\Big(\sup_{0 < t \le 1-\omega} S(\omega, t) > a \;\Big|\; N\Big) \\
&\le \mathrm{P}\Big(\max_{1 \le i \le n-[n\omega]-1} S_n([n\omega]+2, i) > a \;\Big|\; N\Big). \qquad (3.3)
\end{aligned}$$

Note that, for fixed ω, N, $1 < a < N$, and large n, the number of is such that $S_n([n\omega]+2, i) > a$ and $\max_{1 \le i \le n-[n\omega]} S_n([n\omega], i) \le a$ can only be less than or equal to N. The following inequality then follows as a consequence of the Bonferroni inequality and Lemma 3.2:

$$\begin{aligned}
&\mathrm{P}\Big(\max_{1 \le i \le n-[n\omega]-1} S_n([n\omega]+2, i) > a \;\Big|\; N\Big) - \mathrm{P}\Big(\max_{1 \le i \le n-[n\omega]} S_n([n\omega], i) > a \;\Big|\; N\Big) \\
&\quad = \mathrm{P}\Big(\max_{1 \le i \le n-[n\omega]-1} S_n([n\omega]+2, i) > a,\ \max_{1 \le i \le n-[n\omega]} S_n([n\omega], i) \le a \;\Big|\; N\Big) \\
&\quad \le N \max_{1 \le i \le n-[n\omega]} \mathrm{P}(S_n([n\omega]+2, i) > a,\ S_n([n\omega], i) \le a \mid N) \\
&\quad \le \frac{2N^2}{n}.
\end{aligned}$$

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

Hence, for given N and a,

$$\lim_{n\to\infty} \mathrm{P}(S_n([n\omega]) > a \mid N) = \mathrm{P}(S(\omega) > a \mid N) = \lim_{n\to\infty} \mathrm{P}(S_n([n\omega]+2) > a \mid N).$$

This completes the proof.

Remark 3.1. Inequality (3.3) is equivalent to

$$\mathrm{P}(S_n([n\omega]) > a \mid N) \le \mathrm{P}(S(\omega) > a \mid N) \le \mathrm{P}(S_n([n\omega]+2) > a \mid N),$$

and we expect that $\mathrm{P}(S_n([n\omega]) > a \mid N)$ increases monotonically in terms of $[n\omega]$ to $\mathrm{P}(S(\omega) > a \mid N)$, while $\mathrm{P}(S_n([n\omega]+2) > a \mid N)$ decreases monotonically to the exact result. The probability $\mathrm{P}(S_n([n\omega]+1) > a \mid N)$ may be less or greater than the exact value $\mathrm{P}(S(\omega) > a \mid N)$ in a very complex way that may depend on ω, a, n, and N. Furthermore, in view of our proof, for fixed integers $l \ge 0$ and $k \ge 2$, we have

$$\lim_{n\to\infty} \mathrm{P}(S_n([nw]-l) > a \mid N) = \mathrm{P}(S(\omega) > a \mid N) = \lim_{n\to\infty} \mathrm{P}(S_n([n\omega]+k) > a \mid N).$$

The numerical results of Table 1 and Figure 1 in Section 4 will show the rates of $\mathrm{P}(S_n([n\omega]-l) > a \mid N)$ and $\mathrm{P}(S_n([n\omega]+k) > a \mid N)$ converging to $\mathrm{P}(S(\omega) > a \mid N)$.

Remark 3.2. Given s and r, let $D_n(s)$ be the length of the smallest window that contains at least s successes:

$$D_n(s) = \inf\{r\colon S_n(r) \ge s\}.$$

The result of Theorem 3.1 also holds for the scan statistic $D_n(s)$ conditional on a given N in the following sense:

$$\mathrm{P}(S_n(\omega) > s \mid N) = \lim_{n\to\infty} \mathrm{P}(D_n(s) < [n\omega] \mid N).$$

Furthermore, from another viewpoint of the Poisson process, let $\{X_i\}_{i=1}^n$ be a sequence of Bernoulli trials with probability $p_n = \lambda/n$. For given r, it follows that

$$\mathrm{P}(S_n(r) < s) = \sum_{N=0}^{n} \binom{n}{N} p_n^N (1-p_n)^{n-N}\, \mathrm{P}\Big(S_n(r) < s \;\Big|\; \sum_{i=1}^{n} X_i = N\Big).$$

Taking $r = [n\omega]+k$, and since $\sum_{i=1}^n X_i$ converges in the limit of large n to a Poisson random variable with parameter λ, the above equation yields the following result: for sufficiently large n,

$$\lim_{n\to\infty} \mathrm{P}(S_n([n\omega]+k) < s) = \sum_{N=0}^{\infty} \frac{\lambda^N}{N!} \mathrm{e}^{-\lambda}\, \mathrm{P}(S(\omega) < s \mid N). \tag{3.4}$$

This is equivalent to saying that

$$\lim_{n\to\infty} \mathrm{P}(S_n([n\omega]+k) < s) = \mathrm{P}\Big(\sup_{0<t\le 1-\omega} S(\omega, t) < s\Big). \tag{3.5}$$

Taking $p_n = \lambda\Delta t$, it follows from (2.1) and (3.5) that we have

$$\mathrm{P}\Big(\sup_{0<t\le 1-\omega} S(\omega, t) < s\Big) = \lim_{n\to\infty} \boldsymbol{\xi}_0 \boldsymbol{N}_{r,s}^n(p_n)\mathbf{1}^\top, \tag{3.6}$$

where $r = [n\omega]+k$.

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

Remark 3.3. For computing the unconditional probability $P(S(\omega) < s)$, we can use either (3.4) or (3.6). Note that using (3.4) entails summing a large number of terms with each term requiring a nonhomogeneous Markov chain, and, hence, this approach is rather time consuming. On the other hand, (3.6) is a direct consequence of homogeneous Markov chain imbedding, and we expect (3.6) to be more efficient and accurate in computing $P(S(\omega) < s)$.

Remark 3.4. The method could be extended to the case of the two-dimensional scan statistic under a Poisson process. Using the same idea of Lemma 3.1, the two-dimensional scan can be 'outer-and-inner' approximated (i.e. from above and below) by two-dimensional discrete scan statistics, and the finite Markov chain imbedding technique can then be used to compute the distributions of the two-dimensional discrete scan statistics.

4. Numerical results and discussion

The result of Theorem 3.1 also holds for, given $1 \le a \le N$,

$$P(S(\omega) > a \mid N) = \lim_{n\to\infty} P(S_n([n\omega] + k) > a \mid N).$$

Figure 1 and Table 1 show the rate of convergence of the conditional discrete scan statistic to the conditional continuous scan statistic for various parameters λ, ω, and N. They also show the connection between the conditional probability and the unconditional probability for the continuous scan statistic under a Poisson process. Two important phenomena can be observed from Figure 1 and Table 1. (i) The probability $P(S_n([n\omega] - l) > a \mid N)$ is monotonically increasing in integers of $[n\omega]$ as $n \to \infty$, and displays a sawtooth shape for n between the integers $[n\omega]$ and $[n\omega] + 1$, a special characteristic of discrete scan statistics; the $P(S_n([n\omega] + k) > a \mid N)$ behaves in the opposite manner. (ii) For fixed N, the error bounds decrease on the order of $1/n$, while, for fixed n, the bounds increase on the order of N.

Table 2 provides a numerical comparison study for various bounds and approximations to the probabilities for the continuous scan statistic $P(S(\omega) < s)$ under a Poisson process given

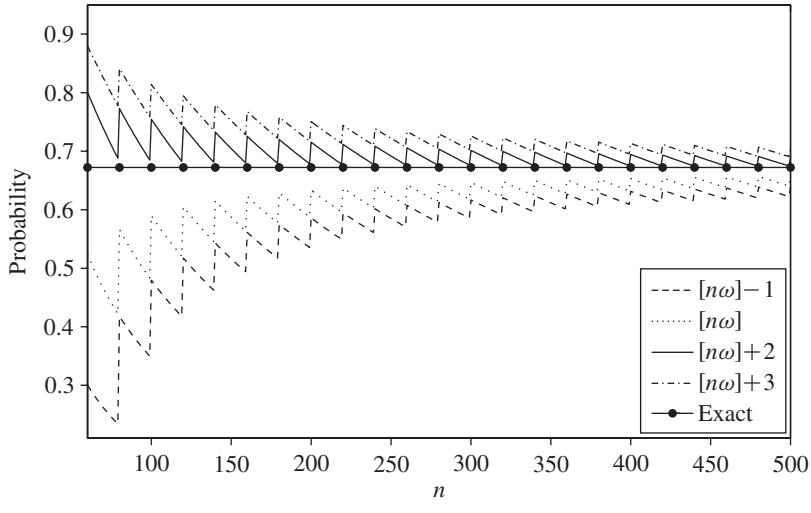

FIGURE 1: The probabilities $P(S_n([n\omega] + k) \ge 2 \mid N = 5)$ with $\omega = 0.05$ and $k = -1, 0, 2, 3$ for n from 60 to 500.

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

TABLE 1: The probabilities $P(S_n([n\omega]) < s \mid N)$ for approximating $P(S(\omega) < s \mid N)$, given various λ, ω, and s.

λ	ω	s	N	n				Exact
				100	300	500	1000	
3	0.1	2	0	1	1	1	1	1
			1	1	1	1	1	1
			2	0.827	0.816	0.813	0.812	0.810
			3	0.548	0.524	0.519	0.516	0.512
			4	0.278	0.252	0.247	0.244	0.240
			5	0.101	0.085	0.082	0.080	0.078
			6	0.024	0.018	0.017	0.016	0.016
			7	0.003	0.002	0.002	0.002	0.002
			$P(S(\omega) < s)$	0.565	0.551	0.548	0.547	0.544
λ	ω	s	N	n				Exact
				100	300	500	1000	
5	0.05	3	0	1	1	1	1	1
			1	1	1	1	1	1
			2	1	1	1	1	1
			3	0.996	0.994	0.994	0.993	0.993
			4	0.986	0.978	0.976	0.974	0.973
			5	0.967	0.947	0.943	0.940	0.937
			6	0.936	0.901	0.894	0.888	0.882
			7	0.893	0.838	0.827	0.818	0.810
			8	0.838	0.760	0.745	0.733	0.722
			9	0.770	0.669	0.650	0.636	0.622
			10	0.692	0.571	0.548	0.532	0.516
			11	0.607	0.470	0.445	0.428	0.411
			12	0.517	0.371	0.347	0.330	0.313
			13	0.428	0.281	0.258	0.242	0.227
			14	0.342	0.203	0.183	0.169	0.155
			15	0.263	0.139	0.122	0.111	0.101
			$P(S(\omega) < s)$	0.940	0.911	0.905	0.901	0.897

a range of parameter values. The upper bound (UB) and lower bound (LB) are obtained using Equations (1.2) and (1.7) of Janson (1984). The values in columns 'Alm' and 'Haiman' are approximations based on the manuscripts of Alm (1999) and Haiman (2000), respectively. The values of the column 'FWL' are calculated using (3.6) with probability of success $p_n = \lambda/n$. The exact values are taken from the table in Neff and Naus (1980); since their table provides only values for $s \geq 3$, the exact values for $s = 2$ (marked with an asterisk) are calculated using equation (3.6) with large n ($n \geq 50\,000$). The numerical results in Table 2 show that all three approximations perform well for large s ($s \geq 5$), while $P(S(\omega) < s)$ is near 1. For small s, it is a completely different story, with bounds and approximations all performing poorly. It is further evident that, even at moderate n ($125 \leq n \leq 1000$), our method performs very well over the entire range of s, including for small s. For larger n, we expect that our method will do extremely well. The reason for the accuracy of our method comes from the fact that, for

https://doi.org/10.1239/jap/1331216842 Published online by Cambridge University Press

TABLE 2: The unconditional probabilities $P(S(\omega) < s)$, given various ω, λ, and s. (NA denotes that a value is not available.)

w	λ	n	s	LB	Alm	Haiman	FWL	Exact	UB
0.2	6	125	2	NA	0.3310	NA	0.0614	0.0686*	0.2525
			3	0.1866	0.4710	0.4220	0.3245	0.3417	0.3788
			4	0.5933	0.7219	0.7014	0.6809	0.6902	0.6732
			5	0.8609	0.9024	0.8987	0.8985	0.8982	0.8855
			6	0.9642	0.9748	0.9744	0.9764	0.9744	0.9703
0.1	9	250	2	NA	0.1513	NA	0.0239	0.0273*	0.0829
			3	0.2199	0.3684	0.3354	0.2770	0.2932	0.3086
			4	0.6646	0.7235	0.7137	0.7017	0.7095	0.6995
			5	0.9123	0.9267	0.9257	0.9266	0.925	0.9206
			6	0.9833	0.9860	0.9859	0.9873	0.98	0.9848
0.06	8	400	2	0.0749	0.1677	NA	0.0958	0.104*	0.1270
			3	0.5827	0.6292	0.7089	0.6018	0.6148	0.6103
			4	0.9176	0.9259	0.9517	0.9239	0.9252	0.9225
			5	0.9896	0.9906	0.9948	0.9910	0.9900	0.9901
			6	0.9990	0.9991	0.9996	0.9992	0.9980	0.9990
0.04	10	600	2	0.0622	0.1197	NA	0.0725	0.079*	0.0934
			3	0.6050	0.6350	0.6288	0.6133	0.626	0.6226
			4	0.9349	0.9394	0.9391	0.9381	0.94	0.9375
			5	0.9932	0.9936	0.9936	0.9940	0.99	0.9934
			6	0.9994	0.9995	0.9995	0.9996	0.99	0.9995
0.02	10	1000	2	0.2016	0.2318	NA	0.2002	0.2139*	0.2187
			3	0.8548	0.8599	0.8594	0.8526	0.86	0.8579
			4	0.9895	0.9898	0.9898	0.9897	0.99	0.9897
			5	0.9995	0.9995	0.9995	0.9995	1	0.9995

every n, (2.1) provides exact probabilities $P(S_n(r) < s)$ of discrete scan statistics and carries a rate of $O(1/n)$ in converging to $P(S(\omega) < s)$.

Acknowledgements

The authors would like to thank the anonymous referee for valuable comments and suggestions. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

ALM, S. E. (1999). Approximations of the distributions of scan statistics of Poisson processes. In *Scan Statistics and Applications*, Birkhäuser, Boston, MA, pp. 113–139.

CHEN, J. AND GLAZ, J. (1997). Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials. In *Advances in the Theory and Practice of Statistics*, John Wiley, New York, pp. 285–298.

FU, J. C. (2001). Distribution of the scan statistics for a sequence of bistate trials. *J. Appl. Prob.* **38,** 908–916.

GLAZ, J. (1989). Approximations and bounds for the distribution of the scan statistics. *J. Amer. Statist. Assoc.* **84,** 560–566.

GLAZ, J. (1992). Approximations for tail probabilities and moments of the scan statistic. *Comput. Statist. Data Anal.* **14,** 213–227.

GLAZ, J. AND BALKRISHNAN, N. (eds) (1999). *Scan Statistics and Applications*. Birkhäuser, Boston, MA.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*. Springer, New York.
Haiman, G. (2000). Estimating the distributions of scan statistics with high precision. *Extremes* **3,** 349–361.
Janson, S. (1984). Bounds on the distributions of extremal values of a scanning process. *Stoch. Process. Appl.* **18,** 313–328.
Karlin, S. and McGregor, J. (1959). Coincidence probabilities. *Pacific J. Math.* **9,** 1141–1164.
Koutras, M. V. and Alexandrou, V. A. (1995). Runs, scans and urn model distributions: a unified Markov chain approach. *Ann. Inst. Statist. Math.* **47,** 743–766.
Naus, J. (1974). Probabilities for a generalized birthday problem. *J. Amer. Statist. Assoc.* **69,** 810–815.
Naus, J. I. (1982). Approximations for distributions of scan statistics. *J. Amer. Statist. Assoc.* **77,** 177–183.
Neff, N. D. and Naus, J. I. (1980). *Selected Tables in Mathematical Statistics*, Vol. VI, American Mathematical Society, Providence, RI.