# Scan statistics viewed as maximum of 1-dependent random variables

George Haiman, Cristian Preda

# Scan statistics viewed as maximum of 1-dependent random variables

George Haiman
LSTA
Université Paris 6, Paris, France
e-mail : george.haiman@upmc.fr

Cristian Preda
Laboratoire Paul Painlevé
Université de Lille, France
e-mail : cpreda@polytech-lille.fr

**Abstract**

A method of approximating the distribution function of the partial maximum sequence generated by a 1-dependent stationary sequence can be applied to estimate the distribution function of one or multi dimensional scan statistics. The method, which provides error bounds for the approximations, was investigated and evaluated in several papers.

**Keywords**. scan statistics, d-dependence, Poisson process.

## 1 Introduction

Let $\{Y_i\}_{1\leq i\leq n}$ be a sequence of length $n$ of identically distributed random variables and let

$$S_n = \max_{1\leq t\leq n-m+1} \sum_{i=t}^{t+m-1} Y_i, n \geq m, \tag{1}$$

be the one dimensional discrete scan statistic with scanning window of length $m$.

We recall that a sequence of r.v.'s $\{X_i\}_{i\geq 1}$ is $d-dependent$, $d \geq 0$ , if for any $t \geq 1$, the $\sigma$-fields generated by $\{X_1, \ldots, X_t\}$ and $\{X_{t+d+1,\ldots,}\}$ are independent.

Let assume that $n = (K+1)(m-1)$, $K \in \mathbb{N}$, and let define for any integer $k, 1 \leq k \leq K$, the random variables $W_k$, the scan statistics over sequences of length $2(m-1)$ :

$$W_k = \max_{(k-1)(m-1)+1 \leq t \leq k(m-1)} \sum_{i=t}^{t+m-1} Y_i. \tag{2}$$

It can be seen that $\{W_k\}$ is a 1-dependent stationary sequence and we have :

$$P(S_n \leq s) = P(\max(W_1, \ldots, W_K) \leq s). \tag{3}$$

The approximation method for the distribution of the scan statistics $S_n$ presented in this work is based on the following result of Haiman (1999):

**Theorem 1** *Let $\{W_i\}_{i \geq 1}$ be a 1-dependent sequence of r.v.'s and let*

$$q_n = q_n(s) = P\{\max(W_1, ..., W_n) \leq s\}. \tag{4}$$

*Then, for any $s$ such that $p_1 = p_1(s) = 1 - q_1(s) \leq .025$ and any integer $n > 3$ such that $3.3np_1^2 \leq 1$ we have*

$$\left| q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n} \right| / q_n \leq p_1^2[3.3n(1+4.7np_1^2)+9+561p_1]. \tag{5}$$

Let

$$q_1 = P(W_1 \leq s) = P(\max_{1 \leq t \leq m-1} \sum_{i=t}^{t+m-1} Y_i \leq s) \tag{6}$$

and

$$q_2 = P(W_1 \leq s, W_2 \leq s) = P(\max_{1 \leq t \leq 2(m-1)} \sum_{i=t}^{t+m-1} Y_i \leq s). \tag{7}$$

Thus, for the values of $s$ such that $1 - q_1(s) \leq .025$, the theorem provides the approximation of the scan statistic distribution

$$P(S_n \leq s) \simeq \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^K} \tag{8}$$

with an error of less than about $3.3(K+1)(1-q_1)^2$.

**Remark 1**

*a) For any $n \geq 2(m-1)$ not a multiple of $(m-1)$, let $K = [\frac{n}{m-1} - 1]$ where $[x]$ =integer part of $x$. Then, we have*

$$P(\max(W_1, \ldots, W_{K+1}) \leq s) \leq P(S_n \leq s) \leq P(\max(W_1, \ldots, W_K) \leq s). \tag{9}$$

*b) The result presented in 1 has been improved by Amarioarei (2012) enlarging the range of values of $q_1 = 1 - p_1$ and providing tighter error bounds.*

As it will be shown in the following sections, the previous method, which requires a prior calculation of $q_1(s)$ and $q_2(s)$, can be adapted to one dimensional continuous scan statistics, to multi dimensional scan statistics, and also, in the discrete case, when the underlying random variables are $d$-dependent, $d \geq 1$.The approximation is particularly efficient to obtain critical values for $P(S_n \leq s)$ such as .95 or .99 with high precision and for large $K$. The application domain of the approximation corresponds to the situation where $1-q_1$ is small and $\frac{2q_1-q_2}{[1+q_1-q_2+2(q_1-q_2)^2]^K} \simeq (1-q_1+q_2)^K \simeq 1-K(q_1-q_2) \simeq$ .95 or .99. Since (see Haiman et al (1998), Proposition 2.1) for $1-q_1$ sufficiently small $q_1 - q_2 \geq \frac{1}{2}(1-q_1)$, the error bound $3.3(K+1)(1-q_1)^2$ is then negligible with respect to the term $K(q_1 - q_2) \simeq .05$ or .01.

# 2 Application to 1-dimensional scan statistics.

## 2.1 1-dimensional discrete scan statistic

For 1-dimensional discrete scan statistics, approximation (8) was investigated for some i.i.d. and 1−dependent underlying $Y_n$'s.

### 2.1.1 Scan statistics for i.i.d. $Y_i$'s

Let the scan statistic be generated by i.i.d. $Y_n$'s. If the $Y_n$ are Bernoulli $B(1, p)$ r.v.'s, Naus (1982) provides exact expressions for $q_1(s)$ and $q_2(s), 0 \leq$

$k \leq m$. In the same paper, based on heuristics supposing a Markov-like behaviour of the sequence $\{W_n\}_{n\geq 1}$, the authors propose the well known approximation

$$P(S_n \leq s) \simeq q_1(s) \left(\frac{q_2(s)}{q_1(s)}\right)^{K-2}, K = [\frac{n}{m-1} - 1] \tag{10}$$

which also depends only on $q_1$ and $q_2$. In Haiman (2007), using the Naus exact expressions of $q_1$ and $q_2$, we illustrate and compare numerically approximations (8) and (10). Notice that Fu (2001) has developed the well known Markov embedding algorithm. This algorithm enables the exact computation of the distribution function of scan statistics generated by Markov chains (in particular i.i.d. sequences) of Bernoulli $Y_n$'s.

Another situation where exact formulas for $q_1$ and $q_2$ are available, is (see Saperstein (1976) and Karve (1993)) when the $Y_n$ take values $0, 1, 2$ (or $-1, 0, 1$). In Haiman (2007) we use the tables established in Karve (1993) for $q_1$ and $q_2$ to compare numerically approximations (8) and (10).

In other cases there are no exact formulas for $q_1$ and $q_2$ . Thus, in order to use approximation (8), these quantities are calculated by Monte Carlo simulation methods. But then, one has to add to the above approximation error the error due to the simulation which depends in particular on the number of iterations used to estimate $q_1$ and $q_2$. This problem will be approached with more details in the further section concerning the application of approximation (9) to two and multi dimensional scan statistics.

For i.i.d. normal $Y_n$'s , Glaz et al. (2012) propose approximations for $P(S_n \leq s)$ based on bounds established in Glaz and Naus (1991). More details about these results are given below.

### 2.1.2 Scan statistics for $d-$*dependent* $Y_n$'s

Let the scan statistic defined in equation (1) be generated by $d-$dependent $Y_n$'s and for any integer $k \geq 1$, let

$$W_k = \max_{(k-1)(m+d-1)+1\leq t\leq k(m+d-1)} \sum_{i=t}^{t+m-1} Y_i. \tag{11}$$

It can be seen that the above sequence $\{W_k\}$ is also stationary, 1-dependent and that, if $n = (K+1)(m+d-1) - d$, $K \geq 1$, we have

$$P(S_n \leq s) = P(\max(W_1, ..., W_K) \leq s). \quad (12)$$

Thus, for any $n \geq 2(m+d-1) - d$ and $K = [\frac{n+d}{m+d-1} - 1]$ we have

$$P\{\max(W_1, ..., W_{K+1}) \leq s\} \leq P(S_n \leq s) \leq P\{\max(W_1, ..., W_K) \leq s\} \quad (13)$$

and we can use again approximation (8) for $P(S_n \leq s)$ with an error of less than about $3,3(K+1)(1-q_1)^2$. Here

$$q_1 = P(W_1 \leq s) = P(\max_{1 \leq t \leq m+d-1} \sum_{i=t}^{t+m-1} Y_i \leq s) \quad (14)$$

and

$$q_2 = P(W_1 \leq s, W_2 \leq s) = P(\max_{1 \leq t \leq 2(m+d-1)} \sum_{i=t}^{t+m-1} Y_i \leq s). \quad (15)$$

For i.i.d. normal $Y_n$'s , Glaz et al. (2012) propose approximations for $P(S_n \leq s)$ based on bounds established in Glaz and Naus (1991). These approximations are similar to approximation (8) and also depend only on $q_1$ and $q_2$. In Haiman and Preda (2013) we use approximation (8) to illustrate numerically the effect of dependence on the scan statistics distribution. We consider scan statistics generated by i.i.d. standard normal $Y_n$'s and by $1-$dependent sequences $\{Y_n\}_{n \geq 1}$ such that $Y_n = aZ_n + \sqrt{1-a^2}Z_{n+1}$, $0 <| a |\leq 1/2$ , where $\{Z_n\}_{n \geq 1}$ is a i.i.d. sequence of standard normal r.v.'s ( if $a = 0, Y_n$ are i.i.d.). The values of $q_1$ and $q_2$ are approximated by their corresponding empirical distributions $q_1^*$ and $q_2^*$ obtained by Monte Carlo simulation. It can then be seen that the total error on $P(S_n \leq s)$ at confidence level .95 is bounded by about $3.3(K+1)(1-q_1^*(s))^2 + 2K \times 1.96\sqrt{\frac{q_1^*(s)(1-q_1^*(s))}{I}}$, where $I$ is the number of replications used to estimate $q_1$ and $q_2$. The numerical results show that the distribution of scan statistic is very sensitive to the dependence parameter $a$.

In Amarioarei and Preda (2014) the authors extend this particular dependence model to block-factor models obtained from i.i.d. sequences in the context of the one and two dimensional scan statistics. For more details see also Amarioarei (2014).

In Haiman (2013) we have developed a model of 1-dependent stationary sequences adapted to any given joint distribution of two consecutive r.v.'s

provided it is sufficiently close to independence. Namely, if the $Y_n$'s are integer valued, the condition is that there exists $\alpha$, $.75 \leq \alpha \leq 1$ such that

$$P(Y_n = i, Y_{n+1} = j) - \alpha P\{Y_1 = i\}P\{Y_1 = j\} \geq 0 \ \forall i, j \in \mathbb{N}^* \ . \tag{16}$$

Notice that the previous condition is weaker than the more classical mixing condition

$$\max_{i,j} \frac{\mid P\{Y_n = i, Y_{n+1} = j\} - P\{Y_1 = i\}P\{Y_1 = j\} \mid}{P\{Y_1 = i\}P\{Y_1 = j\}} \leq .25. \tag{17}$$

The joint distribution of our model of sequence $\{Y_n\}_{n\geq 1}$ is given by the recurrence formula :

$P\{Y_1 = l_1, ..., Y_{n+1} = l_{n+1}\} = p(l_1, ..., l_{n+1}) =$

$$p(l_1, ..., l_n)p(l_{n+1}) + p(l_1, \ldots, l_{n-1}) \times \ [\ p(l_n, l_{n+1}) - p(l_n)p(l_{n+1})] \ , \tag{18}$$

$l_i \in \mathbb{N}, i = 1, \ldots, n+1, n \geq 3$.

For Bernoulli $B(1, p)$ $Y_n$'s, the sequences satisfying condition (16) are members of either one of the following one parameter families **a**) and **b)** : Let $p = P(Y_1 = 1) = 1 - p(0), 0 < p < 1$ be fixed and consider the set of bivariate distributions $p(i, j) = P(Y_n = i, Y_{n+1} = j), i, j \in \{0, 1\}$, such that :

**a**) if $p(0, 0) < p^2(0) = (1 - p)^2$,
$p(0, 0) = (1 - p)^2\nu$, $p(0, 1) = p(1, 0) = 1 - p - (1 - p)^2\nu$ and
$p(1, 1) = 2p - 1 + (1 - p)^2\nu$ where
$1 - \frac{1}{4}(\frac{p}{1-p})^2 \leq \nu < 1$ if $p \leq \frac{1}{2}$ and $\frac{3}{4} \leq \nu < 1$ if $p > \frac{1}{2}$.
**b**) if $p(0, 0) \geq p^2(0)$,
$p(0, 0) = 1 - p - (1 - p)p\nu$, $p(0, 1) = p(1, 0) = (1 - p)p\nu$ and
$p(1, 1) = p - (1 - p)p\nu$, where $\frac{3}{4} \leq \nu < 1$ .

In Haiman and Preda (2013) we compare numerically the distributions of scan statistics generated by the previous 1-dependent model of Bernoulli $Y_n$'s with those generated by Markov chains having the same distribution for two consecutive r.v.'s. For the 1-dependent model we use approximation (8) whereas for the Markov chain we use the Markov embedding algorithm of Fu (2001). For the 1-dependent sequences, the exact values of $q_1$ and $q_2$ are calculated using the recurrence formula (18). The numerical results show in particular that a higher dependence (model b, $\nu = .75$) between consecutive r.v.'s changes significantly the distribution of the scan statistics

as compared to the corresponding i.i.d. case $\nu = 1$.These results also show that the Markov and the associated 1-dependent model generate significantly different distributions of scan statistics.

## 2.2 *One-dimensional continuous scan statistics*

Let $N$ be a Poisson process of intensity $\lambda$ on $\mathbb{R}_+$ and let $u > 0$ and $T > u$ be fixed constants. The one dimensional continuous scan statistic is defined as

$$S_T(u,\lambda) = S_T = \max_{0\le t\le T-u}(N(t+u) - N(t)) \tag{19}$$

Let $T = (K+1)u, K$ integer $\geq 1$, and let

$$W_k = \max_{(k-1)u\le t\le ku}(N(t+u) - N(t)), k = 1, ..., K. \tag{20}$$

As for the discrete scan statistic , $\{W_k\}$ is a 1-dependent stationary sequence and

$$P(S_T \le s) = P(\max(W_1, ..., W_K) \le s), s \in \mathbb{N}. \tag{21}$$

We then also can apply approximation (8), with

$$q_1 = q_1(s) = P(W_1 \le s) = P\{\max_{0\le t\le u}(N(t+u) - N(t)) \le s\}, s \in \mathbb{N} \tag{22}$$

and

$$q_2 = q_2(s) = P(W_1 \le s, W_2 \le s) = P\{\max_{0\le t\le 2u}(N(t+u) - N(t)\} \le s), s \in \mathbb{N}. \tag{23}$$

Huntigton and Naus (1975) give an exact formula for $P(S_T \le s)$ that sums many products of determinants and for large $T$ requires excessive computer time. This formula is used in Neff and Naus (1980) to establish tables for the d.f. of $S_n(1,\lambda)$ (notice that $S_T(u,\lambda) = S_{T/u}(1,\lambda u)$), for several discrete values of $\lambda$ and $n \le 100$. In Haiman (2000) we have applied and compared approximations (8) and (10) with $q_1$ and $q_2$ from Neff and Naus (1980) tables, for several values of $\lambda$ and $n = 1000$.

# 3 Application to multi dimensional scan statistics.

## 3.1 Two dimensional discrete scan statistic

Let $m_1$ and $m_2$ be positive integers and let $n_1$ and $n_2$ be integers such that $1 \leq m_1 \leq n_1$ and $1 \leq m_2 \leq n_2$. Let $\{Y_{i,j}\}_{i,j\geq 1}$ be a family of non negative i.i.d. integer valued r.v.'s and let the two dimensional scan statistic generated by $\{Y_{i,j}\}$ be defined as

$$S = S_{n_1,n_2} = S_{n_1,n_2}(m_1, m_2) = \max_{1\leq u\leq n_1-m_1+1, 1\leq v\leq n_2-m_2+1} \sum_{i=u}^{u+m_1-1} \sum_{j=v}^{v+m_2-1} Y_{i,j}. \tag{24}$$

Let $n_1 = (K+1)m_1 - 1$ and $n_2 = (L+1)m_2 - 1$ where $K$ and $L$ are positive integers.

Let

$$U_k = \max_{(k-1)m_1+1\leq u\leq km_1, 1\leq v\leq n_2-m_2+1} \sum_{i=u}^{u+m_1-1} \sum_{j=v}^{v+m_2-1} Y_{i,j}. \tag{25}$$

Observe that $\{U_k\}, k = 1, ..., K$ is a 1-dependent stationary sequence and for any $s \in \mathbb{N}$

$$P\{S \leq s\} = P\{\max_{k=1,...,K} U_k \leq s\}. \tag{26}$$

Put

$$q_1 = P(U_1 \leq s) = P\{\max_{1\leq u\leq m_1, 1\leq v\leq n_2-m_2+1} \sum_{i=u}^{u+m_1-1} \sum_{j=v}^{v+m_2-1} Y_{i,j} \leq s\} \tag{27}$$

and

$$q_2 = P(U_1 \leq s, U_2 \leq s) = P\{\max_{1\leq u\leq 2m_1, 1\leq v\leq n_2-m_2+1} \sum_{i=u}^{u+m_1-1} \sum_{j=v}^{v+m_2-1} Y_{i,j} \leq s\}. \tag{28}$$

Then, if $1- \ q_1 \ \leq \ .025$, we can apply approximation (8), $P(S \leq s) \simeq \frac{2q_1-q_2}{[1+q_1-q_2+2(q_1-q_2)^2]^K}$, with an error of less than about $3.3(K+1)(1-q_1(s))^2$.

Next, in order to calculate $q_1$ and $q_2$ in the previous approximation, for $l = 1, ..., L$ let

$$V_l = \max_{1\le u\le m_1,(l-1)m_2+1\le v\le lm_2} \sum_{i=u}^{u+m_1-1} \sum_{j=v}^{v+m_2-1} Y_{i,j} \tag{29}$$

and

$$W_l = \max_{1\le u\le 2m_1,(l-1)m_2+1\le v\le lm_2} \sum_{i=u}^{u+m_1-1} \sum_{j=v}^{v+m_2-1} Y_{i,j}. \tag{30}$$

The sequences $\{V_l\}_{l=1,...,L}$ and $\{W_l\}_{l=1,...,L}$ are both stationary 1-dependent and one can again use approximation (8): let $q_{11} = P(V_1 \le s)$, $q_{12} = P(V_1 \le s, V_2 \le s)$, $q_{21} = P(W_1 \le s)$ and $q_{22} = P(W_1 \le s, W_2 \le s)$. Notice that $q_{11} = P(S_{2m_1,2m_2} \le s)$, $q_{12} = P(S_{2m_1,3m_2} \le s)$, $q_{21} = P(S_{3m_1,2m_2} \le s)$ and $q_{22} = P(S_{3m_1,3m_2} \le s)$. If $1 - q_{11} \le .025$ and $1 - q_{21} \le .025$ we have

$$q_1 \simeq (2q_{11} - q_{12})[1 + q_{11} - q_{12} + 2(q_{11} - q_{12})^2]^{-L} \tag{31}$$

with an error of less than about $3.3(L + 1)(1 - q_{11})^2$ and

$$q_2 \simeq (2q_{12} - q_{22})[1 + q_{12} - q_{22} + 2(q_{12} - q_{22})^2]^{-L} \tag{32}$$

with an error of $3.3(L + 1)(1 - q_{21})^2$. Assuming that the $q_{ij}$ are known, $1 - q_{11}$ and $1 - q_{21}$ are small and $L \le K$, it can be seen, substituting the above approximations of $q_1$ and $q_2$ in approximation (9), that the total error on $P(S \le s)$ is bounded by about

$$E_{app} = 3.3(L + 1)(K + 1)[(1 - q_{11})^2 + (1 - q_{21})^2 + (L + 1)(q_{11} - q_{21})^2]. \tag{33}$$

However, in general there are no exact formulas for $q_{11}, q_{12}$ and $q_{22}$ which can only be approximated by Monte Carlo simulation . In Haiman and Preda (2006) we have applied the previous method to binomial and Poisson $Y_{i,j}$ 's. As for one dimensional scan statistics, we calculate the additional, at confidence level .95 simulation error $E_{sim}$ . Here $E_{sim}$ is proportional to $(L + 1)(K + 1) \times .95\sqrt{\frac{1}{I}}$ where $I$ is the number of replications used to estimate the $q_{ij}$. Thus, the total error on $P\{S \le s\}$ is bounded by about $E = E_{app} + E_{sim}$ . We compare numerically our results with results obtained using the product approximation, the Poisson approximation and Bonferroni inequality techniques as presented in Glaz et al (2001). For binary $Y_{i,j}$ 's we compare our values to bounds obtained in Boutsikas and Koutras (2003).

In Amarioarei and Preda (2014), the previous method was adapted to some block factor type dependent models of binary $Y_{i,j}$ 's.

## 3.2 Three dimensional discrete scan statistics

Amarioarei and Preda (2015) have also adapted the above method to he three dimensional discrete scan statistic. Similarly to the two-dimensional, the three dimensional discrete scan statistic $S = S_{n_1,n_2,n_3}(m_1, m_2, m_3)$ is defined as the maximum of sums of r.v.'s $Y_{i,j,k}$ over all three-dimensional paralellipipedic windows of side lengths $m_1, m_2, m_3$ moving inside the parallelepiped of side lengths $n_1, n_2, n_3$, where $m_1 \leq n_1, m_2 \leq n_2$ and $m_3 \leq n_3$. In this case the application of the method requires the estimation of quantities $q_{111}, q_{121}, q_{112}, q_{122}, q_{211}, q_{221}, q_{212}$ and $q_{222}$ which are the d.f.'s of scan statistics $S_{2m_1,2m_2,2m_3}, \ldots, S_{3m_1,3m_2,3m_3}$. In order to obtain reasonable simulation errors for these quantities, as in Amarioarei and Preda (2014), the authors use the *importance sampling method* introduced in Naiman and Priebe (2001).

## 3.3 Two dimensional continuous scan statistics

Let $N$ be a two dimensional Poisson process of intensity $\lambda$ . For fixed real numbers $0 < u < n_1$ and $0 < v < n_2$, let the two dimensional continuous scan statistic generated by $N$ be defined as

$$S = S((u, v), \lambda, n_1, n_2) = \max_{0 \leq t \leq n_1 - u, 0 \leq z \leq n_2 - v} N([t, t + u] \times [z, z + v]). \quad (34)$$

Observing that for any integer $s \geq 0$

$$P\{S((u, v), \lambda, n_1, n_2) \leq s\} = P\{S((1, 1), \lambda uv, n_1/u, n_2/v) \leq s\} \quad (35)$$

there is no loss of generality to assume that $u = v = 1$. Let $n_1 = K + 1$ , $n_2 = L + 1$ where $L$ and $K$ are positive integers and let

$$U_k = \max_{k-1 \leq t \leq k, 0 \leq z \leq L} N([t, t + 1] \times [z, z + 1]), k = 1, ..., K. \quad (36)$$

Then, $\{U_k\}, k = 1, \ldots, K$ is a 1-dependent stationary sequence and for any $s \in \mathbb{N}$ we have also, as previously for the two dimensional discrete scan statistics, $P\{S[(1, 1), \lambda, K+1, L+1] \leq s\} = P\{S \leq s\} = P\{\max_{k=1,\ldots,K} U_k \leq s\}$. Put

$$q_1 = P(U_1 \leq s) = P\{\max_{0 \leq t \leq 1, 0 \leq z \leq L} N([t, t + 1] \times [z, z + 1]) \leq s\} \quad (37)$$

and

$$q_2 = P(U_1 \leq s, U_2 \leq s) = P\{\max_{0\leq t\leq 2, 0\leq z\leq L} N([t,t+1]\times[z,z+1] \leq s\}. \quad (38)$$

Then, if $1- q_1 \leq .025$, we can again apply approximation (8), $P(S \leq s) \simeq \frac{2q_1-q_2}{[1+q_1-q_2+2(q_1-q_2)^2]^K}$, with an error of less than about $3.3(K+1)(1-q_1(s))^2$. Next , in order to calculate $q_1$ and $q_2$ in the previous approximation, for $l = 1, ..., L$ let

$$V_l = \max_{1\leq t\leq 1, l-1\leq z\leq l} N([t,t+1]\times[z,z+1]) \quad (39)$$

and

$$W_l = \max_{1\leq t\leq 2, l-1\leq z\leq l} N([t,t+1]\times[z,z+1]). \quad (40)$$

The sequences $\{V_l\}_{l=1,...,L}$ and $\{W_l\}_{l=1,...,L}$ are both stationary 1-dependent and one can again use approximation (8): let $q_{11} = P(V_1 \leq s)$, $q_{12} = P(V_1 \leq s, V_2 \leq s)$, $q_{21} = P(W_1 \leq s)$ and $q_{22} = P(W_1 \leq s, W_2 \leq s)$. Notice that $q_{11} = P\{S[(1,1),\lambda,2,2] \leq s\}$, $q_{12} = P\{S[(1,1),\lambda,2,3] \leq s\} = q_{21} = P\{S[(1,1),\lambda,3,2] \leq s\}$ and $q_{22} = P\{S[(1,1),\lambda,3,3] \leq s\}$. If $1 - q_{11} \leq .025$ and $1 - q_{21} \leq .025$ we again have

$$q_1 \simeq (2q_{11} - q_{12})[1 + q_{11} - q_{12} + 2(q_{11} - q_{12})^2]^{-L} \quad (41)$$

with an error of less than about $3.3(L+1)(1-q_{11})^2$ and

$$q_2 \simeq (2q_{12} - q_{22})[1 + q_{12} - q_{22} + 2(q_{12} - q_{22})^2]^{-L} \quad (42)$$

with an error of $3.3(L+1)(1-q_{21})^2$. Assuming that the $q_{ij}$ are known, $1-q_{11}$ and $1-q_{21}$ are small and $L \leq K$, as previously, substituting the above approximations of $q_1$ and $q_2$ in approximation (9), it can be shown that the total error on $P(S \leq s)$ is bounded by about

$$E_{app} = 3.3(L+1)(K+1)[(1-q_{11})^2 + (1-q_{21})^2 + (L+1)(q_{11}-q_{21})^2]. \quad (43)$$

However, as for the discrete two dimensional scan statistics, there are no exact formulas for $q_{11}, q_{12}$ and $q_{22}$ which can only be approximated by Monte Carlo simulation . Let $S_{k,l} = S[(1,1),\lambda,k+1,l+1]$, $k,l = 1,2$ and consider the d.f. of the conditional scan statistic given that a fixed number $n$ of points fall in $[0,k+1]\times[0,l+1]$ :

$$q_{k,l}^n(s) = P\{S_{k,l} \leq s \mid N([0,k+1]\times[0,l+1]) = n\},\ 1 \leq s \leq n\ . \qquad (44)$$

Observe that $q_{k,l}^n$ is the d.f. of r.v. $S_{k,l}^n$ = maximum number of points obtained by scanning the rectangle $[0,k+1]\times[0,l+1]$ in which $n$ independent points are drawn uniformly. We then have

$$q_{k,l}(s) = e^{-\lambda kl}(\sum_{j=0}^{s}\frac{[\lambda(k+1)(l+1)]^j}{j!} + \sum_{j=s+1}^{s(k+1)(l+1)} q_{k,l}^j(s)\frac{[\lambda(k+1)(l+1)]^j}{j!}). \qquad (45)$$

In Haiman and Preda (2002) we have developed a particular method of simulation independent replications of r.v.'s $S_{k,l}^n, k,l = 1,2$.We use this method to obtain empirical estimations of $q_{k,l}^n(s)$ from which by formula (45) we deduce the final approximations $q_{k,l}^*$ of $q_{k,l}$ . The empirical estimations of $q_{k,l}^n$ generate additional errors. These errors are bounded at the .95 confidence level by $\varepsilon_{k,l}$ where $\varepsilon_{k,l} \simeq 1,96\sqrt{\frac{q_{k,l}^*(1-q_{k,l}^*)}{I}}$. Here $I$ is the number of replications of r.v.'s $S_{k,l}^n, k,l = 1,2$. The total error on $P(S \leq s)$ is then bounded by about

$$E = E_{app} + (K+1)(L+1)(\varepsilon_{1,1} + \varepsilon_{1,2} + \varepsilon_{2,2}). \qquad (46)$$

Naus (1965) and Neff (1978) give exact formulas for $q_{k,l}^s(s-1)$ and $q_{k,l}^s(s-2)$. In Haiman and Preda (2002) these formulas are used to evaluate the simulation results. Numerical examples for several values of $K, L$ and $\lambda$ are given and the results are compared with approximation formulas proposed in Aldous (1989) and Alm (1997).

# 4 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. Amarioarei A. (2012), *Approximation for the distribution of extremes of one dependent stationary sequences of random variables*, arXiv:1211.5456 [math.PR].

4. Amarioarei A. and Preda C. (2015) *Approximation for the distribution of three-dimensional discrete scan statistic*, Methodology and Computing in Applied Probability, Volume 17, Issue 3, pag. 565-578,

5. Amarioarei A. (2014) *Approximations for Multidimensional Discrete Scan Statistics*, Phd thesis, no. 41498, Université de Lille 1, France. https://alexamarioarei.github.io/Research/docs/Thesis.pdf

6. Amarioarei A. and Preda C. (2014) *Approximation for two-dimensional discrete scan statistic in some block -factor type dependent models,* Journal of Statistical Planning and Inference, vol 151-152: 107-120.

7. 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.

8. Fu J.C. (2001) *Distribution of the scan statistic for a sequence of bistate trials,* Journal of Applied Probability, 38, 59-68.

9. Glaz J., Naus J. (1991) *Tight bounds and approximations for scan statistics probabilities for discrete data.* Annals of Applied Probability 1, 306-318.

10. Glaz J., Naus J., Wallenstein S.(2001) *Scan statistics*, Springer Series in Statistics.

11. Glaz J., Pozdnyakov V., Wallenstein S. (2009) *Scan statistics Methods and Applications ,* Birkhäuser , Statistics for Industry and Technology.

12. Glaz J., Naus J.and Xiao Wang (2012) *Approximations and Inequalities for Moving Sums,* Methodology and Computing in Applied Probability, 14 (3),597-613.

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) Probabilité et Statistiques, 34:4, 481-503.

14. Haiman G. (1999) *First passage time for some stationary processes,* Stoch. Process. Their Appl. 80:231-248.

15. Haiman G. (2000) *Estimating the distribution of scan statistics with high precision,* Extremes, 3, 349-361.

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

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

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

19. Haiman G., Preda C. (2013) *One dimensional scan statistics generated by some dependent stationary sequences ,* Statistics and Probability Letters 83, 1457-1463.

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

21. Naiman D.Q. and Priebe C.E.(2001) *Computing scan statistic p values using importance sampling , with applications to genetics and medical image analysis* . J. Comput. Graph. Statist., vol. 10 : 296-328.

22. Naus J.I. (1965 *) A power comparison of two tests of non-random clustering ,* Technometrics vol.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.D. (1978) *Piecevise 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 of points on a line*, Vol. VI in IMS Series of selected tables in mathematical statistics, American Mathematical Society, Providence.

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