Journal of Statistical Distributions and Applications
a SpringerOpen Journal

RESEARCH Open Access

Approximating the distributions of runs and patterns

Brad C Johnson*† and James C Fu†

*Correspondence:
brad.johnson@umanitoba.ca
†Equal Contributors
Department of Statistics, University of Manitoba, Winnipeg, Canada

Abstract

The distribution theory of runs and patterns has been successfully used in a variety of applications including, for example, nonparametric hypothesis testing, reliability theory, quality control, DNA sequence analysis, general applied probability and computer science. The exact distributions of the number of runs and patterns are often very hard to obtain or computationally problematic, especially when the pattern is complex and n is very large. Normal, Poisson and compound Poisson approximations are frequently used to approximate these distributions. In this manuscript, we (i) study the asymptotic relative error of the normal, Poisson, compound Poisson and finite Markov chain imbedding and large deviation approximations; and (ii) provide some numerical studies to comparing these approximations with the exact probabilities for moderately sized n. Both theoretical and numerical results show that, in the relative sense, the finite Markov chain imbedding approximation performs the best in the left tail and the large deviation approximation performs best in the right tail.

AMS Subject Classification: Primary 60E05; Secondary 60J10

Keywords: Finite Markov chain imbedding; Rate functions; Multi-state trials; Runs and patterns

Introduction and notation

Let $\{X_i\}_{i=1}^n$ be a sequence of m-state trials ($m \geq 2$) taking values in the set $\mathcal{S} = \{s_1, \ldots, s_m\}$ of m symbols. For simplicity, $\{X_i\}_{i=1}^n$ will be denoted $\{X_i\}$ and n will be allowed to be ∞. A *simple pattern* $\Lambda = s_{i_1} s_{i_2} \cdots s_{i_\ell}$, of length ℓ, is the juxtaposition of ℓ (not necessarily distinct) symbols from $\mathcal{S}$. Given a simple pattern Λ, we let $X_n(\Lambda)$ denote the number of either non-overlapping or overlapping occurrences of Λ in the sequence $\{X_i\}_{i=1}^n$, where the method of counting will be made clear by the context. The waiting time $W(\Lambda, x)$ until the x'th occurrence of the simple pattern Λ in $\{X_i\}_{i=1}^n$ is thus defined by

$$W(\Lambda, x) = \inf\{n \in \mathbb{N} : X_n(\Lambda) = x\},$$

and, by convention, the waiting time for the first occurrence is denoted $W(\Lambda) = W(\Lambda, 1)$. Finally, we define the inter arrival times

$$W_i(\Lambda) = W(\Lambda, i) - W(\Lambda, i-1), \qquad \text{for } i = 1, 2, \ldots,$$

where $W(\Lambda, 0) := 0$.

We say that two patterns Λ_1 and Λ_2 are distinct if neither Λ_1 appears in Λ_2 nor Λ_2 appears in Λ_1. If $\Lambda_1, \ldots, \Lambda_r$ are pairwise distinct simple patterns, we define the compound pattern $\Lambda = \bigcup_{i=1}^{r} \Lambda_i$, where an occurrence of any Λ_i is considered an occurrence of Λ. For a compound pattern $\Lambda = \Lambda_1 \cup \cdots \cup \Lambda_r$, we similarly define

$$X_n(\Lambda) = \sum_{j=1}^{r} X_n(\Lambda_j).$$

The waiting times $W(\Lambda, x)$, $W(\Lambda)$ and $W_i(\Lambda)$ are then defined as above, and often referred to as *sooner* waiting times.

From these definitions it is easy to see that, for any simple or compound pattern Λ, x and n, the events $\{X_n(\Lambda) < x\}$ and $\{W(\Lambda, x) > n\}$ are equivalent and hence

$$\mathbb{P}\{X_n(\Lambda) < x\} = \mathbb{P}\{W(\Lambda, x) > n\}, \tag{1}$$

which provides a convenient way of studying the exact and approximate distribution of $X_n(\Lambda)$ through the waiting time distributions of $W(\Lambda, x)$.

Throughout this paper, unless specified otherwise, we assume that the trials $\{X_i\}$ are either independent and identically distributed (i.i.d.) or first order Markov dependent; the pattern Λ is either simple or compound; and the counting of occurrences of Λ is in a non-overlapping fashion.

The distribution of the number of runs and patterns in a sequence of multi-state trials or random permutations of a set of integers have been successfully used in various fields in applied probability, statistics and discrete mathematics. Examples include reliability theory, quality control, DNA sequence analysis, psychology, ecology, astronomy, nonparametric tests, successions, and the Eulerian and Simon-Newcomb numbers (the latter 3 being defined for permutations). Two recent books, Balakrishnan and Koutras (2002) and Fu and Lou (2003), provide some scope of the distribution theory of runs and patterns and Martin et al. (2010) and Nuel et al. (2010) provides some extensions to sets of sequences.

Given a pattern Λ, the exact distribution of $X_n(\Lambda)$ traditionally has been determined using combinatoric analysis on a case by case basis. The formulae for these distributions are often very complex and computationally problematic. Even for many simple patterns, their distributions in terms of combinatoric analysis remains unknown, especially when the $\{X_i\}$ are Markov dependent multi-state trials.

The waiting time $W(\Lambda)$ for the first occurrence of certain types of runs and patterns have been studied by many authors. See, for example, Blom and Thorburn (1982), Gerber and Li (1981), Schwager (1983), and Solov'ev (1966). More recently, Fu and Koutras (1994) developed a method for determining the exact distributions of $X_n(\Lambda)$ and $W(\Lambda)$ for any simple or compound Λ in either i.i.d. or Markov dependent trials (see also Fu and Lou 2003). The method was referred to as the Finite Markov Chain Imbedding (FMCI) technique, which can be easily described as follows: given a simple or compound pattern Λ, there exists a finite Markov chain $\{Y_i\}$ defined on a finite state space, say $\Omega = \{1, \ldots, d, \alpha\}$, with an absorbing state α and transition probability matrix of the form

$$\mathbf{P} = \begin{array}{c} \Omega - \alpha \\ \alpha \end{array} \left[\begin{array}{c|c} \mathbf{N} & \mathbf{c} \\ \hline \mathbf{0} & 1 \end{array}\right], \tag{2}$$

where $\boldsymbol{c}$ is a column vector. The distribution of the waiting time for Λ is given by

$$\mathbb{P}\{W(\Lambda) = n\} = \boldsymbol{\xi}_0 \mathbf{N}^{n-1}(\mathbf{I} - \mathbf{N})\mathbf{1}' \tag{3}$$

where $\boldsymbol{\xi}_0$ is the initial distribution, $\mathbf{N}$ is the *essential transition probability matrix* (i.e. the sub-stochastic matrix consisting of only the transient states of $\{Y_i\}$) as defined in (2), $\mathbf{I}$ is a $d \times d$ identity matrix and $\mathbf{1} = (1, 1, \ldots, 1)$ is a $1 \times d$ row-vector. Furthermore, the random variable $X_n(\Lambda)$, the number of occurrences of Λ in $\{X_i\}$, is also finite Markov chain imbeddable and its distribution is given by

$$\mathbb{P}\{X_n(\Lambda) < x\} = \mathbb{P}\{W(\Lambda, x) > n\} = \boldsymbol{\xi}_0 \mathbf{N}_x^n \mathbf{1}', \tag{4}$$

where the essential transition probability matrix $\mathbf{N}_x$ has the form

$$\mathbf{N}_x = \begin{bmatrix} \mathbf{N} & \mathbf{C} & & & \\ & \mathbf{N} & \mathbf{C} & \mathbf{0} & \\ & & \ddots & \ddots & \\ & \mathbf{0} & & \mathbf{N} & \mathbf{C} \\ & & & & \mathbf{N} \end{bmatrix}, \tag{5}$$

the matrix $\mathbf{N}$ is given by (2), and the matrix $\mathbf{C}$ defines the "continuation" transition probabilities from one occurrence to the next and depends on $\mathbf{c}$ in (2).

If the pattern Λ is long and complex and n is very large, then the computation of $\mathbb{P}\{X_n(\Lambda) = x\}$ can become problematic and, to overcome this problem, various asymptotic approximations have been developed for these probabilities.

In real applications, if the exact distribution is not available or is hard to compute, it is important to know which approximations perform well and are easy to compute. Furthermore, it is important to know how these approximations perform with respect to each other and the exact distribution from both a theoretical and numerical standpoint. The aims of this manuscript are two-fold: (i) we first study the asymptotic relative error of the normal, Poisson (or compound Poisson), and FMCI approximations with respect to the exact distribution; and (ii) we then provide a numerical study of these three approximations with the exact probabilities in cases where x is fixed and $n \to \infty$ and when n is fixed and x varies. As an important byproduct, the FMCI technique allows the normal and Poisson approximations to be applied in more cases, for example, the distribution of compound patterns and patterns in Markov dependent trials.

The approximations

Normal approximation

The normal approximation is one of the most popular for approximating the distribution of the number of runs or patterns $X_n(\Lambda)$ in Statistics. In general, when Λ is simple or compound, the trials are i.i.d., and the counting is non-overlapping, by appealing to (1) and renewal arguments, it has been shown that $X_n(\Lambda)$ is asymptotically normally distributed (cf. Fu and Lou 2007; Karlin and Taylor 1975). The form of the approximation is

$$\lim_{n\to\infty} \mathbb{P}\left\{ \frac{X_n(\Lambda) - n/\mu_W}{\sqrt{n\sigma_W^2 \mu_W^{-3}}} \le u \right\} = \Phi(u), \tag{6}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function and μ_W and σ_W^2 are the mean and variance of $W(\Lambda)$ respectively, which are given by

$$\mu_W = \boldsymbol{\xi}_0(\mathbf{I} - \mathbf{N})^{-1}\mathbf{1}', \quad \text{and} \tag{7}$$

$$\sigma_W^2 = \boldsymbol{\xi}_0(\mathbf{I} + \mathbf{N})(\mathbf{I} - \mathbf{N})^{-2}\mathbf{1}' - \mu_W^2. \tag{8}$$

Given a pattern Λ, it is well known that the mean μ_W and the variance σ_W^2 are difficult to obtain via combinatoric arguments, especially when Λ is a compound pattern or the trials are Markov dependent. For example, as pointed out in Karlin (2005) and Kleffe and Borodovski (1992), approximate values of μ_W and σ_W^2 must sometimes be used. Since $W(\Lambda)$ is finite Markov chain imbeddeble, (7) and (8), provide the exact values.

The limit in (6) is appropriate when the sequence of inter arrival times $\{W_i(\Lambda)\}$ are i.i.d., which is the case for simple and compound patterns when the $\{X_i\}$ are i.i.d. and counting is non-overlapping. When occurrences of Λ correspond to a delayed renewal process, which can occur for Markov dependent trials and/or overlapping counting, we could use the mean and variance of $W_2(\Lambda)$ for the normalizing constants, which are easily obtained by modifying $\boldsymbol{\xi}_0$ in (7) and (8). Even more general cases can be handled by making use of a functional central limit theorem for Markov chains (see, for example, (Meyn and Tweedie 1993, §17.4) and (Asmussen 2003, Theorem 7.2, pg. 30) for the details).

Poisson and compound poisson approximations

It is well known that, in a sequence of Bernoulli(p) trials, if $np \to \lambda$ as $n \to \infty$, then the probability of k successes in n trials can be approximated by a Poisson probability with parameter λ, denoted $\mathscr{P}(\lambda)$. This idea has been extended to certain patterns Λ and, under certain conditions, the distribution of $X_n(\Lambda)$ can be approximated by a Poisson distribution with parameter μ_n in the sense that

$$d_{\mathrm{TV}}(\mathscr{L}(X_n(\Lambda)), \mathscr{P}(\mu_n)) < \varepsilon_n, \tag{9}$$

where $\mathscr{L}(\cdot)$ denotes the distribution (law) of a random variable and $d_{\mathrm{TV}}(\cdot,\cdot)$ denotes the total variation distance.

The primary tool used to obtain μ_n and the bound ε_n is the Stein-Chen method (Chen 1975), and this method has been refined by various authors Arratia et al. (1990), Barbour and Eagleson (1983), Barbour and Eagleson (1984), Barbour and Eagleson (1987), Barbour and Hall (1984), Godbole (1990a), Godbole (1990b), Godbole (1991), Godbole and Schaffner (1993), and Holst et al. (1988). This method has also been extended to compound Poisson approximations for the distributions of runs and patterns and Barbour and Chryssaphinou (2001) provides an excellent theoretical review of these approximations.

In practice, $\mu_n = \mathbb{E}X_n(\Lambda)$ or the expectation of a closely related run statistic is used (cf. Balakrishnan and Koutras 2002, §5.2.3) so that, in the former case,

$$\mathbb{P}\{X_n(\Lambda) = x\} \approx \frac{(\mathbb{E}X_n(\Lambda))^x}{x!} \exp\{-\mathbb{E}X_n(\Lambda)\}. \tag{10}$$

Finding $\mathbb{E}X_n(\Lambda)$ and the bound ε_n is usually done on a case by case basis. For the mathematical details, the books (Barbour et al. 1992a) and (Balakrishnan and Koutras 2002) are recommended.

Let $\mathscr{P}_c(\lambda, \nu)$ denote the compound Poisson distribution, that is, the distribution of the random variable $\sum_{j=1}^{M} Y_j$ where the random variable M has a Poisson distribution with parameter λ and the Y_j are i.i.d. having distribution ν. A compound Poisson distribution

for approximating nonnegative random variables was suggested in Barbour et al. (1992b) (see also Barbour et al. (1995,1996)). The approximation is formulated similarly to the Poisson approximation:

$$d_{\mathrm{TV}}(\mathscr{L}(X_n(\Lambda)), \mathscr{P}_c(\lambda, \nu)) < \varepsilon_n. \tag{11}$$

The distribution of $N_{n,k}$, the number of non-overlapping occurrences of k consecutive successes in n i.i.d. Bernoulli trials, is one of the most important in this area and one of the most studied in the literature. Reversing the roles of S (success) and F (failure), the reliability of consecutive-k-out-of-n system, denoted $C(k, n:F)$, is given by $\mathbb{P}\{N_{n,k} = 0\}$. Even in this simple case (i.e. $\Lambda = SS\cdots S$), there are several ways to apply the Poisson approximation techniques. For example, (Godbole 1991, Theorem 2) shows that approximating $N_{n,k}$ with a $\mathscr{P}(\mathbb{E}N_{n,k})$ distribution works well if certain conditions hold. Godbole and Schaffner (Godbole and Schaffner 1993, pg. 340) suggests an improved Poisson approximation for word patterns.

The primary difficulty in applying the Poisson approximation is the determination of the optimal parameter μ_n, which is higly dependent on the structure of the pattern Λ. In particular, if Λ is long and has several uneven overlapping sub-patterns, then finding μ_n by their method can be very tedious. In the sequel, we show that even the (asymptotic) best choice for μ_n for Poisson approximations does not perform well in the relative sense.

FMCI approximations

Approximations based on the FMCI approach depend on the spectral decomposition of the essential transition probability matrix $\mathbf{N}$.

Let $\mathbf{N}$ be a $w \times w$ essential transition probability matrix associated with a finite Markov chain $\{Y_n : n \geq 0\}$ corresponding to the distribution of the waiting time $W(\Lambda)$. Let $1 > \lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_w|$ denote the ordered eigenvalues of $\mathbf{N}$, repeated according to their algebraic multiplicities, with associated (right) eigenvectors $\boldsymbol{\eta}_1', \boldsymbol{\eta}_2', \cdots, \boldsymbol{\eta}_w'$. When the geometric multiplicity of λ_i is less than its algebraic multiplicity, we will use vectors of 0's for the unspecified eigenvectors. The fact that λ_1 can be taken as a positive real number and that $\boldsymbol{\eta}_1$ can be taken to be non-negative are consequences of the Perron-Frobenious Theorem for non-negative matrices (*cf.* Seneta 1981).

Definition 1. We will say that $\{Y_n : n \geq 0\}$, or equivalently, $\mathbf{N}$, satisfies the *FMCI Approximation Conditions* if

(i) there exists constants $a_1, \ldots, a_w$ such that

$$\mathbf{1}' = \sum_{i=1}^{w} a_i \boldsymbol{\eta}_i', \tag{12}$$

(ii) λ_1 has algebraic multiplicity g and $\lambda_1 > |\lambda_j|$ for all $j > g$.

Verifying these conditions is usually straightforward. They certainly hold if $\mathbf{N}$ is irreducible and aperiodic, but also hold in many other cases as well. For example, (12) requires only that $\mathbf{1}'$ is in the linear space spanned by $\{\boldsymbol{\eta}_1', \boldsymbol{\eta}_2', \cdots, \boldsymbol{\eta}_w'\}$, which can hold even when $\mathbf{N}$ is defective (not diagonizable). Condition (ii) requires that the communication classes corresponding λ_1 are aperiodic. That is, if Ψ is a communication class and $\mathbf{N}[\,\Psi]$ corresponds to the substocastic matrix $\mathbf{N}$ restricted to the states in Ψ, with largest eigenvalue $\lambda_1[\,\Psi]$,

then all Ψ such that $\lambda_1[\Psi] = \lambda_1$ should be aperiodic. We also mention that the algebraic multiplicity of λ_1 is the number of communication classes Ψ such that $\lambda_1[\Psi] = \lambda_1$.

Fu and Johnson (2009) give the following theorem.

Theorem 1. *Let $\{X_i\}$ be a sequence of i.i.d. trials taking values in $\mathcal{S}$, let Λ be a simple pattern of length ℓ with $d \times d$ essential transition probability matrix $\mathbf{N}$ and let $X_n(\Lambda)$ be the number of non-overlapping occurrences of Λ in $\{X_i\}$. If $\mathbf{N}$ satisfies the FMCI approximation conditions then, for any fixed $x \geq 0$,*

$$\mathbb{P}\{X_n(\Lambda) = x\} \sim a^{x+1}\binom{n - x(\ell-1)}{x}(1-\lambda_1)^x\lambda_1^{n-x}, \tag{13}$$

where $a = \sum_{j=1}^{g} a_j(\boldsymbol{\xi}_0\boldsymbol{\eta}_j')$. If $g = 1$, as is usually the case, then $a = a_1(\boldsymbol{\xi}_0\boldsymbol{\eta}_1')$.

Given a pattern Λ, the approximation in (13) requires finding the Markov chain imbedding associated with the waiting time $W(\Lambda)$, the essential transition probability matrix $\mathbf{N}$ as well as its eigenvalues and associated eigenvectors. Usually, these steps are rather simple and can be easily automated together with (13). Even for very large n and large ℓ, say $n = 1{,}000{,}000$ and $\ell = 50$, the CPU time is negligible. Fu and Johnson (2009) also provide details on extending these results to compound patterns, overlapping counting and Markov dependent trials.

For the purpose of comparing these approximations, we prefer to write (13) as

$$\mathbb{P}\{X_n(\Lambda) = x\} \sim a^{x+1}\left(\frac{1-\lambda_1}{\lambda_1}\right)^x\binom{n - x(\ell-1)}{x}\exp\{n\ln\lambda_1\} \tag{14}$$

Note that the approximation havs three parts: a constant part; a polynomial in n of degree x; and a third (dominant) part which converges to 0 exponentially fast as $n \to \infty$.

More precisely, the FMCI approximation in (13) may be written as

$$\begin{aligned}\mathbb{P}\{X_n(\Lambda) = x\} = a^{x+1}&\left(\frac{1-\lambda_1}{\lambda_1}\right)^x\binom{n - x(\ell-1)}{x}\\ &\times\exp\{n\ln\lambda_1\}\left[1 + o\left(\left|\frac{\lambda_{g+1}}{\lambda_1}\right|^{n/(x+1)-\ell}\right)\right].\end{aligned} \tag{15}$$

Since $|\lambda_{g+1}| < \lambda_1$, the term $\left|\lambda_{g+1}/\lambda_1\right|^{n/(x+1)-\ell}$ tends to 0 exponentially as $n \to \infty$ and hence is negligible if $n/(x+1) - \ell$ is moderate or large (say ≥ 50).

Large deviation approximation

Fu et al. (2012) provide the following large deviation approximation for right-tail probabilities for the number of non-overlapping occurrences for simple patterns Λ. The reasons for providing only the right-tail large deviation approximation are (i) all of the above mentioned approximations fail to approximate the extreme right-tail probabilities and (ii) the FMCI approximation provides an accurate approximation for left-tail probabilities.

Theorem 2. *Let $\varepsilon = x\mu_W^2/(1 + x\mu_W)$ and let*

$$\varphi_W(t) = 1 + (e^t - 1)\boldsymbol{\xi}(\mathbf{I} - e^t\mathbf{N})^{-1}\mathbf{1}', \tag{16}$$

be the moment generating function of $W(\Lambda)$. *Then*

$$\mathbb{P}\{X_n(\Lambda) \geq \mathbb{E}X_n(\Lambda) + nx\} = e^{-n\beta(\varepsilon,\Lambda)} \frac{1}{\sqrt{n}} \left\{ b_0 + b_1 n^{-1} + \cdots + b_m n^{-m} + \mathcal{O}(n^{-m-1}) \right\}, \tag{17}$$

where

$$\beta(x,\Lambda) = \left(\frac{1}{\mu_W} + x\right) h(\varepsilon,\tau) = \left(\frac{1}{\mu_W} + x\right) \left[-\frac{\tau \mu_W}{1 + x\mu_W} - \ln \varphi_{W(\Lambda)}(-\tau) \right], \tag{18}$$

$h(\varepsilon,t) = \varepsilon t - \ln \varphi_{\mu_W - W(\Lambda)}(t)$, τ *is the solution to* $h'(\varepsilon,\tau) = 0$, *and*

$$\begin{aligned} b_0 &= \frac{1}{\sigma\tau\sqrt{2\pi(\mu^{-1}+x)}} \\ b_1 &= \frac{1}{\sigma\tau\sqrt{2\pi(\mu^{-1}+x)^3}} \left\{ -\frac{1}{\sigma^2\tau^2} + \frac{h^{(3)}(\varepsilon,\tau)}{2\tau\sigma^4} - \frac{h^{(4)}(\varepsilon,\tau)}{8\sigma^4} - \frac{5(h^{(3)}(\varepsilon,\tau))^2}{24\sigma^6} \right\} \\ \sigma &= \sqrt{-h''(\varepsilon,\tau)}. \end{aligned} \tag{19}$$

Comparisons and relative error

For a given n, x and pattern Λ, we define the relative error of an approximation with respect to the exact probability $\mathbb{P}\{X_n(\Lambda) = x\}$ as

$$R(x : E, A) = \operatorname{sgn}(A - E) \left[\max\left(\frac{E}{A}, \frac{A}{E}\right) - 1 \right],$$

where A stands for the approximate probability and E stands for the exact probability $\mathbb{P}\{X_n(\Lambda) = x\}$. This quantity, $R(x : E, A)$, goes from $-\infty$ to ∞ and treats the importance of overestimation the same as underestimation. It is clear that $R(x : E, A) > 0$ implies that the approximation is overestimating the exact probability and that $R(x : E, A) < 0$ implies that the approximation is underestimating the exact probability. Since, for fixed x, the probability $\mathbb{P}\{X_n(\Lambda) = x\}$ converges to 0 exponentially fast as $n \to \infty$, it follows that $R(x : E, A) \to \pm\infty$ implies that the approximation tends to 0 with the wrong rate. If $R(x : E, A)$ is near 0 then the approximation is close to the exact probability $\mathbb{P}\{X_n(\Lambda) = x\}$.

Note that $R(x : E, A)$ is a function of x, n and the method of approximation used. The following theorem provides the asymptotic relative error for the Normal approximation (N), the Poisson approximation ($P(\mu_n)$) and the finite Markov chain imbedding approximation (F).

Theorem 3. *Let* $\{X_i\}$ *be a sequence of i.i.d. multi-state trials taking values in* $\mathcal{S}$ *and let* Λ *be a simple pattern defined on* $\mathcal{S}$. *Then, for every fixed x, we have,*

$$(i) \qquad \lim_{n\to\infty} R(x : E, F) = 0; \tag{20}$$

$$(ii) \qquad \lim_{n\to\infty} R(x : E, P(\mu_n)) = \begin{cases} \infty, & \text{if } \limsup_n \mu_n/n < -\ln\lambda_1; \\ c(x), & \text{if } \lim_n \mu_n/n = -\ln\lambda_1; \\ -\infty, & \text{if } \liminf_n \mu_n/n > -\ln\lambda_1; \end{cases} \tag{21}$$

$$(iii) \qquad \lim_{n\to\infty} R(x:E,N) = \begin{cases} \infty, & \text{if } \mu_W/2\sigma_W^2 \le -\ln\lambda_1; \\ \\ -\infty, & \text{if } \mu_W/2\sigma_W^2 > -\ln\lambda_1; \end{cases} \tag{22}$$

where the exact probability is computed using (4) and

$$c(x) = a^{x+1}\left(\frac{\lambda_1 - 1}{\lambda_1 \ln \lambda_1}\right)^x - 1.$$

Proof. Given a pattern Λ and x, for the finite Markov chain imbedding approximation we have

$$\lim_{n\to\infty} \frac{\mathbb{P}\{X_n(\Lambda) = x\}}{a^{x+1}\left(\frac{1-\lambda_1}{\lambda_1}\right)^x \binom{n - x(\ell-1)}{x} \exp\{n \ln \lambda_1\}} = 1$$

and hence (i) follows immediately from the definition of $R(x:E,A)$ and Theorem 1.

For the Poisson approximation we have, since $E/F \sim 1$ by (i),

$$\frac{E}{P(\mu_n)} = \frac{E}{F} \times \frac{F}{P(\mu_n)} \sim \frac{F}{P(\mu_n)}$$

and hence

$$\begin{aligned} \frac{E}{P(\mu_n)} &= \frac{\mathbb{P}\{X_n(\Lambda) = x\}}{\frac{\mu_n^x}{x!}\exp\{-\mu_n\}} \\ &\sim \frac{a^{x+1}\left(\frac{1-\lambda_1}{\lambda_1}\right)^x \binom{n - x(\ell-1)}{x} \exp\{n \ln \lambda_1\}}{\frac{\mu_n^x}{x!}\exp\{-\mu_n\}}. \end{aligned} \tag{23}$$

If $\liminf_n \mu_n/n > -\ln\lambda_1$ then $\exp\{n\ln\lambda_1 + \mu_n\}$ tends to 0 exponentially fast which overrides the polynomial term and hence $R(x:E,P(\mu_n)) \to -\infty$ as $n \to \infty$ for all fixed x. Similarly, if $\limsup_n \mu_n/n < -\ln\lambda_1$, then $R(x:E,P(\mu_n)) \to \infty$ as $n \to \infty$ for all fixed x. Furthermore, if $\lim_n \mu_n/n = -\ln\lambda_1$, then the ratio yields

$$\lim_{n\to\infty} R(x:E,P(-n\ln\lambda_1)) = a^{x+1}\left(\frac{\lambda_1 - 1}{\lambda_1 \ln \lambda_1}\right)^x - 1$$

and this completes the proof of (ii). Note also that, if $\limsup_n \mu_n/n > -\ln\lambda_1$ and $\liminf_n \mu_n/n < -\ln\lambda_1$, then $\lim_n R(x:E,P(\mu_n))$ will not exist.

For the normal approximation we have that $X_n(\Lambda)$ is approximately normal with mean n/μ_W and variance $n\sigma_W^2/\mu_W^3$ and hence

$$\mathbb{P}\{X_n(\Lambda) = x\} \approx N = \int_{x-1/2}^{x+1/2} \frac{1}{\sqrt{2\pi n\sigma_W^2\mu_W^{-3}}} \exp\left\{-\frac{(t - n/\mu_W)^2}{2n\sigma_W^2\mu_W^{-3}}\right\} dt$$

Hence, provided $n > \mu_W(x + 1/2)$, we have

$$N \le \frac{1}{\sqrt{2\pi n\sigma_W^2\mu_W^{-3}}} \exp\left\{-\frac{(x + 1/2 - n/\mu_W)^2}{2n\sigma_W^2\mu_W^{-3}}\right\}.$$

Therefore, as in the proof of (ii), we are interested in the asymptotics of F/N, which yields

$$\frac{F}{N} \sim \sqrt{\frac{2\pi n\sigma_W^2}{\mu_W^3}}\, a^{x+1} \left(\frac{1-\lambda_1}{\lambda_1}\right)^x \binom{n-x(\ell-1)}{x} \times \exp\left\{n\ln\lambda_1 + \frac{(x+1/2-n/\mu_W)^2}{2n\sigma_W^2\mu_W^{-3}}\right\}. \tag{24}$$

We may rewrite the argument of the exponential function as

$$n\left[\ln\lambda_1 + \frac{\mu_W}{2\sigma_W^2}\left(\frac{\mu_W(x+1/2)}{n} - 1\right)^2\right],$$

making it clear that (24) converges to ∞ if $\mu_W/2\sigma_W^2 \geq -\ln\lambda_1$ and 0 otherwise. Therefore, $R(x : E, N) \to \infty$ if $\mu_W/2\sigma_W^2 \geq -\ln\lambda_1$ and $R(x : E, N) \to -\infty$ if $\mu_W/2\sigma_W^2 < -\ln\lambda_1$ and the proof of (iii) is complete. □

Theorem 3 (ii) implies that asymptotically (for fixed x and $n \to \infty$), the Poisson approximation performs poorly (in the relative sense) regardless of the value μ_n used. When Λ is simple and does not have overlapping sub-patterns, taking $\mu_n = \mathbb{E}X_n(\Lambda)$ is normally recommended for the Poisson approximation (*cf.* Arratia et al. 1990). In this case, non-overlapping and overlapping counting is equivalent. The following corollary shows that, for fixed x, the Poisson approximation will (asymptotically) always overestimate the exact probability in the following sense.

Corollary 1. *Let Λ be a simple pattern defined on an i.i.d. sequence of multi-state trials. For $\mu_n = \mathbb{E}X_n(\Lambda)$, we have*

$$\lim_{n\to\infty} R(x : E, P(\mu_n)) = \infty$$

for all fixed x.

Proof. Recall that, in this case, $X_n(\Lambda)$ is a renewal process with i.i.d. inter-renewal times with mean $\mu_W = \mathbb{E}W(\Lambda)$ and hence, by the elementary renewal theorem, we have $\mathbb{E}X_n(\Lambda)/n \to 1/\mu_W$ so that $\mathbb{E}X_n(\Lambda) \sim n/\mu_W$. Therefore, by Theorem 3 (ii), it is sufficient to show that $n/\mu_W < -n\ln\lambda_1$ for all sufficiently large n, or

$$e^{-1/\mu_W} > \lambda_1.$$

Now, since $0 < \lambda_1 \in \mathbb{R}$ is a dominant eigenvalue of $\mathbf{N}$, it follows that: $0 < (1-\lambda_1)^{-1} \in \mathbb{R}$ is a dominant eigenvalue of the matrix $(\mathbf{I}-\mathbf{N})^{-1} = \mathbf{A} = (a_{ij})$; $a_{ij} \geq 0$ with at least one $a_{ij} > 0$; and $\mathbf{A}\mathbf{1}' = (\mathbf{I}-\mathbf{N})^{-1}\mathbf{1}' \leq \mu_W\mathbf{1}'$. Hence, by a simple corollary to the Perron-Frobenius Theorem for nonnegative matrices (cf. Karlin and Taylor 1975, Corollary 2.2, pg. 551), we have

$$\frac{1}{1-\lambda_1} = \limsup_{n\to\infty}\left(\max_{i,j}|a_{ij}^{(n)}|\right)^{1/n} \leq \mu_W,$$

where $a_{ij}^{(n)} = (\mathbf{A}^n)_{ij}$. Therefore, provided $\mu_W < \infty$,

$$e^{-1/\mu_W} > 1 - \frac{1}{\mu_W} \geq \lambda_1,$$

which completes the proof. □

Corollary 1 implies that, if $\mu_n \sim \mathbb{E}X_n(\Lambda)$, then the Poisson approximation will always overestimate the exact probability as $n \to \infty$. Together with Theorem 3 (ii), this implies that using $\mu_n \sim -n \ln \lambda_1$ results in the best Poisson approximation as $n \to \infty$.

We also comment that, for the normal approximation, both $\mu_W/2\sigma_W^2 < -\ln \lambda_1$ and $\mu_W/2\sigma_W^2 \geq -\ln \lambda_1$ are possible. As a simple example, suppose we have a sequence of i.i.d. Bernoulli(p) trials and $\Lambda = SSS$. If $p = 1/2$, we obtain

$$\mu_W = 14, \quad \sigma_W^2 = 142 \quad \text{and} \quad \lambda_1 = 0.9196434,$$

and

$$\frac{\mu_W}{2\sigma_W^2} = 0.04929577 < -\ln \lambda_1 = 0.08376932.$$

However, with $p = 0.9$, we obtain

$$\mu_W = 3.717421, \quad \sigma_W^2 = 2.145694 \quad \text{and} \quad \lambda_1 = 0.5419067;$$

and

$$\frac{\mu_W}{2\sigma_W^2} = 0.8662513 > -\ln \lambda_1 = 0.6126614.$$

Thus, $R(x : E, N) \to \pm\infty$ are both possible depending on x, the pattern, and the probability structure of the $\{X_i\}$.

Numerical comparisons

In the previous section we showed that, for fixed x and $n \to \infty$, the approximation based on the finite Markov chain imbedding technique outperforms the Poisson and normal approximations. In practice, however, one is interested in the performance of these approximations not only when x is fixed and $n \to \infty$, but also when n is fixed (at some moderate value) and x varies. The reason we consider only large or moderate n in our numerical study is that, for small n, the FMCI technique easily gives the exact results. In this section we present some numerical experiments to illustrate the advantages (and disadvantages) of the methods discussed.

The approximations we compare are: the finite Markov chain approximation in (13) (FMCI); the Poisson approximation with $\mu_n = n/\mu_W$ ($\sim \mathbb{E}X_n(\Lambda)$) where μ_W is calculated using (7) (Poisson); The normal approximation given in (6) (Normal); and the large deviation approximation given in Theorem 2 (LD), which is only for right-tail probabilities.

Reliability of $C(k, n{:}F)$ systems

A consecutive-k-out-of-n:F system is a system of n independent and linearly connected components, each with common (continuous) lifetime distribution F, in which the system fails if k consecutive components fail. At a given time $t > 0$, the probability a component is working is $p = 1 - F(t)$ and the probability a single component has failed is $q = 1 - p$ and hence the probability the system has failed is equivalent to the probability that k (or more) consecutive components have failed, which is equivalent to the probability of k consecutive failures in a sequence of n Bernoulli trials with success probability p. Barbour et al. (1995) present a table of various bounds for system reliability based on a Poisson approximation and a compound approximation and compare these to bounds found in Fu (1985). Table 1 shows the exact probabilities and relative errors for the FMCI and Poisson

Table 1 Approximation errors for $C(k, n:F)$ systems

n	k	q	Exact	FMCI	Poisson	CP
5	2	0.01	0.99960	0.00000	-0.00010	0.00000
5	2	0.10	0.96309	0.00000	-0.00788	0.00119
5	2	0.25	0.79980	-0.00002	-0.02697	0.04654
10	2	0.01	0.99911	0.00000	-0.00010	0.00000
10	2	0.10	0.91975	0.00000	-0.00728	0.00312
10	2	0.25	0.61180	0.00000	-0.00869	0.12266
10	4	0.01	1.00000	0.00000	0.00000	0.00000
10	4	0.10	0.99936	0.00000	-0.00026	0.00000
10	4	0.25	0.97855	0.00000	-0.00776	0.00038
50	2	0.01	0.99516	0.00000	-0.00010	0.00000
50	2	0.10	0.63633	0.00000	-0.00251	0.01871
50	2	0.25	0.07173	0.00000	0.14441	0.96838
50	4	0.01	1.00000	0.00000	0.00000	0.00000
50	4	0.10	0.99577	0.00000	-0.00026	0.00000
50	4	0.25	0.86897	0.00000	-0.00663	0.00312
100	2	0.01	0.99024	0.00000	-0.00010	0.00000
100	2	0.10	0.40151	0.00000	0.00343	0.03854
100	2	0.25	0.00492	0.00000	0.36933	2.97133
100	4	0.01	1.00000	0.00000	0.00000	0.00000
100	4	0.10	0.99129	0.00000	-0.00026	0.00001
100	4	0.25	0.74908	0.00000	-0.00523	0.00656
500	4	0.20	0.52721	0.00000	-0.00086	0.00611
1,000	4	0.20	0.27696	0.00000	0.00183	0.01232
10,000	5	0.20	0.07710	0.00000	0.00183	0.00560

approximations as well as the compound Poisson approximation in Barbour et al. (1995) (CP).

The FMCI approximation performs very well for the parameters tested here. As expected, the Poisson and compound Poisson approximations perform well when nq^k is relatively small. When the reliability of the system is relatively low, the Poisson and compound Poisson approximations begin to degrade.

Approximating the distribution of $N_{n,k}$

Recall that $N_{n,k}$ is the number of non-overlapping occurrences of k consecutive successes in $\{X_i\}$ (i.e. $N_{n,k} = X_n(\Lambda)$ with $\Lambda = SS\cdots S$ of length k). By reversing the roles of success and failure, the reliability of $C(k, n:F)$ systems can be related to the distribution of $N_{n,k}$. In this section we present some examples of approximating $\mathbb{P}\{N_{n,k} = x\}$ with the approximations FMCI, Normal, Poisson and LD.

Figure 1 shows the relative error $R(x : E, A)$ in these approximations for (a) $N_{2000,4}$; (b) $N_{5000,4}$; and (c) $N_{250000,6}$ when the probability of success is $p = 0.3$. On all of the figures, the top axis is on a standard z-scale making use of the asymptotic mean and variance of $X_n(\Lambda)$ — namely,

$$z = \frac{x - n/\mu_W}{\sqrt{n\sigma_W^2\mu_W^{-3}}}.$$

We notice that the Finite Markov chain imbedding approximation (FMCI) performs very well in the left tail of the distribution in all cases. Its performance degrades as x gets

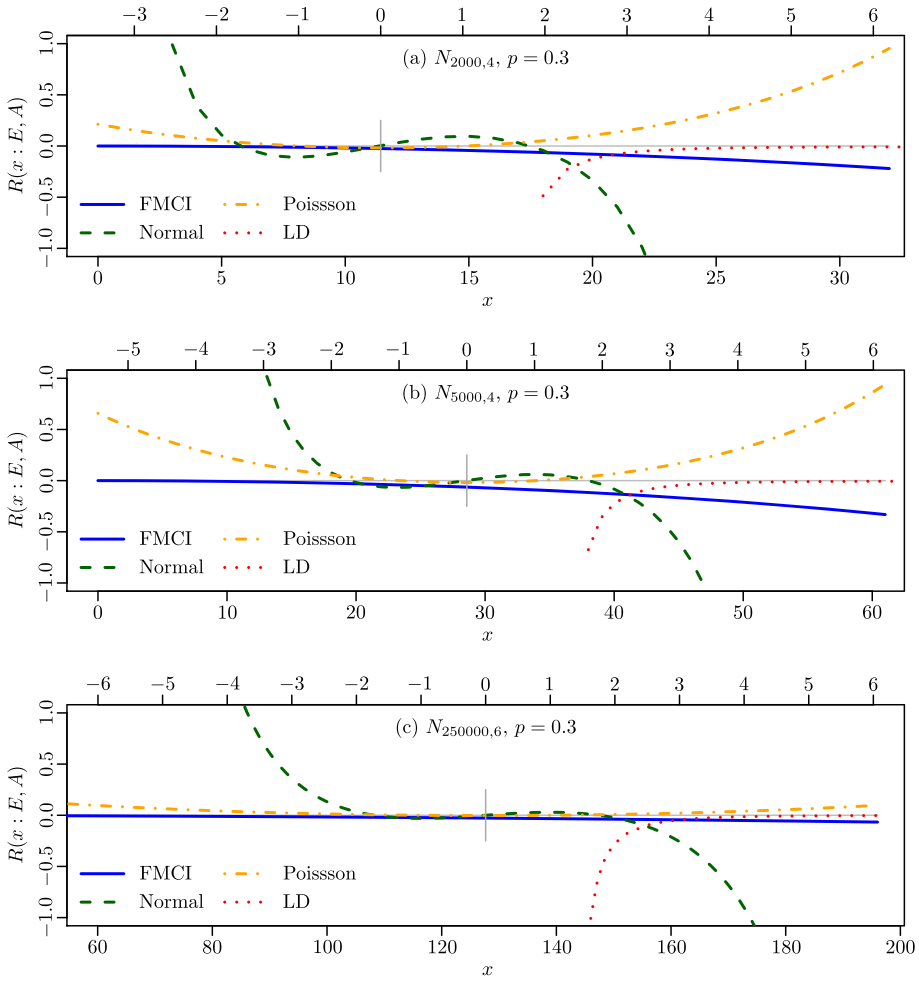

Figure 1 Relative errors of the FMCI, Normal, Poisson and LD approximations $N_{2000,4}$, $N_{5000,4}$ and $N_{250000,6}$ with $p = 0.3$.

large but its performance is more consistent than both the Poisson and Normal approximations in this case. The large deviation approximation performs well in the right tail in all cases. In (c), the FMCI approximation performs very well throughout most of the support. The Poisson approximations also perform well over most of the x considered. The normal approximation performs well in the neighbourhood of $\mathbb{E}X_n(\Lambda)$ but not in the tails.

As the probability of success p increases, the FMCI approximation still performs very well in the left tail, but it's performance tends to degrade more quickly as x increases. The Poisson approximations also quickly degrades as p increases since $\mathbb{E}N_{n,k}$ increases. For larger p, the Normal approximation tends to work better near the mean. In the far left tail, the FMCI approximation is preferred and in the far right tail, the LD approximation is preferred.

Biological sequences

Sequences of DNA nucleotides are of great interest (as are sequences of amino acids and other biological sequences). Figure 2 shows the relative errors for approximating $\mathbb{P}\{X_n(\Lambda) = x\}$ with $\Lambda =$ ACG ($n =$ 1,000 and 10,000) and $\Lambda =$ CATTAG ($n =$ 500,000).

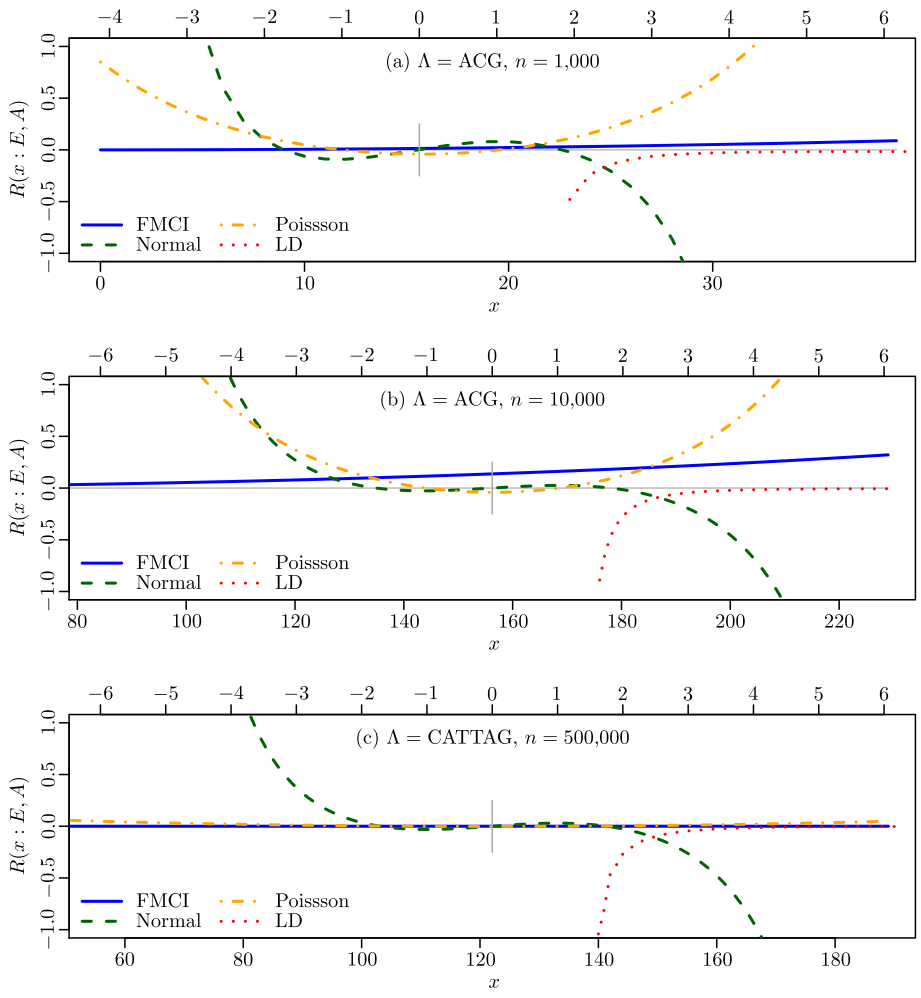

Figure 2 Relative errors of the FMCI, normal, Poisson and large deviation approximations for the patterns $\Lambda = ACG$ with $n = 1{,}000$ and $10{,}000$ and $\Lambda =$ CATTAG with $n = 500{,}000$.

We see that the FMCI approximation again performs very well in the left tail, although, in (b), the performance degrades somewhat as x gets large. The large deviation approximation performs very well in the right tail, especially when x is greater than 3 standard deviations above the mean. While it is difficult to give a rule of thumb, the FMCI approximation seems to perform very well when $x \leq \mathcal{O}(n^{1/2})$. The normal approximation works best within a few standard deviations of the mean and performs best in this region when $\mathbb{E}X_n(\Lambda)$ is relatively large.

Discussion and conclusions

The finite Markov chain imbedding approximations (FMCI and LD) provide an alternative to the usual normal and Poisson approximations for the distributions of runs and patterns. While the FMCI approximation is simple, accurate and fast, it has one disadvantage over the normal and Poisson approximations — it requires the use of the FMCI technique, which is non-traditional and less known in the Statistics community, except in the area of system reliability (*cf.* Cui et al. 2010). On the other hand, the FMCI technique does not require the rather strong conditions necessary for the Poisson techniques, such

as $np^k \to \lambda$. This condition is seldom satisfied in practical applications. For example, in DNA sequence analysis, the probabilities p_A, p_C, p_G and p_T do not tend to 0 as n increases. They may not all be in the neighbourhood of $1/4$ but they are bounded away from 0.

For all of the numeric results in the previous section, the exact probabilities $\mathbb{P}\{X_n(\Lambda) = x\}$ are obtained via the FMCI technique and their CPU times were only a few seconds or less than a minute even in the case of $\Lambda =$ CATTAG and $n =$ 500,000. Based on our experience, if the length of the pattern is less than 20 and n is less than 1,000,000, the exact probability should be computed.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
BJ and JF contributed equally to the mathematical details. BJ performed the numerical comparisons and prepared the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada.

Received: 14 November 2013 Accepted: 7 March 2014
Published: 11 June 2014

References

Arratia, R, Goldstein, L, Gordon, L: Poisson approximation and the Chen-Stein method. Stat. Sci. **5**(4), 403–434 (1990)
Asmussen, S: Applied Probability and Queues. 2nd edn. Applications of Mathematics, vol. 51, p. 438. Springer, New York (2003)
Balakrishnan, N, Koutras, MV: Runs and Scans with Applications. Wiley Series in Probability and Statistics. p. 452. Wiley-Interscience [John Wiley & Sons], New York (2002)
Barbour, AD, Eagleson, GK: Poisson approximation for some statistics based on exchangeable trials. Adv. Appl. Probab. **15**(3), 585–600 (1983)
Barbour, AD, Eagleson, GK: Poisson convergence for dissociated statistics. J. Roy. Statist. Soc. Ser. B. **46**(3), 397–402 (1984)
Barbour, AD, Eagleson, GK: An improved Poisson limit theorem for sums of dissociated random variables. J. Appl. Probab. **24**(3), 586–599 (1987)
Barbour, AD, Hall, P: On the rate of Poisson convergence. Math. Proc. Cambridge Philos. Soc. **95**(3), 473–480 (1984)
Barbour, AD, Chryssaphinou, O: Compound Poisson approximation: a user's guide. Ann. Appl. Probab. **11**(3), 964–1002 (2001)
Barbour, AD, Holst, L, Janson, S: Poisson Approximation. Oxford Studies in Probability (1992a). Oxford Science Publications
Barbour, AD, Chen, LHY, Loh, W-L: Compound Poisson approximation for nonnegative random variables via Stein's method. Ann. Probab. **20**(4), 1843–1866 (1992b)
Barbour, AD, Chryssaphinou, O, Roos, M: Compound Poisson approximation in reliability theory. IEEE T. Reliab. **44**(3), 398–402 (1995)
Barbour, AD, Chryssaphinou, O, Roos, M: Compound Poisson approximation in systems reliability. Naval Res. Logist. **43**(2), 251–264 (1996)
Blom, G, Thorburn, D: How many random digits are required until given sequences are obtained? J. Appl. Probab. **19**(3), 518–531 (1982)
Chen, LHY: Poisson approximation for dependent trials. Ann. Probab. **3**(3), 534–545 (1975)
Cui, L, Xu, Y, Zhao, X: Developments and applications of the finite Markov chain imbedding approach in reliability. IEEE T. Reliab. **59**(4), 685–690 (2010)
Fu, JC: Reliability of a large consecutive-k-out-of-n:F system. IEEE T. Reliab. **R-34**, 120–127 (1985)
Fu, JC, Johnson, BC: Approximate probabilities for runs and patterns in i.i.d. and Markov dependent multi-state trials. Adv. Appl. Probab. **41**(1), 292–308 (2009)
Fu, JC, Koutras, MV: Distribution theory of runs: a Markov chain approach. J. Amer. Statist. Assoc. **89**(427), 1050–1058 (1994)
Fu, JC, Lou, WYW: Distribution Theory of Runs and Patterns and Its Applications. p. 162. World Scientific Publishing Co. Inc, River Edge (2003)
Fu, JC, Lou, WYW: On the normal approximation for the distribution of the number of simple or compound patterns in a random sequence of multi-state trials. Methodol. Comput. Appl. Probab. **9**(2), 195–205 (2007)
Fu, JC, Johnson, BC, Chang, Y-M: Approximating the extreme right-hand tail probability for the distribution of the number of patterns in a sequence of multi-state trials. J. Stat. Plan. Infer. **142**(2), 473–480 (2012)
Gerber, HU, Li, S-YR: The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. Stochastic Process. Appl. **11**(1), 101–108 (1981)
Godbole, AP: Degenerate and Poisson convergence criteria for success runs. Statist. Probab. Lett. **10**(3), 247–255 (1990a)
Godbole, AP: Specific formulae for some success run distributions. Statist. Probab. Lett. **10**(2), 119–124 (1990b)
Godbole, AP: Poisson approximations for runs and patterns of rare events. Adv. Appl. Probab. **23**(4), 851–865 (1991)
Godbole, AP, Schaffner, AA: Improved Poisson approximations for word patterns. Adv. Appl. Probab. **25**(2), 334–347 (1993)
Holst, L, Kennedy, JE, Quine, MP: Rates of Poisson convergence for some coverage and urn problems using coupling. J. Appl. Probab. **25**(4), 717–724 (1988)
Karlin, S: Statistical signals in bioinformatics. Proc. Natl. Acad. Sci. U. S. A. **102**(38), 13355–13362 (2005)

Karlin, S, Taylor, HM: A First Course in Stochastic Processes. 2nd edn., p. 557. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London (1975)
Kleffe, J, Borodovski, M: First and second moment of counts of words in random text generated by Markov chains. Comp Applic Biosci. **8**, 443–441 (1992)
Martin, J, Regad, L, Camproux, A-C, Nuel, G: Finite Markov chain embedding for the exact distribution of patterns in a set of random sequences. In: Skiadas, CH (ed.) Advances in Data Analysis. Statistics for Industry and Technology, pp. 171–180. Birkhäuser, Boston (2010)
Meyn, SP, Tweedie, RL: Markov Chains and Stochastic Stability. Communications and Control Engineering Series (1993)
Nuel, G, Regad, L, Martin, J, Camproux, A-C: Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. Algorithm Mol. Biol. **5**(1), 1–18 (2010)
Schwager, SJ: Run probabilities in sequences of Markov-dependent trials. J. Amer. Statist. Assoc. **78**(381), 168–180 (1983)
Seneta, E: Non-negative Matrices and Markov Chains. 2nd edn. Springer, New York (1981)
Solov'ev, AD: A combinatorial identity and its application to the problem on the first occurrence of a rare event. Teor. Verojatnost. i Primenen. **11**, 313–320 (1966)

doi:10.1186/2195-5832-1-5
Cite this article as: Johnson and Fu: **Approximating the distributions of runs and patterns.** *Journal of Statistical Distributions and Applications* 2014 **1**:5.