Scan statistics for some dependent models. Applications.

Alexandru Amărioarei, Cristian Preda

Faculty of Mathematics, University of Bucharest
Université de Lille and ISMMA Romanian Academy

StatMod2020
Statistical Modeling with Applications

Friday, November 6, 2020, Bucharest, Romania

Outline

Outline

One dimensional discrete scan statistics

The scan statistics

Let $1 \leq m \leq T$ be positive integers, $X_1, X_2, \ldots, X_T$ a sequence of r.v.'s. Then, the one dimensional discrete scan statistics is defined as

$$S_m(T) = \max_{1 \leq i \leq T-m+1} \sum_{j=i}^{i+m-1} X_j$$

Example $\left(T = 26,\ m = 6,\ X_i \sim \mathcal{B}(p),\ Y_i = X_i + \cdots + X_{i+m-1},\ 1 \leq i \leq 21\right)$

1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	1	1	1	1	0

$Y_1 = 4$

Related statistics

Let $X_1, \ldots, X_T$ be a sequence of i.i.d. $0-1$ Bernoulli of parameter p

- $W_{m,k}$ - the **waiting time** until we first observe at least k successes in a window of size m

$$\mathbb{P}(W_{m,k} \leq T) = \mathbb{P}(S_m(T) \geq k)$$

- $D_T(k)$ - the length of the **smallest window** that contains at least k successes

$$\mathbb{P}(D_T(k) \leq m) = \mathbb{P}(S_m(T) \geq k)$$

- L_T - the length of the **longest success** run

$$\mathbb{P}(L_T \geq m) = \mathbb{P}(S_m(T) \geq m) = \mathbb{P}(S_m(T) = m)$$

Problem and approaches

Problem

Find a good aproximation for the distribution of the discrete scan statistic

$$\mathbb{P}\left(S_m(T) \leq s\right).$$

Previous work (i.i.d. model):

- Exact results (Bernoulli)
  - Combinatorial method: [Naus, 1974], [Naus, 1982]
  - Finite Markov chain imbedding: [Fu, 2001], [Fu and Lou, 2003], [Wu, 2013]
  - Conditional generating function: [Ebneshahrashoob and Sobel, 1990], [Gao et al., 2005]
- Approximations
  - Product-type: [Naus, 1982], [Karwe and Naus, 1997]
  - Poisson: [Chen and Glaz, 1997], [Glaz et al., 2001]
- Bounds
  - Product-type: [Glaz and Naus, 1991], [Wang et al., 2012]
  - Bonferroni: [Glaz, 1990]

▸ Product-Type Approximations & Bounds ▸ MCIT

The 1-dependent model

Haiman (2011)

Let $p(x_1, x_2)$ be a bivariate discrete distribution and $p(x_1)$ and $p(x_2)$ its marginals.

Dependence condition : there exists α, $\frac{3}{4} \leq \alpha \leq 1$ such that

$$p(x_1, x_2) - \alpha p(x_1)p(x_2) \geq 0.$$

The 1-dependent model

If the dependence condition holds, then there exists a 1-dependent stationary sequence $\{X_i\}_{i\geq 1}$ having p as distribution for (X_i, X_{i+1}) and is such that the following recurrence holds :

$$\mathbb{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n, X_{n+1} = x_{n+1}) =$$

$$\begin{aligned}
&\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)\,\mathbb{P}(X_{n+1} = x_{n+1}) + \\
&\mathbb{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1})\,[\mathbb{P}(X_n = x_n, X_{n+1} = x_{n+1}) - \\
&\qquad\qquad \mathbb{P}(X_n = x_n)\,\mathbb{P}(X_{n+1} = x_{n+1})].
\end{aligned}$$

The joint distribution of $X_1, \ldots, X_n$ depends only on the stationary distribution (X_1) and the bivariate distribution (X_1, X_2).

The 1-dependent Bernoulli model

$X_1 \sim \mathcal{B}(p)$, $p = \mathbb{P}(X_1 = 1)$.
Denote by $p(i) = \mathbb{P}(X_1 = i)$ and by $p(i,j) = \mathbb{P}(X_1 = i, X_2 = j)$, $i,j \in \{0,1\}$.

The mixing condition is satisfied for the joint distribution of the two parametrized families :

A) if $p(0,0) < p(0)^2$, then

$$p(0,0) = (1-p)^2\nu, \quad p(0,1) = p(1,0) = 1 - p - (1-p)^2\nu,$$
$$p(1,1) = 2p - 1 + (1-p)^2\nu,$$

$$\text{with } \begin{cases} 1 - \frac{1}{4}\left(\frac{p}{1-p}\right)^2 \leq \nu < 1 & \text{if } p \leq \frac{1}{2}, \\ \frac{3}{4} \leq \nu < 1 & \text{if } p > \frac{1}{2}. \end{cases}$$

B) if $p(0,1) \leq p(0)p(1)$, then

$$p(0,0) = 1 - p - (1-p)p\nu, \quad p(0,1) = p(1,0) = (1-p)p\nu,$$
$$p(1,1) = p - (1-p)p\nu,$$

$$\text{with } \frac{3}{4} \leq \nu < 1.$$

Outline

Approximation

Let $X_1, \ldots, X_T$ be a 1-dependent stationary sequence.
Let define the random variables :

$$Z_k = \max_{(k-1)m+1 \le t \le km} \sum_{i=t}^{t+m-1} X_i, \quad k \ge 1$$

Z_k is the scan statistic on the sub-sequence of length $2m-1$,

$$\{X_{(k-1)m}, \ldots, X_{(k+1)m-1}\}.$$

Then, if $T = (K+1)m - 1$, for some $K \in \mathbb{N}^*$,

$$S_m(T) = \max_{1 \le k \le K} Z_k$$

Example. $m = 3$, $n = 11$ $(K = 3)$.

$$\overbrace{\underbrace{X_1 X_2 X_3 X_4 X_5}_{Z_1} X_6 \underbrace{X_7 X_8 X_9 X_{10} X_{11}}_{Z_3}}^{Z_2}$$

Approximation

Observe that, under the model of 1-dependence of $X_i's$ (independence is included),

$$\textit{the sequence } \{Z_1, Z_2, \ldots, Z_K\} \textit{ is 1-dependent}$$

A result of Haiman (1999) approximates

$$\mathbb{P}(S_m(T) \leq s) = \mathbb{P}(\max_{1 \leq k \leq K} Z_k \leq s)$$

using only the distributions of Z_1 and Z_2.

Approximation

More precisely, let denote by

$$q_L = q_L(s) = \mathbb{P}(\max_{1 \le k \le L} Z_k \le s),\ 1 \le L \le K.$$

Then, for s such that $1 - q_1(s) \le 0.025$ and any $L \ge 3$,

$$q_L \approx \frac{2q_1 - q_2}{(1 + q_1 - q_2 + 2(q_1 - q_2)^2)^L},$$

with an error bound of about $3.3L(1 - q_1)^2$.

q_1 AND q_2

$$q_1(s) = \mathbb{P}(Z_1 \leq s) = \mathbb{P}(\max_{1 \leq t \leq \mathbf{m}} \sum_{i=t}^{t+m-1} X_i \leq s).$$

Scanning over : $X_1 X_2 \ldots X_m X_{m+1} \ldots X_{2m-1}$

$$q_2(s) = \mathbb{P}(\max(Z_1, Z_2) \leq s) = \mathbb{P}(\max_{1 \leq t \leq \mathbf{2m}} \sum_{i=t}^{t+m-1} X_i \leq s).$$

Scanning over : $X_1 X_2 \ldots X_m X_{m+1} \ldots X_{2m-1} X_{2m} \ldots X_{3m-1}$

Outline

Computation of q_1 and q_2 for 1-dependent Bernoulli model

$$q_1(s) = \mathbb{P}(Z_1 \leq s) = \sum_{u=0}^{s} \mathbb{P}(Z_1 = u)$$

Let $\Omega(2m-1) = \{0,1\}^{2m-1}$ and $\mathbf{x} \in \Omega(2m-1)$, $\mathbf{x} = x_1 \dots x_{2m-1}$. Then,

$$\mathbb{P}(Z_1 = u) = \sum_{\{\mathbf{x} \in \Omega(2m-1) \,\big|\, Z_1(\mathbf{x}) = u\}} p(\mathbf{x}).$$

The probability $p(\mathbf{x}) = P(X_1 = x_1, \dots, X_{2m-1} = x_{2m-1})$ is computed by the recurrence formula.

Similarly for q_2 : $\Omega(3m-1)$

Comparison of independent, Markov and 1-dependent models for Bernoulli trials

Let $X_1, \ldots, X_T$ a stationary 1-dependent sequence of Bernoulli $\mathcal{B}(p)$ rv's with joint bivariate distribution from family A) or B),

$$P = \begin{bmatrix} p(0,0) & p(0,1) \\ p(1,0) & p(1,1) \end{bmatrix}.$$

- independent models : Naus (1982), Fu et a. (2001, 2003), Haiman (2007)
  $X_i's$ are independent and identically distributed as $\mathcal{B}(p)$.

- Markov chain models : Fu et al. (2003)
  $X_i's$ are identically distributed as $\mathcal{B}(p)$ with transition matrix,

$$M = \begin{bmatrix} \frac{p(0,0)}{p(0)} & \frac{p(0,1)}{p(0)} \\ \frac{p(1,0)}{p(1)} & \frac{p(1,1)}{p(1)} \end{bmatrix}.$$

Numerical results

$p = 0.1$, P_A : $\nu = 0.997$ (family A), P_B : $\nu = 0.75$ (family B).

$X_1 \setminus X_2$	0	1
0	0.81	0.09
1	0.09	0.01

i.i.d.

$X_1 \setminus X_2$	0	1
0	0.80757	0.09243
1	0.09243	0.000757

family A)

$X_1 \setminus X_2$	0	1
0	0.8325	0.0675
1	0.0675	0.0325

family B)

$m = 6$. Exact distribution of Z_1, $q_1(s)$:

s	$i.i.d.$	$1-dep(A)$	$1-dep(B)$	$Markov(A)$	$Markov(B)$
0	0.313811	0.304497	0.410080	0.304522	0.412724
1	0.755469	0.759346	0.700596	0.759614	0.726652
2	0.953724	0.960320	0.909669	0.960177	0.896173
3	0.995232	0.997003	0.977538	0.996862	0.967309
4	0.999746	0.999910	0.996744	0.999889	0.991652
5	0.999994	0.999999	0.999731	0.999998	0.998413
6	1	1	1	1	1

$m = 6$. Exact distribution of $max\{Z_1, Z_2\}$, $q_2(s)$:

s	$i.i.d.$	$1 - dep(A)$	$1 - dep(B)$	$Markov(A)$	$Markov(B)$
0	0.166771	0.158923	0.255779	0.158944	0.258529
1	0.625680	0.629343	0.564735	0.629705	0.599638
2	0.919182	0.929812	0.850219	0.929636	0.832195
3	0.991107	0.994334	0.959838	0.994086	0.943436
4	0.999508	0.999824	0.993871	0.999783	0.984759
5	0.999989	0.999998	0.999473	0.999997	0.996946
6	1	1	1	1	1

- $m = 6,\ T = 125\ \ (K = 20) : \mathbb{P}(S \leq s)$

s	$i.i.d.$	$1 - dep(A)$	$1 - dep(B)$	$Markov(A)$	$Markov(B)$
3	0.919713	0.947506 (± 0.000598)	0.692126 (± 0.03329)	0.945422	0.601693
4	0.995241	0.998277 ($\pm 5.34 \times 10^{-7}$)	0.943565 (± 0.000699)	0.997893	0.868558
5	0.999891	0.999980 ($\pm 6.6 \times 10^{-11}$)	0.994840 ($\pm 4.7 \times 10^{-6}$)	0.999972	0.970896

- $m = 6,\ T = 605\ \ (K = 100) : \mathbb{P}(S \leq s)$

s	$i.i.d.$	$1 - dep(A)$	$1 - dep(B)$	$Markov(A)$	$Markov(B)$
3	0.659705	0.764683 (± 0.002964)	0.161889 (± 0.166445!!)	0.756386	0.081509
4	0.976497	0.991431 ($\pm 2.67 \times 10^{-6}$)	0.749074 (± 0.003498)	0.989533	0.497092
5	0.999460	0.999900 ($\pm 3.3 \times 10^{-10}$)	0.974509 ($\pm 2.38 \times 10^{-5}$)	0.999862	0.863110

Outline

Longest increasing run

Longest increasing run

Let $(Y_n)_{n\geq 1}$ be a sequence of i.i.d. r.v.'s with the common distribution G.

Increasing run

A subsequence $(Y_k, \ldots, Y_{k+l-1})$ forms an *increasing run* of length $l \geq 1$, starting at position $k \geq 1$, if

$$Y_{k-1} > Y_k < Y_{k+1} < \cdots < Y_{k+l-1} > Y_{k+l}$$

Notations

- $M_{\tilde{T}}$ = the length of the longest increasing run among the first $\tilde{T}$ r.v.'s
- $L_{\tilde{T}}$ = the length of the longest run of ones among the first $\tilde{T}$ r.v.'s

The asymptotic distribution was studied

- G continuous distribution: [Pittel, 1981], [Révész, 1983], [Grill, 1987], [Novak, 1992], etc.
- G discrete distribution: [Csaki and Foldes, 1996], [Grabner et al., 2003], [Eryilmaz, 2006], etc.

Longest increasing run

Scan statistics approach

Let $T = \tilde{T} - 1$ and define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_i = \mathbf{1}_{Y_i < Y_{i+1}}$

Example ($Y_i \sim \mathcal{U}(0,1)$, $\tilde{T} = 10$)

Y_i : 0.79 0.31 0.52 0.16 0.60 0.26 0.65 0.68 0.74 0.45

X_i :

We have

$$\mathbb{P}\left(M_{\tilde{T}} \leq m\right) = \mathbb{P}\left(L_T < m\right) = \mathbb{P}\left(S_m(T) < m\right), \text{ for } m \geq 1$$

Longest increasing run

Scan statistics approach

Let $T = \tilde{T} - 1$ and define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_i = \mathbf{1}_{Y_i < Y_{i+1}}$

Example ($Y_i \sim \mathcal{U}(0,1)$, $\tilde{T} = 10$)

Y_i :	0.79	0.31	0.52	0.16	0.60	0.26	0.65	0.68	0.74	0.45
X_i :		0								

We have

$$\mathbb{P}\left(M_{\tilde{T}} \leq m\right) = \mathbb{P}\left(L_T < m\right) = \mathbb{P}\left(S_m(T) < m\right), \text{ for } m \geq 1$$

Longest increasing run

Scan statistics approach

Let $T = \tilde{T} - 1$ and define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_i = \mathbf{1}_{Y_i < Y_{i+1}}$

Example ($Y_i \sim \mathcal{U}(0,1)$, $\tilde{T} = 10$)

Y_i :	0.79	0.31	0.52	0.16	0.60	0.26	0.65	0.68	0.74	0.45
X_i :		0		1						

We have

$$\mathbb{P}\left(M_{\tilde{T}} \leq m\right) = \mathbb{P}\left(L_T < m\right) = \mathbb{P}\left(S_m(T) < m\right), \text{ for } m \geq 1$$

Longest increasing run

Scan statistics approach

Let $T = \tilde{T} - 1$ and define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_i = \mathbf{1}_{Y_i < Y_{i+1}}$

Example ($Y_i \sim \mathcal{U}(0,1)$, $\tilde{T} = 10$)

Y_i : 0.79 0.31 0.52 0.16 0.60 0.26 0.65 0.68 0.74 0.45

X_i : 0 1 0

We have

$$\mathbb{P}\left(M_{\tilde{T}} \leq m\right) = \mathbb{P}\left(L_T < m\right) = \mathbb{P}\left(S_m(T) < m\right), \text{ for } m \geq 1$$

Longest increasing run

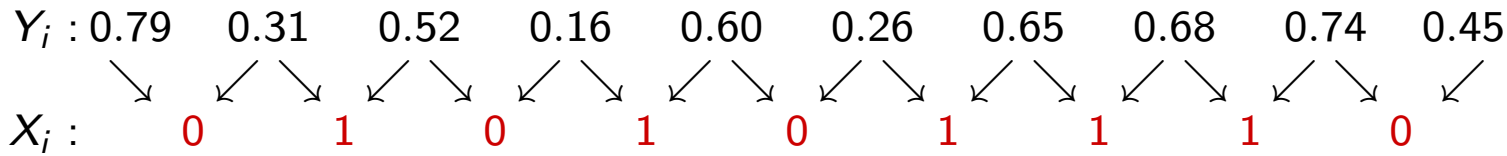
Scan statistics approach

Let $T = \tilde{T} - 1$ and define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_i = \mathbf{1}_{Y_i < Y_{i+1}}$

Example ($Y_i \sim \mathcal{U}(0,1)$, $\tilde{T} = 10$)



We have

$$\mathbb{P}\left(M_{\tilde{T}} \leq m\right) = \mathbb{P}\left(L_T < m\right) = \mathbb{P}\left(S_m(T) < m\right), \text{ for } m \geq 1$$

Longest increasing run

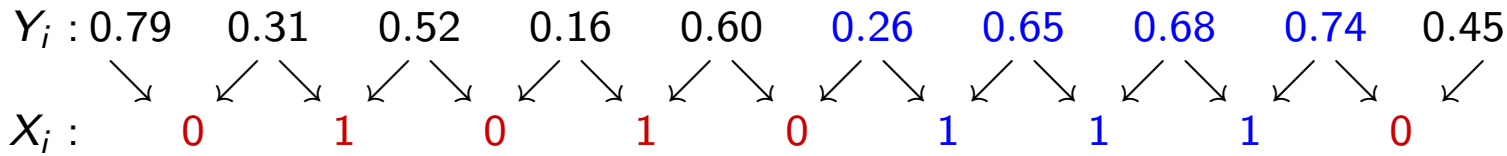
Scan statistics approach

Let $T = \tilde{T} - 1$ and define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_i = \mathbf{1}_{Y_i < Y_{i+1}}$

Example ($Y_i \sim \mathcal{U}(0,1)$, $\tilde{T} = 10$)



We have

$$\mathbb{P}\left(M_{\tilde{T}} \leq m\right) = \mathbb{P}\left(L_T < m\right) = \mathbb{P}\left(S_m(T) < m\right), \text{ for } m \geq 1$$

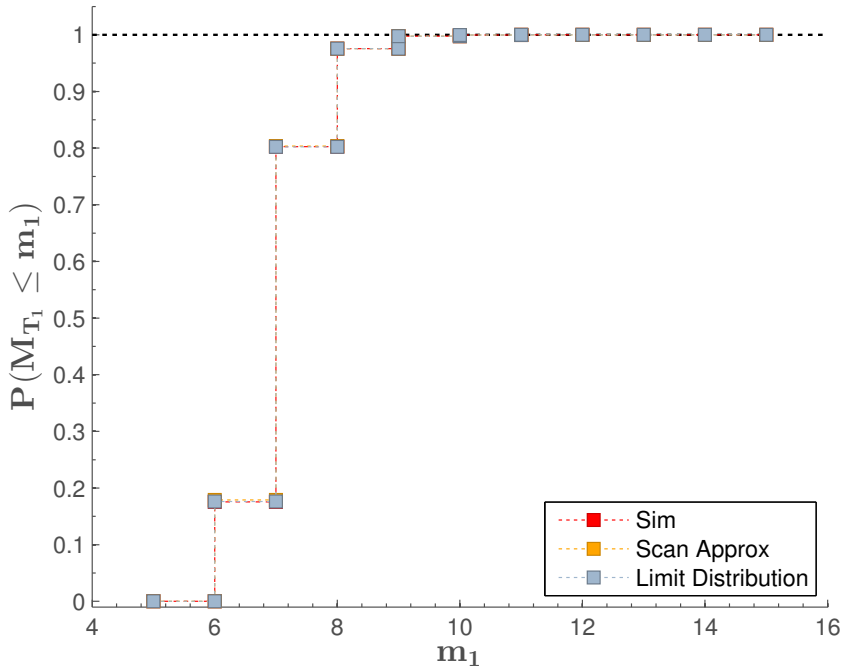
Longest increasing run: numerical results

For $Y_i \sim \mathcal{U}([0,1])$, [Novak, 1992] showed that

$$\max_{1 \le m \le T} \left| \mathbb{P}(M_T \le m) - e^{-T \frac{m+\mathbf{1}}{(m+\mathbf{2})!}} \right| = \mathcal{O}\left(\frac{\ln T}{T}\right)$$

m	Sim	AppH	$E_{total}(1)$	LimApp
5	0.00000700	0.00000733	0.14860299	0.00000676
6	0.17567262	0.17937645	0.01089628	0.17620431
7	0.80257424	0.80362353	0.00110990	0.80215088
8	0.97548510	0.97566460	0.00011579	0.97550345
9	0.99749821	0.99751049	0.00001114	0.99749792
10	0.99977074	0.99977183	0.00000098	0.99977038
11	0.99998075	0.99998083	0.00000008	0.99998073
12	0.99999851	0.99999851	0.00000001	0.99999851
13	0.99999989	0.99999989	0.00000000	0.99999989
14	0.99999999	0.99999999	0.00000000	0.99999999
15	1.00000000	1.00000000	0.00000000	1.00000000

We used $T = 10000$, $p = 0.1$ and $Iter = 10^5$.



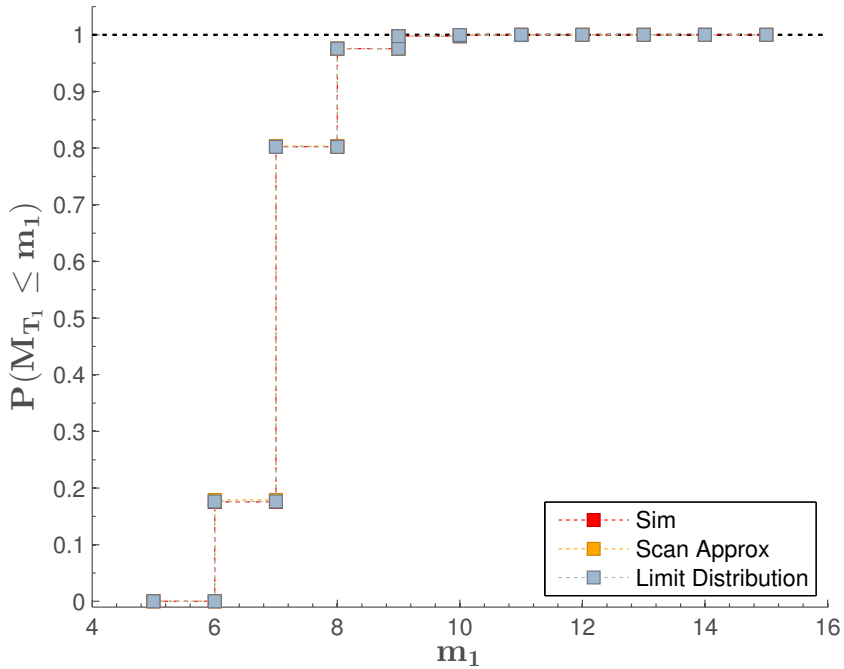
Longest increasing run: numerical results

For $Y_i \sim \mathcal{U}([0,1])$, [Novak, 1992] showed that

$$\max_{1 \le m \le T} \left| \mathbb{P}(M_T \le m) - e^{-T \frac{m+\mathbf{1}}{(m+\mathbf{2})!}} \right| = \mathcal{O}\left(\frac{\ln T}{T}\right)$$

m	Sim	AppH	$E_{total}(1)$	LimApp
5	0.00000700	0.00000733	0.14860299	0.00000676
6	0.17567262	0.17937645	0.01089628	0.17620431
7	0.80257424	0.80362353	0.00110990	0.80215088
8	0.97548510	0.97566460	0.00011579	0.97550345
9	0.99749821	0.99751049	0.00001114	0.99749792
10	0.99977074	0.99977183	0.00000098	0.99977038
11	0.99998075	0.99998083	0.00000008	0.99998073
12	0.99999851	0.99999851	0.00000001	0.99999851
13	0.99999989	0.99999989	0.00000000	0.99999989
14	0.99999999	0.99999999	0.00000000	0.99999999
15	1.00000000	1.00000000	0.00000000	1.00000000

We used $T = 10000$, $p = 0.1$ and $Iter = 10^5$.



Longest increasing run: numerical results

For $Y_i \sim Geom(p)$, [Louchard and Prodinger, 2003] showed that

$$\mathbb{P}(M_T \leq m) \sim \exp(-\exp \eta),$$

$$\eta = \frac{m(m+1)}{2} \log \frac{1}{1-p} + m \log \frac{1}{p} - \log T - \log p + \log D(m),$$

$$D(m) = \prod_{k=1}^{m} \left[1 - (1-p)^k\right] \left[1 - (1-p)^{m+2}\right]$$

m	Sim	AppH	$E_{total}(1)$	LimApp
6	0.56445934	0.56997462	0.00255592	0.56810748
7	0.95295406	0.95325180	0.00018554	0.95294598
8	0.99658057	0.99659071	0.00001214	0.99657969
9	0.99979460	0.99979550	0.00000068	0.99979435
10	0.99998950	0.99998950	0.00000003	0.99998947

We used $T = 10000$, $p = 0.1$ and $Iter = 10^5$.

Longest increasing run: numerical results

For $Y_i \sim Geom(p)$, [Louchard and Prodinger, 2003] showed that

$$\mathbb{P}(M_T \leq m) \sim \exp(-\exp \eta),$$

$$\eta = \frac{m(m+1)}{2} \log \frac{1}{1-p} + m \log \frac{1}{p} - \log T - \log p + \log D(m),$$

$$D(m) = \prod_{k=1}^{m} \left[1 - (1-p)^k\right] \left[1 - (1-p)^{m+2}\right]$$

m	Sim	AppH	$E_{total}(1)$	LimApp
6	0.56445934	0.56997462	0.00255592	0.56810748
7	0.95295406	0.95325180	0.00018554	0.95294598
8	0.99658057	0.99659071	0.00001214	0.99657969
9	0.99979460	0.99979550	0.00000068	0.99979435
10	0.99998950	0.99998950	0.00000003	0.99998947

We used $T = 10000$, $p = 0.1$ and $Iter = 10^5$.

Amărioarei, A. (2012).

Approximation for the distribution of extremes of one dependent stationary sequences of random variables.

arXiv:1211.5456v1, submitted.

Amărioarei, A. (2014).

Approximations for the multidimensional discrete scan statistics.

PhD thesis, University of Lille 1.

Chen, J. and Glaz, J. (1997).

Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials.

Advances in the Theory and Practice of Statistics, 1:285–298.

Csaki, E. and Foldes, A. (1996).

On the length of theh longest monnotone block.

Studio Scientiarum Mathematicarum Hungarica, 31:35–46.

Ebneshahrashoob, M. and Sobel, M. (1990).

Sooner and later waiting time problems for Bernoulli trials: frequency and run quotas.

Statist. Probab. Lett., 9:5–11.

Eryilmaz, S. (2006).

A note on runs of geometrically distributed random variables.

Discrete Mathematics, 306:1765–1770.

Fu, J. (2001).

Distribution of the scan statistic for a sequence of bistate trials.

J. Appl. Probab., 38:908–916.

Fu, J. C. and Lou, W. (2003).

Distribution theory of runs and patterns and its applications. A finite Markov chain imbedding approach.

World Scientific Publishing Co., Inc., River Edge, NJ.

Gao, T., Ebneshahrashoob, M., and Wu, M. (2005).

An efficient algorithm for exact distribution of discrete scan statistics.

Methodol. Comput. Appl. Probab., 7:1423–1436.

Glaz, J. (1990).

A comparison of product-type and Bonferroni-type inequalities in presence of dependence.

In *Symposium on Dependence in Probability and Statistics.*, volume 16 of *IMS Lecture Notes-Monograph Series*, pages 223–235. IMS Lecture Notes.

Glaz, J. and Naus, J. (1991).

Tight bounds and approximations for scan statistic probabilities for discrete data.

Annals of Applied Probability, 1:306–318.

Glaz, J., Naus, J., and Wallenstein, S. (2001).

Scan statistics.

Springer Series in Statistics. Springer-Verlag, New York.

Grabner, P., Knopfmacher, A., and Prodinger, H. (2003).

Combinatorics of geometrically distributed random variables: run statistics.

Theoret. Comput. Sci., 297:261–270.

Grill, K. (1987).

Erdos-Révész type bounds for the length of the longest run from a stationary mixing sequence.

Probab. Theory Relat. Fields, 75:169–179.

Karwe, V. and Naus, J. (1997).

New recursive methods for scan statistic probabilities.

Computational Statistics & Data Analysis, 17:389–402.

Louchard, G. and Prodinger, H. (2003).

Ascending runs of sequences of geometrically distributed random variables: a probabilistic analysis.
Theoret. Comput. Sci., 304:59–86.


Naus, J. (1974).
Probabilities for a generalized birthday problem.
Journal of American Statistical Association, 69:810–815.


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


Novak, S. (1992).
Longest runs in a sequence of m-dependent random variables.
Probab. Theory Relat. Fields, 91:269–281.


Pittel, B. (1981).
Limiting behavior of a process of runs.
Ann. Probab., 9:119–129.


Révész, P. (1983).
Three problems on the llength of increasing runs.
Stochastic Process. Appl., 5:169–179.

Wang, X. (2013).

Scan statistics for normal data.

PhD thesis, University of Connecticut.

Wang, X. and Glaz, J. (2013).

A variable window scan statistic for $MA(1)$ process.

In *Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA 2013)*, pages 905–912.

Wang, X., Glaz, J., and Naus, J. (2012).

Approximations and inequalities for moving sums.

Methodol. Comput. Appl. Probab., 14:597–616.

Wu, T.-L. (2013).

On finite Markov chain imbedding technique.

Methodol Comput Appl Probab, 15:453–465.

Product-type approximation and bounds $d = 1$

- Approximation

$$\mathbb{P}\left(S_{m_1}(T) \leq \tau\right) \approx Q(2m_1)\left[\frac{Q(3m_1)}{Q(2m_1)}\right]^{\frac{T}{m_1}-2},$$

- Lower Bounds

$$\begin{aligned}\mathbb{P}\left(S_{m_1}(T) \leq \tau\right) &\leq \frac{Q(2m_1)}{\left[1+\frac{Q(2m_1-1)-Q(2m_1)}{Q(2m_1-1)Q(2m_1)}\right]^{T-2m_1}}, \quad T \geq 2m_1 \\ &\leq \frac{Q(3m_1)}{\left[1+\frac{Q(2m_1-1)-Q(2m_1)}{Q(3m_1-1)}\right]^{T-3m_1}}, \quad T \geq 3m_1\end{aligned}$$

- Upper Bounds

$$\begin{aligned}\mathbb{P}\left(S_{m_1}(T) \leq \tau\right) &\leq Q(2m_1)\left[1-Q(2m_1-1)+Q(2m_1)\right]^{T-2m_1}, \quad T \geq 2m_1 \\ &\leq Q(3m_1)\left[1-Q(2m_1-1)+Q(2m_1)\right]^{T-3m_1}, \quad T \geq 3m_1\end{aligned}$$

The values $Q(2m_1-1), Q(2m_1), Q(3m_1-1), Q(3m_1)$ are computed using [Karwe and Naus, 1997] algorithm.

Return

Product-type approximation and bounds $d = 2$

- Approximation (Bernoulli)

$$\mathbb{P}(S_{m_1,m_2}(T, T_2) \leq k) \approx \frac{Q(m_1,m_2)^{(T-m_1-1)(T_2-m_2-1)} Q(m_1+1,m_2+1)^{(T-m_1)(T_2-m_2)}}{Q(m_1,m_2+1)^{(T-m_1-1)(T_2-m_2)} Q(m_1+1,m_2)^{(T-m_1)(T_2-m_2-1)}}$$

- Approximation (binomial and Poisson)

$$\mathbb{P}(S_{m_1,m_2}(T, T_2) \leq k) \approx \frac{Q(m_1+1,m_2+1)^{(T-m_1)(T_2-m_2)}}{Q(m_1+1,m_2)^{(T-m_1)(T_2-m_2-1)}} \times \frac{Q(m_1,2m_2-1)^{(T-m_1-1)(T_2-2m_2)}}{Q(m_1,2m_2)^{(T-m_1-1)(T_2-2m_2+1)}}$$

To compute the unknown variables we use

- $Q(m_1, 2m_2 - 1)$ and $Q(m_1, 2m_2)$ - adaptation of [Karwe and Naus, 1997] algorithm
- $Q(m_1 + 1, m_2)$ and $Q(m_1 + 1, m_2 + 1)$ - conditioning

Return

Approach

[Fu, 2001] applied the Markov Chain Imbedding Technique to find the distribution of binary scan statistics.

Main Idea

Express the distribution of the $S_{m_1}(T)$ in terms of the waiting time distribution of a special compound pattern

- define for $0 \leq k \leq m_1$

$$\mathcal{F}_{m_1,k} = \{\Lambda_i | \Lambda = \underbrace{1\ldots1}_{k}, \Lambda_2 = 10\underbrace{1\ldots1}_{k-1}, \ldots, \Lambda_l = \overbrace{\underbrace{1\ldots1}_{k-1}0\ldots01}^{m_1}\}$$

$$|\mathcal{F}_{m_1,k}| = \sum_{j=0}^{m_1-k} \binom{k-2+j}{j}$$

- the compound pattern $\Lambda = \cup_{i=1}^{l} \Lambda_i,\ \Lambda_i \in \mathcal{F}_{m_1,k}$

$$\mathbb{P}(S_{m_1}(T) < k) = \mathbb{P}(W(\Lambda) \geq T+1).$$

$$\mathbb{P}(S_{m_1}(T) < k) = \boldsymbol{\xi} \boldsymbol{N}^T \mathbf{1}^\intercal, \text{where } \boldsymbol{\xi} = (1, 0, \ldots, 0)$$

Return

Example

Consider the i.i.d. two-state sequence $(X_i)_{i\in\{1,2,\ldots,T\}}$ with $p = \mathbb{P}(X_1 = 1)$ and $q = \mathbb{P}(X_1 = 0)$.

- A realisation for $T = 20$

  00101011101101010110

- For $k = 3$ and $m_1 = 4$

  $\mathcal{F}_{4,3} = \{\Lambda = 111, \Lambda_2 = 1011, \Lambda_3 = 1101\}$

- The state space

  $\Omega = \{\emptyset, 0, 1, 10, 11, 101, 110, \alpha_1, \alpha_2, \alpha_3\}$

- the principal matrix:

$$\boldsymbol{N} = \begin{pmatrix} 0 & q & p & 0 & 0 & 0 & 0 \\ 0 & q & p & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & q & p & 0 & 0 \\ 0 & q & 0 & 0 & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & q \\ 0 & 0 & 0 & q & 0 & 0 & 0 \\ 0 & q & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Return

Selected Values for $K(\cdot)$ and $\Gamma(\cdot)$

Table 1: Selected values for $K(\cdot)$ and $\Gamma(\cdot)$

$1-q_1$	$K(1-q_1)$	$\Gamma(1-q_1)$
0.1	38.63	480.69
0.05	21.28	180.53
0.025	17.56	145.20
0.01	15.92	131.43

Return

Selected Values for $K(\cdot)$ and $\Gamma(\cdot)$

Table 1: Selected values for $K(\cdot)$ and $\Gamma(\cdot)$

$1-q_1$	$K(1-q_1)$	$\Gamma(1-q_1)$
0.1	38.63	480.69
0.05	21.28	180.53
0.025	17.56	145.20
0.01	15.92	131.43

Return