Methodol Comput Appl Probab (2007) 9:195–205
DOI 10.1007/s11009-007-9019-5

On the Normal Approximation for the Distribution of the Number of Simple or Compound Patterns in a Random Sequence of Multi-state Trials

James C. Fu · W. Y. Wendy Lou

Received: 12 November 2005 / Revised: 10 February 2006 /
Accepted: 16 July 2006 / Published online: 29 March 2007

Abstract Distributions of numbers of runs and patterns in a sequence of multi-state trials have been successfully used in various areas of statistics and applied probability. For such distributions, there are many results on Poisson approximations, some results on large deviation approximations, but no general results on normal approximations. In this manuscript, using the finite Markov chain imbedding technique and renewal theory, we show that the number of simple or compound patterns, under overlap or non-overlap counting, in a sequence of multi-state trials follows a normal distribution. Poisson and large deviation approximations are briefly reviewed.

Keywords Runs and patterns · Finite Markov chain imbedding · Waiting time distribution

AMS 2000 Subject Classification Primary 60E05 · Secondary 60J10

1 Introduction

Let $\mathcal{S} = \{a_1, \cdots, a_m\}$ be a set containing m ($m \geq 2$) states (or symbols of an alphabet), and let $\{X_i\}$ be a sequence of independent and identically distributed (*i.i.d.*) m-state trials defined on $\mathcal{S}$ with probabilities p_i, $i = 1, 2, \cdots, m$, respectively. Two types of patterns, simple and compound, are considered in this manuscript.

J. C. Fu
Department of Statistics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada
e-mail: fu@cc.umanitoba.ca

W. Y. W. Lou (✉)
Department of Public Health Sciences and Department of Statistics, University of Toronto, Toronto, Ontario M5T 3M7, Canada
e-mail: wendy.lou@utoronto.ca

Definition 1 Λ is a simple pattern of length k if Λ is composed of k specified states; i.e. $\Lambda = b_1 \cdots b_k$, $b_i \in \mathcal{S}$. The length k is fixed, and the states b_i are allowed to be repeated.

We say that two patterns Λ_1 and Λ_2 are distinct if neither is a subsequence of the other, and define $\Lambda = \Lambda_1 \cup \Lambda_2$, the union of Λ_1 and Λ_2, to be the occurrence of either the pattern Λ_1 or Λ_2.

Definition 2 Λ is a compound pattern if it is a union of l distinct simple patterns (l is fixed).

Two important distributions associated with the pattern Λ in a sequence of multi-state trials have been studied in the literature: (1) the distribution of $X_n(\Lambda)$, the number of Λ patterns that occurred in n trials (under non-overlap or overlap counting), and (2) the waiting time distribution of $W(\Lambda)$, the number of trials required until the first occurrence of pattern Λ. Traditionally, the exact distributions of the numbers of runs and patterns in n trials (n is fixed) were studied using the combinatorial method (see MacMahon, 1915, and Riordan, 1958). Recently, the finite Markov chain imbedding technique has been used to study these exact distributions (see Balakrishnan and Koutras, 2002, and Fu and Lou, 2003).

For n large, calculating the probability

$$\alpha_n(\Lambda, x) = P(X_n(\Lambda) < x) \tag{1}$$

is often a hard task. There are three main types of approximations, Poisson, large deviation and normal, that can be used to approximate $\alpha_n(\Lambda, x)$.

Many results on Poisson and compound Poisson approximations of the tail probability $\alpha_n(\Lambda, n)$ for various specified patterns have been obtained using the Stein–Chen method (see Barbour and Chryssaphinou 2001). A result for the longest success run on the large deviation approximation for $\alpha_n(\Lambda, x)$ has been developed using finite Markov chain imbedding together with the Perron–Frobenius theorem and a vector decomposition of the eigenspace (see Fu and Lou 2003).

There are no general results on the asymptotic normality of the distribution of $X_n(\Lambda)$: i.e.

$$\lim_{n\to\infty} P\left(\frac{X_n(\Lambda) - EX_n(\Lambda)}{\sigma(n)} < x\right) = \int_{-\infty}^{x} \phi(z)dz, \tag{2}$$

where $\sigma^2(n) = Var(X_n(\Lambda))$ and $\phi(z)$ is the density function of the standard normal distribution. However, there are some special cases for which Eq. 2 clearly holds: consider, for example, the pattern $\Lambda = S$ and $X_n(\Lambda) = \sum_{i=1}^{n} I(X_i)$, the number of successes in n Bernoulli trials, where the indicator function $I(X) = 1$ if $X = S$ and $I(X) = 0$ if $X = F$.

In Section 2, we will show that the random variable $X_n(\Lambda)$ is asymptotically normally distributed for Λ being either a simple or a compound pattern under non-overlap counting, using the finite Markov chain imbedding technique together with renewal theory. We further show that $X_n^o(\Lambda)$, the number of Λ patterns under overlap counting, is also asymptotically normally distributed if Λ is a simple pattern, using Hoeffding and Robbins (1948) central limit theorem for the sum of $(k-1)$th order dependent random variables. For the case of Λ being a compound pattern

under overlap counting, we conjecture the result of asymptotic normality remains true, but we are not able to prove it mathematically.

In Section 3, we provide a short review on recent developments associated with the three types of approximations for the tail probability $\alpha_n(\Lambda, \alpha)$, and give some discussion regarding their numerical performance.

2 Asymptotic Normality

Let us consider the pattern Λ (simple or compound) and the waiting time $W(\Lambda)$ of the pattern Λ in the *i.i.d.* m-state trials $\{X_t\}_{t=1}^n$. It is known that the waiting time random variable $W(\Lambda)$ is homogeneous finite Markov chain imbeddable. The following results hold.

Theorem 1 *Given a pattern Λ, there exists a homogeneous Markov chain $\{Y_t\}$ on a state space Ω with transition probability matrix*

$$\boldsymbol{M} = \begin{array}{c} \Omega - \alpha \\ \alpha \end{array} \begin{pmatrix} \boldsymbol{N} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{pmatrix},$$

such that

(1)

$$P(W(\Lambda) = n) = \boldsymbol{\xi} \boldsymbol{N}^{n-1} (\boldsymbol{I} - \boldsymbol{N}) \boldsymbol{1}', \tag{3}$$

where $\boldsymbol{\xi}$ is the initial distribution, $\boldsymbol{I}$ denotes an identity matrix, and $\boldsymbol{1}$ is a row vector with each entry equal to one,

(2) *The probability generating function is given by*

$$\varphi_W(s) = 1 + (s-1)\boldsymbol{\xi}(\boldsymbol{I} - s\boldsymbol{N})^{-1}\boldsymbol{1}', \tag{4}$$

(3) *The mean of $W(\Lambda)$ is given by*

$$\mu_W = \varphi_W^{(1)}(1), \tag{5}$$

and the variance of $W(\Lambda)$ by

$$\sigma_W^2 = \varphi_W^{(2)}(1) + \varphi_W^{(1)}(1) - (\varphi_W^{(1)}(1))^2, \tag{6}$$

where $\varphi_W^{(i)}(1) = (\partial/\partial s)^i \varphi_W(s)|_{s=1}$, for $i = 1, 2$.

We omit the proofs of the above results, which can be found in the works by Koutras (1996, 1997), and Fu and Lou (2003, 2006).

For non-overlap counting and given a fixed n, the duality relationship between the number of patterns Λ in n trials and the interarrival times $W_i(\Lambda)$ can be stated as follows:

$$P(X_n(\Lambda) < k) = P(W_1(\Lambda) + \cdots + W_k(\Lambda) > n), \tag{7}$$

for all k, where $W_i(\Lambda)$, $i = 1, \cdots, k$, are *i.i.d.* random variables having the same general geometric distribution as $W(\Lambda)$ given in Theorem 1. Further, it follows from renewal theory that for n large,

$$EX_n(\Lambda) = \frac{n}{\mu_W}(1 + o(1)); \tag{8}$$

see Feller (1968).

Now we are ready to show the asymptotic normality of the random variable $X_n(\Lambda)$ as $n \to \infty$.

Theorem 2 *Under non-overlap counting, the random variable $X_n(\Lambda)$ is asymptotically normally distributed in the sense*

$$\frac{X_n(\Lambda) - EX_n(\Lambda)}{\sqrt{\sigma_W^2 \mu_W^{-3} n}} \xrightarrow{\mathcal{L}} N(0, 1) \tag{9}$$

as $n \to \infty$, where $\xrightarrow{\mathcal{L}}$ stands for convergence in law and μ_W and σ_W^2 are given in Theorem 1.

In view of the above theorem, the result is useful only if μ_W and σ_W^2 can be obtained through the finite Markov chain imbedding (or other) technique. In general, the exact forms of μ_W and σ_W^2 are rather complex, especially when Λ is a compound pattern, a union of l distinct simple patterns.

Proof of Theorem 2 Given $x \in R$, we define

$$k_n = \left[EX_n(\Lambda) + x\sqrt{\sigma_W^2 \mu_W^{-3} n} \right], \tag{10}$$

where $[a]$ stands for the integer part of $a \in R$. It follows from the identity of duality, Eq. 7, that

$$P(X_n(\Lambda) < k_n) = P(W_1(\Lambda) + \cdots + W_{k_n}(\Lambda) > n), \tag{11}$$

where $W_i(\Lambda)$, $i = 1, \cdots, k_n$, are *i.i.d.* interarrival times with common mean μ_W and common variance σ_W^2. From the definition of k_n and Eq. 8, $k_n \to \infty$ as $n \to \infty$, and a simple calculation yields

$$\lim_{n\to\infty} \frac{n - k_n EW(\Lambda)}{\sigma_W \sqrt{k_n}} = -x. \tag{12}$$

It follows from the central limit theorem, Eq. 12, and the symmetry of the standard normal distribution at zero that

$$\begin{aligned} P(X_n(\Lambda) < k_n) &= P\left(\sum_{i=1}^{k_n} W_i(\Lambda) > n \right) \\ &= P\left(\frac{\sum_{i=1}^{k_n} W_i(\Lambda) - k_n EW(\Lambda)}{\sigma_W \sqrt{k_n}} > \frac{n - k_n EW(\Lambda)}{\sigma_W \sqrt{k_n}} \right) \\ &\to \int_{-\infty}^{x} \phi(z) dz, \end{aligned} \tag{13}$$

as $k_n \to \infty$ $(n \to \infty)$, where $\phi(z)$ is the density function of the standard normal distribution with mean zero and variance 1. □

Under overlap counting, the random variable $X_n^o(\Lambda)$ can not be directly linked to a sum of *i.i.d.* interarrival waiting times $W_i(\Lambda)$. Hence in this case we can not adopt the above method to prove the asymptotic normality of $X_n^o(\Lambda)$. In the following, we prove the asymptotic normality of $X_n^o(\Lambda)$ for the case when $\Lambda = b_1 b_2 \cdots b_k$ is a simple pattern of fixed length k. To achieve this goal, we first write $X_n^o(\Lambda)$ as a sum of n $(k-1)$th order dependent index functions, and then apply the central limit theorem for a $(k-1)$th order dependent stationary sequence, due to Hoeffding and Robbins (1948). First we state their result without proof.

Lemma 1 *For fixed k, let $Z_1, \cdots, Z_n$ be a stationary sequence of $(k-1)$th order dependent random variables. If $E|Z_1|^3 < \infty$, then the central limit theorem holds:*

$$\frac{\sum_{i=1}^n (Z_i - EZ_i)}{\sqrt{n}} \xrightarrow{\mathcal{L}} N(0, \sigma^2) \quad as\ n \to \infty,$$

where $\sigma^2 = Var(Z_1) + 2\sum_{i=1}^k (EZ_1 Z_i - EZ_1 EZ_i)$.

Given $\Lambda = b_1 b_2 \cdots b_k$, let us consider the indicator function $\eta_i(\Lambda)$ defined on the sequence $\{X_i\}$ as

$$\eta_i(\Lambda) = \begin{cases} 1 \text{ if } X_{i-k+1} = b_1, X_{i-k+2} = b_2, \cdots, X_i = b_k \\ 0 \text{ otherwise,} \end{cases} \tag{14}$$

for $i = 1, \cdots, n$.

Since X_i are *i.i.d.* multi-state trials, it is easy to see that the indicator functions $\eta_i(\Lambda)$, $i = 1, \cdots, n$, are identically distributed, but also $(k-1)$th order dependent, random variables.

Theorem 3 *Given a simple pattern $\Lambda = b_1 b_2 \cdots b_k$ of length k, it follows that*

$$\frac{X_n^o(\Lambda) - EX_n^o(\Lambda)}{\sqrt{n}} \xrightarrow{\mathcal{L}} N(0, \sigma^2) \quad as\ n \to \infty,$$

where

$$\sigma^2 = \prod_{i=1}^k P(b_i) \left(1 - \prod_{i=1}^k P(b_i)\right) + 2\sum_{l=2}^k Cov(\eta_1(\Lambda), \eta_l(\Lambda)),$$

$P(b_i) = P(X_1 = b_i)$, for $i = 1, 2 \cdots, k$, and

$$Cov(\eta_1(\Lambda), \eta_l(\Lambda)) = \begin{cases} \left(\prod_{i=1}^{l-1} P(b_i)\right)\left(\prod_{i=1}^k P(b_i)\right) - \left(\prod_{i=1}^k P(b_i)\right)^2 \\ \qquad\qquad if\ b_1 \cdots b_{k-l+1} = b_l \cdots b_k, 2 \le l \le k \\ -\left(\prod_{i=1}^k P(b_i)\right)^2 \qquad otherwise. \end{cases} \tag{15}$$

Proof of Theorem 3 Since X_i are *i.i.d.* multi-state trials, the $\eta_i(\Lambda)$ are $(k-1)$th order dependent index functions and are identically distributed. It is easy to check that

(1) $$E\eta_1(\Lambda) = \prod_{i=1}^k P(b_i), \text{ and } E|\eta_1(\Lambda)|^3 = \prod_{i=1}^k P(b_i),$$

(2)

$$Var(\eta_1(\Lambda)) = P(\eta_1(\Lambda) = 1) - (P(\eta_1(\Lambda) = 1))^2 = \prod_{i=1}^{k} P(b_i)\left(1 - \prod_{i=1}^{k} P(b_i)\right),$$

(3) For $2 \le l \le k$,

$$E\eta_1(\Lambda)\eta_l(\Lambda) = P(\eta_1(\Lambda) = 1, \eta_l(\Lambda) = 1) = \begin{cases} \prod_{i=1}^{l-1} P(b_i) \prod_{j=1}^{k} P(b_j) & \text{if } b_1 \cdots b_{k-l+1} = b_l \cdots b_k \\ 0 & \text{otherwise,} \end{cases}$$

$$E\eta_1(\Lambda) E\eta_l(\Lambda) = \left(\prod_{i=1}^{k} P(b_i)\right)^2, \text{ and}$$

(4) $Cov(\eta_1(\Lambda), \eta_l(\Lambda)) \equiv 0$, for $l \ge k+1$.

Further, by the definition of overlap counting, we have

$$X_n^o(\Lambda) = \sum_{i=1}^{n} \eta_i(\Lambda). \tag{16}$$

It follows from the above Results (1)–(4) and Eq. 16 that

$$Var(X_n^o(\Lambda)) = \sum_{i=k}^{n} Var(\eta_i(\Lambda)) + 2\sum_{i=k}^{n}\sum_{l=i+1}^{n} Cov(\eta_i(\Lambda), \eta_l(\Lambda))$$
$$= (n-k+1)\left[\prod_{i=1}^{k} P(b_i)\left(1 - \prod_{i=1}^{k} P(b_i)\right) + 2\sum_{l=2}^{k} Cov(\eta_1(\Lambda), \eta_l(\Lambda))\right],$$

where $Cov(\eta_1(\Lambda), \eta_l(\Lambda))$, $2 \le l \le k$, are given by Eq. 15. Hence it follows from Lemma 1, the Hoeffding and Robbins' central limit theorem for $(k-1)$th order dependent random variables, that, for any $x \in R$,

$$P\left(X_n^o(\Lambda) \le (n-k+1)E\eta_1(\Lambda) + x\sqrt{Var(X_n^o(\Lambda))}\right) \to \Phi(x), \quad \text{as } n \to \infty,$$

where $\Phi(x)$ is the cumulative distribution of the standard normal. □

In view of our proof of Theorem 3, we can extend the above results to show asymptotic normality for the distribution of the number of success runs R_s in a sequence of Bernoulli trials. Let us define the indicator functions

$$I_s(X_i) = \begin{cases} 1 \text{ if } X_i = S \text{ and } X_{i-1} = F \\ 0 \text{ otherwise,} \end{cases}$$

with $P(X_i = S) = p$ and $q = 1 - p$. Note that

$$R_s = \sum_{i=1}^{n} I_s(X_i).$$

Then the following theorem holds.

Theorem 4 *If* $\{X_i\}$ *is a sequence of Bernoulli trials, then*

$$\frac{R_s - ER_s}{\sqrt{n}} \xrightarrow{\mathcal{L}} N(0, \sigma^2), \tag{17}$$

where $\sigma^2 = pq(1 - 2pq)$.

The proof of Theorem 4 is similar to that of Theorem 3, and is omitted here.

In the following, we would like to provide several technical remarks on conditions, extensions, and other alternative proofs.

Remark 1 If the sequence $\{X_i\}$ consists of mth order Markov dependent multi-state trials and Λ is a simple pattern, then with minor modifications to our approach, it can still be shown that the random variable $X_n(\Lambda)$ remains asymptotically normally distributed.

Remark 2 If the size of the transition probability substochastic matrix $\boldsymbol{N}$ of the imbedded Markov chain is so large that the inverse of the matrix $(\boldsymbol{I} - \boldsymbol{N})$ is hard to obtain, then the mean μ_w and variance σ_w^2 used in Theorem 1 can also be obtained through alternative methods, as given by Chang (2005).

Remark 3 The mean and variance of $X_n(\Lambda)$ can also be approximated by various general methods, such as the Stein–Chen method (see Barbour and Chryssaphinou 2001; Barbour et al. 1992; Rinott and Rotar 2000; Stein 1986). The key point of our approach is that the finite Markov chain imbedding technique readily provides the exact probability, the mean, and the variance of $X_n(\Lambda)$ even when Λ is a compound pattern.

Remark 4 The final steps in Theorems 2 and 3 relied on results from renewal theory and from Hoeffding and Robbins, respectively, results which can be replaced by similar ones from other methods. For example, Theorems 2 and 3 can be proved via the invariance principle (see Billingsley 1968, or McLeish 1974).

Remark 5 Note that the number of runs of size 2, under overlap counting, in a sequence of multi-state trials can be written as a sum of n index functions:

$$R_n = 1 + \sum_{i=2}^{n} I(X_i),$$

where the index function $I(X_i) = 1$ if $X_i \neq X_{i-1}$, and $I(X_i) = 0$ otherwise. In view of Theorem 4, it is clear that the technique used in Theorem 3 can also be extended to prove that the distribution of R_n is asymptotically normal in general.

3 Bounds and Approximations

There are many bounds and approximations for the tail probability that were obtained via Poisson or compound Poisson approximations using the Stein–Chen

method (see Chen 1975; Stein 1986; Barbour et al. 1992; Rinott and Rotar 2000). The recent article by Barbour and Chryssaphinou (2001) gave an excellent review of compound Poisson approximations and the Stein-Chen method, focusing on applications to various fields. Basically, given a pattern Λ, the distribution of the number of patterns Λ in n trials, $X_n(\Lambda)$, can be approximated by a Poisson distribution $Po(\lambda)$ or a compound Poisson distribution $CP(\lambda, \mu)$ in the following sense: as $n \to \infty$

$$d(\mathcal{L}(X_n(\Lambda)), Po(\lambda)) \leq h_n(p) \tag{18}$$

or

$$d(\mathcal{L}(X_n(\Lambda)), CP(\lambda, \mu)) \leq h_n(p), \tag{19}$$

where $d(\cdot, \cdot)$ represents some measure of distance, such as Kolmogorov distance $d_K(\cdot, \cdot)$ or total variation distance $d_{TV}(\cdot, \cdot)$, $\mathcal{L}$ stands for the distribution of the random variable X, and $h_n(p)$ denotes the error function of the approximation with $h_n(p) \to 0$ as $n \to \infty$. For instance, Arratia et al. (1990) showed that

$$d_{TV}(\mathcal{L}(Z_n), CP(\lambda, \mu)) \leq \sum_{i=1}^{n} p_i^2,$$

where the sequence of random variables

$$Z_n = \sum_{i=1}^{n} \sum_{j \geq 1} X_{ij}$$

is the sum over a double array of indicator variables X_{ij} with $p_i \to 0$ and $np_i \to \infty$ as $n \to \infty$.

For given a simple or compound pattern Λ, approximating the distribution of $X_n(\Lambda)$ by a Poisson or a compound Poisson distribution often requires a strong condition that the probability of the pattern Λ occurring, $P(\Lambda)$, is extremely small, or more precisely that $P(\Lambda) \to 0$ at a certain rate as $n \to \infty$. For example, the reliability, $R_{n,k}$, of a consecutive-k-out-of-n:F system can be approximated by the Poisson probability $R_{n,k} \sim \exp\{\mu\}$, which requires the condition $np^k \to \mu$ as $n \to \infty$, where p is the failure probability of a component and a function of n. Barbour et al. (1995) provided the following compound Poisson approximation for the reliability $R_{n,k}$:

$$|R_{n,k} - \exp\{-\mu_n\}| \leq \epsilon_n \tag{20}$$

where $\epsilon_n = p^{2k}[(6k-5)n - (k-1)(13k-9)]$,

$$\begin{aligned}\mu_n = {} & (n-k+1)p^k - (n-k)p^{k+1} - 2(n-k+2)p^{2k-1} \\ & + 2(n-2)p^{2k} + \frac{2k(n-k+2)}{2k-1}p^{3k-2} - (n+k-4)p^{3k-1} \\ & + [2k(n-k+2) - 2(2k-1)(n-k+1)p \\ & + 2(k-1)(n-k)p^2]p^{k-1}h(p,k),\end{aligned}$$

$h(q,k) \equiv \sum_{i=k}^{2k-2} q^i/i$, and $q = 1-p$. They also provided numerical comparisons of the compound Poisson approximation given by Eq. 20 and the exponential bound

$$(1-p^k)^{n-k+1} \leq R_{n,k} \leq (1-qp^k)^{n-k} \tag{21}$$

given by Fu (1986); Koutras and Papastavridis (1993); Cai (1994); Muselli (2000); Dwyer (2004), and recently by Chen and Huo (2006). Note that for the consecutive-k-out-of-n:F system, the transition probability substochastic matrix $\boldsymbol{N}$ of the imbedded Markov chain is primitive, and the reliability $R_{n,k}$ converges to zero exponentially in the sense that there exists a constant c such that

$$\lim_{n\to\infty} \frac{R_{n,k}}{c\rho^n} = 1, \tag{22}$$

where ρ is the largest eigenvalue of the transition probability substochastic matrix $\boldsymbol{N}$. The quantity $c\rho^n$ provides an excellent numerical approximation for the reliability $R_{n,k}$, when n is moderate or large.

For a simple pattern Λ, the transition probability matrix $\boldsymbol{N}$ associated with the waiting time random variable $W(\Lambda)$ is also a primitive substochastic matrix, and the tail probability $\alpha_n(\Lambda, x)$ has the following large deviation limit

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n(\Lambda, x) = -\beta(\Lambda, x) \tag{23}$$

where $\beta(\Lambda, x) = -\log \rho$, and $0 < \rho < 1$. Note that the tail probability $\alpha_n(\Lambda, x)$ converges to zero exponentially. Hence it cannot be properly approximated by a normal, Poisson or compound Poisson probability p_n, in the manner

$$\alpha_n(\Lambda, x) \cong p_n + \varepsilon_n, \tag{24}$$

unless the error term ε_n converges to zero at a rate faster than $\alpha_n(\Lambda, x)$, but this is often not the case. Bahadur (1971) gave an example using a sequence of Bernoulli trials with the simple pattern $\Lambda = S$, and showed the difference between the normal and the large deviation approximations using the following relation:

$$P(Y_1 + \cdots + Y_n \geq nx) \cong c_n \exp\{-n\beta(\Lambda, x)\}, \tag{25}$$

where Y_i are *i.i.d.* Bernoulli trials and $0 < p < x < 1$; the normal approximation yields the exponential rate, the wrong one,

$$\beta(\Lambda, x) \cong \frac{(x-p)^2}{2p(1-p)},$$

but the correct rate should be

$$\beta(\Lambda, x) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}. \tag{26}$$

Large deviation approximations have been used to estimate the tail probability for many runs statistics in a variety of applications, especially in the area of bioinformatics and genomic studies (e.g. Waterman, 1995, Dwyer, 2004 and Nuel, 2004). In the recent paper by Nuel (2005), the author developed an efficient algorithm for fast p-value computation and for large deviation approximations using finite Markov chain imbedding with applications to local score and pattern statistics in DNA studies. Based on numerical computations, he argued that the proposed algorithm is far easier than that of Robin and Daudin (1999) to implement, very efficient, more numerically stable, and less memory intensive; he also suggested that there is no need to use approximations whenever the exact probability can be obtained through a fast algorithm.

There are many theoretical results for approximating the tail probability, but there are only limited numerical comparisons for the performance of these types of approximations, especially at the practically important level $\alpha_n(\Lambda, x) \sim 0.05$ or 0.01. We believe a comprehensive numerical comparison study on the performance of the various approximations in realistic applications is urgently needed.

Acknowledgements This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, and the Canada Research Chairs Program.

References

R. Arratia, L. Goldstein, and L. Gordon, "Poisson approximation and the Chen–Stein method," *Statistical Science* vol. 5 pp. 403–434, 1990.
N. Balakrishnan and M. V. Koutras, *Runs and Scans with Applications*, Wiley: New York, 2002.
R. R. Bahadur,"Some limit theorems in statistics," *Regional Conference Series in Applied Mathematics, SIAM*, 1971.
A. D. Barbour and O. Chryssaphinou, "Compound Poisson approximation: A user's guide," *Annals of Applied Probability* vol. 11 pp. 964–1002, 2001.
A. D. Barbour, O. Chryssaphinou, and M. Roos, "Compound Poisson approximation in reliability theory," *IEEE Transactions on Reliability* vol. 44 pp. 393–402, 1995.
A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*, Oxford Studies in Probability, Oxford University Press: New York, 1992.
P. Billingsley, *Convergence of Probability Measures*, Wiley: New York, 1968.
J. Cai, "Reliability of a large consecutive-k-out-of-n:F system with unequal component reliability," *IEEE Transactions on Reliability* vol. 43 pp. 107–111, 1994.
Y. M. Chang, "Distribution of waiting time until the *r*-th occurrence of a compound pattern," *Statistics & Probability Letters* vol. 75 pp. 29–38, 2005.
L. H. Y. Chen, "Poisson approximation for dependent trials," *Annals of Probability* vol. 3 pp. 534–545, 1975.
J. Chen and X. Huo, "Distribution of the length of the longest significance run on a Bernoulli net and its applications," *Journal of the American Statistical Association* vol. 473 pp. 321–331, 2006.
V. M. Dwyer, "The influence of microstructure on the probability of early failure in aluminum-based interconnects," *Journal of Applied Physics* vol. 96 pp. 2914–2922, 2004.
W. Feller, *An Introduction to Probability Theory and its Applications* (Vol. I, 3rd ed.) Wiley: New York, 1968.
J. C. Fu, "Bounds for reliability of large consecutive-*k*-out-of-*n*:F systems with unequal component reliability," *IEEE Transactions on Reliability* vol. 35 pp. 316–319, 1986.
J. C. Fu and W. Y. W. Lou, *Distribution Theory of Runs and Patterns and Its Applications: A Finite Markov Chain Imbedding Approach*, World Scientific: New Jersey, 2003.
J. C. Fu and W. Y. W. Lou, "Waiting time distributions of simple and compound patterns in a sequence of *R*-th order Markov dependent multi-state trials," *Annals of the Institute of Statistical Mathematics* vol. 28 pp. 291–310, 2006.
W. Hoeffding and H. Robbins, "The central limit theorem for dependent random variables," *Duke Mathematical Journal* vol. 15 pp. 773–780, 1948.
M. V. Koutras, "On a waiting time distribution in a sequence of Bernoulli trials," *Annals of the Institute of Statistical Mathematics* vol. 48 pp. 789–806, 1996.
M. V. Koutras and S. G. Papastavridis, "Application of the Stein–Chen method for bounds and limit theorems in the reliability of coherent structures," *Naval Research Logistic* vol. 40 pp. 617–631, 1993.
M. V. Koutras, "Waiting time distributions associated with runs of fixed length in two-state Markov chains," *Annals of the Institute of Statistical Mathematics* vol. 49 pp. 123–139, 1997.
P. A. MacMahon, *Combinatory Analysis*, Cambridge University Press: London, 1915.
D. L. McLeish, "Dependent central limit theorems and invariance principles," *Annals of Probability* vol. 2 pp. 620–628, 1974.
M. Muselli, "New improved bounds for reliability of consecutive-*k*-out-of-*n*:*F* systems," *Journal of Applied Probability* vol. 37 pp. 1164–1170, 2000.

G. Nuel, *Fast p-value Computations Using Finite Markov Chain Imbedding: Application to Local Score and Pattern Statistics*, TR223 Université d'Evry Val d'Essonne, 2005.
G. Nuel, "LD-SPatt: Large deviations statistics for patterns on Markov chains," *Journal of Computational Biology* vol. 11 pp. 1023-1033, 2004.
Y. Rinott and V. Rotar, "Normal approximations by Stein's method," *Decisions in Economics and Finance* vol. 23 pp. 15–29, 2000.
J. Riordan, *An Introduction to Combinatorial Analysis*, Wiley: New York, 1958.
S. Robin and J. J. Daudin, "Exact distribution of word occurrences in a random sequence of letters," *Journal of Applied Probability* vol. 36 pp. 179–193, 1999.
C. Stein, "Approximate computation of expectations," Institute of Mathematical Statistics Lecture Notes—Monograph Series 7, *Institute of Mathematical Statistics*, Hayward, CA, 1986.
M. S. Waterman, *Introduction to Computational Biology*, Chapman and Hall: New York, 1995.