Approximations for two-dimensional discrete scan statistics in some dependent models

Alexandru Amărioarei
Cristian Preda

Laboratoire de Mathématiques Paul Painlevé
Département de Probabilités et Statistique
Université de Lille 1, INRIA Modal Team

Applied Stochastic Models and Data Analysis
25-28 June, 2013, Mataró (Barcelona), Spain

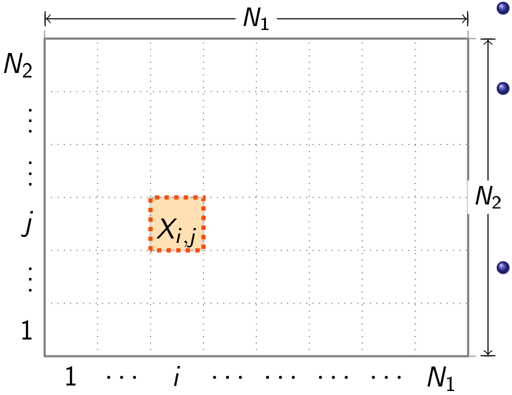
Outline

1. Introduction
   - Framework and Previous Work
   - Block-Factor Type Model
2. Description of the method
   - Main Idea and Tools
   - The Approximation
3. Error Bound
   - Approximation Error
   - Simulation Error
4. Illustrative Example
   - Description of the Example
5. References

Outline

1. Introduction
   - Framework and Previous Work
   - Block-Factor Type Model
2. Description of the method
   - Main Idea and Tools
   - The Approximation
3. Error Bound
   - Approximation Error
   - Simulation Error
4. Illustrative Example
   - Description of the Example
5. References

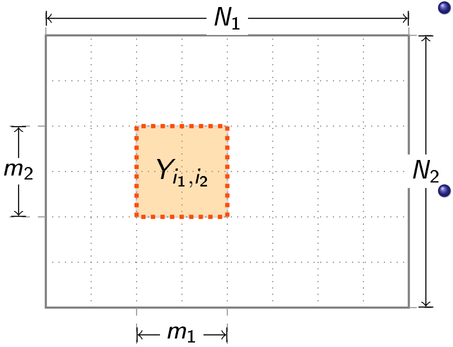
The Two-Dimensional Discrete Scan Statistic



Let N_1, N_2 be positive integers

- Rectangular region $\mathcal{R} = [0, N_1] \times [0, N_2]$
- $(X_{ij})_{\substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}}$ integer r.v.'s
  - Bernoulli($\mathcal{B}(1, p)$)
  - Binomial($\mathcal{B}(n, p)$)
  - Poisson($\mathcal{P}(\lambda)$)
- X_{ij} number of observed events in the elementary subregion $r_{ij} = [i-1, i] \times [j-1, j]$

The Two-Dimensional Discrete Scan Statistic



Let m_1, m_2 be positive integers

- Define for $1 \leq i_j \leq N_j - m_j + 1$,

$$Y_{i_1 i_2} = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} X_{ij}$$

- The two dimensional scan statistic,

$$S_{m_1,m_2}(N_1, N_2) = \max_{\substack{1 \leq i_1 \leq N_1 - m_1 + 1 \\ 1 \leq i_2 \leq N_2 - m_2 + 1}} Y_{i_1 i_2}$$

- Used for testing the null hypotheses of randomness against the alternative hypothesis of clustering

Problem and related results

Problem

Approximate the distribution of two dimensional discrete scan statistic

$$\mathbb{P}\left(S_{m_1,m_2}(N_1, N_2) \leq n\right).$$

- The i.i.d. model:
  - No exact formulas
  - For Bernoulli case:
    - product type approximations (Boutsikas and Koutras 2000)
    - Poisson approximations (Chen and Glaz 1996)
    - bounds (Boutsikas and Koutras 2003)
  - For binomial and Poisson cases: (Glaz 2009)
    - Product type approximation
    - Lower bound
  - Approximation and error bounds (Haiman 2006)
- Dependent model: **no results !**

Outline

1. Introduction
   - Framework and Previous Work
   - Block-Factor Type Model
2. Description of the method
   - Main Idea and Tools
   - The Approximation
3. Error Bound
   - Approximation Error
   - Simulation Error
4. Illustrative Example
   - Description of the Example
5. References

Introducing the Model

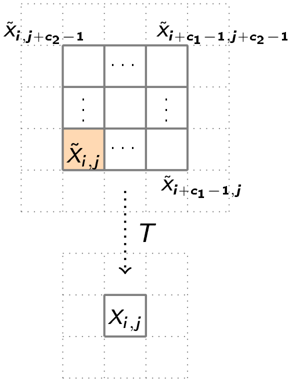
Let $1 \leq c_s \leq \tilde{N}_s$, $s \in \{1, 2\}$ integers

- $\left(\tilde{X}_{ij}\right)_{\substack{1 \leq i \leq \tilde{N}_1 \\ 1 \leq j \leq \tilde{N}_2}}$ i.i.d. r.v.'s
- configuration matrix in (i, j)

$$C_{(i,j)} = \left(C_{(i,j)}(k, l)\right)_{\substack{1 \leq k \leq c_2 \\ 1 \leq l \leq c_1}}$$

$$C_{(i,j)}(k, l) = \tilde{X}_{i+l-1, j+c_2-k}$$

- transformation $T : \mathcal{M}_{c_2, c_1}(\mathbb{R}) \to \mathbb{R}$

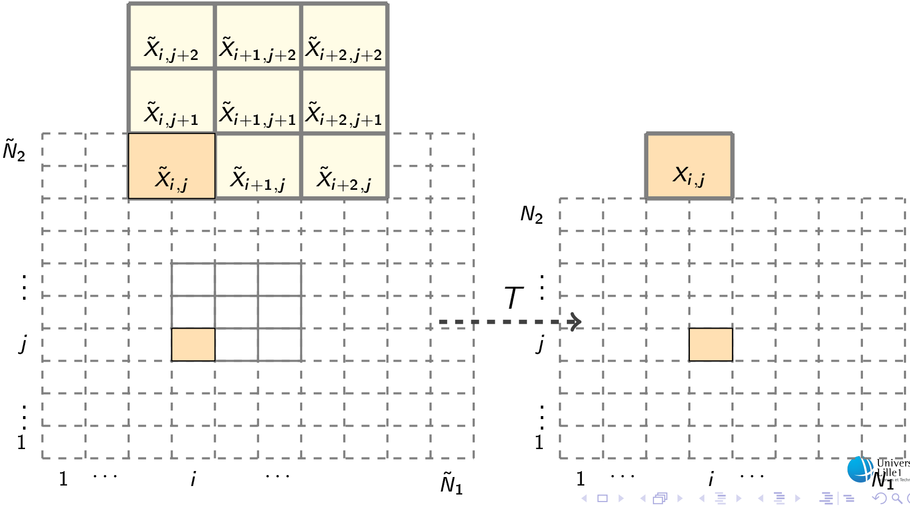


Define the block-factor model, $N_1 = \tilde{N}_1 - c_1 + 1$, $N_2 = \tilde{N}_2 - c_2 + 1$

$$X_{i,j} = T\left(C_{(i,j)}\right), \quad \substack{1 \leq i \leq N_1 \\ 1 \leq j \leq N_2}$$

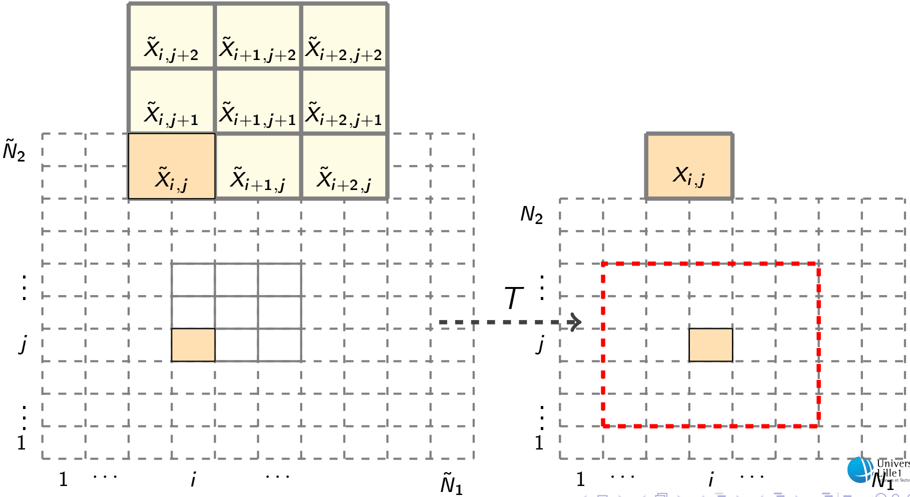
Model: case $c_1 = c_2 = 3$

- To simplify the presentation we consider $c_1 = c_2 = 3$

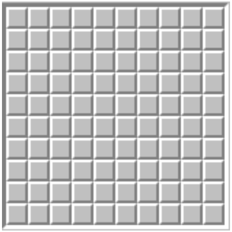


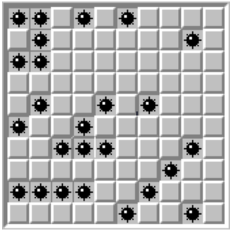
Model: case $c_1 = c_2 = 3$

- To simplify the presentation we consider $c_1 = c_2 = 3$



Example: A game of minesweeper





Model:

- $\tilde{X}_{i,j} \sim \mathcal{B}(p)$ (presence, absence of a mine)
- number of neighboring mines

$$T\left(C_{(i,j)}\right) = \sum_{\substack{(s,t)\in\{0,1,2\}^2 \\ (s,t)\neq(1,1)}} \tilde{X}_{i+s,j+t}$$

- $X_{i,j} = T\left(C_{(i,j)}\right)$

Outline

1. Introduction
   - Framework and Previous Work
   - Block-Factor Type Model
2. Description of the method
   - Main Idea and Tools
   - The Approximation
3. Error Bound
   - Approximation Error
   - Simulation Error
4. Illustrative Example
   - Description of the Example
5. References

Key Idea

Haiman(2000) proposed a different approach

Main Observation

The scan statistic r.v. can be viewed as a maximum of a sequence of 1-dependent stationary r.v..

- The idea:
  - discrete and continuous one dimensional scan statistic: Haiman (2000,2007)
  - discrete and continuous two dimensional scan statistic: Haiman and Preda (2002,2006)
  - discrete three dimensional scan statistic: Amarioarei (2013)

Writing the Scan as an Extreme of 1-Dependent R.V.'s

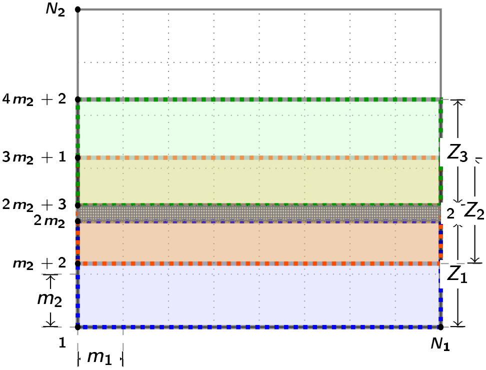
Let $N_j = (L_j + 1)(m_j + 1) - 2$, $j \in \{1, 2\}$ positive integers

- Define for $l \in \{1, 2, \dots, L_2\}$

$$Z_l = \max_{\substack{1 \le i_1 \le L_1(m_1+1) \\ (l-1)(m_2+1)+1 \le i_2 \le l(m_2+1)}} Y_{i_1 i_2}$$

- $(Z_l)_l$ is 1-dependent and stationary
- Observe

$$S_{m_1,m_2}(N_1, N_2) = \max_{1 \le l \le L_2} Z_l$$



Main Tool

Let $(Z_j)_{j\geq 1}$ be a strictly stationary 1-dependent sequence of r.v.'s and let $q_m = q_m(x) = \mathbb{P}(\max(Z_1, \ldots, Z_m) \leq x)$, with $x < \sup\{u|\mathbb{P}(Z_1 \leq u) < 1\}$.

Main Theorem (Haiman 1999, Amarioarei 2012)

For x such that $\mathbb{P}(Z_1 > x) = 1 - q_1 \leq \alpha < 0.1$ and $m > 3$ we have

$$\left|q_m - \frac{6(q_1 - q_2)^2 + 4q_3 - 3q_4}{(1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2)^m}\right| \leq \Delta_1(1 - q_1)^3,$$

$$\left|q_m - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^m}\right| \leq \Delta_2(1 - q_1)^2,$$

- $\Delta_1 = \Delta_1(\alpha, q_1, m) = \Gamma(\alpha) + mK(\alpha)$
- $\Delta_2 = mE(\alpha, q_1, m) = m\left[1 + \frac{3}{m} + K(\alpha)(1 - q_1) + \frac{\Gamma(\alpha)(1-q_1)}{m}\right].$

Selected values for $K(\alpha)$ and $\Gamma(\alpha)$

Université Lille1 Sciences et Technologies

Outline

1 Introduction
- Framework and Previous Work
- Block-Factor Type Model

2 Description of the method
- Main Idea and Tools
- The Approximation

3 Error Bound
- Approximation Error
- Simulation Error

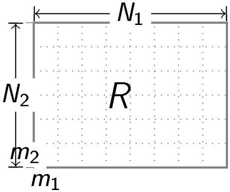
4 Illustrative Example
- Description of the Example

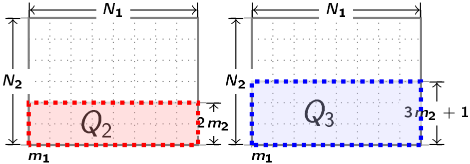
5 References

First Step Approximation

Using Main Theorem we obtain

- Define
$$Q_2 = \mathbb{P}(Z_1 \le k)$$
$$Q_3 = \mathbb{P}(Z_1 \le k, Z_2 \le k)$$
- If $1 - Q_2 \le \alpha_1 < 0.1$ the (first) approximation
$$\mathbb{P}(S \le k) \approx \frac{2Q_2 - Q_3}{[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{L_2}}$$
where $S = S_{m_1,m_2}(N_1, N_2)$
- Approximation error
$$L_2 E(\alpha_1, L_2)(1 - Q_2)^2$$





Second Step Approximation

Q_2 :

- For $s \in \{1, 2, \ldots, L_1\}$

$$Z_s^{(2)} = \max_{\substack{(s-1)(m_1+1)+1 \leq i_1 \leq s(m_1+1) \\ 1 \leq i_2 \leq m_2+1}} Y_{i_1 i_2}$$

- $$Q_2 = \mathbb{P}\left(\max_{1 \leq s \leq L_1} Z_s^{(2)} \leq k\right)$$

- Define

$$Q_{22} = \mathbb{P}(Z_1^{(2)} \leq k)$$

$$Q_{32} = \mathbb{P}(Z_1^{(2)} \leq k, Z_2^{(2)} \leq k)$$

- Approximation $(1 - Q_{22} \leq \alpha_2)$

$$Q_2 \approx \frac{2Q_{22} - Q_{32}}{\left[1 + Q_{22} - Q_{32} + 2(Q_{22} - Q_{32})^2\right]^{L_1}}$$

- Error

$$L_1 E(\alpha_2, L_1)(1 - Q_{22})^2$$

Q_3 :

- For $s \in \{1, 2, \ldots, L_1\}$

$$Z_s^{(3)} = \max_{\substack{(s-1)(m_1+1)+1 \leq i_1 \leq s(m_1+1) \\ 1 \leq i_2 \leq 2(m_2+1)}} Y_{i_1 i_2}$$

- $$Q_3 = \mathbb{P}\left(\max_{1 \leq l \leq L_1} Z_s^{(3)} \leq k\right)$$

- Define

$$Q_{23} = \mathbb{P}(Z_1^{(3)} \leq k)$$

$$Q_{33} = \mathbb{P}(Z_1^{(3)} \leq k, Z_2^{(3)} \leq k)$$

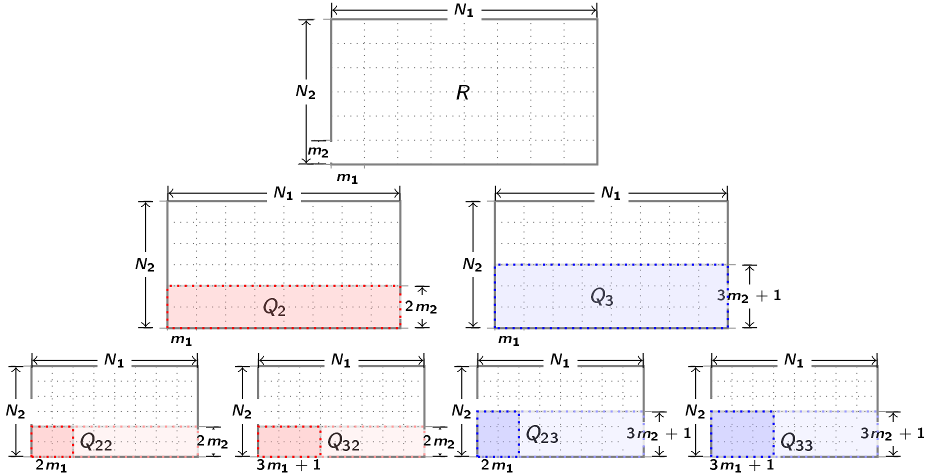
- Approximation $(1 - Q_{23} \leq \alpha_2)$

$$Q_3 \approx \frac{2Q_{23} - Q_{33}}{\left[1 + Q_{23} - Q_{33} + 2(Q_{23} - Q_{33})^2\right]^{L_1}}$$

- Error

$$L_1 E(\alpha_2, L_1)(1 - Q_{23})^2$$

Illustration of the Approximation Process



Outline

1. Introduction
   - Framework and Previous Work
   - Block-Factor Type Model
2. Description of the method
   - Main Idea and Tools
   - The Approximation
3. Error Bound
   - Approximation Error
   - Simulation Error
4. Illustrative Example
   - Description of the Example
5. References

Theoretical Approximation Error

Define for $s \in \{2, 3\}$

$$H(x, y, m) = \frac{2x - y}{[1 + x - y + 2(x - y)^2]^m}, \ \alpha_1 = 1 - Q_3, \ \alpha_2 = 1 - Q_{23},$$
$$E_1 = E(\alpha_2, L_1), \ E_2 = E(\alpha_1, L_2), \ R_s = H\left(Q_{2s}, Q_{3s}, L_1\right),$$

The approximation error

$$\boxed{E_{app} = L_2 F_2 B_2^2 + L_1 L_2 F_1 \left[(1 - Q_{22})^2 + (1 - Q_{23})^2\right]}$$

where B_2 is given by

$$\boxed{B_2 = 1 - R_2 + L_1 F_1 (1 - Q_{22})^2}$$

Outline

1. Introduction
   - Framework and Previous Work
   - Block-Factor Type Model
2. Description of the method
   - Main Idea and Tools
   - The Approximation
3. Error Bound
   - Approximation Error
   - Simulation Error
4. Illustrative Example
   - Description of the Example
5. References

Simulation Error for Approximation Formula

If *ITER* is the number of simulations, we can say, at 95% confidence level,

$$\left|Q_{rt} - \hat{Q}_{rt}\right| \leq 1.96\sqrt{\frac{\hat{Q}_{rt}(1-\hat{Q}_{rt})}{ITER}} = \beta_{rt}, \;\; r, t \in \{2, 3\}$$

where $\hat{Q}_{rt}$ is the simulated value.
Define for $r \in \{2, 3\}$,

$$\hat{Q}_r = H\left(\hat{Q}_{2r}, \hat{Q}_{3r}, L_1\right)$$

The simulation error corresponding to the approximation formula

$$E_{sf} = L_1 L_2 \left(\beta_{22} + \beta_{23} + \beta_{32} + \beta_{33}\right)$$

Simulation Error for Approximation Error

Introducing

$$C_{2r} = 1 - \hat{Q}_{2r} + \beta_{2r}, \quad r \in \{2,3\},$$
$$C_2 = 1 - \hat{Q}_2 + L_1(\beta_{22} + \beta_{32}) + L_1 F_1 C_{22}^2,$$

The simulation error corresponding to the approximation

$$E_{sapp} = L_2 F_2 C_2^2 + L_1 L_2 F_1 \left[C_{22}^2 + C_{23}^2\right]$$

The total error

$$E_{total} = E_{app} + E_{sf} + E_{sapp}$$

Outline

1. Introduction
   - Framework and Previous Work
   - Block-Factor Type Model
2. Description of the method
   - Main Idea and Tools
   - The Approximation
3. Error Bound
   - Approximation Error
   - Simulation Error
4. Illustrative Example
   - Description of the Example
5. References

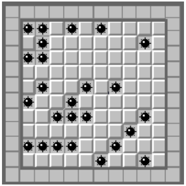
A Game of Minesweeper - Part 2

Recall the model:

- $\tilde{X}_{i,j} \sim \mathcal{B}(p)$ i.i.d. representing the absence/presence of a mine
- $X_{i,j}$ - number of neighboring mines corresponding to (i,j)

$$X_{i,j} = T\left(C_{(i,j)}\right) = \sum_{\substack{(s,t)\in\{0,1,2\}^2 \\ (s,t)\neq(1,1)}} \tilde{X}_{i+s,j+t}$$

$\tilde{X}_{i,j}$:



T

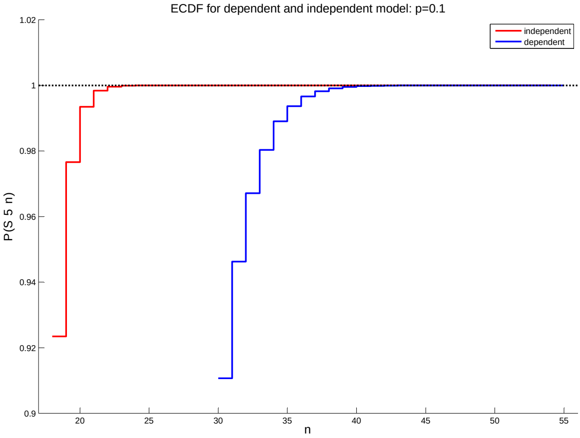
$X_{i,j}$:

Numerical Results for $\mathbb{P}(S_{m_1,m_2}(N_1, N_2) \leq n)$

Table 1 : $m_1 = 3, m_2 = 3, N_1 = 42, N_2 = 42, \mathbf{p} = \mathbf{0.1}, ITER = 10^7$

n	Sim Dep	Approx Dep	E_{app}	E_{sim}	E_{total}	Sim Indep	Approx Indep
30	0.88289	0.91068	0.00280	0.01489	0.01770	1	1
31	0.92769	0.94628	0.00085	0.00999	0.01084	1	1
32	0.95632	0.96713	0.00027	0.00725	0.00753	1	1
33	0.97356	0.98033	0.00009	0.00544	0.00553	1	1
34	0.98516	0.98909	0.00002	0.00396	0.00399	1	1
35	0.99161	0.99366	0.00000	0.00298	0.00299	1	1
36	0.99548	0.99663	0.00000	0.00216	0.00216	1	1
37	0.99760	0.99825	0.00000	0.00157	0.00157	1	1
38	0.99864	0.99911	0.00000	0.00110	0.00110	1	1
39	0.99926	0.99955	0.00000	0.00080	0.00080	1	1
40	0.99963	0.99978	0.00000	0.00056	0.00056	1	1
41	0.99987	0.99989	0.00000	0.00037	0.00037	1	1
42	0.99996	0.99994	0.00000	0.00023	0.00023	1	1
43	0.99998	0.99997	0.00000	0.00016	0.00016	1	1
44	0.99998	0.99999	0.00000	0.00009	0.00009	1	1
45	0.99999	0.99999	0.00000	0.00005	0.00005	1	1
46	0.99999	0.99999	0.00000	0.00003	0.00003	1	1

Graphical Illustration: $p = 0.1$

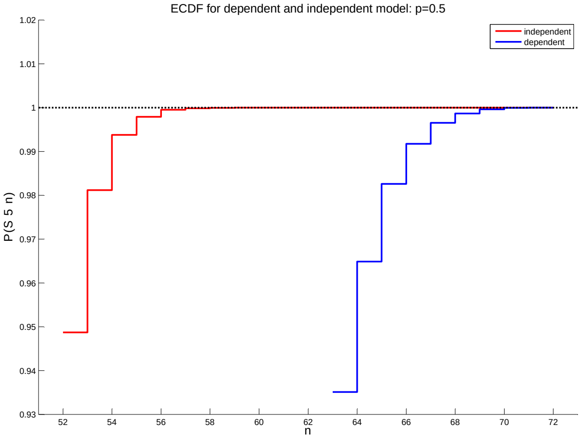


Numerical Results for $\mathbb{P}(S_{m_1,m_2}(N_1, N_2) \leq n)$

Table 2 : $m_1 = 3, m_2 = 3, N_1 = 42, N_2 = 42, \mathbf{p} = \mathbf{0.5}, ITER = 10^7$

n	*Sim Dep*	*Approx Dep*	E_{app}	E_{sim}	E_{total}	*Sim Indep*	*Approx Indep*
62	0.82484	0.88863	0.00487	0.01859	0.02346	1	1
63	0.89706	0.93509	0.00132	0.01139	0.01272	1	1
64	0.94327	0.96484	0.00032	0.00751	0.00784	1	1
65	0.97135	0.98256	0.00007	0.00510	0.00517	1	1
66	0.98668	0.99173	0.00001	0.00339	0.00340	1	1
67	0.99426	0.99650	0.00000	0.00222	0.00222	1	1
68	0.99796	0.99865	0.00000	0.00136	0.00136	1	1
69	0.99929	0.99958	0.00000	0.00077	0.00077	1	1
70	0.99979	0.99992	0.00000	0.00034	0.00034	1	1
71	0.99995	0.99998	0.00000	0.00017	0.00017	1	1
72	1	1	0.00000	0.00000	0.00000	1	1

Graphical Illustration: $p = 0.5$

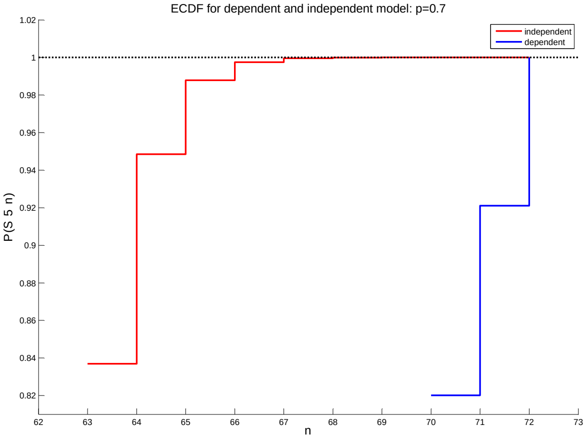


Numerical Results for $\mathbb{P}(S_{m_1,m_2}(N_1, N_2) \leq n)$

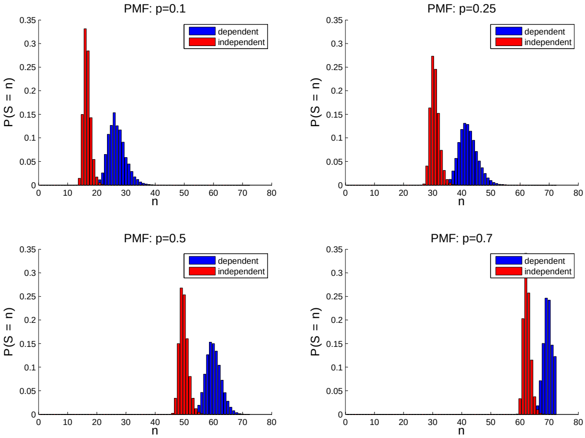
Table 3 : $m_1 = 3, m_2 = 3, N_1 = 42, N_2 = 42, \mathbf{p} = \mathbf{0.7}, ITER = 10^7$

n	*Sim Dep*	*Approx Dep*	E_{app}	E_{sim}	E_{total}	*Sim Indep*	*Approx Indep*
70	0.73026	0.82012	0.01490	0.03271	0.04761	1	0.99999
71	0.87721	0.92103	0.00194	0.01291	0.01485	1	1
72	1	1	0.00000	0.00000	0.00000	1	1

Graphical Illustration: $p = 0.7$



Dependence Effect



- Glaz, J., Naus, J., Wallenstein, S.: Scan statistic. *Springer* (2001).
- Glaz, J., Pozdnyakov, V., Wallenstein, S.: Scan statistic: Methods and Applications. *Birkhauser* (2009).
- Amarioarei, A.: *Approximation for the distribution of extremes of one dependent stationary sequences of random variables*, arXiv:1211.5456v1 (submitted)
- Amarioarei, A.: *Approximation for the Distribution of Three-dimensional Discrete Scan Statistic*, rXiv:1303.3775 (submitted)
- Boutsikas, M.V., Koutras, M.: Reliability approximations for Markov chain imbeddable systems. *Methodol Comput Appl Probab* **2** (2000), 393–412.
- Boutsikas, M. and Koutras, M. *Bounds for the distribution of two dimensional binary scan statistics*, Probability in the Engineering and Information Sciences, 17, 509–525, 2003.

- Chen, J. and Glaz, J. *Two-dimensional discrete scan statistics*, Statistics and Probability Letters 31, 59–68, 1996.
- Haiman, G.: First passage time for some stationary sequence. *Stoch Proc Appl* **80** (1999), 231–248.
- Haiman, G.: Estimating the distribution of scan statistics with high precision. *Extremes* **3** (2000), 349–361.
- Haiman, G., Preda, C.: A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process. *Methodol Comput Appl Probab* **4** (2002), 393–407.
- Haiman, G., Preda, C.: Estimation for the distribution of two-dimensional scan statistics. *Methodol Comput Appl Probab* **8** (2006), 373–381.
- Haiman, G.: Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *J. Stat Plan Infer* **137** (2007), 821–828.

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

α	$K(\alpha)$	$\Gamma(\alpha)$
0.1	38.63	480.69
0.05	21.28	180.53
0.025	17.56	145.20
0.01	15.92	131.43

Table 4 : Selected values for $K(\alpha)$ and $\Gamma(\alpha)$

Return