JIRSS (2005)
Vol. 4, No. 1, pp 51-56

A Further Note on Runs in Independent Sequences

R. T. Smythe

Department of Statistics, Oregon State University, 44 Kidder Hall, Corvallis, OR 97331.

Abstract. Given a sequence of letters generated independently from a finite alphabet, we consider the case when more than one, but not all, letters are generated with the highest probability. The length of the longest run of any of these letters is shown to be one greater than the length of the longest run in a particular state of an associated Markov chain. Using results of Foulser and Karlin (1987), a conjecture of a previous paper (Smythe, 2003) concerning the expectation of this length is verified.

1 Introduction

Let $\mathcal{A}$ be an alphabet of k letters, $a_1, \ldots, a_k$, and assume that we have an independent sequence of n letters chosen from $\mathcal{A}$ with probabilities $p_1, \ldots, p_k$ respectively. In the case when all p_i are equal, we presented in this journal (Smythe (2003)) some results on the longest run of *any* letter in $\mathcal{A}$. An exact relation, valid for all n, was given between the longest run of any letter and the longest run of a specific letter. An extension was given to the case where the sequence is generated by

Received: August 2004, Accepted: November 2004
Keywords and phrases: Finite alphabet, independent sequences, longest runs.

a Markov chain. When the letters have different probabilities, it was shown that the length of the longest run of any letter is asymptotically the same as the length of the longest run of the letter with highest probability, when this letter is unique.

Left unresolved in that paper was the case when the alphabet has size greater than two, and exactly r letters $a_1, \ldots, a_r$, $2 \leq r < k$, are taken with the highest probability among the p_i. A conjecture was made, based on a heuristic argument, relating the asymptotic expectation of the length of longest run of any letter among $a_1, \ldots, a_r$, to the longest expected run of a_1. The purpose of this note is first to present an exact relation between the longest run of any of $a_1, \ldots, a_r$ and the longest run in state 1 of a simple Markov chain with states {0,1,2}. Then, using a previously unnoticed (by the author) paper of Foulser and Karlin (1987), we are able to verify the conjecture in equation (1) below.

In a slight departure from the notation of the previous paper, let

$$R_{i,n} := \text{ length of the longest run of a given letter } a_i,$$

$$L_n := \text{ length of the longest run of any letter,}$$

$$L_n^* := \text{ length of the longest run of any letter among } a_1, \ldots, a_r.$$

The conjecture was made in Smythe (2003) that

$$E(L_n) = E(R_{1,n}) + log_{1/p_1}(r) + o(1). \qquad (1)$$

As shown in that paper, the longest run of $a_1, \ldots, a_r$ will dominate the longest run of any other letters in this case, so the conjecture in (1) is equivalent to

$$E(L_n^*) = E(R_{1,n}) + log_{1/p_1}(r) + o(1). \qquad (1')$$

In Section 2 we give an exact result relating $E(L_n^*)$ to the expected length of the longest run in a state of a related Markov chain. Section 3 then applies the results of Foulser and Karlin (1987) to this Markov chain and gives a verification of $(1')$.

2 The Exact Result

Assume for simplicity that there are just two letters, a_1, a_2, taken with the highest probability p. (The extension to the general case

is straightforward.) Define a mapping T from sequences $\{s_i\}$ with letters in $\mathcal{A}$ into {0,1,2} as follows:

$$T(s_1) = 0,$$
$$T(s_i) = 0 \; if \; s_i = a_1 \; or \; a_2, \; s_i \neq s_{i-1},$$
$$T(s_i) = 1 \; if \; s_i = s_{i-1} = a_1 \; or \; a_2,$$
$$T(s_i) = 2 \;\; otherwise.$$

It is straightforward to verify that the resulting sequence $\{T(s_i) : i \geq 2\}$ is a Markov chain on {0,1,2} with initial distribution $\pi(0) = 2p(1-p), \; \pi(1) = 2p^2, \; \pi(2) = 1 - 2p$. The transition matrix of this Markov chain is

$$\begin{pmatrix} p & p & 1-2p \\ p & p & 1-2p \\ 2p & 0 & 1-2p \end{pmatrix}$$

and the stationary distribution is given by the initial measure.

Let L_n^* denote the longest run of a_1 or a_2 in the initial sequence. If R_{n-1}^* denotes the longest run of 1's in the derived Markov chain, then clearly

$$L_n^*(\{s_i\}) = 1 + R_{n-1}^*(T(\{s_i\})). \qquad (2)$$

Both sides of equation (2) are random variables defined on the probability space of sequences $\{s_1, \ldots, s_n\}$ with the product measure, so the expected value of the longest run of a_1 or a_2 is one greater than the expected value of the longest run of 1's in the Markov chain.

3 Relation to asymptotic results

Foulser and Karlin (1987) prove a general result for semi-Markov chains that can be used to derive the asymptotic distribution of R_{n-1}^* and of $R_{1,n}$ in our problem. The limiting distribution of L_n^* will of course follow from that of R_{n-1}^*, permitting verification of the conjecture of (1′). (In fact, the result of Foulser and Karlin is sufficiently general to permit calculation of the limiting distribution of L_n^* directly, providing a second route to establishing (1′).)

The aymptotic distributions of $R_{1,n}$ and R_{n-1}^* are not difficult to compute from Section 4 of Foulser and Karlin. Our case is simple enough that their Corollary 2 includes the result we need, but some preparations are necessary to present this result.

Let the Markov chain $X_1, X_2, \ldots$ on the states $1, 2, \ldots K$ have transition probability matrix P. Let $\Sigma = \{(\sigma_1, s_1), \ldots, (\sigma_j, s_j)\}$ be the set of transitions regarded as successes, where σ_i and s_i denote states of the chain. (In our Markov chain of Section 2, a successful transition would be from state 1 to itself.) All other transitions are considered failure transitions. Let $\phi_{ij} = 1$ iff (i, j) is a success transition in Σ, and $\phi_{ij} = 0$ otherwise, and let $\mathbf{\Phi}(\mathbf{\Sigma})$ denote the matrix with entries $\phi_{ij}, 1 \leq i, j \leq K$.

Denote by the matrix $\mathbf{S}$ the success transition probabilities of the grouped transition runs, and by $\mathbf{R}$ the remaining failure transitions, so that

$$\mathbf{S} = \mathbf{\Phi}(\mathbf{\Sigma}) \circ \mathbf{P}, \quad \mathbf{R} = \mathbf{P} - \mathbf{S},$$

where the matrix composition $\mathbf{A} \circ \mathbf{B}$ is the Schur matrix product $\parallel a_{ij} b_{ij} \parallel$. The matrix $\mathbf{S}$ is substochastic, hence has principal eigenvalue satisfying $0 \leq \lambda \leq 1$. In our application (and generally in nontrivial cases) we have $\lambda < 1$.

The success transitions of the Markov chain induce a semi-Markov process on the state space, with transition probability matrix $\mathbf{Q}$, where q_{ij} gives the probability of a success run starting at state i and making an eventual failure transition to j; thus

$$\mathbf{Q} = (\mathbf{I} - \mathbf{S})^{-\mathbf{1}} \mathbf{R}.$$

Let $\mathbf{w}$ be the stationary distribution for $\mathbf{Q}$. Let the distribution tail $1 - F_{ij}(m)$ be defined as the probability that a success transition run exceeds length m, conditioned on starting in state i and ending in state j, and let $a_{ij} = \sum_0^\infty (1 - F_{ij}(m))$ (assumed to be finite). The matrix $\mathbf{F}$ then gives the sojourn duration distribution for the induced semi-Markov process, and we have

$$\mathbf{Q} \circ (\mathbf{E} - \mathbf{F}(m)) = \mathbf{S^m Q},$$

where $\mathbf{E}$ is the matrix of all 1's. The matrix $\mathbf{A} = \parallel a_{ij} \parallel$ thus satisfies

$$\mathbf{Q} \circ \mathbf{A} = \sum_{\mathbf{m} \geq \mathbf{0}} \mathbf{S^m Q} = (\mathbf{I} - \mathbf{S})^{-\mathbf{1}} \mathbf{Q}.$$

The growth rate of the largest relevant tails depends on the eigenvalue λ of the matrix $\mathbf{S}$. The normalizing function for our case is λ^{-m}, the reciprocal of the decay of $\mathbf{S^m}$. The matrix of normalized tail distributions then satisfies

$$\mathbf{Q} \circ \lambda^{-\mathbf{m}} (\mathbf{E} - \mathbf{F})(m) == \lambda^{-m} \mathbf{S^m Q}.$$

The key to the limit distribution is the normalized tail quantity

$$\rho(m) = \frac{\langle \mathbf{w}, \mathbf{Q} \circ \lambda^{-\mathbf{m}}(\mathbf{E}-\mathbf{F})(m)\mathbf{e}\rangle}{\langle \mathbf{w}, \mathbf{Q} \circ \mathbf{A}\mathbf{e}\rangle} = \frac{\langle \mathbf{w}, \lambda^{-\mathbf{m}}\mathbf{S}^{\mathbf{m}}\mathbf{e}\rangle}{\langle \mathbf{w}, (\mathbf{I}-\mathbf{S})^{-\mathbf{1}}\mathbf{e}\rangle},$$

where $\mathbf{e}$ is the vector $(1, 1, \ldots, 1)$. If $\mathbf{S}$ is aperiodic and irreducible, as in our case,

$$lim_{m\to\infty}\lambda^{-m}\mathbf{S}^m \equiv \mathbf{U} \;\; exists, \; and \;\; \mathbf{S}^m = \lambda^m \mathbf{U}(1 + O(\mu^m)),$$

where $|\mu| < 1$. This allows simplification of $\rho(m)$ as $m \to \infty$:

$$lim_{m\to\infty}\rho(m) = \frac{\langle \mathbf{w}, \mathbf{U}\mathbf{e}\rangle}{\langle \mathbf{w}, (\mathbf{I}-\mathbf{S})^{-\mathbf{1}}\mathbf{e}\rangle.}$$

Let $M_\sigma(t)$ denote the maximal sojourn duration (longest run, in our case). Corollary 2 of Foulser and Karlin (1987) for this case then reads as follows:

Corollary 3.2. *Suppose that all relevant distribution functions with largest order tail growth have expponential tail behavior, e.g., $max_{relevant(i,j)}\{1 - F_{ij}(x)\} = \lambda^x$, for some $0 < \lambda < 1$. Then $\hat{\rho} = lim_{x\to\infty}\rho(x)$ exists and the maximal sojourn duration random variable has the limiting distribution*

$$lim_{u\to\infty} Pr\Big\{M_\sigma(u) - \frac{ln(u)}{-ln\lambda} < z\Big\} = exp(-\hat{\rho}\lambda^z).$$

Applying this result to our case, for $R_{1,n}$ we derive $\hat{\rho} = p(1-p)$, whereas for R^*_{n-1} we get $\hat{\rho} = 2p^2(1-p)$. This gives the asymptotic distributions

$$P(R_{1,n} - log_{1/p}(n) < z) \;=\; exp\{-p(1-p)p^z\} + o(1),$$

$$P(R^*_{n-1} - log_{1/p}(n-1) < z) \;=\; exp\{-2p^2(1-p)p^z\} + o(1).$$

These asymptotic distributions are both location-scale shifts of an extreme value distribution (with distribution function $F(x) = e^{e^{-x}}$). The asymptotic means are then given by

$$E(R_{1,n}) = log_{1/p}(n) + \frac{\gamma + ln(p(1-p))}{ln(1/p)} + o(1),$$

$$E(R^*_{n-1}) = log_{1/p}(n-1) + \frac{\gamma + ln(2p^2(1-p))}{ln(1/p)} + o(1),$$

where γ is Euler's constant (cf. David (1981), p. 260). Then

$$E(L_n^*) = E(R_{n-1}^* + 1) = E(R_{1,n}) + log_{1/p}(2) + o(1),$$

as asserted by $(1')$ in the case r = 2.

In the case when $r > 2$ letters are taken with the same maximal probability p, we can use the same 3-state derived Markov chain, now with a transition matix of

$$\begin{pmatrix} (r-1)p & p & 1-rp \\ (r-1)p & p & 1-rp \\ rp & 0 & 1-rp \end{pmatrix}.$$

In this case $\rho(x) = rp^2(1-p)$, and computing the asymptotic distribution of the longest run of 1's in this Markov chain gives the general form of $(1')$.

References

David, H. A. (1982), Order Statistics. 2nd ed., New York: Wiley.

Foulser, D. E. and Karlin, S. (1987), Maximal success durations for a semimarkov process. Stochastic Processes and their Applications, **24**, 203-224.

Smythe, R. T. (2003), On runs in independent sequences. J. Iranian Statist. Soc., **2**, 43-52.