Adv. Appl. Prob. **41,** 292–308 (2009)
Printed in Northern Ireland

APPROXIMATE PROBABILITIES FOR RUNS AND PATTERNS IN I.I.D. AND MARKOV-DEPENDENT MULTISTATE TRIALS

JAMES C. FU [*] AND
BRAD C. JOHNSON,[*] [**] *University of Manitoba*

Abstract

Let $X_n(\Lambda)$ be the number of nonoverlapping occurrences of a simple pattern Λ in a sequence of independent and identically distributed (i.i.d.) multistate trials. For fixed k, the exact tail probability $\mathrm{P}\{X_n(\Lambda) < k\}$ is difficult to compute and tends to 0 exponentially as $n \to \infty$. In this paper we use the finite Markov chain imbedding technique and standard matrix theory results to obtain an approximation for this tail probability. The result is extended to compound patterns, Markov-dependent multistate trials, and overlapping occurrences of Λ. Numerical comparisons with Poisson and normal approximations are provided. Results indicate that the proposed approximations perform very well and do significantly better than the Poisson and normal approximations in many cases.

Keywords: Finite Markov chain imbedding; rate functions; multistate trial

2000 Mathematics Subject Classification: Primary 60E05
Secondary 60J10

1. Introduction

The distribution theory of runs and patterns has been successfully applied to various areas of science, for example, statistics, reliability, quality control, genomic studies, bioinformatics, and social science. It has a long history which dates back to at least the sixteenth century. There was a considerable amount of research at the end of the nineteenth century and the early twentieth century; see, for example, Wishart and Hirschfeld (1936), Mood (1940), Wald and Wolfowitz (1940), and Wolfowitz (1943). Traditionally, the exact distribution of runs and patterns were studied via combinatorial analysis. The book by Riordan (1958) provides an excellent review of early developments in this area. Recently, Fu and Koutras (1994) developed the finite Markov chain imbedding (FMCI) technique which provided an alternative way to study the exact distributions for runs and patterns. The method is rather simple, especially for finding distributions of runs and patterns in sequences of non-i.i.d. or Markov-dependent multistate trials or random permutations.

Let $\mathcal{A} = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$ be a set of m symbols, and let $\{X_i\}$ be a sequence of i.i.d. trials taking values in $\mathcal{A}$ with corresponding probabilities p_i, $i = 1, \dots, m$. To avoid trivialities, we assume that $m \geq 2$ and that $p_i > 0$ for $i = 1, \dots, m$. We say that Λ is a simple pattern (or word) of length ℓ defined on $\mathcal{A}$ if $\Lambda = \alpha_{i_1} \cdots \alpha_{i_\ell}$, where $\alpha_{i_j} \in \mathcal{A}$. The length ℓ of the pattern is a fixed positive integer and the symbols α_{i_j} are allowed to repeat. Let $X_n(\Lambda)$ denote the number of nonoverlapping occurrences of the simple pattern Λ in $X_1, \dots, X_n$. To fix these

Received 24 June 2008; revision received 8 December 2008.

[*] Postal address: Department of Statistics, University of Manitoba, Winnipeg, Canada R3T 2N2.

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

[**] Email address: brad_johnson@umanitoba.ca

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

ideas, suppose that $\mathcal{A} = \{S, F\}$ and $\Lambda = SS$. Then the realization

$$X_1 X_2 \cdots X_{14} = \underline{S\,S}\,F\,\underline{S\,S}\,S\,F\,\underline{S\,S}\,\underline{S\,S}\,S\,F\,S \tag{1.1}$$

has $X_{14}(\Lambda) = 4$ nonoverlapping occurrences of Λ (which are underlined). We also define the waiting time until the kth occurrence of Λ as

$$W(k, \Lambda) = \min\{n\colon X_n(\Lambda) = k\}.$$

For brevity, we denote $W(1, \Lambda)$ simply by $W(\Lambda)$. A realization of $\{X_i\}$ starting as in (1.1) yields, for example,

$$W(\Lambda) = W(1, \Lambda) = 2, \qquad W(2, \Lambda) = 5, \qquad W(3, \Lambda) = 9, \quad \text{and} \quad W(4, \Lambda) = 11.$$

Let Λ_1 and Λ_2 be two distinct simple patterns (i.e. neither Λ_1 nor Λ_2 is a sub-pattern of the other). We define the union of Λ_1 and Λ_2, $\Lambda = \Lambda_1 \cup \Lambda_2$, as the occurrence of either Λ_1 or Λ_2. A pattern Λ is called a compound pattern if it is a union of r distinct simple patterns $\Lambda_1, \dots, \Lambda_r$, i.e. $\Lambda = \bigcup_{i=1}^r \Lambda_i$. It follows that if Λ is a compound pattern then the waiting time $W(\Lambda)$ is defined as

$$W(\Lambda) = \min_{1 \le i \le r} \{W(\Lambda_1), \dots, W(\Lambda_r)\}.$$

Given n, the probability that the number of occurrences of Λ in n multistate trials is less than k,

$$\alpha_n(k, \Lambda) = \mathrm{P}\{X_n(\Lambda) < k\},$$

is often very hard to compute numerically, especially when the exact formula is expressed in terms of complex combinatorics and ℓ and n are large. For large n, there are mainly three ways to approximate the tail probabilities $\alpha_n(k, \Lambda)$. First, these probabilities can be approximated by a Poisson distribution $\mathcal{P}(\lambda_n)$ or a compound Poisson distribution $\mathcal{P}_{\mathrm{c}}(\lambda_n, \mu_n)$ in the sense that, provided certain conditions are satisfied, the total variation distance between the distribution of $X_n(\Lambda)$ and $\mathcal{P}(\lambda_n)$ (or $\mathcal{P}_{\mathrm{c}}(\lambda_n, \mu_n)$) tends to 0 as $n \to \infty$ and, hence, the Poisson (or compound Poisson) distribution provides a good approximation for the tail probabilities. See, for example, Arratia *et al.* (1990), Godbole (1991), Godbole and Schaffner (1993), and Barbour *et al.* (1996).

The tail probability $\alpha_n(k, \Lambda)$ can also be approximated by a normal distribution. If we consider the occurrences of Λ in $\{X_i\}$ as a renewal process (which is always possible since the X_i are i.i.d. and we are considering nonoverlapping counting), then $X_n(\Lambda)$ is the number of renewals by time n and the central limit theorem for renewal processes yields

$$\frac{X_n(\Lambda) - n/\mu_W}{\sqrt{n\sigma_W^2/\mu_W^3}} \xrightarrow{\mathcal{L}} N(0, 1) \quad \text{as } n \to \infty, \tag{1.2}$$

where μ_W and σ_W^2 are respectively the mean and variance of the interarrival times of Λ in $\{X_i\}$, which can easily be obtained via FMCI (see Fu and Lou (2003, p. 73) for the details). For Markov-dependent trials and/or overlapping counting, we have a delayed renewal process and similar results hold.

It is well known (see Fu and Lou (2003, pp. 49–96)) that the random variables $X_n(\Lambda)$ and $W(\Lambda)$ can be imbedded in a finite Markov chain. The transition probability matrix for the imbedding of $W(\Lambda)$ has the form

$$\boldsymbol{P} = \left[\begin{array}{c|c} \boldsymbol{N} & \boldsymbol{c}^\top \\ \hline \boldsymbol{0} & 1 \end{array}\right],$$

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

where $\boldsymbol{N}$ is the $d \times d$ essential transition probability matrix (the substochastic matrix corresponding to the transient states only), $\boldsymbol{0}$ represents a matrix, row vector, or column vector of 0s depending on the context, and '$^\top$' denotes the transpose. The essential transition probability matrix for determining $\mathrm{P}\{X_n(\Lambda) < k\}$ depends on k and has the form

$$\boldsymbol{N}_k = \begin{bmatrix} \boldsymbol{N} & \boldsymbol{C} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{N} & \boldsymbol{C} & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \boldsymbol{C} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{N} \end{bmatrix}_{dk \times dk},$$

where there are k copies of $\boldsymbol{N}$ on the diagonal and $\boldsymbol{C} = [\boldsymbol{c}^\top \mid \boldsymbol{0}]$ is the 'continuation' matrix. Note that $\boldsymbol{C}$ may take a different form for Markov-dependent trials and/or overlapping counting.

Given the matrices $\boldsymbol{N}$ and $\boldsymbol{N}_k$, we have

$$\mathrm{P}\{W(\Lambda) > n\} = \boldsymbol{\xi}_0 \boldsymbol{N}^n \boldsymbol{1}^\top \tag{1.3}$$

and

$$\mathrm{P}\{X_n(\Lambda) < k\} = \boldsymbol{\xi}_0 \boldsymbol{N}_k^n \boldsymbol{1}^\top, \tag{1.4}$$

where $\boldsymbol{\xi}_0 = (1, 0, \dots, 0)$ and $\boldsymbol{1} = (1, \dots, 1)$ are row vectors of appropriate length. For fixed k, these probabilities tend to 0 exponentially as $n \to \infty$ and, for large ℓ and very large n, the computation of $\mathrm{P}\{X_n(\Lambda) < k\}$ by (1.4) can be problematic, and, hence, approximations are of general interest. In this paper, our main goal is to develop a large deviation type approximation for the tail probabilities in (1.4) by making use of (1.3) and its approximation.

The remainder of this paper is organized as follows. In Section 2 we give the main results under some specific assumptions about the structure of $\boldsymbol{N}$. We then extend these results to more general cases. Section 3 provides some numerical results and comparisons with the Poisson and normal approximations. Numerically, the approximations developed work very well for tail probabilities less than 0.05, a critical probability for hypothesis testing.

2. Main results

Here and throughout, we make use of the following notation.

(i) $\{Y_n : n \geq 0\}$ will denote the Markov chain induced by Λ with transition probability matrix $\boldsymbol{P}$ and $d \times d$ essential transition probability matrix $\boldsymbol{N}$.

(ii) The eigenvalues for $\boldsymbol{N}$ will be denoted by $\lambda_1, \dots, \lambda_d$, repeated according to their algebraic multiplicities and ordered such that $1 > \lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_d| \geq 0$ (such an ordering is possible by the Perron–Frobenius theorem for nonnegative matrices). For simple patterns of length ℓ, we have $d = \ell$.

(iii) The right eigenvectors of $\boldsymbol{N}$ associated with $\lambda_1, \dots, \lambda_d$ will be denoted by $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_d$. If the geometric multiplicity of an eigenvalue is less than its algebraic multiplicity, we use vectors of 0s for the unspecified eigenvectors. We also take $\boldsymbol{\eta}_1$ to be an eigenvector associated with λ_1 with nonnegative entries (Perron–Frobenius again).

(iv) The linear space spanned by $\{\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_d\}$ will be denoted by $\mathcal{E}_N$ and its orthogonal complement will be denoted by $\mathcal{E}_N^\perp$.

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

(v) For $f(n)$ and $g(n)$, $n \in \mathbb{N}$, we say that $f(n) = o(g(n))$ if $|f(n)/g(n)| \to 0$ as $n \to \infty$, $f(n) = \mathcal{O}(g(n))$ if there exists a constant $C > 0$ such that $|f(n)| \leq C|g(n)|$ for all $n > 0$, and $f(n) \sim g(n)$ if $f(n)/g(n) \to 1$ as $n \to \infty$.

Before stating and proving the main result, we introduce some notation and prove a lemma. Let n and k be positive integers with $n \geq k$, let $\mathcal{C}_{n,k}$ denote the nonnegative integer solutions to $n_1 + \cdots + n_k = n$, and recall (cf. Riordan (1958, p. 124)) that

$$|\mathcal{C}_{n,k}| = \binom{n+k-1}{k-1},$$

where, for sets, $|\cdot|$ denotes cardinality. Furthermore, for $\boldsymbol{\ell} = (\ell_1, \dots, \ell_k)$, let $\mathcal{C}_{n,k}(\boldsymbol{\ell})$ denote the nonnegative integer solutions to $n_1 + \cdots + n_k = n$ such that $n_i \geq \ell_i$ for each $i = 1, \dots, k$, and note that

$$|\mathcal{C}_{n,k}(\boldsymbol{\ell})| = \binom{n - \sum_{i=1}^{k} \ell_i + k - 1}{k-1}.$$

Lastly, we define the complement of $\mathcal{C}_{n,k}(\boldsymbol{\ell})$ by $\overline{\mathcal{C}_{n,k}(\boldsymbol{\ell})} = \mathcal{C}_{n,k} \setminus \mathcal{C}_{n,k}(\boldsymbol{\ell})$.

Lemma 2.1. *Let $\{f_n\}$ be a sequence of positive integers such that $f_n \to \infty$ and $f_n/n \to 0$, and let $g\colon \mathcal{C}_{n,k} \to (K_1, K_2)$ for some positive constants $0 < K_1 \leq K_2 < \infty$. Then, for any fixed k,*

$$\lim_{n\to\infty} \frac{\sum_{\boldsymbol{n} \in \overline{\mathcal{C}_{n,k}(f_n)}} g(\boldsymbol{n})}{\sum_{\boldsymbol{n} \in \mathcal{C}_{n,k}(f_n)} g(\boldsymbol{n})} = 0.$$

Proof. Clearly, we have

$$\frac{K_1 |\overline{\mathcal{C}_{n,k}(f_n)}|}{K_2 |\mathcal{C}_{n,k}(f_n)|} \leq \frac{\sum_{\boldsymbol{n} \in \overline{\mathcal{C}_{n,k}(f_n)}} g(\boldsymbol{n})}{\sum_{\boldsymbol{n} \in \mathcal{C}_{n,k}(f_n)} g(\boldsymbol{n})} \leq \frac{K_2 |\overline{\mathcal{C}_{n,k}(f_n)}|}{K_1 |\mathcal{C}_{n,k}(f_n)|}.$$

However,

$$\frac{|\mathcal{C}_{n,k}|}{|\mathcal{C}_{n,k}(f_n)|} = \left(\frac{n+k-1}{n-kf_n+k-1}\right)\left(\frac{n+k-2}{n-kf_n+k-2}\right)\cdots\left(\frac{n}{n-kf_n}\right) \to 1 \quad \text{as } n \to \infty,$$

and the result now follows from the definition of $\overline{\mathcal{C}_{n,k}(\ell)}$.

Theorem 2.1. *Let $X_1, X_2, \dots$ be a sequence of i.i.d. trials taking values in $\mathcal{A}$, let Λ be a simple pattern of length ℓ with $d \times d$ essential transition probability matrix $\boldsymbol{N}$, and let $X_n(\Lambda)$ be the number of nonoverlapping occurrences of Λ in $X_1, \dots, X_n$. If*

(i) *λ_1 has algebraic multiplicity m and $\lambda_1 > |\lambda_j|$ for all $j > m$, and*

(ii) *there exist constants $a_1, \dots, a_d$ such that $\boldsymbol{1}^\top = \sum_{j=1}^{d} a_j \boldsymbol{\eta}_j^\top$ and $a_1(\boldsymbol{\xi}_0 \boldsymbol{\eta}_1^\top) > 0$,*

then, for any fixed $k \geq 0$,

$$\mathrm{P}\{X_n(\Lambda) = k\} \sim a^{k+1} \binom{n - k(\ell-1)}{k} (1-\lambda_1)^k \lambda_1^{n-k}, \tag{2.1}$$

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

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

Proof. Let W_i denote the waiting time (or interarrival time) between the $(i-1)$th and the ith occurrence of Λ in $X_1, X_2, \dots$ so that

$$\mathrm{P}\{X_n(\Lambda) = k-1\} = \mathrm{P}\{W_1 + \cdots + W_{k-1} \le n,\ W_1 + \cdots + W_k > n\}.$$

Summing over all possible values of $W_1, \dots, W_k$ (i.e. over all $\boldsymbol{n} = (n_1, \dots, n_k) \in \mathcal{C}_{n,k}(\boldsymbol{\ell})$ with $\boldsymbol{\ell} = (\ell, \dots, \ell, 0)$) and noting that, with nonoverlapping counting, these W_i are i.i.d., yields

$$\begin{aligned}\mathrm{P}\{X_n(\Lambda) = k-1\} &= \sum_{\boldsymbol{n} \in \mathcal{C}_{n,k}(\boldsymbol{\ell})} \mathrm{P}\{W_1 = n_1, \dots, W_{k-1} = n_{k-1},\ W_k > n_k\} \\ &= \sum_{\boldsymbol{n} \in \mathcal{C}_{n,k}(\boldsymbol{\ell})} \biggl(\prod_{i=1}^{k-1} \mathrm{P}\{W_i = n_i\}\biggr) \mathrm{P}\{W_k > n_k\}. \qquad (2.2)\end{aligned}$$

Now, since $\lambda_1 = \cdots = \lambda_m$ by assumption, we have, letting $a = \sum_{j=1}^{m} a_j(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top)$,

$$\mathrm{P}\{W(\Lambda) > n\} = \boldsymbol{\xi}_0 \boldsymbol{N}^n \boldsymbol{1}^\top = \boldsymbol{\xi}_0 \boldsymbol{N}^n \sum_{j=1}^{d} a_j \boldsymbol{\eta}_j^\top = a\lambda_1^n \biggl(1 + \sum_{j=m+1}^{d} \frac{a_j(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top)}{a} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^{n}\biggr)$$

and

$$\begin{aligned}\mathrm{P}\{W(\Lambda) = n\} &= \mathrm{P}\{W(\Lambda) > n-1\} - \mathrm{P}\{W(\Lambda) > n\} \\ &= \boldsymbol{\xi}_0 \boldsymbol{N}^{n-1} (\boldsymbol{I} - \boldsymbol{N}) \boldsymbol{1}^\top \\ &= \sum_{j=1}^{d} a_j (1-\lambda_j)(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top) \lambda_j^{n-1} \\ &= a(1-\lambda_1)\lambda_1^{n-1} \biggl(1 + \sum_{j=m+1}^{d} \frac{a_j(1-\lambda_j)(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top)}{a(1-\lambda_1)} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^{n-1}\biggr).\end{aligned}$$

Substituting these into (2.2) yields

$$\begin{aligned}&\mathrm{P}\{X_n(\Lambda) = k-1\} \\ &\quad = \sum_{\boldsymbol{n} \in \mathcal{C}_{n,k}(\boldsymbol{\ell})} \biggl(\prod_{i=1}^{k-1} a(1-\lambda_1)\lambda_1^{n_i-1} \biggl[1 + \sum_{j=m+1}^{d} \frac{a_j(1-\lambda_j)(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top)}{a(1-\lambda_1)} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^{n_i-1}\biggr]\biggr) \\ &\qquad\qquad \times a\lambda_1^{n_k} \biggl(1 + \sum_{j=m+1}^{d} \frac{a_j(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top)}{a} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^{n_k}\biggr) \\ &\quad = a^k (1-\lambda_1)^{k-1} \lambda_1^{n-k+1} \sum_{\boldsymbol{n} \in \mathcal{C}_{n,k}(\boldsymbol{\ell})} \psi(\boldsymbol{n}, \ell),\end{aligned}$$

where

$$\psi(\boldsymbol{n}, \ell) = \biggl(1 + \sum_{j=m+1}^{d} \frac{a_j(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top)}{a} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^{n_k}\biggr) \prod_{i=1}^{k-1} \biggl(1 + \sum_{j=m+1}^{d} \frac{a_j(1-\lambda_j)(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top)}{a(1-\lambda_1)} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^{n_i-1}\biggr).$$

Now, for any sequence of integers $\{f_n\}$ such that $f_1 > \ell$, $f_n \to \infty$, and $f_n/n \to 0$ as $n \to \infty$, we have

$$\lim_{n\to\infty} \inf_{\boldsymbol{n}\in\mathcal{C}_{n,k}(f_n)} \psi(\boldsymbol{n},\ell) = \lim_{n\to\infty} \sup_{\boldsymbol{n}\in\mathcal{C}_{n,k}(f_n)} \psi(\boldsymbol{n},\ell) = 1.$$

Furthermore,

$$\sum_{\boldsymbol{n}\in\mathcal{C}_{n,k}(f_n)} \psi(\boldsymbol{n},\ell) \le \sum_{\boldsymbol{n}\in\mathcal{C}_{n,k}(\boldsymbol{\ell})} \psi(\boldsymbol{n},\ell) \le \sum_{\boldsymbol{n}\in\mathcal{C}_{n,k}} \psi(\boldsymbol{n},\ell),$$

and, since

$$\sum_{\boldsymbol{n}\in\mathcal{C}_{n,k}} \psi(\boldsymbol{n},\ell), = \sum_{\boldsymbol{n}\in\mathcal{C}_{n,k}(f_n)} \psi(\boldsymbol{n},\ell)\left(1 + \frac{\sum_{\boldsymbol{n}\in\overline{\mathcal{C}_{n,k}(f_n)}} \psi(\boldsymbol{n},\ell)}{\sum_{\boldsymbol{n}\in\mathcal{C}_{n,k}(f_n)} \psi(\boldsymbol{n},\ell)}\right),$$

we have, by Lemma 2.1,

$$\sum_{\boldsymbol{n}\in\mathcal{C}_{n,k}(\boldsymbol{\ell})} \psi(\boldsymbol{n},\ell) \sim \binom{n-(k-1)(\ell-1)}{k-1}. \tag{2.3}$$

Combining this with the above, we have

$$\mathrm{P}\{X_n(\Lambda) = k-1\} \sim a^k \binom{n-(k-1)(\ell-1)}{k-1} (1-\lambda_1)^{k-1}\lambda_1^{n-k+1}.$$

Replacing $k-1$ with k completes the proof.

The accuracy of the approximation in (2.1) depends primarily on the approximation in (2.3), which performs well when $n/\ell(k+1)$ is relatively large and $\lambda_{m+1}/|\lambda_1|$ is not too close to 1. It also depends indirectly on ℓ in a more complicated way. An alternate way of writing (2.1) is, for fixed $k \ge 0$,

$$\lim_{n\to\infty}\left[\mathrm{P}\{X_n(\Lambda) = k\} \Big/ a^{k+1}\left(\frac{1-\lambda_1}{\lambda_1}\right)^k \binom{n-k(\ell-1)}{k} \exp\{n\log\lambda_1\}\right] = 1. \tag{2.4}$$

This shows that (a) for fixed k, $\mathrm{P}\{X_n(\Lambda) = k\} \to 0$ exponentially with rate $-\log\lambda_1$ as $n \to \infty$, and (b) for fixed but large n, our approximation resembles a Poisson probability with an adjustment constant that is independent of n. This follows from the fact that, for large n,

$$a^{k+1}\left(\frac{1-\lambda_1}{\lambda_1}\right)^k \binom{n-k(\ell-1)}{k} \exp\{n\log\lambda_1\} \sim c_k \frac{(-n\log\lambda_1)^k}{k!} \exp\{n\log\lambda_1\},$$

where

$$c_k = a^{k+1}\left(\frac{1-\lambda_1}{\lambda_1 \log 1/\lambda_1}\right)^k.$$

Note that, for binary trials with $\mathcal{A} = \{S, F\}$, $\Lambda = S$, and $p = p_S$, it is easy to see that $\lambda_1 = q = 1-p$ and $a = 1$ so that the approximation in (2.1) becomes an equality and

$$\mathrm{P}\{X_n(\Lambda) = k\} = \binom{n}{k} p^k q^{n-k},$$

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

as expected. Furthermore, for $k = 0$, we have

$$\mathrm{P}\{W(\Lambda) > n\} = \mathrm{P}\{X_n(\Lambda) = 0\} \sim a\lambda_1^n,$$

which is a result in Fu *et al.* (2003, p. 354).

For cumulative tail probabilities, we have

$$\mathrm{P}\{X_n(\Lambda) < k\} \sim \sum_{j=0}^{k-1} a^{j+1} \binom{n - j(\ell - 1)}{j} (1 - \lambda_1)^j \lambda_1^{n-j}. \tag{2.5}$$

If we note that $\mathrm{P}\{X_n(\Lambda) < 1\} = \mathrm{P}\{X_n(\Lambda) = 0\}$ and, for $k > 1$,

$$\frac{\mathrm{P}\{X_n(\Lambda) = k - 2\}}{\mathrm{P}\{X_n(\Lambda) = k - 1\}} = \mathcal{O}\biggl(\frac{1}{n}\biggr),$$

we see that, in many cases, the approximation

$$\mathrm{P}\{X_n(\Lambda) < k\} \sim a^k \binom{n - (k - 1)(\ell - 1)}{k - 1} (1 - \lambda_1)^{k-1} \lambda_1^{n-k+1}$$

is sufficient provided that n is very large. The examples will show that, for $k > 1$, an approximation based the last two terms in (2.5) is a good compromise provided that $n/\ell(k+1)$ is relatively large.

2.1. Generalizations and extensions

Before moving on to some numerical examples we discuss how these results can be extended to Markov-dependent trials and compound patterns.

2.1.1. *Markov-dependent trials.* In the (first-order) Markov-dependent case, we will use the notation $p_{\alpha_i\alpha_j} = \mathrm{P}\{X_{n+1} = \alpha_j \mid X_n = \alpha_i\}$ for all $\alpha_i, \alpha_j \in \mathcal{A}$ and all $n \geq 0$. The state space $\Omega_{\boldsymbol{P}}$ for the Markov chain induced by the simple pattern $\Lambda = \alpha_{i_1}\alpha_{i_2} \cdots \alpha_{i_\ell}$ of length ℓ will have the same form as for the i.i.d. case except that, in order to incorporate the Markov dependency, the state ϕ is replaced by $\{\phi_{\alpha_1}, \dots, \phi_{\alpha_m}\}$ so that

$$\Omega_{\boldsymbol{P}} = \{\phi_{\alpha_1}; \dots; \phi_{\alpha_m}; \alpha_{i_1}; \alpha_{i_1}\alpha_{i_2}; \dots; \alpha_{i_1} \cdots \alpha_{i_{\ell-1}}; \alpha_{i_1} \cdots \alpha_{i_\ell}\}.$$

Similarly, the state space for the essential transition probability matrix $\boldsymbol{N}$ has the form

$$\Omega_{\boldsymbol{N}} = \Omega_{\boldsymbol{P}} \setminus \{\Lambda\} = \{\phi_{\alpha_1}; \dots; \phi_{\alpha_m}; \alpha_{i_1}; \alpha_{i_1}\alpha_{i_2}; \dots; \alpha_{i_1} \cdots \alpha_{i_{\ell-1}}\}.$$

See Section 3.3 for a worked out example.

The arguments are entirely analogous to those in the previous section. The only difficulty is the definition of the initial state vector $\boldsymbol{\xi}_0$ and the fact that this is no longer constant for each $W_i(\Lambda)$. Indeed, for any initial distribution $\boldsymbol{\xi}_0$, we have

$$\mathrm{P}\{W_1(\Lambda) = n\} = \boldsymbol{\xi}_0 \boldsymbol{N}^{n-1}(\boldsymbol{I} - \boldsymbol{N})\boldsymbol{1}^\top \quad \text{and} \quad \mathrm{P}\{W_1(\Lambda) > n\} = \boldsymbol{\xi}_0 \boldsymbol{N}^n \boldsymbol{1}^\top,$$

as before. However, for $i > 1$, we now have

$$\mathrm{P}\{W_i(\Lambda) = n\} = \boldsymbol{\xi}_\Lambda \boldsymbol{N}^{n-1}(\boldsymbol{I} - \boldsymbol{N})\boldsymbol{1}^\top \quad \text{and} \quad \mathrm{P}\{W_i(\Lambda) > n\} = \boldsymbol{\xi}_\Lambda \boldsymbol{N}^n \boldsymbol{1}^\top,$$

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

where $\boldsymbol{\xi}_\Lambda$ is a row vector of appropriate length with a 1 in the position of state $\phi_{\alpha_{i_\ell}}$ and 0s elsewhere.

When $k = 1$, we obtain

$$\mathrm{P}\{X_n(\Lambda) = 0\} = \boldsymbol{\xi}_0 \boldsymbol{N}^n \mathbf{1}^\top,$$

and, for $k > 1$, we have

$$\mathrm{P}\{X_n(\Lambda) = k-1\} = \sum_{\boldsymbol{n} \in \mathcal{C}_{n,k}(\boldsymbol{\ell})} \boldsymbol{\xi}_0 \boldsymbol{N}^{n_1 - 1}(\boldsymbol{I} - \boldsymbol{N})\mathbf{1}^\top \Bigg(\prod_{i=2}^{k-1} \boldsymbol{\xi}_\Lambda \boldsymbol{N}^{n_i - 1}(\boldsymbol{I} - \boldsymbol{N})\mathbf{1}^\top \Bigg) \boldsymbol{\xi}_\Lambda \boldsymbol{N}^{n_k} \mathbf{1}^\top.$$

Applying the arguments in the proof of Theorem 2.1 and simplifying yields the following theorem.

Theorem 2.2. *Let $X_1, X_2, \dots$ be a sequence of (first-order) Markov-dependent trials taking values in $\mathcal{A}$, let Λ be a simple pattern of length ℓ with $d \times d$ essential transition probability matrix $\boldsymbol{N}$, and let $X_n(\Lambda)$ be the number of nonoverlapping occurrences of Λ in $X_1, \dots, X_n$. If*

(i) *λ_1 has algebraic multiplicity m and $\lambda_1 > |\lambda_j|$ for all $j > m$, and*

(ii) *there exist constants $a_1, \dots, a_d$ such that $\mathbf{1}^\top = \sum_{j=1}^d a_j \boldsymbol{\eta}_j^\top$ and $a_1(\boldsymbol{\xi}_0 \boldsymbol{\eta}_1^\top) > 0$,*

then, for any fixed $k \geq 0$,

$$\mathrm{P}\{X_n(\Lambda) = k\} \sim ab^k \binom{n - k(\ell - 1)}{k} (1 - \lambda_1)^k \lambda_1^{n-k}, \tag{2.6}$$

where

$$a = \sum_{j=1}^m a_j(\boldsymbol{\xi}_0 \boldsymbol{\eta}_j^\top) \quad \textit{and} \quad b = \sum_{j=1}^m a_j(\boldsymbol{\xi}_\Lambda \boldsymbol{\eta}_j^\top). \tag{2.7}$$

If $m = 1$, as is usually the case, then $a = a_1(\boldsymbol{\xi}_0 \boldsymbol{\eta}_1^\top)$ and $b = a_1(\boldsymbol{\xi}_\Lambda \boldsymbol{\eta}_1^\top)$.

By the same token, the tail probability can be approximated by

$$\mathrm{P}\{X_n(\Lambda) < k\} \sim \sum_{j=0}^{k-1} ab^j \binom{n - j(\ell - 1)}{j} (1 - \lambda_1)^j \lambda_1^{n-j},$$

where a and b are defined in (2.7). Note that the constants a, b, and λ_1 are independent of n and depend only on the structure of $\boldsymbol{N}$. In other words, this approximation is available only when $W(\Lambda)$ is finite Markov chain imbeddable.

Equation (2.6) also holds for counting overlapping occurrences of Λ in both the i.i.d. and (first-order) Markov-dependent cases. The only requirement is the identification of the 're-start' vector $\boldsymbol{\xi}_\Lambda$ for each case. For example, with $\mathcal{A} = \{A, C, G, T\}$ and $\Lambda = ACAC$, the state space for the essential transition probability matrix in the i.i.d. case is

$$\Omega_{\boldsymbol{N}} = \{\phi, A, AC, ACA\}.$$

For overlapping counting, we have $\boldsymbol{\xi}_\Lambda = (0, 0, 1, 0)$ (for nonoverlapping counting, $\boldsymbol{\xi}_\Lambda = \boldsymbol{\xi}_0 = (1, 0, 0, 0)$ and $a = b$ so that (2.6) reduces to (2.1)). If the trials are (first-order) Markov dependent, we have

$$\Omega_{\boldsymbol{N}} = \{\phi_A, \phi_C, \phi_G, \phi_T, A, AC, ACA\}.$$

For nonoverlapping counting, $\boldsymbol{\xi}_\Lambda = (0, 1, 0, 0, 0, 0, 0)$, and, for overlapping counting, $\boldsymbol{\xi}_\Lambda = (0, 0, 0, 0, 0, 1, 0)$.

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

2.1.2. *Compound patterns.* Furthermore, we would like to point out that, if $\Lambda = \bigcup_{i=1}^{r} \Lambda_i$ is a compound pattern and the trials are i.i.d., then the approximation given in (2.1) remains applicable. The difficulty is in specifying ℓ in the approximation. A reasonable (but not optimal) choice is to take ℓ as the minimum length of the Λ_i. For large n, this makes little difference since, for any fixed k, ℓ, and u,

$$\lim_{n\to\infty} \binom{n-(k-1)(\ell-1)}{k-1} \bigg/ \binom{n-u}{k-1} = 1.$$

For compound patterns in Markov-dependent trials, the extensions such as those given in Section 2.1.1 are also obvious. Of course, the transition probability matrix $\boldsymbol{N}$ and associated constants depend on the dependence structure of the $\{X_i\}$ and the structure of the compound pattern. No mathematical detail will be provided here.

When $\{X_i\}$ is a Markov-dependent sequence and Λ is a compound pattern, the exact distribution for overlapping counting can be obtained, but the complexity of the approximation method grows. We leave this case as an open problem.

2.1.3. *When* $1^\top \notin \mathcal{E}_N$. When $\mathbf{1}^\top \notin \mathcal{E}_N$, we have two options. The first involves choosing and decomposing $\boldsymbol{\xi}_0$ (and $\boldsymbol{\xi}_\Lambda$ if required), and the second involves the Jordan canonical form for $\boldsymbol{N}_k$.

Let $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_d$ denote the left eigenvectors of $\boldsymbol{N}$. If we are willing to choose a $\boldsymbol{\xi}_0$ such that there exist constants $c_1, \dots, c_d$ such that $\boldsymbol{\xi}_0 = \sum_{j=1}^{d} c_j \boldsymbol{\zeta}_j$ and, if required, we can find constants $u_1, \dots, u_d$ such that $\boldsymbol{\xi}_\Lambda = \sum_{j=1}^{d} u_j \boldsymbol{\zeta}_j$, then we may proceed as in the proof of Theorem 2.1 except that we write

$$\mathrm{P}\{W(\Lambda) > n\} = \boldsymbol{\xi}_0 \boldsymbol{N}^n \mathbf{1}^\top = \sum_{j=1}^{d} c_j \boldsymbol{\zeta}_j \boldsymbol{N}^n \mathbf{1}^\top = c\lambda_1^n \biggl(1 + \sum_{j=m+1}^{d} \frac{c_j(\boldsymbol{\zeta}_j \mathbf{1}^\top)}{c} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^n\biggr),$$

where $c = \sum_{j=1}^{m} c_j(\boldsymbol{\zeta}_j \mathbf{1}^\top)$ and, similarly,

$$\mathrm{P}\{W(\Lambda) = n\} = \boldsymbol{\xi}_0 \boldsymbol{N}^{n-1}(\boldsymbol{I} - \boldsymbol{N})\mathbf{1}^\top = c(1-\lambda_1)\lambda_1^{n-1} \biggl(1 + \sum_{j=m+1}^{d} \frac{c_j(\boldsymbol{\zeta}_j \mathbf{1}^\top)}{c(1-\lambda_1)} \biggl(\frac{\lambda_j}{\lambda_1}\biggr)^n\biggr),$$

and similarly for any terms involving $\boldsymbol{\xi}_\Lambda$. The development is entirely analogous to that in Theorem 2.1 and the details are left to the reader.

If the required constants $\{c_j\}$ and (if required) $\{u_j\}$ do not exist, then we may still appeal to the Jordan canonical form $\boldsymbol{J}$ for $\boldsymbol{N}_k$ to obtain a similar result (i.e. there exists a nonsingular $\boldsymbol{V}$ and Jordan matrix $\boldsymbol{J}$ such that $\boldsymbol{V}\boldsymbol{J}\boldsymbol{V}^{-1} = \boldsymbol{N}_k$). If we let q denote the size of the largest Jordan block of $\boldsymbol{J}$ associated with λ_1, then it is possible to show that there exists a constant $c_q > 0$, independent of n, such that

$$\boldsymbol{\xi}_0 \boldsymbol{N}_k^n \mathbf{1}^\top = \boldsymbol{\xi}_0 \boldsymbol{V} \boldsymbol{J}^n \boldsymbol{V}^{-1} \mathbf{1}^\top \sim c_q \binom{n}{q-1} \lambda_1^n.$$

The reader will notice the connection between this argument and that presented in Section 2. We comment that the only pattern we have encountered that requires this technique is $\mathcal{A} = \{S, F\}$ with $p_S = \frac{1}{2}$ and $\Lambda = SF$, and, in this case, powers of the 2 × 2 essential transition probability matrix are entirely trivial to calculate exactly. For large ℓ and moderate k, this method is computationally more expensive since it requires us to find the generalized eigenvectors of $\boldsymbol{N}_k$, which is of dimension $k\ell \times k\ell$ (for simple patterns).

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

3. Numerical examples and comparisons

In this section we provide some numerical examples of the approximations in Section 2 and give some comparisons to the Poisson approximation and to the normal approximation given in (1.2). In order to fairly compare the approximations, we make use of the comparison function

$$\rho(a, e) = \left(1 - \min\left\{\frac{a}{e}, \frac{e}{a}\right\}\right) \operatorname{sgn}(a - e),$$

where a is an approximate probability and e is the exact probability. This measure is always in $[-1, 1]$; it is symmetric in the sense that it treats over and under estimation equivalently; it goes to 1 as e/a gets small, indicating severe over estimation; it goes to -1 as a/e gets small, indicating severe under estimation; and it is 0 when $a = e$.

In addition to k, n, and the exact probabilities (calculated using (1.4) and the FMCI technique), the tables in this section contain one or more of the following columns, each reporting the value of $\rho(a, e)$.

- F: the FMCI approximation for $\mathrm{P}\{X_n(\Lambda) = k\}$ given in (2.1).
- F(k): the FMCI approximation for $\mathrm{P}\{X_n(\Lambda) < k\}$ using all of the terms in (2.5).
- F(2): the FMCI approximation for $\mathrm{P}\{X_n(\Lambda) < k\}$ using the last two terms in (2.5).
- F(1): the FMCI approximation for $\mathrm{P}\{X_n(\Lambda) < k\}$ using only the last term in (2.5).
- POI: the Poisson approximation for both $\mathrm{P}\{X_n(\Lambda) = k\}$ and $\mathrm{P}\{X_n(\Lambda < k\}$ calculated as in Godbold and Schnaffner (1993).
- CLT: the normal approximation given in (1.2) with continuity correction applied for both $\mathrm{P}\{X_n(\Lambda) = k\}$ and $\mathrm{P}\{X_n(\Lambda < k\}$.

Unless stated otherwise, for these examples, we assume that the $\{X_i\}$ are i.i.d. taking values in $\mathcal{A} = \{A, C, G, T\}$ with equal probability.

3.1. Approximating $\mathbf{P}\{X_n(\Lambda) = k\}$

Before presenting the results of these comparisons we can make the following few remarks.

1. While we expect the normal approximation to work reasonably well in the neighborhood of $\mathrm{E}[X_n(\Lambda)]$, especially when this is large, we do not expect that the normal approximation will be able to capture the tail behavior of $X_n(\Lambda)$ very well. We expect that the Poisson approximation will be better at capturing this tail behavior, especially when the distribution of $X_n(\Lambda)$ is moderately skewed.
2. It is usually the case that $\mathrm{E}[X_n(\Lambda)] > \mathrm{var}[X_n(\Lambda)]$ and, hence, even with an appropriate choice of λ, we expect the Poisson approximation to over estimate probabilities in the tails and under estimate probabilities in the neighborhood of $\mathrm{E}[X_n(\Lambda)]$. The severity of this phenomenon will depend, primarily, on $\mathrm{E}[W(\Lambda)]$—large $\mathrm{E}[W(\Lambda)]$ indicates that Λ is ‘rare’ and the Poisson approximation should work quite well, especially when n is not too large.

This behavior is illustrated in Figure 1, in which we plot $\rho(a, e)$ for three patterns, of lengths 2, 4, and 6, and the three approximations F (solid line), POI (dashed line), and CLT

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

(dotted line). The top axes show the (approximate) standard z-scores making use of (1.2):

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

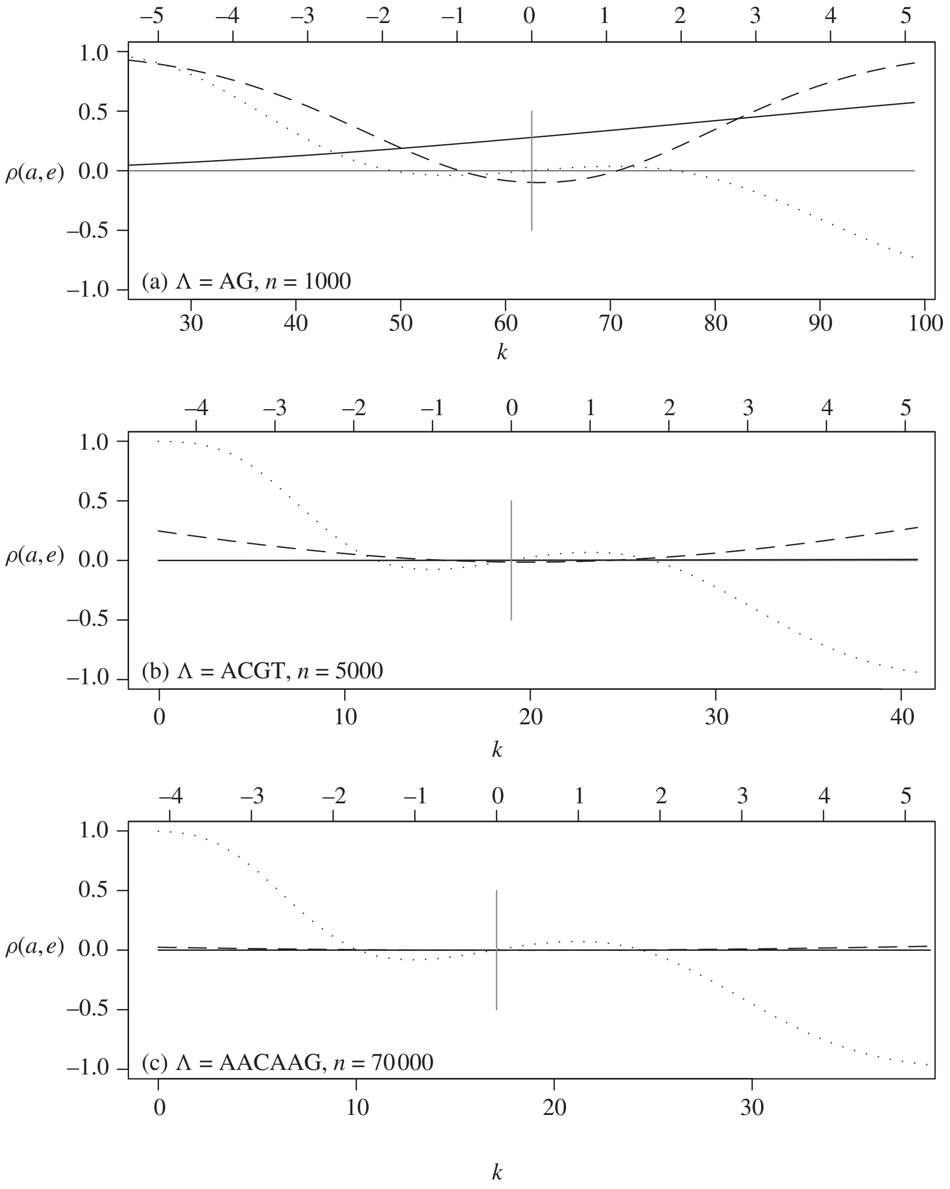

FIGURE 1: The measure $\rho(a, e)$ for three patterns, fixed n and varying k (see text). The dotted line is for the normal approximation, the dashed line is for the Poisson approximation, and the solid line is for the FMCI approximation proposed in Theorem 2.1.

Figure 1(a) ($\Lambda = AG$, $n = 1000$) highlights the above comments the most dramatically. While the proposed FMCI approximation (F) works well in the extreme left tail, its performance degrades as k gets larger. The normal approximation works reasonably well in the neighborhood of $\mathrm{E}[X_n(\Lambda)]$, but performs very poorly in both tails, as expected. The pattern AG cannot be considered rare, occurring on average once in about every $\mathrm{E}[W(\Lambda)] = 16$ trials and, as expected, the Poisson approximation performs poorly for this Λ.

Figure 1(b) ($\Lambda = ACGT$, $n = 5000$) and (c) ($\Lambda = AACAAG$, $n = 70\,000$) show that (i) the FMCI approximation works extremely well in these cases (since $n \gg k$ over most of the k considered) and (ii) the Poisson approximation is still problematic in Figure 1(b) (where $\mathrm{E}[W(\Lambda)] = 256$), but performs better in Figure 1(c) (where $\mathrm{E}[W(\Lambda)] = 4096$). As expected, the performance of the normal approximation actually degrades in these cases since $\mathrm{E}[X_n(\Lambda)] < 20$ in both examples.

We would also like to point out that these tail probabilities can be quite small. Nevertheless, it is sometimes important to be able to estimate these accurately, especially in applications involving likelihood ratios and relative risk assessments.

Table 1 shows the exact probabilities $\mathrm{P}\{X_n(\Lambda) = k\}$ and $\rho(a, e)$ for the FMCI approximation (F) in Theorem 2.1, the Poisson approximation (POI), and the normal approximation (CLT). In this table we report results for the pattern $\Lambda = AACAAG$ in Figure 1(c) and $0 \le k < 40$.

In application, the Poisson approximation is frequently used to estimate $\mathrm{P}\{X_n(\Lambda) = 0\}$ when Λ is rare and $\mathrm{E}[X_n(\Lambda)]$ is small. In Table 2 we present a few comparisons of the FMCI approximation in Theorem 2.1 and the Poisson approximation. The n where chosen such that

$$n = \min\{j\colon\ \mathrm{P}\{X_j(\Lambda) = 0\} < 0.95\},$$

which would be typical in a hypothesis testing situation.

Finally, Godbole and Schaffner (1993) gave an example with $\Lambda = ABRACADABRA$ defined on $\mathcal{A} = \{A, B, C, D, R\}$, which has overlaps. With each letter equally likely, they gave the mean of the approximating Poisson random variable as

$$\lambda = \frac{(n - 20)(5^{10} - 5^3 - 1)}{5^{21}}.$$

With $n = 250\,000$, this yields $\lambda = 0.005\,119\,524$, which is very close to

$$\frac{n}{\mathrm{E}[W(\Lambda)]} = 0.005\,119\,934.$$

For $k = 0, \dots, 4$, the exact and approximate probabilities are given in Table 3.

While both approximations perform well, the proposed approximation does slightly better than the Poisson approximation.

3.2. Approximating $\mathrm{P}\{X_n(\Lambda) < k\}$

Table 4 shows the performance of the approximation(s) given in (2.5) for $\mathrm{P}\{X_n(\Lambda) < k\}$ and that of the normal and Poisson approximations. While it is clear that using all of the terms ($\mathrm{F}(k)$) is preferred, we see that, for large $n/\ell(k+1)$, using only the last two (F(2)) also performs well provided that $n/\ell(k + 1)$ is large enough so that k is in the far-left tail. Using a single term requires much larger n for the approximation to become accurate. Of course, once λ_1 and $\boldsymbol{\eta}_1$ have been determined, there is little additional effort required to make use of all the terms in the sum, and this is recommended. As expected, the CLT approximation does not perform that well at all. Note also that Table 4 illustrates the convergence in (2.4) for each fixed k.

TABLE 1: Exact values of $\mathrm{P}\{X_n(\Lambda) = k\}$ for the simple pattern $\Lambda = AACAAG$ and the values of $\rho(a, e)$ for the approximation proposed in Theorem 2.1 (F), the Poisson approximation (POI), and the normal approximation (CLT).

k	Exact	F	POI	CLT
0	$3.702\,71\times10^{-08}$	0.000	0.024	0.998
1	$6.344\,47\times10^{-07}$	0.000	0.021	0.987
2	$5.434\,66\times10^{-06}$	0.000	0.019	0.957
3	$3.103\,06\times10^{-05}$	0.000	0.016	0.894
4	$1.328\,62\times10^{-04}$	0.000	0.014	0.795
5	$4.550\,22\times10^{-04}$	0.000	0.012	0.663
6	$1.298\,42\times10^{-03}$	0.000	0.010	0.512
7	$3.175\,28\times10^{-03}$	0.000	0.008	0.357
8	$6.793\,43\times10^{-03}$	0.000	0.007	0.214
9	$1.291\,74\times10^{-02}$	0.000	0.005	0.095
10	$2.210\,22\times10^{-02}$	0.000	0.004	0.005
11	$3.437\,44\times10^{-02}$	0.000	0.002	–0.050
12	$4.899\,79\times10^{-02}$	0.000	0.001	–0.077
13	$6.445\,98\times10^{-02}$	0.000	0.001	–0.082
14	$7.873\,14\times10^{-02}$	0.000	–0.000	–0.073
15	$8.973\,77\times10^{-02}$	0.000	–0.001	–0.054
16	$9.587\,49\times10^{-02}$	0.000	–0.001	–0.029
17	$9.639\,12\times10^{-02}$	0.000	–0.001	–0.000
18	$9.151\,20\times10^{-02}$	0.000	–0.001	0.027
19	$8.229\,42\times10^{-02}$	0.000	–0.001	0.050
20	$7.029\,36\times10^{-02}$	0.000	–0.001	0.066
21	$5.717\,47\times10^{-02}$	0.000	–0.001	0.073
22	$4.438\,34\times10^{-02}$	0.000	–0.000	0.069
23	$3.295\,06\times10^{-02}$	0.000	0.001	0.052
24	$2.343\,98\times10^{-02}$	0.000	0.001	0.020
25	$1.600\,47\times10^{-02}$	0.000	0.002	–0.031
26	$1.050\,61\times10^{-02}$	0.000	0.004	–0.096
27	$6.640\,09\times10^{-03}$	0.000	0.005	–0.174
28	$4.046\,17\times10^{-03}$	0.000	0.006	–0.262
29	$2.380\,16\times10^{-03}$	0.000	0.008	–0.356
30	$1.353\,24\times10^{-03}$	0.000	0.010	–0.451
31	$7.444\,53\times10^{-04}$	0.000	0.012	–0.544
32	$3.966\,81\times10^{-04}$	0.000	0.014	–0.631
33	$2.049\,34\times10^{-04}$	0.000	0.016	–0.710
34	$1.027\,43\times10^{-04}$	0.000	0.018	–0.778
35	$5.003\,02\times10^{-05}$	0.000	0.021	–0.835
36	$2.368\,15\times10^{-05}$	0.000	0.024	–0.881
37	$1.090\,48\times10^{-05}$	0.000	0.027	–0.917
38	$4.888\,53\times10^{-06}$	0.000	0.030	–0.944
39	$2.134\,95\times10^{-06}$	0.000	0.033	–0.963

TABLE 2: Exact and approximate values of $\mathrm{P}\{X_n(\Lambda) < 0\}$ for various Λ and n.

Λ	n	Exact	F	POI
$AACAAG$	215	0.949 957 90	0.949 957 90	0.951 182 98
$AACAAGAAT$	13 454	0.949 999 25	0.949 999 25	0.950 029 82
$AACAAGAATT$	53 794	0.949 999 48	0.949 999 48	0.950 008 08

TABLE 3: Exact and approximate probabilities for the pattern $\Lambda = ABRACADABRA$ with $n = 250\,000$ and $0 \le k \le 4$.

k	Exact	F	POI
0	$9.948\,934\times10^{-01}$	$9.948\,934\times10^{-01}$	$9.948\,936\times10^{-01}$
1	$5.093\,587\times10^{-03}$	$5.093\,587\times10^{-03}$	$5.093\,382\times10^{-03}$
2	$1.303\,780\times10^{-05}$	$1.303\,780\times10^{-05}$	$1.303\,785\times10^{-05}$
3	$2.224\,628\times10^{-08}$	$2.224\,628\times10^{-08}$	$2.224\,919\times10^{-08}$
4	$2.846\,657\times10^{-11}$	$2.846\,656\times10^{-11}$	$2.847\,632\times10^{-11}$

TABLE 4: Left tail probabilities $\mathrm{P}\{X_n(\Lambda) < k\}$ and $\rho(a, e)$ for $\Lambda = ACGT$ and various k and n.

Λ	k	n	Exact	F(k)	F(2)	POI	CLT
$ACGT$	1	500	$1.396\,79\times10^{-01}$	0.000		0.038	0.042
	1	1 000	$1.927\,79\times10^{-02}$	0.000		0.064	0.521
	1	2 500	$5.068\,09\times10^{-05}$	0.000		0.137	0.962
	1	5 000	$2.537\,97\times10^{-09}$	0.000		0.247	1.000
	3	100	$9.942\,21\times10^{-01}$	0.000	–0.685	–0.001	0.005
	3	250	$9.286\,10\times10^{-01}$	0.000	–0.405	–0.001	0.013
	3	1 000	$2.499\,27\times10^{-01}$	0.000	–0.077	0.023	–0.058
	3	2 500	$3.100\,47\times10^{-03}$	0.000	–0.016	0.093	0.663
	5	500	$9.548\,90\times10^{-01}$	0.000	–0.725	–0.001	0.013
	5	2 500	$3.255\,13\times10^{-02}$	0.000	–0.095	0.057	0.256
	5	5 000	$2.098\,23\times10^{-05}$	0.000	–0.027	0.168	0.926
	5	10 000	$7.602\,86\times10^{-13}$	0.000	–0.007	0.363	1.000
$AACAAG$	1	100	$9.770\,44\times10^{-01}$	0.000		0.001	0.022
	1	250	$9.418\,64\times10^{-01}$	0.000		0.001	0.021
	1	10 000	$8.685\,91\times10^{-02}$	0.000		0.004	0.186
	1	30 000	$6.537\,08\times10^{-04}$	0.000		0.011	0.887
	3	2 500	$9.761\,62\times10^{-01}$	0.000	–0.557	–0.000	0.016
	3	5 000	$8.754\,66\times10^{-01}$	0.000	–0.337	0.000	0.002
	3	30 000	$2.306\,97\times10^{-02}$	0.000	–0.028	0.006	0.379
	3	50 000	$4.335\,12\times10^{-04}$	0.000	–0.011	0.013	0.839
	5	5 000	$9.918\,37\times10^{-01}$	0.000	–0.883	–0.000	0.007
	5	30 000	$1.451\,58\times10^{-01}$	0.000	–0.159	0.003	0.019
	5	50 000	$6.528\,78\times10^{-03}$	0.000	–0.066	0.009	0.520
	5	100 000	$4.283\,65\times10^{-07}$	0.000	–0.018	0.024	0.984

3.3. Markov-dependent trials

In this section we consider an example where the X_i are (first-order) Markov dependent. Suppose that $\mathcal{A} = \{A, C, G, T\}$ and that $\{X_i : i \geq 0\}$ is a Markov chain with transition probability matrix given by

$$\boldsymbol{P}_X = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \left[\begin{array}{cccc} 0.2 & 0.3 & 0.2 & 0.3 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.1 & 0.3 & 0.3 & 0.3 \\ 0.2 & 0.4 & 0.2 & 0.2 \end{array}\right] \end{array}.$$

Note that the stationary distribution for this chain is

$$\boldsymbol{\pi} = \left(\tfrac{16}{90}, \tfrac{27}{90}, \tfrac{20}{90}, \tfrac{27}{90}\right).$$

The state space $\Omega_{\boldsymbol{P}}$ is similar to that for the i.i.d. case except that we replace ϕ with $\phi_A, \phi_C, \phi_G, \phi_T$. For $\Lambda = CAA$, we have

$$\Omega_{\boldsymbol{P}} = \{\phi_A, \phi_C, \phi_G, \phi_T, C, CA, CAA\}$$

and

$$\Omega_{\boldsymbol{N}} = \{\phi_A, \phi_C, \phi_G, \phi_T, C, CA\}.$$

The essential transition probability matrix is given by

$$\boldsymbol{N} = \begin{array}{c|cccccc} & \phi_A & \phi_C & \phi_G & \phi_T & C & CA \\ \hline \phi_A & p_{AA} & 0 & p_{AG} & p_{AT} & p_{AC} & 0 \\ \phi_C & p_{CA} & 0 & p_{CG} & p_{CT} & p_{CC} & 0 \\ \phi_G & p_{GA} & 0 & p_{GG} & p_{GT} & p_{GC} & 0 \\ \phi_T & p_{TA} & 0 & p_{TG} & p_{TT} & p_{TC} & 0 \\ C & 0 & 0 & p_{AG} & p_{AT} & p_{CC} & p_{CA} \\ CA & 0 & 0 & p_{CG} & p_{CT} & p_{AC} & 0 \end{array}.$$

Table 5 shows the exact tail probabilities and the approximation ratios for the simple pattern $\Lambda = CAA$ and various k and n. The exact probabilities (and the approximations) were calculated by taking

$$\xi_0 = \xi_\Lambda = (1, 0, 0, 0, 0, 0)$$

so that the interarrival times W_i are i.i.d. and the CLT is easier to apply. Normally, we would assume that the $\{X_i\}$ chain was stationary and take

$$\xi_0 = \left(\tfrac{16}{90}, \tfrac{27}{90}, \tfrac{20}{90}, \tfrac{27}{90}, 0, 0\right).$$

As in the i.i.d. case, we see that the approximation $\mathrm{F}(k)$ works very well and the two-term approximation $\mathrm{F}(2)$ performs well for large $n/\ell(k+1)$. Results for $\mathrm{P}\{X_n(\Lambda) = k\}$ are analogous to those in the i.i.d. case and are not tabulated. The Poisson approximations for this case were not tabulated since CAA is not a 'rare' pattern. In general, we expect the results for the Poisson approximation in Markov-dependent trials to mirror those for the i.i.d. case.

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press

Table 5: Left tail probabilities $\mathrm{P}\{X_n(\Lambda) < k\}$ and $\rho(a, e)$ for $\Lambda = CAA$ and various k and n when the $\{X_i\}$ are first-order Markov dependent (see text).

k	n	Exact	F(k)	F(2)	CLT
1	250	$4.646\,59\times10^{-02}$	–0.000	–0.000	0.319
1	500	$2.105\,31\times10^{-03}$	–0.000	–0.000	0.795
1	1 000	$4.321\,91\times10^{-06}$	–0.000	–0.000	0.986
1	2 500	$3.739\,02\times10^{-14}$	–0.000	–0.000	1.000
5	500	$2.800\,53\times10^{-01}$	0.001	–0.205	–0.058
5	1 000	$6.486\,29\times10^{-03}$	0.001	–0.062	0.492
5	2 500	$1.815\,78\times10^{-09}$	0.000	–0.011	0.998
5	5 000	$9.991\,37\times10^{-22}$	0.000	–0.003	1.000
10	1 000	$2.364\,13\times10^{-01}$	0.003	–0.353	–0.034
10	2 500	$4.522\,55\times10^{-06}$	0.001	–0.066	0.920
10	5 000	$7.407\,14\times10^{-17}$	0.001	–0.017	1.000
10	10 000	$4.759\,92\times10^{-41}$	0.000	–0.004	1.000
20	2 500	$1.917\,17\times10^{-02}$	0.005	–0.313	0.201
20	5 000	$2.644\,92\times10^{-10}$	0.003	–0.081	0.992
20	10 000	$1.669\,14\times10^{-31}$	0.001	–0.020	1.000
20	20 000	$1.488\,25\times10^{-79}$	0.001	–0.005	1.000

4. Concluding comments

In this paper we have developed an approximation for $\mathrm{P}\{X_n(\Lambda) = k\}$ and $\mathrm{P}\{X_n(\Lambda) < k\}$ based on the FMCI of Λ that work very well for fixed k and large n. The approximations are appropriate for both i.i.d. and Markov-dependent multistate trials; both overlapping and nonoverlapping counting; and both simple and compound patterns. The proposed approximations perform very well and, in many cases, outperform the typical normal and Poisson approximations.

Acknowledgement

The authors would like to thank Alexandre Leblanc for his insightful comments on an earlier version of this paper.

References

Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen–Stein method. *Statist. Sci.* **5,** 403–434.

Barbour, A. D., Chryssaphinou, O. and Roos, M. (1996). Compound Poisson approximation in systems reliability. *Naval Res. Logistics* **43,** 251–264.

Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: a Markov chain approach. *J. Amer. Statist. Assoc.* **89,** 1050–1058.

Fu, J. C. and Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns and Its Applications*. World Scientific, River Edge, NJ.

Fu, J. C., Wang, L.and Lou, W. Y. W. (2003). On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state Markov dependent trials. *J. Appl. Prob.* **40,** 346–360.

Godbole, A. P. (1991). Poisson approximations for runs and patterns of rare events. *Adv. Appl. Prob.* **23,** 851–865.

Godbole, A. P. and Schaffner, A. A. (1993). Improved Poisson approximations for word patterns. *Adv. Appl. Prob.* **25,** 334–347.

Mood, A. M. (1940). The distribution theory of runs. *Ann. Math. Statist.* **11,** 367–392.

Riordan, J. (1958). *An Introduction to Combinatorial Analysis*. John Wiley, New York.

Wald, J. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.* **11,** 147–162.

Wishart, J. and Hirschfeld, H. O. (1936). A theorem concerning the distribution of joins between line segments. *J. London Math. Soc.* **11,** 227–235.

Wolfowitz, J. (1943). On the theory of runs with some applications to quality control. *Ann. Math. Statist.* **14,** 280–288.

https://doi.org/10.1239/aap/1240319586 Published online by Cambridge University Press