8

1-Dependent Stationary Sequences and Applications to Scan Statistics

George Haiman[1] **and Cristian Preda**[2]

[1] *UFR de Mathématiques, Université de Lille 1, Lille, France*
[2] *Faculté de Médecine, Université de Lille 2, Lille, France*

Abstract: A new method of estimating the distribution function of scan statistics was presented and studied by the authors in a series of papers. This method is based on the application of some results concerning the distribution function of the partial maximum sequence generated by a 1-dependent stationary sequence. We present a review of our results and compare the method with other existing methods.

Keywords and phrases: Scan statistic, 1-dependence, Poisson process

8.1 Introduction

Let N be a Poisson process of intensity λ on the real line and let $u > 0$ and $T > u$ be fixed constants. Let $\nu_t = N(t+u) - N(t)$ be the number of points in the interval $[t, t+u[$, $t \in [0, T-u]$.

The one-dimensional continuous scan statistic is defined [see Glaz *et al.* (2001)] as

$$S = S(u, \lambda, T) = \max_{0 \le t \le T-u} \nu_t. \tag{8.1}$$

Let $T = \tau u$, $\tau \in \mathbf{N}$ and let

$$X_n = \max_{(n-1)u \le t < nu} \nu_t, \;\; n = 1, \ldots, \tau - 1. \tag{8.2}$$

It can be easily seen that $\{X_n\}$ forms a 1-dependent stationary sequence and

$$S = S_\tau = \max_{1 \le n \le \tau - 1} X_n. \tag{8.3}$$

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_8,

Then, in order to approximate the distribution function (d.f.) of S, we can apply either one of the following equivalent versions of Haiman (1999), Theorems 3 and 4.

Let $\{X_n\}$ be a general 1-dependent stationary sequence of random variables (r.v.'s) and let

$$q_n = q_n(x) = \mathbf{P}\left\{\max(X_1, \ldots, X_n) \leq x\right\}, \ \ n \geq 1.$$

Theorem 8.1.1 *For any x such that $1 - q_1(x) \leq 0.025$ and any integer $n > 3$ such that $88n(1-q_1)^3 \leq 1$, we have*

$$\left| q_n - \frac{4q_3 - 3q_4 + 6(q_1 - q_2)^2}{(1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2)^n} \right| \Big/ q_n \leq (1-q_1)^3[88n(1 + 124n(1-q_1)^3) + 561]. \tag{8.4}$$

Theorem 8.1.2 *For any x such that $1 - q_1(x) \leq 0.025$ and any integer $n > 3$ such that $3.3n(1-q_1)^2 \leq 1$, we have*

$$\left| q_n - \frac{2q_1 - q_2}{(1 + q_1 - q_2 + 2(q_1 - q_2)^2)^n} \right| \Big/ q_n \leq (1-q_1)^2[3.3n(1 + 4.7n(1-q_1)^2) + 9 + 561(1-q_1)]. \tag{8.5}$$

From Theorem 8.1.1 and Theorem 8.1.2 we deduce, respectively, the approximations

$$\mathbf{P}(S_\tau \leq x) \approx \frac{4q_3 - 3q_4 + 6(q_1 - q_2)^2}{(1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2)^{\tau-1}} \tag{8.6}$$

with a relative error bound of about $88\tau(1-q_1)^3$ and

$$\mathbf{P}(S_\tau \leq x) \approx \frac{2q_1 - q_2}{(1 + q_1 - q_2 + 2(q_1 - q_2)^2)^{\tau-1}} \tag{8.7}$$

with a relative error bound of about $3.3\tau(1-q_1)^2$.

The approximations (8.6) and (8.7) for the d.f. of continuous scan statistics have been introduced and studied in Haiman (2000). A characteristic of these approximations is that they depend on a prior knowledge of $q_i = q_i(x) = \mathbf{P}(S_{i+1} \leq x)$, $i = 1, \ldots, 4$, respectively, $i = 1, 2$.

Let $Z_1, \ldots, Z_N$ be a sequence of integer-valued r.v.'s that are independent and identically distributed (i.i.d.), typically Bernoulli $\mathcal{B}(1, p)$. Let $1 \leq m \leq N$ be a fixed positive integer, let

$$\mu_t = \sum_{i=t}^{t+m-1} Z_i, i \leq t \leq N - m + 1, \tag{8.8}$$

and define the one-dimensional discrete scan statistic [see Glaz *et al.* (2001)] by

$$S = S(m, p, N) = \max_{1 \le t \le N-m+1} \mu_t. \tag{8.9}$$

Let $N = \tau m$, $\tau \in \mathbf{N}$, $\tau \ge 1$, and let

$$Y_n = \max_{(n-1)m+1 \le t \le nm+1} \mu_t, n \in \mathbf{N}, n \ge 1. \tag{8.10}$$

Then $\{Y_n\}$ similarly forms a stationary 1-dependent sequence, $S = S_\tau = \max_{1 \le n \le \tau-1} Y_n$ and the d.f. of S can again be approximated by either one of the corresponding versions of approximations (8.6) and (8.7).

This type of approximation for the d.f. of discrete scan statistics was introduced and studied in Haiman (2007).

In Section 8.2 we present and discuss the main aspects related to the application of approximations (8.6) and (8.7) to continuous and discrete one-dimensional scan statistics.

Let N be a two-dimensional Poisson process of intensity λ. For fixed positive u and v, let $\nu_{t,s}(u, v)$ be the number of points in the rectangle $[t, t+u) \times [s, s+u)$, i.e.,

$$\nu_{t,s} = \nu_{t,s}(u, v) = N([t, t + u) \times [s, s + v)). \tag{8.11}$$

For $0 < u < L$ and $0 < v < K$, the two-dimensional continuous scan statistic

$$S = S((u, v), \lambda, L, K) = \max_{\substack{0 \le t \le L-u \\ 0 \le s \le K-v}} \nu_{t,s} \tag{8.12}$$

represents the largest number of points in any rectangle of dimension $u \times v$ within the rectangular region $[0, L] \times [0, K]$. Observing that for any $0 < u < L$ and $0 < v < K$ we have

$$\mathbf{P}\left(S((u, v), \lambda, L, K) \le k\right) = \mathbf{P}\left(S((1, 1), \lambda uv, \frac{L}{u}, \frac{K}{v}) \le k\right),$$

we now suppose that $u = v = 1$.

Let K and L be positive integers and let

$$X_k = \max_{\substack{0 \le t \le L-1 \\ k-1 \le s \le k}} \nu_{t,s}, k = 1, \ldots, K-1. \tag{8.13}$$

We first observe that $\{X_k\}$ is a stationary 1-dependent sequence and

$$S = S_{L,K} = \max_{1 \le k \le K-1} X_k.$$

Then, a first application of Theorem 8.1.2 (under the required conditions) leads to the approximation

$$\mathbf{P}(S \leq n) \approx (2q_1 - q_2)(1 + q_1 + q_2 + 2(q_1 - q_2)^2)^{-(K-1)}, n \in \mathbf{N}. \quad (8.14)$$

with an error bound of about $3.3(K-1)(1-q_1)^2$. Here $q_1 = \mathbf{P}(X_1 \leq n)$ and $q_2 = \mathbf{P}(X_1 \leq n, X_2 \leq n)$. In order to obtain the final approximation of $\mathbf{P}(S \leq n)$, q_1 and q_2 are replaced in (8.14) by their approximations obtained using again Theorem 8.1.2. Indeed,

$$Y_l = \max_{\substack{l-1 \leq t \leq l \\ 0 \leq s \leq 1}} \nu_{t,s}, l = 1, \ldots, L-1 \quad (8.15)$$

is a 1-dependent stationary sequence and

$$q_1 = \mathbf{P}\left(\max_{0 \leq l \leq L-1} Y_l \leq n\right). \quad (8.16)$$

Analogously,

$$Z_l = \max_{\substack{l-1 \leq t \leq l \\ 0 \leq s \leq 2}} \nu_{t,s}, l = 1, \ldots, L-1$$

is also a 1-dependent stationary sequence and

$$q_2 = \mathbf{P}\left(\max_{0 \leq l \leq L-1} Z_l \leq n\right). \quad (8.17)$$

Then, Theorem 8.1.2 provides the approximations

$$q_1 \approx (2q_{2,2} - q_{2,3})(1 + q_{2,2} + q_{2,3} + 2(q_{2,2} - q_{2,3})^2)^{-(L-1)}, \quad (8.18)$$

and

$$q_2 \approx (2q_{3,2} - q_{3,3})(1 + q_{3,2} + q_{3,3} + 2(q_{3,2} - q_{3,3})^2)^{-(L-1)}, \quad (8.19)$$

where $q_{2,2} = \mathbf{P}(S_{2,2} \leq n)$, $q_{2,3} = q_{3,2} = \mathbf{P}(S_{2,3} \leq n)$ and $q_{3,3} = \mathbf{P}(S_{3,3} \leq n)$.

Thus, in the two-dimensional case, the final approximation of $\mathbf{P}(S_{L,K} \leq n)$ depends on a prior knowledge of $q_{2,2}$, $q_{2,3}$ and $q_{3,3}$. If $q_{2,2}$, $q_{2,3}$ and $q_{3,3}$ are known and $L \leq K$, it can be shown that the resulting error on the approximation of $\mathbf{P}(S \leq n)$ is bounded by about

$$e = 3.3(L-1)(K-1)\left((1-q_{2,2})^2 + (1-q_{3,2})^2 + (L-1)(q_{2,2} - q_{3,2})^2\right). \quad (8.20)$$

The main difficulty in the two-dimensional case arises from the fact that currently there are no, exact formulas for $q_{2,2}$, $q_{3,2}$ and $q_{3,3}$. This type of approximation for the d.f. of two-dimensional scan statistics generated by a Poisson process was introduced and studied in Haiman and Preda (2002).

As in the one-dimensional case, the method can be adapted to the two-dimensional discrete scan statistics defined as follows.

Let N_1 and N_2 be positive integers and $\{X_{i,j}; 0 \le i \le N_1 - 1, 0 \le j \le N_2 - 1\}$ be a family of i.i.d. nonnegative integer valued r.v.'s from some specified distribution (typically $\mathcal{B}(n,p)$ or Poisson(λ)). For $\le i \le N_1 - 1$ and $0 \le j \le N_2 - 1$, $X_{i,j}$ represents the number of some events observed in the elementary square subregion $[i, i+1] \times [j, j+1]$. Let m_1, m_2 be positive integers, $1 \le m_1 \le N_1$, $1 \le m_2 \le N_2$. For $0 \le t \le N_1 - m_1$, $0 \le s \le N_2 - m_2$, let

$$\nu_{t,s} = \nu_{t,s}(m_1, m_2) = \sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} X_{i,j}. \tag{8.21}$$

The two-dimensional discrete scan statistic is defined as the largest number of events in any $m_1 \times m_2$ rectangular scanning window within the rectangular region $[0, N_1] \times [0, N_2]$, i.e.,

$$S = S_{N_1,N_2}(m_1, m_2) = \max_{\substack{0 \le t \le N_1 - m_1 \\ 0 \le s \le N_2 - m_2}} \nu_{t,s}. \tag{8.22}$$

Let $N_1 = Lm_1$, $N_2 = Km_2$, with L and K integers, $L > 3$, $K > 3$. In this case, the same arguments and formulas as those leading to the approximation for the continuous scan statistics can be used with the following changes:

- X_k in formula (8.13) is now

$$X_k = \max_{\substack{0 \le t \le (L-1)m_1 \\ (k-1)m_2 \le s \le km_2}} \nu_{t,s}, \quad k = 1, \ldots, K-1,$$

- Y_l in formula (8.15) becomes

$$Y_l = \max_{\substack{(l-1)m_1 \le t \le lm_1 \\ 0 \le s \le m_2}} \nu_{t,s}, \quad l = 1, \ldots, L-1,$$

- and Z_l is

$$Z_l = \max_{\substack{(l-1)m_1 \le t \le lm_1 \\ 0 \le s \le 2m_2}} \nu_{t,s}, \quad l = 1, \ldots, L-1.$$

This type of approximation for the d.f. of discrete two-dimensional scan statistics was studied in Haiman and Preda (2006).

The main aspects related to the application of this method to two-dimensional scan statistics are presented and discussed in Section 8.3.

8.2 Application of the Approximations (8.6) and (8.7) to One-Dimensional Scan Statistics

The approximations (8.6) and (8.7) require a prior knowledge of $q_i = q_i(x) = \mathbf{P}(S_{i+1} \leq x)$, $i = 1, 2, 3, 4$, respectively, $i = 1, 2$. In the one-dimensional case, for both continuous and discrete scan statistics, there are exact formulas for q_n (see the references in Subsections 8.2.1 and 8.2.2). However, these formulas become rapidly intractable as n becomes large. There are also bounds and approximations as those mentioned below. The approximation formulas are based on heuristics, and their accuracy is evaluated using simulation results only in some particular configurations.

Our formulas include error bounds from which one can characterize completely their domain of applicability. A typical application of our method is the following. Suppose we want to establish, for a large $\tau \in \mathbf{N}$, the value of $x_{0.95}$ such that $q_{\tau-1}(x_{0.95}) \approx 0.95$. In the case of the continuous scan statistic, $x_{0.95}$ represents the critical value for testing the intensity λ of the underlying Poisson distribution at the 5% level of significance, i.e., reject the null hypothesis $(\lambda = \lambda_0)$ if $S_{\tau-1} > x_{0.95}(\lambda_0)$.

Under the condition "large τ", $1 - q_1$ is then necessarily small with respect to 0.05 and condition $1 - q_1 \leq 0.025$ is satisfied. Indeed, we then have

$$\begin{aligned} q_{\tau-1}(x_{0.95}) &\approx (2q_1 - q_2)\left(1 + (q_1 - q_2) + 2(q_1 - q_2)^2\right)^{-(\tau-1)} \\ &\approx 1 - (\tau - 1)(q_1 - q_2) \approx 0.95 \text{ as } q_1 \to 1. \end{aligned} \tag{8.23}$$

By Haiman *et al.* (1998), Proposition 2.1, page 490, if $1-q_1$ is sufficiently small, we have $1 - q_1 \leq 2(q_1 - q_2)$.

Thus,

$$1 - q_{\tau-1}(x_{0.95}) \approx 0.05 \approx (\tau - 1)(q_1 - q_2) \geq 2(\tau - 1)(1 - q_1), \tag{8.24}$$

whereas the error bound is about $3.3\tau(1 - q_1)^2$, thus very small with respect to the approximated value of 0.05. When q_3 and q_4 are available, the approximation (8.6) is more accurate (error of order $(1-q_1)^3$ instead of $(1-q_1)^2$), but generally, the approximation (8.7) appears to be sufficiently precise.

We now examine separately the application of the method to continuous and discrete scan statistics.

8.2.1 Application to one-dimensional continuous scan statistics

Let $S = S(u, \lambda, T)$ be the scan statistic generated by a Poisson process as defined in (8.1). Huntington and Naus (1975) give an exact formula for $\mathbf{P}(S \leq n)$ for $n \geq 0$ and $T \geq u$ that sums many products of determinants and for large

T requires excessive computation time. This formula is used in Neff and Nauss (1980) to establish tables for the d.f. of $S(1, \lambda, \tau)$ (notice that $S(u, \lambda, T) = S(1, \lambda u, \frac{T}{u})$) for several discrete values of λ and $\tau \leq 100$.

In Haiman (2000) we have applied the approximation (8.7) with q_1 and q_2 from Neff and Naus tables and $\tau = 1000$. Notice that when we mention a numerical application of the approximations (8.6) or (8.7), it means that we also provide the corresponding error bound.

Naus (1982), making a reasoning based on the hypothesis of a Markov-like behavior of the sequence $\{X_n\}$ defined in (8.2), proposes the approximation

$$q_\tau = q_{\tau(x)} = \mathbf{P}(S_\tau \leq x) \approx q_1 \left(\frac{q_2}{q_1}\right)^{\tau-2}, \tau > 2. \tag{8.25}$$

He shows, using the exact formula, that for λ and τ ranging in a certain domain, and also compared to other existing approximations, approximation (8.25) is remarkably accurate. This fact is not surprising: if we denote, respectively, by q_τ^H and q_τ^N the approximations in (8.7) and (8.25), it can be shown that for τ sufficiently large we have $|q_\tau^H - q_\tau^N| \leq 5(1 - q_1)^2$. Table 8.1 presents some numerical examples of these approximations and illustrates this fact.

Another scan statistic of interest generated by a Poisson process is defined as

$$S^* = S^*(u, \lambda, T) = \min_{0 \leq t \leq T-u} \nu_t, \tag{8.26}$$

where, as in (8.1), $\nu_t = N(t + u) - N_t$.

Let $T = \tau u$, $\tau \in \mathbf{N}$, $\tau > 0$ and let

$$X_k^* = - \min_{(k-1)u \leq t \leq ku} \nu_t. \tag{8.27}$$

Then $\{X_k^*\}$ forms a 1-dependent stationary sequence and

$$\overline{q}_\tau^*(n) = \mathbf{P}(S_\tau^* > n) = \mathbf{P}\left(\max_{1 \leq k \leq \tau-1} X_k^* < -n\right), \ n \geq 0. \tag{8.28}$$

Theorems 8.1.1 and 8.1.2 can also be applied here and corresponding versions of the approximations (8.6) and (8.7) can be used to estimate $\overline{q}_\tau^*(n)$. The values

Table 8.1. Approximations for $\mathbf{P}(S \leq x)$ by approximations (8.25) and (8.7). $T = 1001$.

x	λ	Naus (1982)	Haiman (2000)	Error
4	0.1	0.985399334	0.9854	2×10^{-6}
6	0.5	0.930142831	0.9302	2.5×10^{-5}
9	1.3	0.940503808	0.9405	1.7×10^{-5}

of $\overline{q}_i^*(n)$, $i = 1, 2, 3, 4$, or $i = 1, 2$ used in these approximations can be obtained from the exact formulas established in Huntington (1978) (these exact formulas also become intractable as τ becomes large).

Janson (1984) gives upper and lower bounds for q_τ and $\overline{q}_\tau^*$. In Haiman (2000) we have shown that the approximation (8.7) and Janson's bounds have similar precision.

The waiting time until the first occurrence of n points within an interval of length u, W_n, is an r.v. whose distribution is important in several applications (see Naus (1982)). For $n \geq 1$ and $t \geq 2$ we have

$$\mathbf{P}(W_n > t) = \mathbf{P}\left(\max_{0 \leq s \leq t-u} \nu_s < n\right). \tag{8.29}$$

Let W_n^* be the corresponding discretized waiting time defined as

$$W_n^* = \left[\frac{\min\{s \geq 0 : \nu_s = n\}}{u}\right] u, \tag{8.30}$$

where $[\,\cdot\,]$ stands for integer part.

For $n \geq 1$ and $\tau = 2, 3, \ldots$, we have

$$\mathbf{P}(W_n^* > \tau) = \mathbf{P}(S_\tau \leq n-1) = q_{\tau-1}(n-1). \tag{8.31}$$

We then can apply the approximations (8.6) or (8.7) to estimate the expected waiting time, $\mathbf{E}(W_n^*)$. Details about this application and a numerical example are given in Haiman (2000).

Let $M(T)$ be the number of subintervals, each of length u, dropped so that their midpoints are the occurrence points of a homogenous Poisson process N in the interval $[0, T]$. We say that a point x is covered by a subinterval with midpoint y if $y - \frac{u}{2} \leq x \leq y + \frac{u}{2}$. The calculation of the probability of the event $E_n =$ *"all points of the interval [0,T] are covered by at least n subintervals"* is of interest in several applications [see Glaz and Naus (1978)]. Let $T = \tau u$, $\tau = 2, 3, \ldots$. In Haiman (2000) we use the fact that the calculation of $\mathbf{P}(E_n)$ is related to the calculation of $\overline{q}_\tau^*$. Thus, via the approximation (8.7) we obtain an approximation formula for $\mathbf{P}(E_n)$.

Let F_n be the event *"there does not exist a subarc of length $u = 1$ of a circle with circumference τ, $\tau = 2, 3, \ldots$, that contains n points."* Using similar arguments, in Haiman (2000) we obtain an approximation formula for $\mathbf{P}(F_n)$.

8.2.2 Application to one-dimensional discrete scan statistics

Let $Z_1, \ldots, Z_N$ be a sequence of integer-valued r.v.'s that are i.i.d. and consider the discrete scan statistic S defined in (8.9). Exact formulas for $\mathbf{P}(S \geq k)$ exist, and some of them are tractable only in a limited number of situations. The Bernoulli case ($Z_i \sim \mathcal{B}(1, p)$) plays an important role in the applications. In

this case, exact formulas have been obtained by Naus (1982) for $N = 2m$ and $N = 3m$, i.e., for q_1 and q_2. As for continuous scan statistics, Naus uses q_1 and q_2 to estimate $q_\tau = q_\tau(k) = \mathbf{P}(S_\tau \leq k)$ ($N = \tau m$, $\tau \geq 3$) by formula (8.25).

Fu (2001) employed a finite Markov chain embedding method to derive exact formulas for $\mathbf{P}(S_\tau \leq k)$. However, this method involves quite complicated computations, and it may become difficult to use for large or very large values of m and $\tau = \frac{N}{m}$.

Thereby, various approximation methods and bounds for $\mathbf{P}(S \leq k)$ have been proposed by several authors. However, the quality of these approximations and bounds can be evaluated for a limited number of particular configurations. An overview of these results as well as a complete bibliography on the subject are given in Glaz *et al.* (2001). In Haiman (2007) we have illustrated by several numerical examples the application of our approximation (8.7) in parallel with formula (8.25) of Naus. In these examples, $m = 30$, $p = 0.1$ and N ranges from 256×30 to 1024×30. As for continuous scan statistics and for a similar reason (see Section 8.2.1) the approximations (8.7) and (8.25) give very close results.

Let $V(N)$ denote the *length of the longest success run* in N Bernoulli $\mathcal{B}(1, p)$ trials (1 = success, 0 = failure). We then have

$$\mathbf{P}(V(N) \geq m) = \mathbf{P}(S \geq m) = \mathbf{P}(S = m). \tag{8.32}$$

Thus, if $N = \tau m$, $\tau \geq 2$,

$$\mathbf{P}(V(N) \geq m) = \mathbf{P}(S_\tau \geq m) = 1 - q_{\tau-1}(m). \tag{8.33}$$

An exact formula for $\mathbf{P}(V(N) \geq m)$ of Bateman (1948) allows in this case an easy calculation of $q_i(m)$, $i = 1, 2, 3, 4$, from which $\mathbf{P}(V(N) \geq m)$ can be approximated by either one of the approximations (8.6) or (8.7). In Haiman (2007) we have used numerical examples to illustrate and compare these two approximations. It appears that the approximations (8.6) and (8.7) provide very close results. Table 8.2 presents some numerical examples of these approximation and illustrates this fact.

Fu *et al.* (2003) have used the finite Markov chain embedding to obtain the exact distribution of $V(N)$. They also obtained a large deviation approximation of the above distribution [in relationship to this problem, see also Lou (1996), Vaggelatou (2003) and the references quoted in these papers].

In Haiman (2007), we also compare the approximation (8.6) and exact values of $V(N)$ calculated in Fu *et al.* (2003).

Let k and m, $1 \leq k \leq m$, be positive integers and define the waiting time, until "$k - in - m$ quota" by

$$T = T_{k,m} = \inf\{t \geq 1 : \mu_t \geq k\}, \tag{8.34}$$

where μ_t is defined in (8.8).

Table 8.2. Approximations for $\mathbf{P}(S \leq x)$ by Haiman (2007) and Naus (1982), $X_i \sim \mathcal{B}(1,p)$, $p = 0.1$, $m = 30$.

x	9	10	11
$\mathbf{P}(S(30, 256 \times 30) \leq x)$:			
App. (8.7)	0.5161	0.85979	0.970613
Error	0.008	0.0023	10^{-6}
App. (8.25)	0.5172	0.86028	0.970726
$\mathbf{P}(S(30, 512 \times 30) \leq x)$:			
App. (8.7)	0.2658	0.73888	0.941997
Error	0.017	0.00046	0.000017
App. (8.25)	0.2663	0.739295	0.9421067

Huntington (1974) derives an exact and quite complicated formula for $\mathbf{E}(T)$, in terms of ratios of determinants of some matrices. Naus (1982), using the fact that

$$\mathbf{E}(T_{k,m}) = \sum_{N=0}^{\infty} (1 - \mathbf{P}(S_N < k)),$$

uses the approximation (8.25) to obtain the approximation

$$\mathbf{E}(T_{k,m}) \approx 2m + \frac{q_2}{(1 - \frac{q_2}{q_1})^{\frac{1}{m}}}. \tag{8.35}$$

In Haiman (2007), we similarly use the approximation (8.7) to establish upper and lower bounds for $\mathbf{E}(T_{k,m})$ and give some numerical examples.

Let now r.v. Z_i, $i = 1, \ldots, N$ take values -1, 0 and 1. The corresponding discrete scan statistic S is associated to the "charge problem." Exact results for $\mathbf{P}(S \leq k)$ have been obtained in this case by Saperstein (1976) for $N \leq 2m$ and by Karwe (1993) for $N \in \{2m-1, 2m \text{ (thus } q_1), 3m-1 \text{ and } 3m \text{ (thus } q_2)\}$. In Haiman (2007) we give numerical examples and compare the approximations (8.7) and (8.25) using values of q_1 and q_2 provided in Karwe (1993).

8.3 Application of the Method to Two-Dimensional Scan Statistics

As mentioned in Section 8.1, the main difficulty in applying the method to both, continuous and discrete two-dimensional scan statistics arises from the fact that at present there are no exact formulas allowing us to calculate $q_{i,j}$, $i, j = 2, 3$.

There are some approximation formulas (see references below) based on heuristics; their accuracy is evaluated using simulation results only in some particular configurations. As in the one-dimensional case, the characteristic of our approximation formulas is that they include error bounds.

8.3.1 Application to continuous scan statistics

Let S be defined in (8.12) and for $u = v = 1$ and K, L integers, $K, L > 3$, put

$$S_{L,K} = S = S((1,1), \lambda, L, K). \tag{8.36}$$

Previously, Aldous (1989) and Alm (1997) have established approximation formulas for the d.f. of $S_{L,K}$.

Let

$$q^n_{L,K}(k) = \mathbf{P}\left(S_{L,K} \le k \middle| N([0,L] \times [0,K]) = n\right), 1 \le k \le n \tag{8.37}$$

denote the d.f. of the conditional scan statistic, i.e., the scan statistic given that a fixed number n of points fall in $[0,L] \times [0,K]$. Notice that $q^n_{L,K}$ is the d.f. of the r.v. $S^n_{L,K}$ = *maximum number of points obtained by scanning with the* $[0,1] \times [0,1]$ *window a rectangle* $[0,L] \times [0,K]$ *in which n independent points are drawn uniformly.*

We then have

$$\begin{aligned} q_{L,K}(k) &= \mathbf{P}\left(S_{L,K} \le k\right) \\ &= e^{-\lambda LK}\left(\sum_{j=0}^{k} \frac{(\lambda LK)^j}{j!} + \sum_{j=k+1}^{kLK} q^j_{L,K}(k)\frac{(\lambda LK)^j}{j!}\right). \end{aligned} \tag{8.38}$$

In Haiman and Preda (2002) we have developed a method of "perfect" simulation of independent replications of r.v.'s $S^n_{i,j}$, $i, j = 2, 3$. We construct (Theorem 2) a stopping time T with respect to the filtration generated by a sequence $\{Z_n\}_{n\ge1}$ of Bernoulli $\mathcal{B}(1, \frac{1}{2})$ i.i.d. r.v's together with functions $f_t(z_1, \ldots, z_t)$ such that the r.v. $\mathcal{S}^n_{i,j} = f_T(Z_1, \ldots, Z_T)$ has the same distribution as $S^n_{i,j}$. We use this method to obtain via formula (8.38) empirical estimations of $q^n_{i,j}(k)$, $i, j = 2, 3$ and then we calculate (see Section 8.1) the final approximation of $q_{L,K}(k)$.

The empirical estimation of $q^n_{i,j}$ generates additional errors. These errors are bounded at the 95% confidence level by $\varepsilon_{i,j}$, where $\varepsilon_{i,j} \approx 1.96\sqrt{\frac{q^n_{i,j}(1-q^n_{i,j})}{M}}$. M is the number of replications of r.v.'s $S^n_{i,j}$, $i, j = 2, 3$. The total error on $\mathbf{P}(S_{L,K} \le k)$ is then bounded by about

$$E = e + LK(\varepsilon_{2,2} + \varepsilon_{2,3} + \varepsilon_{3,3}), \tag{8.39}$$

with e, the error bound when $q_{i,j}$ are known, given in (8.20). Naus (1965) and Neff (1978) give exact formulas for $q^m_{L,K}(m-1)$ and $q^m_{L,K}(m-2)$. In Haiman

Table 8.3. Approximation for $\mathbf{P}(S \leq n)$. $L = 500$, $K = 500$, $\lambda = 0.01$.

n	App. (8.14)	Error	Alm (1997)	Aldous (1989)
2	0.69318103	0.008570775	0.7839302629	0.8484459199
3	0.998401542	6.37679E-05	0.9987785770	0.9990759644

and Preda (2002) we use these formulas for $L, K = 2, 3$ to evaluate our simulation results. We then give numerical examples for several values of L, K and λ ($L, K = 10, 50, 100, 1000$, $\lambda = 0.01, 0.05, 0.1, 1$) and compare our results with corresponding results obtained by other approximation formulas in Aldous (1989) and Alm (1997).

In order to obtain error bounds $\varepsilon_{i,j}$ such that their contribution to the total error E has the same order of magnitude as e, we use in our examples up to 10^7 replications of r.v.'s $S^n_{i,j}$, $i, j = 2, 3$.

Table 8.3 presents some numerical examples of application of our method and the corresponding results obtained using the methods of Aldous and Alm.

8.3.2 Application to discrete scan statistics

Let $S = S_{N_1,N_2}$ be defined in (8.22) where the underlying $X_{i,j}$ are binomial $\mathcal{B}(n, p)$ or Poisson $\mathcal{P}(\lambda)$. Since there are no exact formulas for $\mathbf{P}(S \leq k)$, various methods of approximation and bounds have been proposed by several authors. An overview of these methods as well as a complete bibliography on the subject are given in Glaz *et al.* (2001). In particular, the case where $X_{i,j}$ are binary variables, with application to reliability (two-dimensional $r - within\ m_1 \times m_2 - out - of\ N_1 \times N_2$) has received considerable research interest during the last years. In this framework, several approximations and bounds have been proposed and studied in the literature [see, e.g., Chen and Glaz (1996), Boutsikas and Koutras (2003) and references therein].

Let $N_1 = Lm_1$ and $N_2 = Km_2$ with L and K integers, $L, K > 3$. In Haiman and Preda (2006) we have applied our approximation method of $\mathbf{P}(S_{N_1,N_2} \leq k)$ using, similarly to the previous continuous case, empirical estimations of

$$q_{i,j}(k) = \mathbf{P}(S_{im_1,jm_2} \leq k)$$

obtained by simulating i.i.d. replications of r.v.'s S_{im_1,jm_2}, $i, j = 2, 3$. The error bound due to simulation, e_{sim}, is then also proportional to $\frac{LK}{\sqrt{M}}$, where M is the number of replications and the total error bound, as in (8.39), is

$$E = e + e_{sim}.$$

For $X_{i,j}$ binomial and Poisson we give numerical examples and compare our results with those obtained using the product-type approximation, the Poisson approximation and Bonferroni inequality techniques, as presented in Glaz *et al.* (2001).

Table 8.4. Approximation for $\mathbf{P}(S \leq x) : X_{i,j} \sim \text{Poisson}(0.25)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $M = 10^9$.

x	$\hat{\mathbf{P}}(S \leq x)$	P-T	Bonferroni	Poisson	H-P	Error
15	0.8596	0.8374	0.7700	0.8292	0.860427482	0.067409646
16	0.9402	0.9351	0.9130	0.9314	0.940749305	0.010867255
17	0.9783	0.9764	0.9691	0.9750	0.977260378	0.001546897
18	0.9930	0.9920	0.9896	0.9916	0.991966851	0.000217233

Table 8.5. Approximation for $\mathbf{P}(S \leq x) : X_{i,j} \sim \mathcal{B}(5, 0.05)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $M = 10^9$.

x	$\hat{\mathbf{P}}(S \leq x)$	P-T	Bonferroni	Poisson	H-P	Error
15	0.8932	0.8830	0.8387	0.8768	0.896135764	0.035108915
16	0.9617	0.9577	0.9441	0.9554	0.960112719	0.004770939
17	0.9868	0.9862	0.9819	0.9854	0.986256278	0.000584065
18	0.9948	0.9958	0.9946	0.9956	0.995633424	8.08015E-05

For binary $X_{i,j}$ we compare our approximations with bounds obtained in Boutsikas and Koutras (2003).

In all these examples we use up to $M = 10^9$ replications of r.v. S_{im_1,jm_2}, $i, j = 2, 3$. Tables 8.4 and 8.5 present some numerical examples of the application of our method and the corresponding results obtained using the product-type (P-T), the Poisson and the Bonferroni approximation methods. $\hat{\mathbf{P}}(S \leq x)$ denotes the empirical estimation of $\mathbf{P}(S \leq x)$ using 10,000 trials [see Glaz *et al.* (2001)].

For binomial $X_{i,j}$, and in particular Bernoulli, the current work of the authors consists in constructing computer algorithms allowing one to obtain, without using simulations, exact values or sufficiently accurate (with respect to the method) approximations of $q_{i,j}(m)$, $i, j = 2, 3$.

References

1. Aldous, D. (1989). *Probability Approximation via the Poisson Clumping Heuristic*, Springer-Verlag, New York.

2. Alm, S.E. (1997). On the distribution of scan statistics of two-dimensional Poisson processes, *Advances Applied Probability*, **29**, 1–18.

3. Bateman, G.I. (1948). On the power function of the longest run as a test for randomness in a sequence of alternatives, *Biometrika*, **35**, 97–112.

4. Boutsikas, M. and Koutras, M. (2003). Bounds for the distribution of two dimensional binary scan statistics, *Probability in the Engineering and Information Sciences*, **17**, 509–525.

5. Chen, J. and Glaz, J. (1996). Two-dimensional discrete scan statistics, *Statistics and Probability letters*, **31**, 59–68.

6. Fu, J.C. (2001). Distribution of the scan statistic for a sequence of bistate trials, *Journal of Applied Probability*, **38**, 4, 908–916.

7. Fu, J.C., Wang, L. and Lou, W. (2003). On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state Markov dependent trials, *Journal of Applied Probability*, **40**, 2, 346–360.

8. Glaz, J. and Naus, J.I. (1978). Multiple coverage on the line, *Annals of Probability* **7**, 900–906.

9. Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer Series in Statistics, Springer-Verlag, New York.

10. Haiman, G. (1999). First passage time for some stationary processes, *Stochastic Processes and Their Applications*, **80**, 231–248.

11. Haiman, G. (2000). Estimating the distribution of scan statistics with high precision, *Extremes*, **3:4**, 349–361.

12. Haiman, G. (2007). Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences, *Journal of Statistical Planning and Inference*, **137:3**, 821–828.

13. Haiman, G., Mayeur, N., Nevzorov, V. and Puri, M.L. (1998) Records and 2-block records of 1-dependent stationary sequences under local dependence, *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, **34:4**, 481–503.

14. Haiman, G. and Preda, C. (2002). A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process, *Methodology and Computing in Applied Probability*, **4**, 393–407.

15. Haiman, G. and Preda, C. (2006). Estimation for the distribution of two-dimensional discrete scan statistics, *Methodology and Computing in Applied Probability*, **8**, 373–382.

16. Huntington, R.J. (1974). Distributions and expectations for clusters in continuous and discrete cases, with applications, *Ph.D. Thesis*, Rutgers University.

17. Huntington, R.J. (1978). Distribution of the minimum number of points in a scanning interval on the line. *Stochastic Processes and Their Applications*, **7**, 73–78.

18. Huntington, R.J. and Naus, J.I. (1975). A simpler expression for kth nearest-neighbor coincidence probabilities, *Annals of Probability*, **3**, 894–896.

19. Janson, S. (1984). Bounds on the distribution of extremal values of a scanning process, *Stochastic Processes and Their Applications*, **18**, 313–328.

20. Karwe, V.V. (1993). The distribution of the supremum of integer moving average processes with applications to the maximum net charge in DNA sequences, *Ph.D. Thesis*, Rutgers University.

21. Lou, W. (1996). On runs and longest run tests: a method of finite Markov chain imbedding, *J. Amer. Statist. Assoc.* **91**, 1595–1601.

22. Naus, J.I. (1965). A power comparison of two tests of non-random clustering, *Technometrics*, **8**, 493–517.

23. Naus, J.I. (1982). Approximations for distributions of scan statistics, *Journal of the American Statistical Association*, **77**, 177–183.

24. Neff, N. (1978) Piecewise polynomials for the probability of clustering on the unit interval, *Unpublished PhD. dissertation*, Rutgers University.

25. Neff N.D. and Naus, J.I. (1980). The distribution of the size of the maximum cluster points on a line, In *IMS Series Selected Tables in Mathematical Statistics*, Vol. **IV**, American Mathematical Society, Providence, RI.

26. Saperstein, B. (1976). The analysis of attribute moving averages: MIL-STD-105D reduced inspection plan, *Sixth Conference Stochastic Processes and Applications*, Tel Aviv.

27. Vaggelatou, E. (2003). On the length of the longest run in a multi-state Markov chain, *Statistics & Probability Letters*, **62:3**, 211–221.