Statistica Sinica 3(1993), 313-320

ON NUMBER OF OCCURRENCES OF SUCCESS RUNS OF SPECIFIED LENGTH IN A TWO-STATE MARKOV CHAIN

Katuomi Hirano and Sigeo Aki

The Institute of Statistical Mathematics and Osaka University

Abstract: Let $X_1, X_2, \ldots, X_n$ be a time-homogeneous $\{0,1\}$-valued Markov chain. The probability distribution of number of runs of "1" of length at least k in the sequence $X_1, X_2, \ldots, X_n$ is studied. The probability generating function and some characteristics of the distribution are given in a simple form. Another distribution of number of runs of "1" of length k in the sequence by a different way of counting is also investigated.

Key words and phrases: Probability generating function, discrete distributions, Markov chain, binomial distribution of order k, sequence matching.

1. Introduction

Let n and k be fixed positive integers such that $n \geq k$. In the usual Bernoulli trials, the distributions of the number of success runs of length k were studied by many authors. Feller (1968, Chapter XIII) defined a way of counting the number of runs exactly of length k as counting the number from scratch every time a run occurs. For example, the sequence $SSS|SFSSS|SSS|F$ contains 3 success runs of length 3. By adopting this definition, the distribution of the number of success runs of length k until the nth trial could be studied as an application of renewal theory. Exact and asymptotic properties of the distribution have been derived even in the case of dependent trials (Feller (1968), Rajarshi (1974), Aki (1985), Hirano (1986), Philippou and Makri (1986), Aki (1992), and Aki and Hirano (1993)). A problem on the reliability of a system which is called "consecutive-k-out-of-$n\colon F$ system" is closely related to this distribution (cf. Aki (1985) and Hirano (1986)).

We note, however, that there can be different ways of counting the number of success runs of length k. In the classical literature a "success run of length k" meant an uninterrupted sequence of either exactly k, or of at least k, successes (cf. Feller (1968)). Though Feller's way of counting is suitable for deriving theoretical results, it depends on the statistical problem which way of counting should be adopted. It seems that the distribution of the number of success runs of length k or more until the nth trial is not yet known. This problem is important in molecular biology because the distribution gives the probability that in matching two sequences of DNA one observes a long region where both sequences agree. In Goldstein (1990), a Poisson approximation of it was proposed. The number of success runs of length 3 or more in the sequence $SSSS|FSSSSSS|F$ is 2. In Section 2, we obtain the exact distribution of the number of success runs of length k or more until the nth trial.

Recently, Ling (1988) obtained a recurrence relation for the probability generating function (p.g.f.) of a new distribution of the number of success runs of length k until the nth trial by another way of counting (see Section 3). The p.g.f. of the distribution was given explicitly by Hirano, Aki, Kashiwagi and Kuboki (1991). By Ling's way of counting, the above sequence $SSSSFSSSSSSF$ contains 6 success runs of length 3. In Section 3, we consider the corresponding distribution based on dependent trials.

Throughout the paper, we study the distributions of numbers of success runs until the nth trial of the following time-homogeneous two-state Markov chain. Let $X_0, X_1, X_2, \dots$ be a time-homogeneous $\{0,1\}$-valued Markov chain defined by $P(X_0 = 0) = p_0$ $(0 < p_0 < 1)$, $P(X_0 = 1) = p_1 = 1 - p_0$,

$$\begin{aligned} &P(X_{i+1} = 0|X_i = 0) = p_{00}, \quad P(X_{i+1} = 1|X_i = 0) = p_{01}, \\ &P(X_{i+1} = 0|X_i = 1) = p_{10}, \quad P(X_{i+1} = 1|X_i = 1) = p_{11}, \quad \text{for } i = 0, 1, 2, \dots \end{aligned}$$

where $p_{00}+p_{01} = 1$ and $p_{10}+p_{11} = 1$. This contains the strongly stationary model as a special case. Indeed, if $p_0 = p_{10}/(p_{01} + p_{10})$ and $p_1 = p_{01}/(p_{01} + p_{10})$, then the Markov chain is strongly stationary (e.g., see Edwards (1960)). Of course, the Markov chain contains the usual independent trials as a special case. The distribution of the number of success runs until nth trial based on this Markov chain by the usual (Feller's) way of counting has been studied by Aki and Hirano (1993).

In this paper we deal with two distributions. One is the distribution of the number of success runs of length k or more until the nth trial. The other is the distribution of the number of success runs of length k until the nth trial by Ling's overlapping way of counting.

2. Number of Success Runs of Length k or More

Let X be the number of runs of "1" of length k or more in the sequence $X_1, X_2, \dots, X_n$. First we give the probability function heuristically. The proof is straightforward. Here $[a]$ denotes the largest integer not exceeding a.

Theorem 1. *For* $x = 0, 1, 2, \ldots, [(n+1)/(k+1)]$,

$$\begin{aligned} P(X = x) &= p_0 p_k(x; n) + p_1 \sum_{i=0}^{k-1} p_{11}^i p_{10} p_k(x; n-1-i) \\ &\quad + p_1 \sum_{i=k}^{n-1} p_{11}^i p_{10} p_k(x-1; n-1-i) + p_1 p_{11}^n \delta_{1x} \end{aligned} \tag{1}$$

where

$$\begin{aligned} &p_k(x; n) = P(X = x | X_0 = 0) \\ &= \sum_{m=0}^{k-1} \sum_{\substack{x_1 + 2x_2 + \cdots + nx_n = n - m \\ x_{k+1} + \cdots + x_n = x}} \binom{x_1 + \cdots + x_n}{x_1, \ldots, x_n} \\ &\quad \times (p_{00})^{x_1} (p_{01}p_{10})^{x_2} \cdots (p_{01}p_{11}^{n-2}p_{10})^{x_n} (p_{01}p_{11}^{m-1}) \\ &\quad + \sum_{m=k}^{n} \sum_{\substack{x_1 + 2x_2 + \cdots + nx_n = n - m \\ x_{k+1} + \cdots + x_n = x - 1}} \binom{x_1 + \cdots + x_n}{x_1, \ldots, x_n} \\ &\quad \times (p_{00})^{x_1} (p_{01}p_{10})^{x_2} \cdots (p_{01}p_{11}^{n-2}p_{10})^{x_n} (p_{01}p_{11}^{m-1}), \end{aligned} \tag{2}$$

and δ_{1x} *is Kronecker's delta.*

Though Theorem 1 gives the probability function of the distribution of X, the formula is not necessarily convenient for computation or for deriving characteristics of the distribution analytically. So, we give a recurrence relation of the conditional probabilities of X.

Let $B_k^i(n, x)$ for $i = 0, 1$ be the conditional probability of x runs of "1" of length k or more in the sequence $X_1, \ldots, X_n$ given that $X_0 = i$. Conventionally we define $B_k^i(0, x) = 0$ for $x > 0$. Considering where the first 0 occurs, we have

Proposition 1. *The above conditional probabilities satisfy the following recurrence relation with* $\alpha = 0$ *and* $\beta = 0$ *:*

$$\begin{cases} B_k^0(n, 0) = 1 \quad \text{if } \ 0 \le n < k \\ B_k^1(n, 0) = 1 \quad \text{if } \ 0 \le n < k \\ B_k^0(n, 0) = p_{00} B_k^0(n-1, 0) + \sum_{m=0}^{k-2} p_{01} p_{11}^m p_{10} B_k^0(n-m-2, 0) \quad \text{if } \ n \ge k \\ B_k^1(n, 0) = p_{10} B_k^0(n-1, 0) + \sum_{m=0}^{k-2} p_{11}^{m+1} p_{10} B_k^0(n-m-2, 0) \quad \text{if } \ n \ge k \\ B_k^0(n, x) = p_{00} B_k^0(n-1, x) + \sum_{m=0}^{k-2} p_{01} p_{11}^m p_{10} B_k^0(n-m-2, x) \\ \qquad + \sum_{m=k-1}^{n-2} p_{01} p_{11}^m p_{10} B_k^0(n-m-2, x-1-\alpha) \\ \qquad + p_{01} p_{11}^{n-1} B_k^1(0, x-1-\beta) \quad \text{if } n \ge k \text{ and } x = 1, 2, \ldots, [\frac{n+1}{k+1}] \\ B_k^1(n, x) = p_{10} B_k^0(n-1, x) + \sum_{m=0}^{k-2} p_{11}^{m+1} p_{10} B_k^0(n-m-2, x) \\ \qquad + \sum_{m=k-1}^{n-2} p_{11}^{m+1} p_{10} B_k^0(n-m-2, x-1-\alpha) \\ \qquad + p_{11}^n B_k^1(0, x-1-\beta) \quad \text{if } n \ge k \text{ and } x = 1, 2, \ldots, [\frac{n+1}{k+1}]. \end{cases} \tag{3}$$

The result is easily checked.

Next we give the conditional probability generating functions (p.g.f.). Set

$$\psi_n(t) = \sum_{x=0}^{[\frac{n+1}{k+1}]} B_k^0(n,x)t^x \quad \text{and} \quad \xi_n(t) = \sum_{x=0}^{[\frac{n+1}{k+1}]} B_k^1(n,x)t^x.$$

Then, (3) implies

Proposition 2. *The conditional* p.g.f.'s $\psi_n(t)$ *and* $\xi_n(t)$ *satisfy the recurrence relation with* $\alpha = 0$ *and* $\beta = 0$:

$$\begin{cases} \psi_n(t) = 1 \quad \textit{if} \;\; 0 \le n < k, \\ \xi_n(t) = 1 \quad \textit{if} \;\; 0 \le n < k, \\ \psi_n(t) = 1 - p_{01}p_{11}^{k-1} + p_{01}p_{11}^{k-1}t \quad \textit{if} \;\; n = k, \\ \xi_n(t) = 1 - p_{11}^k + p_{11}^k t \quad \textit{if} \;\; n = k, \\ \psi_n(t) = p_{00}\psi_{n-1}(t) + \sum_{m=0}^{k-2} p_{01}p_{11}^m p_{10}\psi_{n-m-2}(t) \\ \qquad + \sum_{m=k-1}^{n-2} p_{01}p_{11}^m p_{10}t^{1+\alpha}\psi_{n-m-2}(t) + p_{01}p_{11}^{n-1}t^{1+\beta} \quad \textit{if} \;\; n > k, \\ \xi_n(t) = p_{10}\psi_{n-1}(t) + \sum_{m=0}^{k-2} p_{11}^{m+1}p_{10}\psi_{n-m-2}(t) \\ \qquad + \sum_{m=k-1}^{n-2} p_{11}^{m+1}p_{10}t^{1+\alpha}\psi_{n-m-2}(t) + p_{11}^n t^{1+\beta} \quad \textit{if} \;\; n > k. \end{cases} \tag{4}$$

By using (4), we can show an explicit form of conditional p.g.f. $\psi_n(t)$. Define

$$\Psi(z)(= \Psi(t,z)) \equiv \sum_{n=0}^{\infty} \psi_n(t)z^n \quad \text{and} \quad \Xi(z)(= \Xi(t,z)) \equiv \sum_{n=0}^{\infty} \xi_n(t)z^n.$$

Then, the following theorem is useful.

Theorem 2. *The generating functions of the conditional* p.g.f's $\Psi(z)$ *and* $\Xi(z)$ *can be written as*

$$\Psi(z) = \frac{1 + az + (t-1)p_{01}p_{11}^{k-1}z^k}{1 - (1-a)z - az^2 - (t-1)p_{01}p_{11}^{k-1}p_{10}z^{k+1}},$$

and

$$\Xi(z) = \frac{p_{10}z + p_{10}p_{11}^k z^{k+1}(t-1)}{1 - p_{11}z}\Psi(z) + \frac{1 - (1-t)p_{11}^k z^k}{1 - p_{11}z}$$

where $a = 1 - p_{00} - p_{11}$.

Proof. From (4) we have $\Psi(z) = B(t,z)/A(t,z)$ where

$$A(t,z) = 1 - p_{00}z - p_{01}p_{10}\sum_{m=0}^{k-2} p_{11}^m z^{m+2} - p_{01}p_{11}^{k-1}p_{10}tz^{k+1}\sum_{l=0}^{\infty} p_{11}^l z^l$$

and

$$\begin{aligned} B(t,z) &= \sum_{n=0}^{k-1} z^n + (1-p_{01}p_{11}^{k-1})z^k + p_{01}p_{11}^{k-1}z^k t - p_{00}z\sum_{j=0}^{k-1} z^j \\ &\quad -p_{01}p_{10}\sum_{m=0}^{k-2} p_{11}^m z^{m+2}\sum_{j=0}^{k-m-2} z^j + p_{01}tz\sum_{n=k+1}^{\infty} p_{11}^{n-1}z^{n-1}. \end{aligned}$$

After some algebla, we have the desired result.

By using Theorem 2, we can immediately show an explicit form of $\psi_n(t)$.

Corollary. *The* p.g.f. *of the conditional distribution* $\psi_n(t)$ *is represented as*

$$\begin{aligned} \psi_n(t) &= \sum_{r=0}^{[\frac{n}{k}]}\sum_{s=0}^{n-rk}(t-1)^r(p_{01}p_{11}^{k-1}p_{10})^r a^s \\ &\quad \times\Bigg(\frac{(n-rk-s)!}{r!s!(n-r(k+1)-2s)!}(1-a)^{n-r(k+1)-2s} \\ &\quad +\frac{(n-rk-s-1)!}{r!s!(n-r(k+1)-2s-1)!}a(1-a)^{n-r(k+1)-2s-1} \\ &\quad +\frac{(n-(r+1)k-s)!}{r!s!(n-(r+1)k-2s-r)!}(t-1)p_{01}p_{11}^{k-1}(1-a)^{n-(r+1)k-2s-r}\Bigg). \end{aligned}$$

Proof. Note that $\psi_n(t)$ is the coefficient of z^n in the first equation of Theorem 2. Thus we have the result.

Remark. If $X_1, X_2, \ldots, X_n$ are independent trials, the p.g.f. of the distribution can be written in a simple form. By setting $a = 0$, $p_{01} = p_{11} = p$ and $p_{10} = p_{00} = q$, we have

$$\psi_n(t) = \sum_{r=0}^{[n/k]}(t-1)^r p^{kr} q^r\left\{\binom{n-rk}{r} + \binom{n-k(r+1)}{r}(t-1)p^k\right\}. \quad (5)$$

Goldstein (1990) discussed a Poisson approximation to the distribution. We can derive some characteristics of the distribution. Let X be a random variable with this distribution. Then $P(X = x)$ and a recurrence relation of it are given by (2) and (3) with $p_{01} = p_{11} = p$ and $p_{10} = p_{00} = q$, respectively. From (5) the mean and the variance are also easily given by

$$E(X) = p^k\{1+(n-k)q\}, \quad n \geq k,$$

$$\mathrm{Var}(X) = \begin{cases} p^k + qp^k(n-k) - p^{2k}\{1+(n-k)q\}^2 & \text{if } \left[\dfrac{n}{k}\right] = 1 \\ (n-2k)(n-2k-1)q^2p^{2k} + 2(n-2k)qp^{2k} & \\ +(n-k)qp^k + p^k - p^{2k}\{1+(n-k)q\}^2 & \text{if } \left[\dfrac{n}{k}\right] \geq 2, \end{cases}$$

respectively, the first of which was also derived by Goldstein (1990). Further, (5) is useful for calculation of $P(X = x)$.

3. Number of Success Runs of Length k by Ling's Way of Counting

Ling (1988,1989) derived distributions related to number of success runs in independent trials in a different way of counting. Fix positive integers n and k. For $i = 1, 2, \ldots, n-k+1$, we define $Y_i = \prod_{j=i}^{i+k-1} X_j$. Let $M_n^{(k)}$ be the sum of Y_i's i.e. $M_n^{(k)} = Y_1 + Y_2 + \cdots + Y_{n-k+1}$. Then $M_n^{(k)}$ means the number of runs of "1" of length k in $X_1, X_2, \ldots, X_n$ by Ling's way of counting. If $X_1, X_2, \ldots, X_n$ are independent trials, then the corresponding distribution of $M_n^{(k)}$ is called the type II binomial distribution of order k (cf. Ling (1988) and Hirano, Aki, Kashiwagi and Kuboki (1991)). First we give the distribution of $M_n^{(k)}$.

Set

$$LB_k^0(n,x) = P(M_n^{(k)} = x|X_0 = 0) \quad \text{and} \quad LB_k^1(n,x) = P(M_n^{(k)} = x|X_0 = 1).$$

Then, we obtain the recurrence relation (3) of LB_k^0 and LB_k^1 with $\alpha = m-k+1$ and $\beta = n-k$, by replacing B_k^0 and B_k^1 by LB_k^0 and LB_k^1, respectively. Note that the support of the distribution is $\{0, 1, \ldots, n-k+1\}$; and $LB_k^0(n,x) = LB_k^1(n,x) = 0$ if $0 \leq n < k$ and $x \neq 0$.

Define

$$\phi_0(n,t) = \begin{cases} 1 & \text{if } 0 \leq n < k \\ \displaystyle\sum_{x=0}^{n-k+1} LB_k^0(n,x)t^x & \text{if } n \geq k \end{cases}$$

and

$$\phi_1(n,t) = \begin{cases} 1 & \text{if } 0 \leq n < k \\ \displaystyle\sum_{x=0}^{n-k+1} LB_k^1(n,x)t^x & \text{if } n \geq k. \end{cases}$$

Then from the recurrence relation of LB_k^0 and LB_k^1, we have the recurrence relation (4) of ϕ_0 and ϕ_1 with $\alpha = m-k+1$ and $\beta = n-k$, by replacing ψ_n and ξ_n by ϕ_0 and ϕ_1, respectively.

Set

$$\Phi_0(z) = \sum_{n=0}^{\infty} \phi_0(n,t)z^n \quad \text{and} \quad \Phi_1(z) = \sum_{n=0}^{\infty} \phi_1(n,t)z^n.$$

Similarly as in Section 2, we can also get explicit representations of the generating functions of the conditional p.g.f.'s $\Phi_0(z)$ and $\Phi_1(z)$.

Theorem 3. *$\Phi_0(z)$ and $\Phi_1(z)$ can be written as*

$$\Phi_0(z) = \frac{1 + z(p_{01} - p_{11}t) + (1-t)p_{01}\sum_{j=1}^{k-2} p_{11}^j z^{j+1}}{1 - (p_{00} + p_{11}t)z - (p_{01}p_{10} - p_{00}p_{11}t)z^2 - \sum_{m=1}^{k-2} p_{01}p_{11}^m p_{10}(1-t)z^{m+2}}$$

and

$$\Phi_1(z) = \frac{1}{p_{01}}\Big[\{p_{11} + (p_{01} - p_{11})z\}\Phi_0(z) + (p_{01} - p_{11})\Big].$$

Proof. Multiplying both sides of the recurrence relation of ϕ_0 and ϕ_1 by z^n and summing over $n \geq k+1$, we have $\Phi_0(z)$ and $\Phi_1(z)$, respectively.

Acknowledgements

This research was partially supported by the ISM Cooperative Research Program (91-ISM-CRP-5) of the Institute of Statistical Mathematics. The authors are grateful to the referees for their helpful comments on an earlier version of the paper.

References

Aki, S. (1985). Discrete distributions of order k on a binary sequence. *Ann. Inst. Statist. Math.* **37**, A, 205–224.

Aki, S. (1992). Waiting time problems for a sequence of discrete random variables. *Ann. Inst. Statist. Math.* **44**, 363–378.

Aki, S. and Hirano, K. (1993). Discrete distributions related to succession events in a two-state Markov chain. To appear in *Statistical Sciences and Data Analysis; Proceedings of the 3rd Pacific Area Statistical Conference* (Edited by K. Matusita et al.), VSP International Science Publishers, Zeist.

Edwards, A. W. F. (1960). The meaning of binomial distribution. *Nature* **186**, 1074.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edition. John Wiley, New York.

Goldstein, L. (1990). Poisson approximation and DNA sequence matching. *Comm. Statist. Theory Methords* **19**, 4167–4179.

Hirano, K. (1986). Some properties of the distributions of order k. *Fibonacci Numbers and Their Applications* (Edited by A. N. Philippou, G. E. Bergum and A. F. Horadam), 43–53, Reidel, Dordrecht.

Hirano, K., Aki, S., Kashiwagi, N. and Kuboki, H. (1991). On Ling's binomial and negative binomial distributions of order k. *Statist. Probab. Lett.* **11**, 503–509.

Ling, K. D. (1988). On binomial distributions of order k. *Statist. Probab. Lett.* **6**, 247–250.

Ling, K. D. (1989). A new class of negative binomial distributions of order k. *Statist. Probab. Lett.* **7**, 371–376.

Philippou, A. N. and Makri, F. S. (1986). Successes, runs and longest runs. *Statist. Probab. Lett.* **4**, 101–105.

Rajarshi, M. B. (1974). Success runs in a two-state Markov chain. *J. Appl. Probab.* **11**, 190–192.

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan.
Department of Mathematical Science, Faculty of Engineering Science, Osaka University, 1-1 Machikaneyama, Toyonaka, Osaka 560, Japan.

(Received March 1992; accepted December 1992)