

Modelling claim exceedances over thresholds*

M.V. Boutsikas and M.V. Koutras

Department of Statistics and Insurance Science, University of Piraeus

Abstract

In this article we consider a simple risk model and study the occurrences of clusters of threshold exceedances by the individual claims. The statistic used to study the model is the discrete multiple scan statistic.

A compound Poisson approximation is established and certain asymptotic results are obtained for both the risk model and a similar in nature financial problem. Finally, we review two typical examples from areas of applied science where the outcomes of this article may have beneficial impact.

Keywords: risk model, claim exceedances, compound Poisson approximation, multiple scan statistic, discrete scan statistic, Poisson convergence, Weibull limit theorem.

1 Introduction

Let $Y_i, i = 1, 2, \dots$ be independent and identically distributed (i.i.d) random variables denoting the claim sizes in a specific portfolio, $u \geq 0$ a given threshold and

$$X_i = I_{(u, \infty)}(Y_i) = \begin{cases} 1 & \text{if } Y_i > u \\ 0 & \text{if } Y_i \leq u \end{cases}, i = 1, 2, \dots \quad (1)$$

the associated Bernoulli variables (indicating whether the i -th claim exceeds threshold u or not) with success probabilities $p = E(X_i) = P(Y_i > u)$. Let us assume that, due to delays in claim settlements, there are always k unprocessed claims. The event of having r or more threshold exceedances in the batch of unprocessed claims may signal a risky situation for the portfolio in the sense that a penalty could be placed to the company by the inspection authorities, or a reinsurance plan should be considered. Manifestly, the waiting time (number of claims) till the company experiences the m -th risky situation can be described as

$$V_m = \min\left\{n : \sum_{i=1}^{n-k+1} I_{[r, \infty)}(S_i) = m\right\}$$

where

$$S_i = \sum_{j=i}^{i+k-1} X_j, i = 1, 2, \dots$$

*Research supported by the Greek General Secretariat for Research and Technology (GRST) under grant PENED 99

is a k -scan process enumerating the number of exceedances of threshold u in a moving window of size k .

A random variable closely associated to V_m , is the number W_n of risky situations till the appearance of the n -th claim (n is a fixed integer), namely

$$W_n = W_{n,r,k,p} = \sum_{i=1}^{n-k+1} I_{[r,\infty)}(S_i).$$

Apparently, $P(W_n < m) = P(V_m > n)$, and the investigation of the distribution of either of the random variables W_n, V_n , would be quite useful for understanding the aforementioned model's behaviour and maintaining an effective policy for the company.

Another instance where the statistic W_n is of special importance arises when the null hypothesis of uniformity in claim exceedances is to be tested against the following alternative hypothesis of clustering: there exists at least one subsequence of k consecutive claims $Y_i, i = i_0, i_0 + 1, \dots, i_0 + k - 1$ with $P(Y_i > u) = p_0 > p$. As Glaz and Naus (1991) indicated, the generalized likelihood ratio test for checking the above hypothesis, rejects the null hypothesis of uniformity whenever $\max_{1 \leq i \leq n-k+1} S_i \geq r$, i.e. when $W_n \neq 0$, the value of r being determined from the significance level of the test. Since the exact size of the statistic W_n is expected to be more informative for detecting departures from the null hypothesis, it seems plausible to consider critical regions of the form $W_n \geq c$ and adjust r and c so that the desired significance level is attained. This approach offers increased flexibility in the test procedure, due to the presence of an additional parameter to play with.

It is of interest to note that

$$P\left(\max_{1 \leq i \leq n-k+1} S_i < r\right) = P(W_n = 0) = P(V_1 > n)$$

i.e., the probability mass function of W_n at 0 determines the cumulative distribution of the maximum number of successes in a fixed length window of size k . Theoretical developments related to the distribution of the statistic $\max_{1 \leq i \leq n-k+1} S_i$, which is known in the statistical literature as *discrete scan statistic* or *maximum generalized run*, can play an important role in the analysis of the so-called "maximum drawdown" problem in finance. More generally, the statistics W_n and V_m could be fruitfully exploited in the study of risk-related financial data (e.g. exchange rates or stock market prices) where the question of maximal loss (or gain) over a fixed length time window is of crucial importance. A discussion on this subject along with a list of applications in other disciplines is presented in Sections 4 and 5.

The statistic W_n , in the form of a random variable enumerating the (overlapping) moving windows of fixed length k which contain at least r successes in a prespecified number n of Bernoulli trials, has been recently introduced in the statistical literature under the name *multiple scan statistic*. Although quite accurate approximations are available by now for the probability mass function of W_n at 0 (for a review see Chen and Glaz (1999)) when the question comes to the whole distribution of W_n , the problem becomes extremely complex. Koutras and Alexandrou (1995) have described a method to obtain the exact distribution of W_n by invoking a Markov chain imbedding technique; however, this approach becomes unwieldy for k and r of moderate size while its computational complexity for large k, r and n renders the whole procedure as non-feasible. As indicated by Chen and Glaz (1997), it seems extremely difficult to establish product-type approximations

or bounds for $P(W_n \leq m), m \geq 1$; on the other hand, since windows with high concentration of successes tend to occur in clumps (that is, they appear in the form of local groups), one expects intuitively that ordinary Poisson approximations would naturally give poor results.

The last observation reveals that, a compound Poisson approximation for W_n 's distribution, may be more suitable than a standard Poisson approximation. This is the primary object of the present work which is organized as follows: In Section 2 we introduce all necessary notation along with a general result on compound Poisson approximation which will be used for the investigation of the distribution of W_n . In Section 3, an upper bound is derived for the (Kolmogorov) distance between the distribution of W_n and an appropriately selected compound Poisson distribution, and the asymptotic behaviour of W_n as $n \rightarrow \infty$ is examined. Extensive numerical experimentation was also carried out to investigate the quality of the approximation and bounds. In Section 4, motivated by a financial problem, we establish an additional asymptotic theorem, and apply it in two special cases, which lead to appealing Weibull and Erlang convergence results. Finally, in Section 5, two attractive applications from the areas of Reliability and DNA sequencing are described in some detail.

2 Preliminaries

The Kolmogorov distance between the distributions of two random variables X and Y is defined as

$$d(X, Y) = \sup_w |P(X \leq w) - P(Y \leq w)|$$

and offers a very efficient tool for establishing convergence in distribution; a sequence of random variables converges to Y if the corresponding sequence of distances converges to 0. The following two elementary properties of the Kolmogorov distance will be proved useful in the sequel

$$d(X, Z) \leq d(X, Y) + d(Y, Z) \tag{2}$$

$$d(X, Y) \leq \max\{P(X < Y), P(Y < X)\} \leq P(X \neq Y). \tag{3}$$

By the term compound Poisson distribution with parameter λ and compounding distribution F , we shall refer to the distribution of a random sum of the form $\sum_{i=1}^N Z_i$ with N being a Poisson random variable with $\lambda = E(N)$ and Z_i being i.i.d random variables (also independent of N) whose distribution function is F .

The main result of the next section, which consists the milestone for the investigation of the claim exceedances model introduced in this article, is an application of a general theorem on compound Poisson approximation published recently by Boutsikas and Koutras (2001). For the purposes of the present exposition, we shall retain a simplified version of their result which is more than adequate to meet our needs.

Consider first a sequence of non-negative random variables $Z_a, a = 1, 2, \dots$. For each $a = 2, 3, \dots$ introduce a subset B_a of $\{1, 2, \dots, a - 1\}$ (left neighborhood of dependence of Z_a) so that Z_a is independent of all $Z_b, b \in \{1, 2, \dots, a - 1\} \setminus B_a$. The next theorem provides an upper bound for the Kolmogorov distance between the distribution of the sum $\sum_{a=1}^{\nu} Z_a$ (ν a fixed positive integer) and a compound Poisson distribution $CP(\lambda, F)$ with suitably chosen λ, F .

Theorem 1 (Boutsikas and Koutras (2001)). If $Z_a, a = 1, 2, \dots, n$ is a sequence of non-negative random variables, then

$$d\left(\sum_{a=1}^{\nu} Z_a, CP(\lambda, F)\right) \leq \sum_{a=2}^{\nu} \left(P(Z_a > 0, \sum_{b \in B_a} Z_b > 0) + P(Z_a > 0)P(\sum_{b \in B_a} Z_b > 0) \right) + \frac{1}{2} \sum_{i=1}^{\nu} P(Z_i > 0)^2 \quad (4)$$

where $\lambda = \sum_{a=1}^{\nu} \lambda_a$, and $F(x) = \frac{1}{\lambda} \sum_{a=1}^{\nu} \lambda_a P(Z_a \leq x | Z_a > 0)$, $x \in \mathbf{R}$, $\lambda_a = P(Z_a > 0)$.

In common language, Theorem 1 states that, if the random variables $Z_a, a = 1, 2, \dots$ are "weakly" dependent and the masses of their distributions are concentrated on 0, then $\sum_{a=1}^{\nu} Z_a$ can be satisfactorily approximated by an appropriate compound Poisson distribution.

In the next sections, the standard notations $\sim, o(\cdot), O(\cdot)$ will assume their usual meaning and $I_A(\cdot)$ will denote the indicator function of the set A , i.e.

$$f(t) \sim g(t) \text{ as } t \rightarrow t_0 \text{ if } \lim_{t \rightarrow t_0} \frac{f(t)}{g(t)} = 1, \quad f(t) = o(g(t)) \text{ as } t \rightarrow t_0 \text{ if } \lim_{t \rightarrow t_0} \frac{f(t)}{g(t)} = 0,$$

$$f(t) = O(g(t)) \text{ if } \frac{f(t)}{g(t)} \text{ is bounded, } I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

In addition, summations of the form $\sum_{i=a}^b x_i$ with $a > b$, will be assumed to vanish.

3 The compound Poisson approximation

In this section we assume that $X_i, i = 1, 2, \dots, n$ is a sequence of i.i.d binary random variables with fixed success probabilities $p = P(X_i = 1) = 1 - P(X_i = 0) = 1 - q$, $i = 1, 2, \dots, n$. Our primary objective is to approximate the distribution of the multiple scan statistic

$$W_n = \sum_{i=1}^{n-k+1} I_{[r, \infty)}(S_i) = \sum_{i=1}^{n-k+1} I_{[r, \infty)}\left(\sum_{j=i}^{i+k-1} X_j\right) \quad (5)$$

by an appropriate compound Poisson distribution, and establish asymptotic results for large sequences ($n \rightarrow \infty$) and small success probabilities p .

As already mentioned in the introduction, the need to resort to compound Poisson distribution (instead of ordinary Poisson) arises from the fact that, the windows of size k containing at least r successes (i.e. generalized runs), tend to occur in clumps; hence, should a specific window include at least r success, it is highly probable that the neighboring windows will share the same property as well. For example, suppose that $r = 3, k = 4$ and that a generalized run begins at position $i = 7$ as indicated in the next realization:

$$\begin{array}{l} X_i: \quad \mathbf{001000101110100000100101} \dots \\ i: \quad \quad \quad 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12 \ 13 \ 14 \ 15 \ 16 \ 17 \ 18 \ 19 \ 20 \ 21 \ 22 \ 23 \ 24 \ \dots \end{array}$$

Then, with probability p there will also be a generalized run beginning at position $i + 1 = 8$, with probability $2p - p^2$ a generalized run beginning at position $i + 2 = 9$ and so forth. Note that, in this case, the unconditional probability that a generalized run occurs at any given position equals $4p^3 - 3p^4$, which is much smaller than p and $2p - p^2$ for $p \rightarrow 0$.

The next theorem exploits an efficient declumping technique and the outcome of Theorem 1 to establish a compound Poisson approximation for W_n and develop an upper bound for the error (in terms of Kolmogorov distance) incurred by it.

Theorem 2 Let $\lambda = (n - k + 1) \binom{k-1}{r-1} p^r q^{k-r+1}$, and F be the distribution function of a discrete random variable on $\{0, 1, \dots, k\}$, defined by

$$F(x) = 1 - \sum_{j=\max\{0, r-x-1\}}^{\min\{k-x-1, r-1\}} \frac{\binom{x}{x-r+j+1} \binom{k-x-1}{j}}{\binom{k-1}{r-1}} \quad (6)$$

$$\times \left(\binom{x}{x-r+j+1} q^{x-r+j+1} p^{r-j-1} + \left(1 - \frac{(x+1)q}{x-r+j+2}\right) \sum_{i=0}^{x-r+j} \binom{x}{i} q^i p^{x-i-1} \right)$$

for $x = 1, \dots, k-1$, $F(0) = 0$, $F(k) = 1$. Then

$$d(W_n, CP(\lambda, F)) \leq (\lambda + 1) \sum_{i=r}^k \binom{k}{i} p^i q^{k-i} + (\lambda(3k-1) + k-1) \binom{k-1}{r-1} p^r q^{k-r+1}$$

$$+ (n-k) \sum_{b=2}^{k-1} \sum_{i=\max\{0, r-k+b-1\}}^{\min\{r-2, b-2\}} \binom{k-b}{r-i-1} \binom{b-2}{i} \binom{k-b}{r-i-2} \quad (7)$$

$$\times p^{2r-i-1} q^{2k-b-2r+i+3} \equiv UB$$

Proof. A standard declumping procedure (see e.g. Barbour *et al.* (1992)) for W_n is offered by expressing it as $W_n = \sum_{a=1}^{n-k+1} C_a^*$ where

$$C_1^* = \sum_{j=1}^{n-k+1} \prod_{m=1}^j I_{[r, \infty)}(S_m),$$

$$C_a^* = (1 - I_{[r, \infty)}(S_{a-1})) \sum_{j=1}^{n-a-k+2} \prod_{m=a}^{a+j-1} I_{[r, \infty)}(S_m), \quad a = 2, 3, \dots, n-k+1$$

enumerate the size of a clump (consecutive windows with at least r successes in each one) starting at position $a = 1, 2, \dots, n-k+1$. Nevertheless, the declumping technique to follow appears to be preferable on grounds of simplicity, since a more easily calculated bound for $d(W_n, CP(\lambda, F))$ pops up if this technique is launched instead of the standard one.

Let us first expand the original sequence of i.i.d variables X_i to the whole set of integers by assuming that

$$p = P(X_i = 1) = 1 - P(X_i = 0) = 1 - q, \quad i \in \mathbf{Z}.$$

Define next the "truncated" declumping variables

$$C_a = (1 - I_{[r, \infty)}(S_{a-1})) \sum_{j=1}^k \prod_{m=a}^{a+j-1} I_{[r, \infty)}(S_m), \quad a = 1, 2, \dots, n-k+1$$

and observe that the random variables W_n and $\sum_{a=1}^{n-k+1} C_a$ differ only on the event of experiencing one of the following three outcomes:

i) A clump of at least $k + 1$ consecutive windows of length k (containing at least r successes each) has started at the first $n - 2k + 1$ trials. In this case we have $W_n > \sum_{a=1}^{n-k+1} C_a$ and the corresponding probability is upper bounded by

$$\sum_{i=1}^{n-2k+1} P(S_{i-1} < r, S_j \geq r, j = i, \dots, i+k) \leq \sum_{i=1}^{n-2k+1} P(S_{i-1} < r, S_i \geq r)P(S_{i+k} \geq r)$$

ii) A clump of length greater than $n - k + 2 - a$ has commenced at trial a , for $a = n - 2k + 3, \dots, n - k + 1$. In this case $W_n < \sum_{a=1}^{n-k+1} C_a$ while an upper bound for the overall probability associated with these situations is given by

$$\sum_{i=n-2k+3}^{n-k+1} P(S_{i-1} < r, S_j \geq r, j = i, \dots, n - k + 2) \leq \sum_{i=n-2k+3}^{n-k+1} P(S_{i-1} < r, S_i \geq r)$$

iii) A clump has started before trial 1 and was extended at least till the window starting at trial 1. In this case we have $W_n > \sum_{a=1}^{n-k+1} C_a$ with the associated probability given by

$$P(S_0 \geq r, S_1 \geq r) \leq P(S_1 \geq r).$$

The foregone analysis may be used in conjunction with inequality (3) to write

$$\begin{aligned} d(W_n, \sum_{a=1}^{n-k+1} C_a) &\leq P(W_n \neq \sum_{a=1}^{n-k+1} C_a) \\ &\leq ((n - 2k + 1)P(S_1 \geq r) + k - 1)P(C_1 > 0) + P(S_1 \geq r). \end{aligned} \quad (8)$$

Let us next apply Theorem 1 for the declumping random variables $C_a, a = 1, 2, \dots, n - k + 1$. On choosing the left neighborhoods of dependence as $B_a = \{\max\{1, a - 2k + 1\}, \dots, a - 1\}, a = 2, 3, \dots, n - k + 1$ it can be easily verified that C_a is independent of $C_b, b \in \{1, 2, \dots, a - 1\} \setminus B_a$ and since all $C_a, a = 1, 2, \dots$ are identically distributed we conclude that $\lambda_a = P(C_a > 0) = P(C_1 > 0)$ for every $a = 1, 2, \dots, n - k + 1$. Hence,

$$\lambda = (n - k + 1)P(C_1 > 0), \quad F(x) = P(C_1 \leq x | C_1 > 0) = 1 - \frac{P(C_1 \geq x + 1)}{P(C_1 > 0)}, \quad x = 0, 1, \dots, k \quad (9)$$

and inequality (4) takes on the form

$$\begin{aligned} d\left(\sum_{a=1}^{n-k+1} C_a, CP(\lambda, F)\right) &\leq \sum_{a=2}^{n-k+1} \sum_{b=\max\{1, a-2k+1\}}^{a-1} (P(C_a > 0, C_b > 0) + P(C_a > 0)P(C_b > 0)) \\ &\quad + \frac{1}{2}(n - k + 1)P(C_1 > 0)^2 \\ &\leq (n - k) \sum_{b=1}^{2k-2} P(C_{2k} > 0, C_b > 0) + 2k(n - k + 1)P(C_1 > 0)^2. \end{aligned} \quad (10)$$

By virtue of the triangle inequality (2) we may write

$$d(W_n, CP(\lambda, F)) \leq d(W_n, \sum_{a=1}^{n-k+1} C_a) + d(\sum_{a=1}^{n-k+1} C_a, CP(\lambda, F))$$

and an upper bound for $d(W_n, CP(\lambda, F))$ is obtained by summing up the right hand sides (RHS) of inequalities (8) and (10).

Therefore, in order to complete the proof of the theorem, we need to derive explicit expressions for the quantities $P(C_1 \geq c)$, $c = 1, 2, \dots, k$ and $\sum_{b=1}^{2k-2} P(C_{2k} > 0, C_b > 0)$. Elementary arguments reveal that, for $c = 1, 2, \dots, k$,

$$\begin{aligned} & P(C_1 \geq c) \\ &= P(S_0 < r, S_m \geq r, m = 1, 2, \dots, c) = P\left(\sum_{i=0}^{k-1} X_i < r, \sum_{i=m}^{m+k-1} X_i \geq r, m = 1, 2, \dots, c\right) \\ &= \sum_{x=0}^{k-c} P\left(\sum_{i=0}^{c-1} X_i < r-x, \sum_{i=m}^{c-1} X_i + \sum_{i=k}^{m+k-1} X_i \geq r-x, m = 1, 2, \dots, c\right) P\left(\sum_{i=c}^{k-1} X_i = x\right) \\ &= \sum_{x=0}^{k-c} P\left(\sum_{i=0}^{c-1} (1-X_i) \geq c-r+x+1, \sum_{i=m}^{m+c-1} (1-X_i) < c-r+x+1, m = 1, 2, \dots, c\right) \\ &\quad \times P\left(\sum_{i=c}^{k-1} X_i = x\right) \\ &= \sum_{x=0}^{k-c} f_{c-r+x+1, c, q}(2c) \binom{k-c}{x} p^x q^{k-c-x} \end{aligned}$$

where $f_{s, c, q}(\cdot)$ denotes the probability mass function of the waiting time until the sum of the $1-X_i$'s in a window of length c is as large as s . Replacing $f_{c-r+x+1, c, q}(2c)$ by the aid of Glaz and Naus (1991) formula

$$f_{s, c, q}(2c) = \frac{q}{s} \binom{c-1}{s-1} q^{s-1} p^{c-s} (sp \binom{c-1}{s-1} q^{s-1} p^{c-s} + (s-cq) \sum_{i=0}^{s-2} \binom{c-1}{i} q^i p^{c-1-i}), 1 \leq s \leq c$$

($f_{s, c, q}(2c) = 0$ if $s > c$ or $s < 1$), and carrying out straightforward algebraic manipulations, the next expression for $P(C_1 \geq c)$, $c = 1, 2, \dots, k$ results

$$\begin{aligned} P(C_1 \geq c) &= p^{r-1} q^{k-r+1} \sum_{x=\max\{0, r-c\}}^{\min\{k-c, r-1\}} \binom{c-1}{c-r+x} \binom{k-c}{x} \\ &\quad \times \left[\binom{c-1}{c-r+x} q^{c-r+x} p^{r-x} + \left(1 - \frac{cq}{c-r+x+1}\right) \sum_{i=0}^{c-r+x-1} \binom{c-1}{i} q^i p^{c-1-i} \right]. \end{aligned} \quad (11)$$

The desired formulae for λ and $F(x)$ are direct consequences of formulae (9) and (11) (note also that $P(C_1 \geq 0) = 1$, $P(C_1 \geq k+1) = 0$ and $P(C_1 > 0) = \binom{k-1}{r-1} p^r q^{k-r+1}$).

It remains to identify an explicit expression for the sum appearing in the RHS of (10). Observe first that

$$\sum_{b=1}^{k-1} P(C_b > 0, C_{2k} > 0) = \sum_{b=1}^{k-1} P(S_{b-1} < r, S_b \geq r) P(S_{2k-1} < r, S_{2k} \geq r) = (k-1)P(C_1 > 0)^2$$

and argue in a way similar to the one employed for the analysis of $P(C_1 \geq c)$, to prove that

$$\begin{aligned} & \sum_{b=k}^{2k-2} P(C_b > 0, C_{2k} > 0) \\ = & \sum_{b=1}^{k-1} P(C_b > 0, C_{k+1} > 0) = \sum_{b=1}^{k-1} P(S_{b-1} < r, S_b \geq r, S_k < r, S_{k+1} \geq r) \\ = & \sum_{b=1}^{k-1} P(X_{b-1} = 0, \sum_{j=b}^{b+k-2} X_j = r-1, X_{b+k-1} = 1, X_k = 0, \sum_{j=k+1}^{2k-1} X_j = r-1, X_{2k} = 1) \\ = & \sum_{b=2}^{k-1} \sum_{i=0}^{\min\{r-1, b-2\}} P(X_{b-1} = 0, \sum_{j=b}^{k-1} X_j = r-i-1, X_k = 0, \sum_{j=k+1}^{k+b-2} X_j = i, X_{b+k-1} = 1, \\ & , \sum_{j=k+b}^{2k-1} X_j = r-i-2, X_{2k} = 1) \\ = & \sum_{b=2}^{k-1} \sum_{i=\max\{0, r-k+b-1\}}^{\min\{r-2, b-2\}} \binom{k-b}{r-i-1} \binom{b-2}{i} \binom{k-b}{r-i-2} p^{2r-i-1} q^{2k-b-2r+i+3}. \end{aligned}$$

This concludes the proof. ■

It is worth mentioning that for $r = k$ Theorem 2 implies that

$$d(W_n, CP(\lambda, F)) \leq (\lambda + 1)p^k + (\lambda(3k-1) + k-2)p^k q$$

where $\lambda = (n-k+1)p^k q$, and $F(x) = 1 - p^x$ for $x = 1, \dots, k-1$, $F(0) = 0$, $F(k) = 1$. This offers an upper bound for the Kolmogorov distance between the distribution of the number of overlapping success runs of length k in a sequence of n Bernoulli trials and a compound Poisson with geometric compounding distribution (Polya-Aeppli distribution). For comparable bounds established by the aid of the celebrated Stein-Chen method (for both Poisson and compound Poisson approximations) the interested reader might consult the excellent monograph by Barbour *et al.* (1992).

An immediate consequence of Theorem 2 is the following asymptotic result which describes W_n 's weak convergence to the compound Poisson distribution as $n \rightarrow \infty, p \rightarrow 0$ with k, r kept fixed.

Theorem 3 *Assume that k, r are kept fixed while $n \rightarrow \infty, p \rightarrow 0$ so that $\lambda_n = (n-k+1) \binom{k-1}{r-1} p^r q^{k-r+1} \rightarrow \lambda \in (0, \infty)$. Then W_n converges weakly to a compound Poisson distribution with parameter λ and compounding distribution*

$$F(x) = 1 - \frac{\binom{k-1-x}{r-1}}{\binom{k-1}{r-1}}, \quad x = 1, 2, \dots, k-r, \quad F(0) = 0, F(k-r+1) = 1.$$

Proof. Under the assumptions made, it can be easily verified that, for the upper bound UB of Theorem 2, we have

$$UB \sim (\lambda + 1) \binom{k}{r} p^r + (\lambda(3k - 1) + k - 1) \binom{k-1}{r-1} p^r + \lambda \sum_{b=r}^{k-1} \frac{(k-b) \binom{b-2}{r-2}}{\binom{k-1}{r-1}} p$$

with the last term vanishing for $r = k$. Hence UB is asymptotically equal to

$$(3\lambda + 1) k p^k = O(p^k)$$

for $r = k$ while

$$UB \sim \lambda \sum_{b=r}^{k-1} \frac{(k-b) \binom{b-2}{r-2}}{\binom{k-1}{r-1}} p = O(p)$$

for $r < k$. In both cases, the upper bound UB converges to 0 as $p \rightarrow 0$ and the proof is completed by observing that

$$\lim_{p \rightarrow 0} F(x) = 1 - \frac{\binom{k-x-1}{r-1}}{\binom{k-1}{r-1}}, \quad x = 1, 2, \dots, k-r, \quad \lim_{p \rightarrow 0} F(k-r+1) = 1$$

■

Employing the Pascal formula for binomial coefficients, it follows that the probability mass function corresponding to $F(x)$ admits the expression

$$f(x) = \frac{\binom{k-x-1}{r-2}}{\binom{k-1}{r-1}}, \quad x = 1, 2, \dots, k-r+1. \quad (12)$$

It is of interest to note that, in the special case $r = 2 \leq k$, $f(x)$ reduces to the discrete uniform distribution on the integers $1, 2, \dots, k-1$. Also, for $k = r$, the compounding distribution becomes degenerate (with all its mass being concentrated at 1) and the limiting compound Poisson law $CP(\lambda, F)$ reduces to an ordinary Poisson distribution.

For the benefit of the practical minded reader, we mention that, the numerical calculation of the asymptotic probability mass function $f_W(x) = \lim_{n \rightarrow \infty} P(W_n = x)$ could be performed by launching the following efficient recursive scheme

$$\begin{aligned} f_W(0) &= e^{-\lambda}, \\ f_W(x) &= \frac{\lambda k}{rx} \binom{k}{r} \sum_{i=1}^x i \binom{k-i-1}{r-2} f_W(x-i), \quad x = 1, 2, \dots \end{aligned} \quad (13)$$

(see e.g. Bowers *et al.* (1997), Theorem 12.4.3).

Chen and Glaz (1999), following the approach in Roos (1993) and Glaz *et al.* (1994), suggested several compound Poisson approximations for the distribution of W_n . The starting point for their approach was an approximation of the probability $P(W_n = 0)$ by an expression of the form

$$P(W_n = 0) = 1 - P(W_n \geq 1) \approx \exp \left(- \sum_{i \geq 1} \lambda_i^* \right) \quad (14)$$

with the parameters λ_i^* being appropriately chosen (such an approximation can be demonstrated rigorously by theoretical arguments contained in Barbour *et al.* (1992) and Roos (1993), (1994)). Then, they suggested extending this formula to the whole distribution of W_n , by considering approximations of the form

$$P(W_n \geq x) \approx 1 - \sum_{j=0}^{x-1} \left(\sum_{\sum_{i=1}^{2k-1} i\beta_i = j} \prod_{i=1}^{2k-1} \frac{(\lambda_i^*)^{\beta_i}}{\beta_i!} \right) \exp \left(- \sum_{i=1}^{2k-1} \lambda_i^* \right) \quad (15)$$

with λ_i^* being exactly the same as the ones engaged in formula (14). As indicated by Chen and Glaz (1999), if the interest is focused on $P(W_n > 0) = 1 - P(W_n = 0)$, product-type approximations are preferable (as compared to compound Poisson approximations) on grounds both of accuracy and simplicity. However, when the question comes to the evaluation of the whole distribution, i.e. the quantities $P(W_n > x), x = 1, 2, \dots$ product-type approximations are extremely complex (if feasible); this remark seems to have led Chen and Glaz to coin the "heuristic" compound Poisson approximation (15), which however has the disadvantage that no error estimates are available for achieving an assessment of the maximum discrepancy between true and approximate values. This deficiency can be resolved by the use of Theorem 2. In addition, extensive numerical experimentation revealed that, the approximation suggested by Theorem 2, performs better than the approximations provided by (15). In Table 1, numerical results are presented for three of the four approximations tabulated by Chen and Glaz (1999) (the fourth one was omitted, since its performance is very poor): CG1, CG2, CG3 refer to approximations of the form (15) with λ_i^* defined respectively by (2.32), (2.34), (2.36)-(2.38) therein. Here, the same choice of values of the parameters was used as in Chen and Glaz (1999), while the simulated estimates were recalculated in order to arrive at more accurate results (10^6 iterations were performed for each entry). The column labeled as "CP(λ, F)" contains the values of the approximation described in Theorem 2, and the last column the uniform upper bounds (7); note that, in the first block, the upper bound UB is useless, while in the rest of them it could be profitably used to gain interval estimates for the survival function (or cumulative distribution function) of W_n . Finally, the last row in each block offers the mean relative error observed in each approximation, namely

$$MRE = \frac{1}{5} \sum_{x=1}^5 \frac{|P_{sim}(W_n \geq x) - P_{appr}(W_n \geq x)|}{P_{sim}(W_n \geq x)}.$$

A graphical illustration of the numerical results detailed above is provided in Figure 1.

In order to investigate the convergence of the aforementioned approximations to the asymptotic distribution described in Theorem 3, we prepared Tables 2 and 3 where k, r and p were assigned specific values ($k = 10, r = 2, 3$ and $p = 10^{-i}, i = 1, 2, 3, 4$) while the value of n was computed by the formula

$$n = \left\lceil \lambda_0 \left(\binom{k-1}{r-1} p^r (1-p)^{k-r+1} \right)^{-1} \right\rceil + k - 1$$

so as the convergence condition $\lambda_n \rightarrow \lambda_0$ is acquired. From the tabulated numerical results, it is clear that, the approximation provided by Theorem 2 performs extremely well especially when p becomes small and n sufficiently large.

Let us return back to the original claim model described in the introduction. The practical meaning of Theorem 3 is that, should we wish to have a non-degenerate result for the total number

W_n of blocks of k consecutive claims with at least r claims in each block exceeding the threshold u , we have to increase $u = u_n$ with n in such a way that $n(P(Y_i > u_n))^r$ converges to a positive number.

Consider for example the case where the claims Y_i follow a Pareto distribution, namely

$$P(Y_i \leq x) = 1 - \left(\frac{c}{x}\right)^a, \quad x \geq c \quad (a > 0, c > 0).$$

If we choose u_n so that $u_n \sim c\beta^{-\frac{1}{ar}}n^{\frac{1}{ar}}$ ($\beta > 0$), we obtain

$$n(P(Y_i > u_n))^r = n \left(\frac{c}{u_n}\right)^{ar} = \beta \left(\frac{c\beta^{-\frac{1}{ar}}n^{\frac{1}{ar}}}{u_n}\right)^{ar} \rightarrow \beta$$

and

$$\lambda = (n - k + 1) \binom{k-1}{r-1} (P(Y_i > u_n))^r (P(Y_i \leq u_n))^{k-r+1} \rightarrow \binom{k-1}{r-1} \beta.$$

Hence, by virtue of Theorem 3, W_n converges to $CP(\lambda, F)$ with $\lambda = \binom{k-1}{r-1} \beta$ and compounding distribution with probability mass function given by (12).

If the claims were distributed according to the Weibull law

$$P(Y_i \leq x) = 1 - e^{-(\theta x)^a}, \quad x \geq 0 \quad (a > 0, \theta > 0)$$

then it can be easily verified that

$$u_n = \frac{1}{\theta} \left(\frac{1}{r} \ln \left(\frac{n}{\beta} + o(n) \right) \right)^{\frac{1}{a}}, \quad \beta > 0$$

is an appropriate selection for the claim thresholds, so as a non-degenerate asymptotic result is deduced. In this case we have

$$n(P(Y_i > u_n))^r = \frac{n}{\frac{n}{\beta} + o(n)} \rightarrow \beta$$

and it is immediate that W_n converges now to $CP(\lambda, F)$ with $\lambda = \binom{k-1}{r-1} \beta$.

Apparently, the machinery developed in this section, could be effectively exploited as well if, instead of Bernoulli random variables $X_i = I_{(u, \infty)}(Y_i)$, $i = 1, 2, \dots$ we considered sequences of the form $X_i = I_A(Y_i)$, $i = 1, 2, \dots$ for any Borel set A . A typical example, which is of intrinsic interest in reinsurance studies, is offered by the choice of a layer of the form $D = (D_1, D_2]$; see e.g. Embrechts *et al.* (1997).

4 An asymptotic result related to a financial problem

As mentioned in the introduction, the multiple scan statistic W_n can play a remarkable role in stochastic financial analysis. Binswanger and Embrechts (1994) point out that the asymptotic results on the longest head run in coin tossing, or equivalently the longest success run in Bernoulli trials, have been recently applied to the so-called "maximum drawdown" problem in finance. In

the sequel we shall formulate a more general model where the assessment of the loss (or gain) over a fixed length time window is couched on the multiple scan statistic W_n .

Let us consider a trader at the stock exchange who is interested in identifying a drop-down trend at the exchange rates within a certain time horizon of n days (the choice of "day" as a time unit is arbitrary here and could be replaced by any other unit). We assume that the exchange rates Y_1, Y_2, \dots, Y_n at the n days are i.i.d random variables whose cumulative distribution function can be expressed in the form

$$P(Y_i \leq y) = (\theta y)^a + o(y^a), \quad y > 0 \quad (16)$$

for $y \rightarrow 0$, where θ and a are positive parameters. There are quite a few distributions, in common use in actuarial science and financial mathematics, which fall within the family described by (16); for example, the exponential, Weibull, Gamma, Uniform are some typical examples.

For fixed $u > 0$, the k -scan process $\sum_{j=i}^{i+k-1} I_{[0,u]}(Y_j)$, $i = 1, 2, \dots, n - k + 1$ can serve as an index of "local" drop-down trend in the sense that should its values tend to exceed a fixed level r , a strong evidence is provided for drawdown movement of the exchange rates. Therefore, a reasonable statistic for an overall assessment of a drawdown tendency, is offered by

$$W_n = \sum_{i=1}^{n-k+1} I_{[r,\infty)} \left(\sum_{j=i}^{i+k-1} I_{[0,u]}(Y_j) \right) \quad (17)$$

and the risk of experiencing this unfavorable event would be measured by

$$G_n(u) = P(W_n \geq m). \quad (18)$$

It is not difficult to verify that $G_n(\cdot)$ is a right-continuous increasing function with $\lim_{u \rightarrow \infty} G_n(u) = 1$; hence it defines a cumulative density function of a random variable say $Y^{(n)}$, that is $P(Y^{(n)} \leq u) = G_n(u)$. It can be easily verified that the random variable $Y^{(n)} = Y_{k,r,m}^{(n)}$ can be viewed as the smallest threshold u such that $W_n = W_{n,u,k,r} \geq m$.

We propose now to investigate the asymptotic behaviour of $Y^{(n)}$ as $n \rightarrow \infty$. Manifestly, for fixed u the asymptotic distribution of $Y^{(n)}$ is degenerate and therefore it seems sensible to look for a suitable normalization that will give rise to a non degenerate limiting distribution law. An answer to this question is offered by the next theorem, which is a direct consequence of Theorem 3.

Theorem 4 *If $H(\cdot; k, r, \theta, a, u)$ is the cumulative distribution function of a $CP(\lambda, F)$, with $\lambda = \binom{k-1}{r-1}(\theta u)^{ar}$ and compounding distribution*

$$F(x) = 1 - \frac{\binom{k-1-x}{r-1}}{\binom{k-1}{r-1}}, \quad x = 1, 2, \dots, k - r, \quad F(0) = 0, F(k - r + 1) = 1,$$

then the asymptotic distribution of $Y^{(n)}$ (suitably adjusted by the use of appropriate normalization constants) can be expressed as follows

$$\lim_{n \rightarrow \infty} P(Y^{(n)} > n^{-1/ra} u) = H(m - 1; k, r, \theta, a, u) \quad (19)$$

Proof. By the definition of $Y^{(n)}$ we may write

$$P(Y^{(n)} > n^{-1/ra} u) = P(W_n < m) \quad (20)$$

where

$$W_n = \sum_{i=1}^{n-k+1} I_{[r,\infty)} \left(\sum_{j=i}^{i+k-1} X_j \right)$$

and $X_j = I_{[0, n^{-1/ra}u]}(Y_j)$ are binary random variables with success probabilities (see also (16))

$$E(X_j) = P(Y_j \leq n^{-1/ra}u) = (\theta u)^a n^{-1/r} + o(n^{-1/r})$$

(as $n \rightarrow \infty$). It follows that

$$n(E(X_j))^r = (\theta u)^{ar} + n o\left(\frac{1}{n}\right)$$

and direct calculations on the sequence

$$\lambda_n = (n - k + 1) \binom{k-1}{r-1} (E(X_j))^r (1 - E(X_j))^{k-r+1}$$

yield

$$\lambda = \lim_{n \rightarrow \infty} \lambda_n = \binom{k-1}{r-1} (\theta u)^{ar}.$$

The proof of the theorem is easily completed by virtue of (20) and the outcome of Theorem 3. ■

It should be mentioned that the evaluation of the cumulative distribution function $H(\cdot; k, r, \theta, a, u)$ could be easily carried out by initiating a recursive scheme similar to (13) to facilitate the computation of the corresponding values of the probability mass function.

In closing, let us proceed to the identification of the asymptotic distribution of $Y^{(n)}$ in the special cases $m = 1$ and $k = r$. The meaning of the first case in the financial model described earlier, is that a drawdown trend is declared if, in the time horizon under inspection, there exist k consecutive days with the exchange rate being below u in at least r of them. We have then, by virtue of (19),

$$\lim_{n \rightarrow \infty} P(Y^{(n)} > n^{-1/ra}u) = H(0; k, r, \theta, a, u) = e^{-\binom{k-1}{r-1}(\theta u)^{ar}} \quad (21)$$

which reveals that the random variable $n^{\frac{1}{ra}}Y^{(n)}$ follows asymptotically a Weibull distribution.

Consider next the case $k = r$, with the financial model signaling a drawdown trend on the occurrence of m (overlapping) runs of k consecutive days with the exchange rate being below u in all k days. Now the limiting compound Poisson law $CP(\lambda, F)$ reduces to an ordinary Poisson distribution with $\lambda = (\theta u)^{ar}$. Thus, by virtue of (19), we obtain

$$\lim_{n \rightarrow \infty} P(Y^{(n)} > n^{-1/ra}u) = e^{-(\theta u)^{ar}} \sum_{x=0}^{m-1} \frac{[(\theta u)^{ar}]^x}{x!}$$

the last formula revealing that $n(Y^{(n)})^{ar}$ follows asymptotically an Erlang distribution with parameters m and θ^{ar} .

5 Further Applications

In the analysis of the experimental trials whose outcomes can be classified in two exclusive categories (success/failure, accept/reject, defective/non-defective etc.) a question that comes in naturally, is whether reasonable criteria providing evidence of clustering of any of the two categories could be established. One of the most commonly used statistic in such situations is the run statistic either in the classical form (see e.g. Gibbons and Chakraborti (1992)) or in the form of success run of fixed length. The random variable W_n studied in the previous sections, is offering an efficient and fascinating alternative test statistic in a variety of fields where the classical run criteria have been traditionally in use. We shall now proceed to review two typical examples which are of intrinsic interest.

Molecular Biology. When studying amino acid sequences, various classification schemes are in common use, including a chemical alphabet of 8 letters, a functional alphabet of 4 letters, a charge alphabet of 3 letters etc. In order to introduce quantitative means for assessing and interpreting genomic inhomogeneities between sequences of different species or sequences subject to different chemical infections and/or several levels of corruption, molecular biologists look for long aligned subsequences that match in most of their positions and try to specify what is an unusually long match. Following Glaz and Naus (1991), let $Z_{i1}, Z_{i2}, i = 1, 2, \dots, n$ be two amino acid sequences from a finite alphabet. The two sequences will be said to match in position i if $Z_{i1} = Z_{i2}$, in which case we let X_i be 1 (and 0 otherwise). It is evident that the occurrence of frequent "almost perfect" matches between the two sequences, corresponds to "dense" scans in the sense that the associated k -scan process $S_i = \sum_{j=i}^{i+k-1} X_j, i = 1, 2, \dots$ takes on values which are very close to k . Therefore, if $r \leq k$ is a positive integer not very far away from k , the multiple scan statistic $W_n = \sum_{i=1}^{n-k+1} I_{[r, \infty)}(S_i)$ could serve as a local correlation index between the sequences Z_{i1} and Z_{i2} . Moreover, should the interest be in establishing a procedure for testing the null hypothesis that matches are independent with constant matching probabilities $p = P(Z_{i1} = Z_{i2}) = P(X_i = 1), i = 1, 2, \dots, n$ (against the alternative that there exist subsequences of length k where the matching probability increases), a rational critical (rejection) region may be established by an inequality of the form $W_n > c$. The theoretical results presented in this article are then of major importance in the determination of constant c so as a prespecified significance level is attained. Since the primary interest in this situation is in extremely long sequences, the asymptotic outcomes derived in the previous sections are expected to play a significant role in analyzing and interpreting real amino acid data.

Reliability Theory. A class of reliability systems, which has attracted considerable research interest during the last decades is the so called "consecutive systems". A typical example is the consecutive- k -out-of- n :F system which fails whenever at least k consecutive components fail. A natural extension of this structure is offered by the r -within-consecutive- k -out-of- n :F system whose failure requires the existence of a strand of k consecutive components containing at least r failed ones among them. A slightly more general structure pops up if system's failure was experienced upon the appearance of at least m strands (probably overlapping) of the type described above. For fixed time $u > 0$, we may visualize the structure as a sequence of n Bernoulli trials $X_i = I_{[0, u]}(Y_i), i = 1, 2, \dots, n$ (Y_i is the i -th component's lifetime); if $Y^{(n)}$ denotes system's lifetime, the reliability $R_n(u)$ can accordingly be expressed as

$$R_n(u) = P(Y^{(n)} > u) = P(W_n < m) = 1 - G_n(u)$$

where W_n and $G_n(u)$ are given by (17), (18). For large i.i.d systems of this type, it can be readily

checked that, if time u is fixed, $R_n(u)$ tends to 0 as n tends to infinity. It is natural then to ask whether proper normalizing terms $a_n > 0$ and b_n could be identified so as the random variable $(Y^{(n)} - b_n)/a_n$ converges in law to a non-degenerate distribution. Exploiting Theorem 4 we may state that, if components' lifetimes satisfy (16) the raised question could be answered to the positive and an appropriate selection for a_n, b_n is $a_n = n^{1/ra}, b = 0, n = 1, 2, \dots$. For alternative approaches of this problem in the special case $m = 1$, see Papastavridis and Koutras (1993) and references therein.

Finally, a few additional areas where the notion of multiple scan statistic arises in quite a natural way include quality control (scan-based control charts, Greenberg (1970), Saperstein (1973), sampling inspection systems, Schmuelli and Cohen (2000)), signal processing (study of detectors using moving window scanners, Glaz (1983)), ecology (study of spatial patterns in the diffusion of species or spread of diseases), educational psychology (investigation of efficiency of transfer and learning procedures) and non-parametric statistical inference (as an alternative to the pure run criterion which is in common use in tests of randomness).

References

- [1] Barbour, A.D., Holst, L. and Janson, S. (1992). *Poisson Approximation*. Clarendon Press, Oxford.
- [2] Binswanger, K. and Embrechts P. (1994). Longest runs in coin tossing. *Insurance: Mathematics and Economics* 15, 139-149.
- [3] Boutsikas, M.V. and Koutras, M.V. (2001): Compound Poisson approximation for sums of dependent random variables. In *Probability and Statistical Models with Applications: A volume in honor of Prof. T. Cacoullos* (Eds. Ch.A. Charalambides, M.V. Koutras, N. Balakrishnan), 63-86, Chapman and Hall/CRC press.
- [4] Bowers, N.L., Gerber, H.U., Hickman, J., Jones, D.A. and Nesbitt, C.J. (1997). *Actuarial Mathematics* (2nd Edition). The Society of Actuaries, Illinois.
- [5] Chen, J. and Glaz, J. (1997). Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials. In *Advances in the theory and practice of Statistics: A volume in honor of S. Kotz* (Eds. N. L. Johnson and N. Balakrishnan), 285-298. Wiley, NY.
- [6] Chen, J. and Glaz, J. (1999). Approximations for the distribution and the moments of discrete scan statistics. In *Scan Statistics and Applications* (Eds. J. Glaz and N. Balakrishnan), 27-66, Birkhauser.
- [7] Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for insurance and finance*. Springer-Verlag.
- [8] Gibbons, J.D. and Chakraborti, S. (1992) *Nonparametric Statistical Inference*. Marcel Dekker Inc, N.Y.

- [9] Glaz, J. (1983). Moving window detection for discrete data. *IEEE Transactions on Information Theory* 29, 457-462.
- [10] Glaz, J. and Naus, J.I. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *The Annals of Applied Probability* 1, 306-318.
- [11] Glaz, J., Naus, J., Roos, M. and Wallenstein, S. (1994). Poisson approximation for the distribution and moments of ordered m -spacings. *Journal of Applied Probability* 31 (special volume in honor of L. Takacs), 271-281.
- [12] Greenberg (1970). The first occurrence of n successes in N trials. *Technometrics* 12, 627-634.
- [13] Koutras, M.V. and Alexandrou, V. (1995) Runs, scans and urn model distributions: A unified Markov chain approach. *Annals of the Institute of Statistical Mathematics* 47, 743-766.
- [14] Papastavridis, S.G. and Koutras, M.V. (1993). Consecutive k -out-of- n systems. In *New Trends in Systems Reliability Evaluation* (Ed. K.B. Misra). Elsevier Science, 228-248.
- [15] Roos, M. (1993). *Stein - Chen method for compound Poisson approximation*. PHD thesis, University of Zurich.
- [16] Roos, M. (1994). Stein's method for compound Poisson approximation. *Annals of Applied Probability* 4, 1177-1187.
- [17] Saperstein, B. (1973). On the occurrence of n successes within N Bernoulli trials. *Technometrics* 15, 809-818.
- [18] Schmuelli, G., and Cohen, A. (2000). Run-related probability functions applied to sampling inspection. *Technometrics* 42, 188-202.

Table 1. Compound Poisson approximations for the discrete multiple scan statistic

n	k	p	r	x	$Sim(10^6)$	CG1	CG2	CG3	$CP(\lambda, F)$	UB
100	10	0.05	2	1	0.7483	0.7130	0.8474	0.7390	0.7248	1.047
				2	0.7177	0.6520	0.7185	0.6836	0.6874	
				3	0.6843	0.5946	0.6371	0.6314	0.6492	
				4	0.6487	0.5407	0.5763	0.5824	0.6104	
				5	0.6102	0.4904	0.5230	0.5367	0.5710	
				MRE		0.126523	0.091407	0.071981	0.04964	
100	10	0.05	3	1	0.2263	0.2372	0.2489	0.2393	0.2379	0.0982
				2	0.1908	0.1834	0.1843	0.1857	0.1942	
				3	0.1582	0.1417	0.1424	0.1441	0.1563	
				4	0.1282	0.1094	0.1099	0.1118	0.1238	
				5	0.1013	0.0843	0.0847	0.0868	0.0963	
				MRE		0.101143	0.108085	0.088873	0.032954	
100	10	0.05	4	1	0.0311	0.0328	0.0329	0.0328	0.0328	0.0074
				2	0.0224	0.0213	0.0213	0.0213	0.0226	
				3	0.0156	0.0138	0.0138	0.0139	0.0151	
				4	0.0105	0.0090	0.0090	0.0090	0.0098	
				5	0.0066	0.0058	0.0058	0.0059	0.0061	
				MRE		0.096645	0.097288	0.092332	0.047613	
100	20	0.05	4	1	0.1710	0.1853	0.1919	0.1859	0.1854	0.1448
				2	0.1523	0.1586	0.1589	0.1594	0.1604	
				3	0.1354	0.1358	0.1360	0.1366	0.1387	
				4	0.1201	0.1163	0.1164	0.1170	0.1199	
				5	0.1060	0.0995	0.0997	0.1003	0.1037	
				MRE		0.044181	0.052046	0.04444	0.037026	
500	10	0.01	2	1	0.3272	0.3275	0.3971	0.3386	0.3321	0.0382
				2	0.2999	0.2752	0.2843	0.2871	0.3025	
				3	0.2721	0.2309	0.2359	0.2436	0.2725	
				4	0.2431	0.1935	0.1976	0.2067	0.2421	
				5	0.2136	0.1619	0.1654	0.1756	0.2113	
				MRE		0.136153	0.162302	0.10198	0.007999	
500	10	0.05	4	1	0.1541	0.1647	0.1653	0.1649	0.1647	0.0212
				2	0.1150	0.1111	0.1111	0.1113	0.1171	
				3	0.0830	0.0749	0.0749	0.0751	0.0813	
				4	0.0579	0.0504	0.0505	0.0507	0.0548	
				5	0.0385	0.0340	0.0340	0.0342	0.0357	
				MRE		0.089341	0.089775	0.086696	0.046759	
500	20	0.05	5	1	0.1977	0.2262	0.2270	0.2263	0.2262	0.0667
				2	0.1705	0.1851	0.1851	0.1852	0.1865	
				3	0.1461	0.1514	0.1514	0.1516	0.1537	
				4	0.1250	0.1238	0.1238	0.1240	0.1267	
				5	0.1062	0.1012	0.1012	0.1014	0.1043	
				MRE		0.064549	0.065358	0.064345	0.064302	
500	20	0.05	6	1	0.0354	0.0397	0.0397	0.0397	0.0397	0.0076
				2	0.0282	0.0297	0.0297	0.0297	0.0300	
				3	0.0221	0.0223	0.0223	0.0223	0.0226	
				4	0.0175	0.0167	0.0167	0.0167	0.0171	
				5	0.0137	0.0125	0.0125	0.0125	0.0129	
				MRE		0.063403	0.063403	0.063403	0.057835	

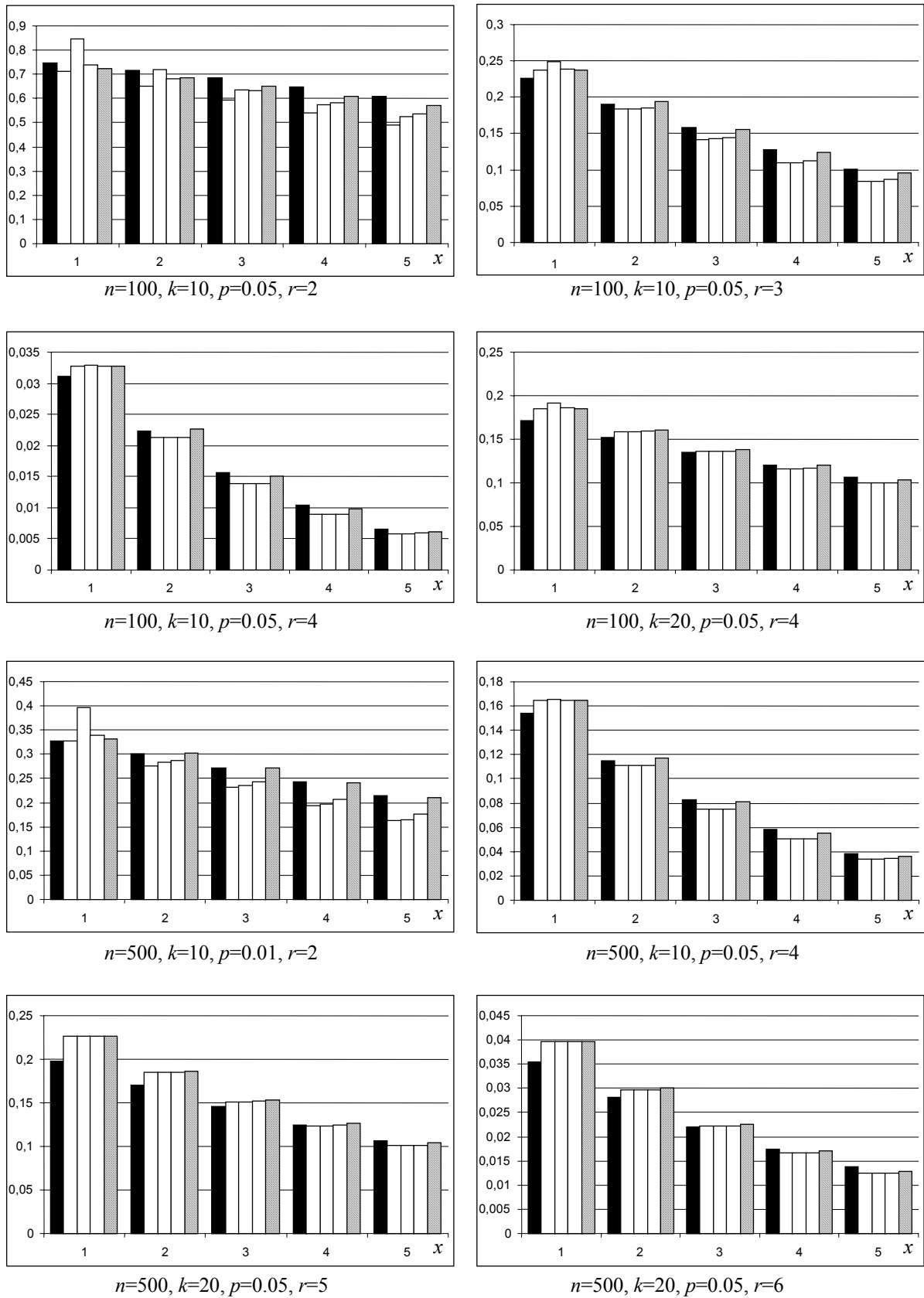
Table 2. Convergence of approximations to the asymptotic compound Poisson distribution: $\lambda_0 = 1, r = 3, k = 10$

x	p=0.1, n=73			p=0.01, n=30112			p=0.001, n=28001012			p=0.0001, n=27800010012							
	CGI	CG2	CG3	CP(λ, F)	CGI	CG2	CG3	CP(λ, F)	CGI	CG2	CG3	CP(λ, F)	CGI	CG2	CG3	CP(λ, F)	CP(λ_0, F)
1	.6259	.6756	.6342	.6291	.6316	.6407	.6335	.6321	.6317	.6392	.6333	.6321	.6317	.6391	.6333	.6321	.6321
2	.5439	.5583	.5542	.5555	.5241	.5269	.5266	.5512	.5215	.5238	.5236	.5505	.5212	.5235	.5233	.5504	.5504
3	.4711	.4817	.4830	.4894	.4323	.4346	.4352	.4721	.4278	.4297	.4303	.4700	.4273	.4292	.4298	.4698	.4698
4	.4068	.4160	.4200	.4300	.3546	.3566	.3580	.3962	.3489	.3506	.3518	.3923	.3484	.3500	.3512	.3919	.3919
5	.3501	.3583	.3646	.3768	.2895	.2911	.2932	.3249	.2831	.2845	.2863	.3191	.2825	.2838	.2856	.3185	.3184
6	.3004	.3076	.3161	.3290	.2353	.2366	.2392	.2597	.2287	.2298	.2321	.2519	.2280	.2291	.2314	.2511	.2511
7	.2570	.2633	.2737	.2859	.1904	.1915	.1946	.2018	.1839	.1848	.1875	.1926	.1832	.1841	.1868	.1917	.1916
8	.2192	.2247	.2367	.2467	.1535	.1544	.1578	.1527	.1472	.1480	.1510	.1430	.1466	.1474	.1503	.1420	.1419
9	.1865	.1912	.2047	.2108	.1232	.1240	.1278	.1136	.1175	.1181	.1214	.1046	.1169	.1175	.1207	.1038	.1037
10	.1581	.1622	.1769	.1775	.0986	.0992	.1032	.0858	.0933	.0938	.0973	.0794	.0928	.0933	.0968	.0788	.0787
	UB=0.957723				UB=0.0253364				UB=0.00234069				UB=0.000233393				

Table 3. Convergence of approximations to the asymptotic compound Poisson distribution: $\lambda_0 = 0.5, r = 2, k = 10$

x	p=0.1 n=23			p=0.01 n=617			p=0.001 n=56067			p=0.0001 n=5560567							
	CGI	CG2	CG3	CP(λ, F)	CGI	CG2	CG3	CP(λ, F)	CGI	CG2	CG3	CP(λ, F)	CGI	CG2	CG3	CP(λ, F)	CP(λ_0, F)
1	.3656	.5931	.4093	.3862	.3882	.4656	.4006	.3934	.3890	.4566	.3998	.3935	.3891	.4557	.3998	.3935	.3935
2	.3247	.3861	.3712	.3563	.3292	.3418	.3429	.3600	.3281	.3387	.3400	.3598	.3280	.3384	.3398	.3598	.3598
3	.2879	.3107	.3369	.3286	.2788	.2859	.2935	.3261	.2764	.2826	.2892	.3252	.2761	.2823	.2888	.3252	.3251
4	.2547	.2705	.3060	.3030	.2357	.2416	.2513	.2916	.2324	.2376	.2460	.2898	.2320	.2372	.2455	.2896	.2896
5	.2249	.2386	.2783	.2793	.1990	.2040	.2152	.2565	.1950	.1995	.2093	.2534	.1946	.1990	.2087	.2530	.2530
6	.1981	.2103	.2533	.2575	.1676	.1719	.1845	.2208	.1634	.1671	.1782	.2160	.1630	.1667	.1776	.2155	.2155
7	.1741	.1850	.2309	.2373	.1409	.1445	.1583	.1845	.1366	.1398	.1519	.1777	.1361	.1393	.1513	.1770	.1769
8	.1524	.1623	.2108	.2187	.1181	.1212	.1360	.1476	.1139	.1166	.1296	.1384	.1135	.1161	.1290	.1374	.1373
9	.1330	.1419	.1927	.2015	.0987	.1014	.1170	.1101	.0947	.0970	.1108	.0981	.0943	.0965	.1102	.0968	.0967
10	.1156	.1235	.1765	.1856	.0823	.0845	.1009	.0720	.0785	.0804	.0949	.0568	.0781	.0800	.0943	.0552	.0550
	UB=1.3022				UB=0.044577				UB=0.0022654				UB=0.00020267				

Figure 1. Compound Poisson Approximations for the discrete multiple scan statistic



Solid bars: simulated value, *blank bars:* CG1, CG2, CG3, *shaded bars:* CP(λ, F)