Ann. Inst. Statist. Math.
Vol. 54, No. 4, 719–730 (2002)

# THE EXACT AND LIMITING DISTRIBUTIONS FOR THE NUMBER OF SUCCESSES IN SUCCESS RUNS WITHIN A SEQUENCE OF MARKOV-DEPENDENT TWO-STATE TRIALS

JAMES C. FU[1], W. Y. WENDY LOU[2], ZHI-DONG BAI[3] AND GANG LI[4]

[1]*Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2*
[2]*Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada M5S 1A8*
[3]*Department of Statistics and Applied Probability, National University of Singapore, Singapore 119260, Singapore*
[4]*Biometrics, Organon, Inc., West Orange, NJ 07052, U.S.A.*

(Received April 16, 2001; revised October 26, 2001)

**Abstract.** The total number of successes in success runs of length greater than or equal to $k$ in a sequence of $n$ two-state trials is a statistic that has been broadly used in statistics and probability. For Bernoulli trials with $k$ equal to one, this statistic has been shown to have binomial and normal distributions as exact and limiting distributions, respectively. For the case of Markov-dependent two-state trials with $k$ greater than one, its exact and limiting distributions have never been considered in the literature. In this article, the finite Markov chain imbedding technique and the invariance principle are used to obtain, in general, the exact and limiting distributions of this statistic under Markov dependence, respectively. Numerical examples are given to illustrate the theoretical results.

*Key words and phrases*: Finite Markov chain imbedding, transition probability matrix, runs and patterns.

## 1. Introduction

One of the simplest and most fundamental models of outcome measurement in statistics and probability is a sequence of two-state trials $\{X_t\}_{t=1}^n$. For instance, the particular dichotomized outcomes may be 1 and 0, success (S) and failure (F), or acceptance and rejection. The total number of successes in $n$ trials,

$$S_n(1) = X_1 + X_2 + \cdots + X_n, \tag{1.1}$$

has a binomial distribution when the random variables $\{X_t\}_{t=1}^n$ are independent and identically distributed (i.i.d.) with $P(X_t = 1) = p$ and $P(X_t = 0) = q$. In general, let $S_n(k)$ be the total number of successes in all success runs of length greater than or equal to $k$ ($k \geq 1$) in a sequence of $n$ two-state trials; i.e.

$$S_n(k) = \sum_{i=k}^{n} iR_n(i), \tag{1.2}$$

where $R_n(i)$, $i = k, \dots, n$, is the number of success runs with length exactly equal to $i$ in the sequence $\{X_t\}_{t=1}^n$. For $k = 1$, it is easy to see that (1.2) and (1.1) are equivalent. The statistic $S_n(k)$ has been broadly used in various areas of statistics and applied probability, such as estimation, hypothesis testing, and DNA sequencing, especially when $k = 1$. For example, Benson (1999) used $S_n(k)$ to form the basis for a sophisticated algorithm to detect DNA tandem repeats, which are segments within a DNA sequence that are repeated at least once in a contiguous fashion, and which have been implicated in the causation of several genetic diseases, such as Huntington's disease (Huntington's Disease Collaborative Research Group (1993)). Tandem repeats are subject to random mutations, so that typically only approximate copies are present. The idea of his approach involves finding matching $k$-tuples between two adjacent DNA segments by aligning one on top of the other, and then converting them into one success-failure sequence of matches. The statistic $S_n(k)$, defined in this application as the sum of matches in matching runs of length $k$ or longer, is then tested for statistical significance to search for potential repeats within DNA sequences.

Markov-dependent two-state trials, only the limiting distribution of $S_n(1)$ has been examined (Nagaev (1957)). There are several difficulties encountered in computing the exact distributions for $S_n(k)$ with $k \geq 2$. The two main difficulties are (i) the probabilities of sequences of outcomes, even among those with the same number of successes and failures, are quite different due to the Markov dependence of the trials, and (ii) the complexity of the joint distribution of success runs $R_n(i)$ with run sizes $i \geq k$. The traditional combinatorial approach is not efficient in dealing with Markov-dependent trials. Currently both exact and limiting distributions of $S_n(k)$ for $k \geq 2$ under Markov-dependent trials remain unknown. In order to avoid the above-mentioned difficulties, we adopt the finite Markov chain imbedding approach (see Fu and Koutras (1994), and Fu (1986, 1996)) to study the exact distributions of $S_n(k)$, and show that the exact probabilities of $S_n(k) = x$ can be expressed concisely in terms of transition probability matrices of an imbedded Markov chain.

Using the invariance principle, we are able to show that

$$\frac{1}{\sqrt{n}}(S_n(k) - ES_n(k)) \to N(0, V(k)) \quad \text{as} \quad n \to \infty \tag{1.3}$$

for $k = 1, 2, \ldots$, where the variance $V(k)$ is determined from the transition probability matrices.

Numerical examples are presented in Section 5, and show that the exact distributions are highly skewed towards the right, and converge to normal distributions rather slowly, especially for large $k$. The mean and variance do not provide complete information for the distribution of $S_n(k)$. Since the computation we propose herein for the exact distribution is simple and efficient, we would like to suggest that, in practice, the limiting distribution be used only when $n$ is very large.

## 2. Notation and preliminary results

Let $\{X_t\}_{t=1}^n$ be a sequence of two-state (1 and 0 or $S$ and $F$) homogeneous Markov-dependent trials with the transition probability matrix

$$A = \begin{pmatrix} p_1 & 1 - p_1 \\ p_2 & 1 - p_2 \end{pmatrix}, \tag{2.1}$$

where $P(X_t = 1 \mid X_{t-1} = 1) = p_1$, $P(X_t = 0 \mid X_{t-1} = 1) = q_1 = 1 - p_1$, $P(X_t = 1 \mid X_{t-1} = 0) = p_2$, and $P(X_t = 0 \mid X_{t-1} = 0) = q_2 = 1 - p_2$, with $0 < p_1 < 1$, $0 < p_2 < 1$, and initial distribution $\pi_1 = (P(X_1 = 1) = p_0, P(X_1 = 0) = q_0)$. The eigenvector matrix $B$ associated with the transition probability matrix $A$ has the form

$$B = \begin{pmatrix} 1 & -q \\ 1 & p \end{pmatrix},$$

where $p = p_2/(1 - p_1 + p_2)$, and $q = (1 - p_1)/(1 - p_1 + p_2)$. It then follows that

$$B^{-1}AB = \begin{pmatrix} 1 & 0 \\ 0 & p_1 - p_2 \end{pmatrix}$$

and

$$\pi_1 A^{n-1} = (p - a(p_1 - p_2)^{n-1}, q + a(p_1 - p_2)^{n-1}),$$

where $a = -p_0 q + q_0 p$.

Let $L_j$, for $j \geq 2$, be the length of the success run located between the $(j-1)$-th and $j$-th failures in the sequence $\{X_t\}_{t=1}^{\infty}$, with $L_1 = 0$ if the first trial is a failure and $L_1 = l$ if the first $l$ trials are successes and the $(l+1)$-th trial is a failure. For given $t$, let $m_t$ be the number of failures in the subsequence $X_1, X_2, \ldots, X_t$, and let $L_t^\star$ represent the number of successes that occur after the $m_t$-th failure in this subsequence. It is clear that $0 \leq L_t^\star \leq t$ and $0 \leq L_t^\star \leq L_{m_t+1}$. Moreover, $S_t(k)$, as defined by (1.2), can also be written as

$$S_t(k) = \sum_{j=1}^{m_t} L_j(k) + L_t^\star(k), \tag{2.2}$$

where

$$L_j(k) = L_j \cdot I_{\{L_j \geq k\}}, \quad \text{and} \quad L_j^\star(k) = L_j^\star \cdot I_{\{L_j^\star \geq k\}}. \tag{2.3}$$

Here $I_{\{L_j \geq k\}}$ is an indicator function equal to one if $L_j \geq k$ and zero otherwise.

Further, we define a new sequence of random variables

$$Y_t = (S_t(k), E_t(k)), \qquad t = 1, 2, \ldots, \tag{2.4}$$

where $E_t(k) = L_t^\star \cdot (1 - I_{\{L_t^\star \geq k\}}) + k^+ I_{\{L_t^\star \geq k\}}$ is the ending-block random variable, and $k^+$ is a symbol indicating that $L_t^\star$ is greater than or equal to $k$. In words, $E_t(k)$ represents the length of the success run counting backward from the $t$-th trial with $E_t(k) = 0$ if the $t$-th trial is a failure and $E_t(k) = k^+$ if the length is greater than or equal to $k$. If the distribution of $Y_t$ is known, then the distribution of $S_n(k)$ can be obtained by projecting the distribution of $Y_t$ onto $S_t(k)$. The finite Markov chain imbedding technique will be used to accomplish this goal in the next section. The representation of $S_t(k)$ given in (2.2) facilitates a direct application of the invariance principle to obtain the limiting distribution.

Note that for each given $t$, the two components of the vector $Y_t = (S_t(k), E_t(k))$ contain the vital information for the sequence $X_1, X_2, \ldots, X_t$:

(i) $S_t(k)$ indicates the total number of successes in success runs of length greater than or equal to $k$ in the first $t$ trials.

(ii) $E_t(k)$ is the ending-block of the first $t$ trials, and also provides essential information about the transition probability from $Y_t$ to $Y_{t+1}$.

(iii) Given the sequence $X_1, \ldots, X_n$, the sequence $\{Y_t : t = 1, \ldots, n\}$ is uniquely determined.

In view of these facts, we are able to obtain the exact distribution of $S_n(k)$, via the projection of the joint distribution of $S_n(k)$ and $E_n(k)$ onto $S_n(k)$, using the finite Markov chain imbedding technique. The details of this approach are provided in the following section.

## 3. The exact distribution

Given $n$, the total number of successes $S_n(k)$ can take on the possible values of $0, k, k+1, \ldots, n$. The random variable $S_n(k)$, by itself, is not a Markov chain, even in the case where $X_1, \ldots, X_n$ are i.i.d. Bernoulli trials. However, we can show that $S_n(k)$ is finite Markov chain imbeddable in the sense of Fu and Koutras (1994): (i) there exists a Markov chain $\{Y_t : t = 1, \ldots, n\}$ defined on a state space $\Omega$ with initial probability $\xi_0$ and transition probability matrices $M_t$, $t = 1, \ldots, n$, and (ii) there exists a finite partition $\{C_x; x = 0, k, \ldots, n\}$ on the state space $\Omega$ such that

$$P(S_n(k) = x) = P(Y_n \in C_x \mid \xi_0)$$

for all $x = 0, k, \ldots, n$. For the sake of clarity in the following discussion, the two states 1 and 0 will also be denoted by success ($S$) and failure ($F$), respectively.

For $1 \leq t \leq n$ with $n$ given, a possible ending block (the $m_t$-th $F$ and the last success run of the subsequence $X_1, \ldots, X_t$) can only be one of the following cases: $\{F, FS, \ldots, FS\cdots S, \text{ or } S\cdots S(\text{if } m_t = 0)\}$. The random variable $E_t(k)$ is the number of successes in the ending block $FS\cdots S$ if it is less than $k$, and $E_t(k) = k^+$ if it is equal to or greater than $k$. We define a state space

$$\Omega = \{(u,v) : u = 0, k, \ldots, n-1, n; v = 0, 1, \ldots, k-1, k^+\} \tag{3.1}$$

with size $d = Card(\Omega) = (n-k+2)(k+1)$.

In our counting procedure, the sequence of random vectors

$$\{Y_t = (S_t(k), E_t(k)), t = 1, 2, \ldots, n\}$$

defined on $\Omega$ obeys the following rules:

(i) Given $Y_{t-1} = (x, 0)$, then $Y_t = (x, 0)$ with probability $q_2$ if the outcome of the $t$-th trial is $F$, and $Y_t = (x, 1)$ with probability $p_2$ if the outcome of the $t$-th trial is $S$.

(ii) Given $Y_{t-1} = (x, y)$ and $1 \leq y \leq k-2$, then $Y_t = (x, 0)$ with probability $q_1$ if the outcome of the $t$-th trial is $F$, and $Y_t = (x, y+1)$ with probability $p_1$ if the outcome of the $t$-th trial is $S$.

(iii) Given $Y_{t-1} = (x, k-1)$, then $Y_t = (x, 0)$ with probability $q_1$ if the outcome of the $t$-th trial is $F$, and $Y_t = (x+k, k^+)$ with probability $p_1$ if the outcome of the $t$-th trial is $S$.

(iv) Given $Y_{t-1} = (x, k^+)$, then $Y_t = (x, 0)$ with probability $q_1$ if the outcome of the $t$-th trial is $F$, and $Y_t = (x+1, k^+)$ with probability $p_1$ if the outcome of the $t$-th trial is $S$.

In view of our construction, the sequence $\{Y_t = (S_t(k), E_t(k)); t = 1, 2, \ldots, n\}$ forms a homogeneous Markov chain with transition probability matrix

$$M = \left(p_{(x,y),(u,v)}\right)_{d \times d},$$

where the paired states $(\cdot,\cdot)$ are lexicographically ordered, and the transition probabilities $p_{(x,y),(u,v)}$ can be specified explicitly as follows. Given $(x,y) \in \Omega$,

$$p_{(x,y),(u,v)} = \begin{cases} q_2 & \text{if } v = y = 0 \text{ and } u = x, \\ p_2 & \text{if } y = 0, v = 1, \text{ and } u = x, \\ q_1 & \text{if } y \neq 0, v = 0, \text{ and } u = x, \\ p_1 & \text{if } 1 \leq y < k-1, v = y+1, \text{ and } u = x, \\ & \text{or if } y = k-1, v = k, \text{ and } u = x+k, \\ & \text{or if } y = k^+, v = k^+, \text{ and } u = x+1, \\ 1 & \text{if } v = y \text{ and } u = x = n, \\ 0 & \text{otherwise.} \end{cases}$$

Hence the random variable $S_n(k)$ is finite Markov chain imbeddable, and thence the exact probabilities can be obtained by

$$P(S_n(k) = x) = \xi_0 M^{n-1} U'(C_x), \qquad x = 0, k, \ldots, n, \tag{3.2}$$

where the $1 \times d$ vector $\xi_0 = (q_0, p_0, 0, \ldots, 0)$ is the initial distribution of $Y_t$, the partition $\{C_x\}$ is defined as

$$C_x = \{(x,y) : y = 0, 1, 2, \ldots, k^+\}, \qquad x = 0, k, \ldots, n,$$

and $U'(C_x)$ is the transpose of the $(1 \times d)$ row vector $U(C_x) = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$ with ones at the coordinates corresponding to states $(x,y) \in C_x$.

The moments of $S_n(k)$ can be computed from

$$ES_n^r(k) = \sum_x \xi_0 M^{n-1} U'(x^r), \qquad r = 1, 2, \ldots, \tag{3.3}$$

where $\sum_x$ is the sum over $x = 0, k, \ldots, n$, and $U'(x^r)$ is the transpose of the $(1 \times d)$ row vector $U(x^r) = (0, \ldots, 0, x^r, \ldots, x^r, 0, \ldots, 0)$ with $x^r$ at the coordinates corresponding to states $(x,y)$ in $C_x$.

## 4. Limiting distribution

The sequence $\{X_t\}_{t=1}^{\infty}$ is separated into blocks of consecutive successes by failures. The size of the $j$-th block is denoted by $L_j$, and $L_j = 0$ if the $(j-1)$-th failure is followed immediately by another failure. It follows from the definition of $L_j$ given in Section 2 that $P(L_1 = l) = p_0 p_1^{l-1} q_1$ if $l \geq 1$ and that $P(L_1 = 0) = q_0$. Further, it can be easily calculated that for every $i$,

$$P(L_{j+1} = l \mid L_j = i) = p_2 p_1^{l-1} q_1 \quad \text{ if } \quad l \geq 1,$$

and

$$P(L_{j+1} = 0 \mid L_j = i) = q_2.$$

Hence $\{L_j, j \geq 2\}$ is a sequence of i.i.d. random variables. This crucial fact provides the foundation for finding the limiting distribution of $S_n(k)$ when the sequence $\{X_t\}_{t=1}^{\infty}$ has a Markov-dependent structure (note that the distribution of $L_1$ differs from that of $L_j$, $j \geq 2$). As we will see, the distribution of $S_n(k)$ is closely related to the partial sums of $\{L_j(k)\}$ and $\{L_j\}$. To obtain the limiting distribution, we need the following lemma:

LEMMA 4.1. *Given $k \geq 1$,*
(i) $\mu_k = E(L_1(k)) = p_2 p_1^{k-1}(kq_1 + p_1)q_1^{-1}$,
(ii) $\sigma_k^2 = \mathrm{Var}(L_1(k)) = [p_2 p_1^{k-1}(kq_1 + p_1)^2(1 - p_2 p_1^{k-1}) + p_2 p_1^k]q_1^{-2}$,
(iii) $\sigma_{1k} = \mathrm{Cov}(L_1, L_1(k)) = [p_2 p_1^{k-1}((kq_1 + p_1)^2 + p_1) - p_2^2 p_1^{k-1}(kq_1 + p_1)]q_1^{-2}$.

PROOF. Results (i) and (ii) may be obtained directly, with

$$\mu_k = EL_1(k) = \sum_{i=k}^{\infty} i p_2 q_1 p_1^{i-1} = p_2 p_1^{k-1}(kq_1 + p_1)q_1^{-1}, \tag{4.1}$$

and

$$\begin{aligned} E(L_1^2(k)) &= \sum_{i=k}^{\infty} i^2 p_2 q_1 p_1^{i-1} \\ &= [p_2 q_1 p_1 k(k-1) p_1^{k-2} q_1^2 + 2k p_1^{k-1} q_1 + 2 p_1^k] q_1^{-2} + \mu_k \\ &= p_2 p_1^{k-1}((kq_1 + p_1)^2 + p_1) q_1^{-2}. \end{aligned} \tag{4.2}$$

Note that $L_1 = L_1(1)$ and $E(L_1 L_1(k)) = E(L_1^2(k))$. Therefore, Result (iii) also follows from (4.1) and (4.2). This completes the proof.

From Lemma 4.1, it can also be shown that

$$\begin{aligned} E(L_1^3(k)) &= \sum_{i=k}^{\infty} i^3 p_2 q_1 p_1^{i-1} \\ &= p_2 p_1^{k-1}[(kq_1 + p_1)^3 - 2q_1^2 p_1 + 3kq_1 p_1 + 3q_1 p_1 + 3q_1 p_1^2 + 5p_1^3] q_1^{-3}. \end{aligned}$$

THEOREM 4.1. *If the sequence $\{X_t\}$ is a homogeneous Markov chain with transition probability matrix $A$ as given by* (2.1), *then*

$$\frac{1}{\sqrt{n}}(S_n(k) - ES_n(k)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V(k)) \quad \text{as} \quad n \to \infty,$$

*where*

$$V(k) = q(1, -q\mu_k)\Sigma_k(1, -q\mu_k)',$$
$$q = q_1(1 - p_1 + p_2)^{-1}, \quad \text{and} \quad \Sigma_k = \begin{pmatrix} \sigma_k^2 & \sigma_{1k} \\ \sigma_{1k} & \sigma_1^2 \end{pmatrix}.$$

PROOF. From the invariance principle (see Billingsley (1968) or Mcleish (1974)), it follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{[n\xi]} \begin{pmatrix} L_i(k) - \mu_k \\ L_i - \mu_1 \end{pmatrix} \to \Sigma_k^{1/2} \begin{pmatrix} W_1(\xi) \\ W_2(\xi) \end{pmatrix} \tag{4.3}$$

on $\xi \in [0, 1]$ as $n \to \infty$, where $W_1(\xi)$ and $W_2(\xi)$ are two independent standard Brownian motions.

It follows from equation (0.4) of Nagaev (1961) that $m_n/n \to q$ in probability. Then by (4.3) and noticing $m_n + 1$ is a stopping time, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{m_n}\begin{pmatrix} L_i(k) - \mu_k \\ L_i - \mu_1 \end{pmatrix} \to \mathcal{N}(0, q\Sigma_k). \tag{4.4}$$

Note that $n = m_n + \sum_{i=1}^{m_n} L_i + L_n^\star$. It follows from the definitions of $L_j$ and $L_j^\star$, and from $EL_1 = \mu_1 = p_2/q_1$, that

$$\begin{aligned} m_n - nq &= np - \sum_{i=1}^{m_n} L_i - L_n^\star \\ &= -\sum_{i=1}^{m_n}(L_i - \mu_1) + \mu_1(nq - m_n) - L_n^\star \\ &= -q\sum_{i=1}^{m_n}(L_i - \mu_1) - qL_n^\star. \end{aligned} \tag{4.5}$$

Since $E(S_n(k)) = E(m_n)\mu_k + E(L_n^\star(k)) = nq\mu_k + O(1)$ and $0 \le L_n^\star \le L_{m_n+1}$, the next result follows immediately from (4.4) and (4.5):

$$\begin{aligned} \frac{1}{\sqrt{n}}(S_n(k) - ES_n(k)) &= \frac{1}{\sqrt{n}}(S_n(k) - m_n\mu_k) + \frac{1}{\sqrt{n}}(m_n - nq)\mu_k + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{m_n}(L_i(k) - \mu_k) \\ &\quad -q\mu_k \times n^{-1/2}\sum_{i=1}^{m_n}(L_i - \mu_1) + o_p(1) \\ &\xrightarrow{L} \mathcal{N}(0, V(k)), \end{aligned} \tag{4.6}$$

as $n \to \infty$. This completes the proof.

For $k = 1$, $S_n(1)$ is the total number of successes and $q(1, -p)\Sigma_1(1, -p)' = pq$. In this case our results reduce to the well-known fact

$$n^{-1/2}(S_n(1) - np) \xrightarrow{L} \mathcal{N}(0, pq).$$

Numerical examples for $k > 1$ are given in the following section, and comparisons between the exact and limiting distributions are also provided.

## 5. Numerical examples

It can be seen from the construction of the imbedded finite Markov chain described in Section 3, and from Theorem 4.1 given in Section 4, that both the exact and limiting distributions of $S_n(k)$ are functions of $k$, $p_1$ and $p_2$. To gain further insight on the effects of these parameters, the distributions of $S_n(k)$ for some selected parameters are graphically presented in Figs 1–3 for $n = 15$, 30, and 60, respectively, where the initial distributions $\xi_0 = (1, 0, \dots, 0)$ are assumed. For purposes of comparison, the expectations are also included.

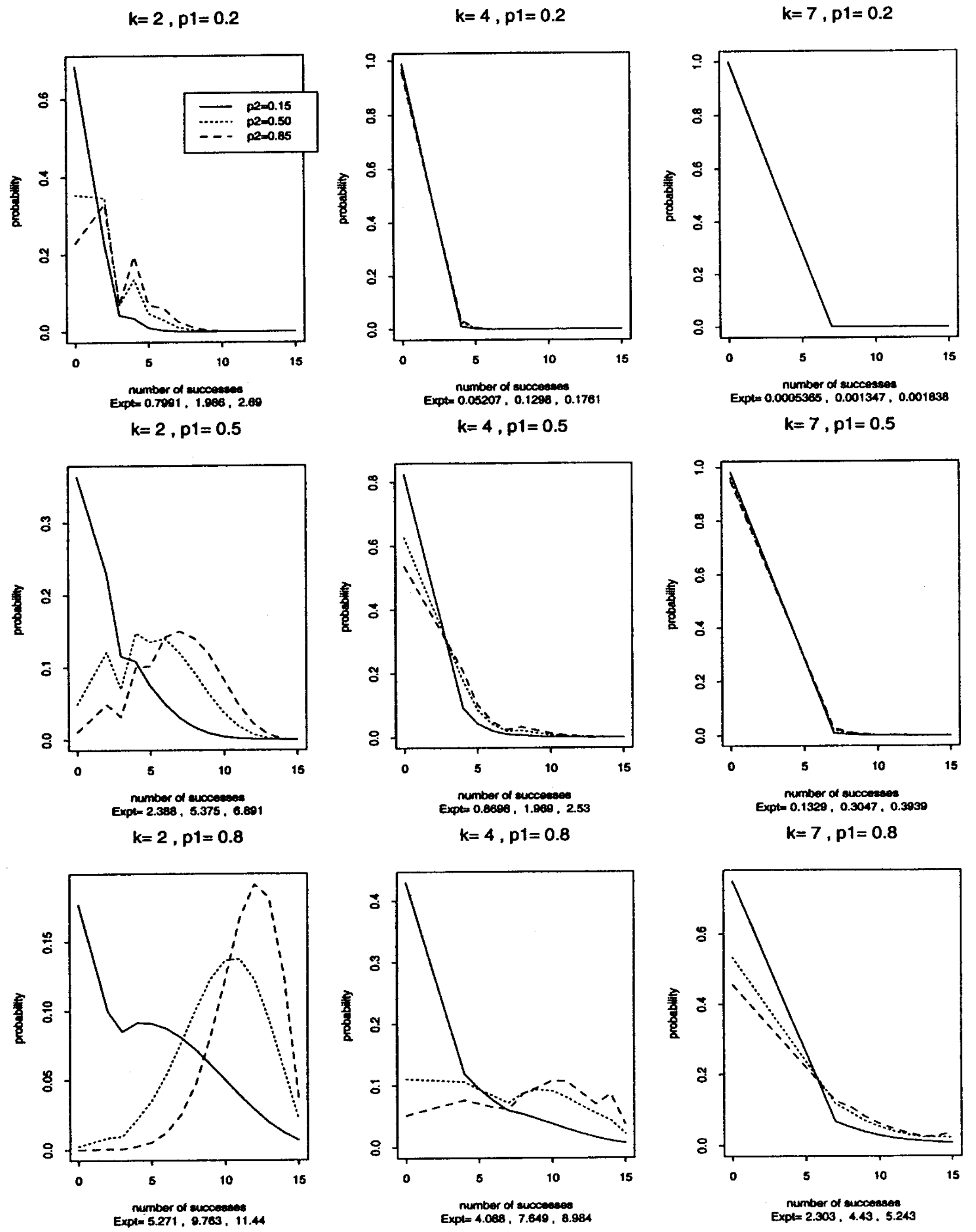

Fig. 1. Distributions of $S_n(k)$ with $n = 15$ and $k = 2, 4, 7$ for some selected $p_1$ $(= 0.2, 0.5, 0.8)$ and $p_2$ $(= 0.15, 0.5, 0.85)$. $ES_n(k)$=Expt.

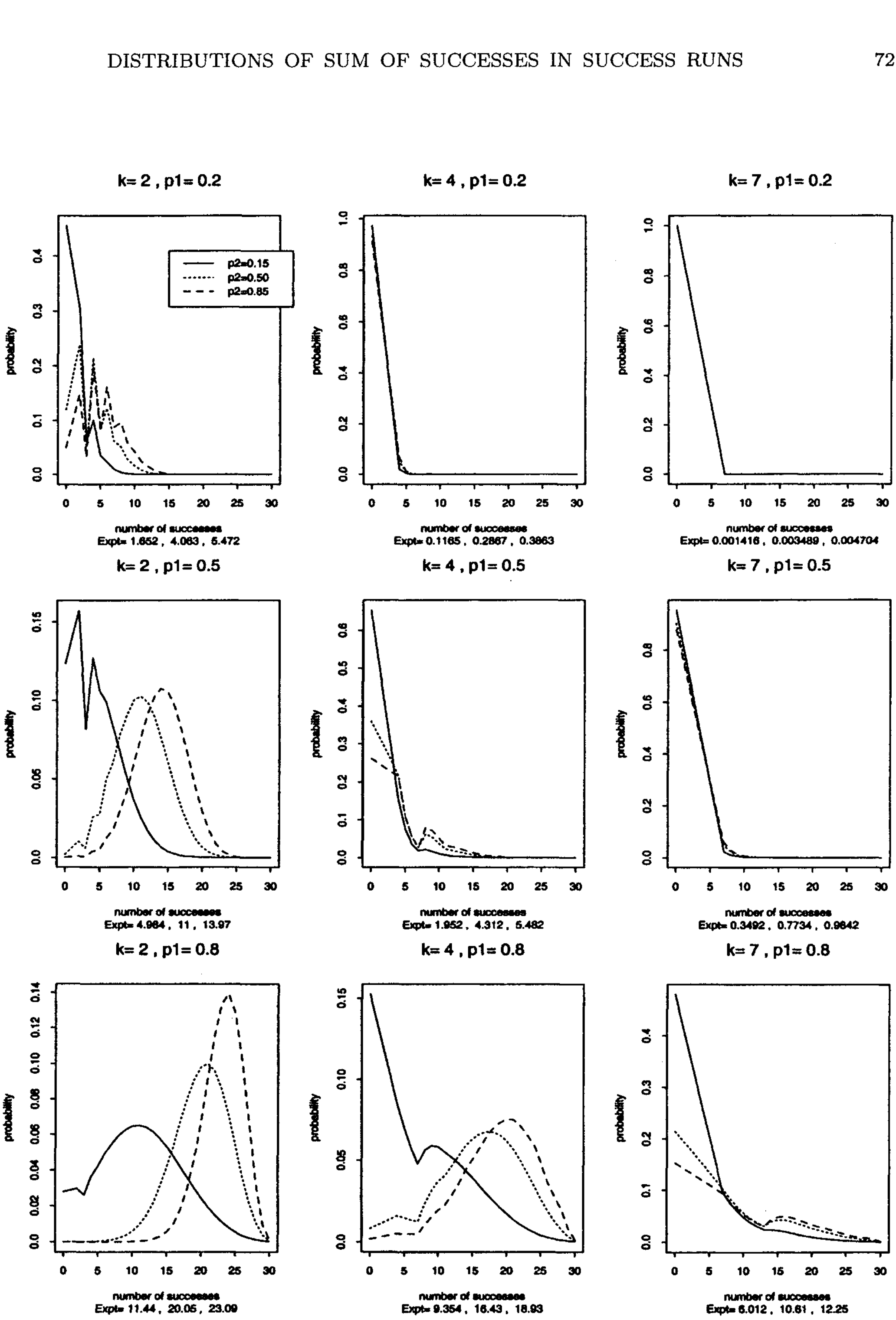

Fig. 2. Distributions of $S_n(k)$ with $n = 30$ and $k = 2, 4, 7$ for some selected $p_1$ $(= 0.2, 0.5, 0.8)$ and $p_2$ $(= 0.15, 0.5, 0.85)$. $ES_n(k)$=Expt.

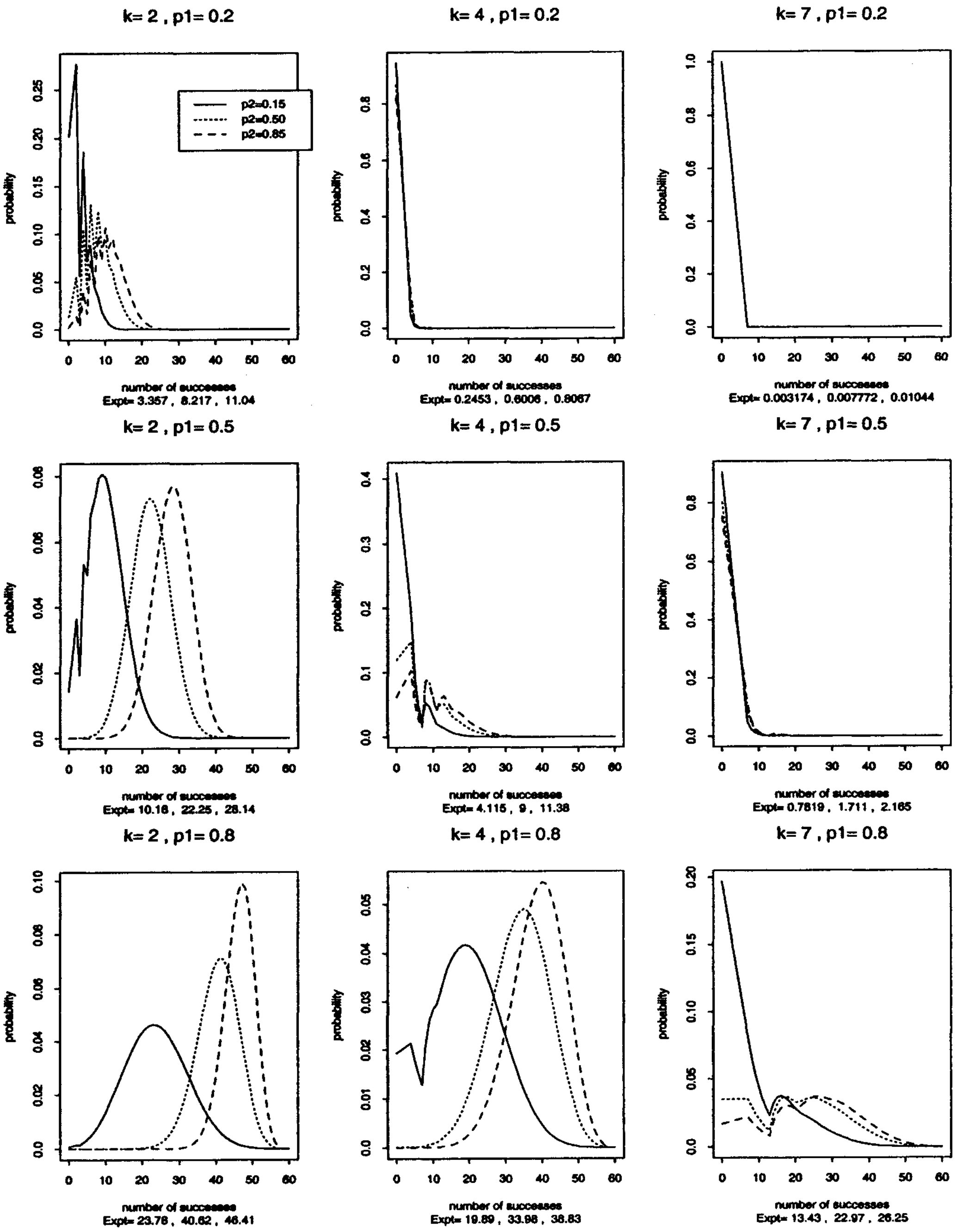

Fig. 3. Distributions of $S_n(k)$ with $n = 60$ and $k = 2, 4, 7$ for some selected $p_1$ $(= 0.2, 0.5, 0.8)$ and $p_2$ $(= 0.15, 0.5, 0.85)$. $ES_n(k)$=Expt.

Given $n$, the effect of $k$ and $p_1$ can be summarized as follows. For small $k$ (e.g. $k = 2$), the distribution becomes smoother and bell-shaped as $p_1$ increases, and this tendency is amplified for larger values of $p_2$. The distributions are highly skewed to the right and drift away from a normal shape as $k$ increases, even though the changes are somewhat less pronounced for larger $p_1$. When $n$ increases, as from Figs 1 to 2 to 3 ($n = 15,30,60$), the distributions are better approximated by a normal distribution for small $k$ and large $p_1$, and less well for large $k$ (e.g. $k = 7$) and small $p_1$ (e.g. $p_1 = 0.2$).

In view of Figs 1–3, it appears that the distribution of $S_n(k)$ is predominantly determined by $p_1^{k-1}$ and seems to have a linear relationship with $p_2$. These findings based on the exact distribution are consistent with the results for the limiting distribution.

## 6. Discussion

In our numerical studies, almost all distributions of $S_n(k)$ are highly skewed, especially when $k$ is large. Here we provide an intuitive explanation of this phenomenon.

Let $\eta_i = L_i(k) - q\mu_k L_i$. From (4.6), $S_n(k) - ES_n(k) = (1/\sqrt{n})\sum_{i=1}^{m_n} \eta_i + o_p(1)$. In order to understand the asymptotic normality of $S_n(k)$, we study the Edgeworth expansion remainder term (see Petrov (1975)) of the distribution $F_n$ of $[1/\sqrt{nV(k)}](S_n(k) - ES_n(k))$. In this case,

$$\|F_n(x) - \Phi(x)\| \le C_r[\beta/\sqrt{Em_n}\phi(x) + B_r Em_n^{-(r-2)/2}/(1+|x|)^r], \tag{6.1}$$

where $\Phi$ is the standard normal distribution, $\beta$ is the skewness of $\eta_1$,

$$\beta = E\eta_1^3/[(E|\eta_1|^2)^{3/2}],$$

and

$$B_r = E|\eta_1|^r/(E|\eta_1|^2)^{r/2}$$

for any $r > 2$. $B_r$ is an index of the tail probability of $\eta_1$ and is expected to be bounded for large $r$. We choose $r$ large, so that the right-hand side of (6.1) is dominated by the skewness of $\eta_1$.

Roughly speaking, the skewness of $\eta_1$ is determined by the skewness of $L_1(k)$ and $L_1$. The skewness of $L_1(k)$ is tabulated in Table 1 for $k = 1, 4, 7$, and 15 with $p = p_1 = p_2 = 0.1$ to 0.9. It is clear that, when $k$ is large, there are no common values of $p$ such

Table 1. Coefficient of skewness $\beta$ for $L_1(k)$, with $k = 1, 4, 7, 15$ and $p$=0.1 to 0.9.

| $p$ | $k$ | | | |
|---|---|---|---|---|
| | 1 | 4 | 7 | 15 |
| 0.10 | 2.15035 | 24.2097 | 444.15 | 2092118.26 |
| 0.20 | 0.71554 | 5.7268 | 38.43 | 11447.58 |
| 0.30 | −0.14606 | 2.2325 | 9.01 | 540.39 |
| 0.40 | −0.82219 | 0.8776 | 3.08 | 61.46 |
| 0.50 | −1.41421 | 0.0831 | 1.15 | 11.25 |
| 0.60 | −1.96231 | −0.5576 | 0.21 | 2.71 |
| 0.70 | −2.48608 | −1.2185 | −0.48 | 0.63 |
| 0.80 | −2.99633 | −2.0274 | −1.26 | −0.28 |
| 0.90 | −3.49990 | −3.0662 | −2.48 | −1.38 |

that the distributions of $L_1(1)$ and $L_1(k)$ are both not skewed. This is the main reason why the normal approximation is not appropriate when $k$ is large with moderate values of $n$.

With respect to computational time, the finite Markov chain imbedding technique for obtaining the exact distribution is very efficient. As can be seen easily from (3.2), it involves only the construction of the transition probability matrices and their multiplication. For all the examples considered, the CPU time is usually less than a fraction of a second using the S-Plus software package on a SUN Ultra-Sparc Unix workstation.

## Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (Fu and Lou). The authors wish to thank the editor and referees for their helpful comments and suggestions.

## References

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences, *Nucleic Acids Research*, **27**, 573–580.

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

Fu, J. C. (1986). Reliability of consecutive-k-out-of-n F system with (k-1)-step Markov dependence, *IEEE Transactions on Reliability*, **R 35**, 602–606.

Fu, J. C. (1996). Distribution theory of runs and patterns associated with a sequence of multi-state trials, *Statist. Sinica*, **6**, 957–974.

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

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes, *Cell*, **72**, 971–983.

Mcleish, D. L. (1974). Dependent central limit theorem and invariance principles, *Ann. Probab.*, **2**, 620–628.

Nagaev, S. V. (1957). Some limit theorems for stationary Markov chains, *Theory Probab. Appl.*, **2**, 378–406.

Nagaev, S. V. (1961). More exact statements of limit theorems for homogeneous Markov chains, *Theory Probab. Appl.*, **6**, 62–81.

Petrov, V. V. (1975). *Sums of Independent Random Variables*, Springer, Berlin.