JIRSS (2003)
Vol. 2, No. 1, pp 43 - 52

On Runs in Independent Sequences

R. T. Smythe

Department of Statistics, Oregon State University, 44 Kidder Hall, Corvallis, OR 97331. (smythe@stat.orst.edu)

Abstract. Given an i.i.d. sequence of n letters from a finite alphabet, we consider the length of the longest run of *any* letter. In the equiprobable case, results for this run turn out to be closely related to the well-known results for the longest run of a given letter. For coin-tossing, tail probabilities are compared for both kinds of runs via Poisson approximation.

1 Introduction

The extensive literature on the behavior of the longest head run in a sequence of independent coin tosses goes back at least to the classic result of Erdös and Rényi, 1970. Their approach has been generalized and used extensively in sequence matching problems in statistical genetics (see, for example, Chapter 11 of the book of Waterman, 1995).

The problem of the longest run of heads *or* tails, by contrast, seems to have been largely neglected, although it is of some interest at least in the context of DNA sequences (Ewens and Grant, 2001,

Received: January 2003
Key words and phrases: Finite alphabet, i.i.d. sequences, success runs, Poisson approximation

p. 162). The purpose of this note is to examine the behavior of the longest run of any letter in a sequence of letters from a finite alphabet. It turns out that there is a very easy way to make the connection between this problem and that of the longest head run for coin-tossing (and more generally, for equiprobable models). A natural extension of the equiprobable results to sequences generated by a Markov chain is also presented. The final section goes beyond the equiprobable model to treat the more difficult general case for finite alphabets.

2 The Equiprobable Model

Let $\mathcal{A}$ be an alphabet of k letters, $a_1, \ldots, a_k$, and assume we have a sequence of n letters from $\mathcal{A}$, chosen with equal probabilities $1/k$. Let

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

$$L_{k,n} \;:=\; \text{length of the longest run of any letter.}$$

Heuristically, to approximate $L_{k,n}$ one can use the same reasoning used to approximate $R_{k,n}$ (cf. Waterman, 1995, p. 264). A run of a_i of length m has probability k^{-m}, and there are roughly n possible starting positions for this run. Thus

$$E[\text{number of runs of } a_i \text{ of length } m] \;\approx\; nk^{-m},$$

and if the longest run is unique, its length $R_{k,n}$ should satisfy $1 \approx nk^{-R_{k,n}}$, giving $R_{k,n} \approx log_k(n)$. This intuition can be made rigorous (Gordon et al, 1986) to give

$$\frac{E(R_{k,n})}{log_k(n)} \longrightarrow 1 \text{ as } n \to \infty.$$

Analogous heuristic reasoning with the longest run of any letter gives

$$E[\text{number of runs of length } m] \;\approx\; nkk^{-m} \;=\; nk^{-(m-1)},$$

leading to

$$L_{k,n} \;\approx\; log_k(n) \;+\; 1 \;\approx\; R_{k,n} \;+\; 1.$$

With this intuition, the proposition that follows is perhaps not surprising.

Proposition 2.1. $E(L_{k,n}) = E(R_{k,n-1}) + 1.$

Proof. Let T map sequences $\{s_j\}$ from $\mathcal{A}$ into dyadic $(0,1)$ sequences $\{\hat{s}_j\}$ by letting $T(s_1) = 0$ and, for $j \geq 1$,

$$T(s_j) = 1(s_j = s_{j-1}),$$

where $1(.)$ denotes the indicator function.

Lemma 2.1. $\{\hat{s}_j\}_{j\geq 2}$ *is a Bernoulli sequence with* $p = 1/k$. □

Proof of Lemma 1.1 We have

$$P(\hat{s}_j = 1) = \sum_{i=1}^{k} P(\hat{s}_j = 1 | s_{j-1} = a_i)(1/k) = (1/k) \sum_{i=1}^{k} 1/k = 1/k.$$

Now for $i \geq 2$, consider $P(\hat{s}_i = \epsilon_i, \ldots, \hat{s}_{i+l} = \epsilon_{i+l})$ where $\epsilon_j = 0$ *or* 1. This probability equals

$$P(s_i \ \Delta_i \ s_{i-1}, \ \ldots, s_{i+l} \ \Delta_{i+l} \ s_{i+l-1})$$

where Δ_j means $=$ if $\epsilon_j = 1$ and $\neq$ if $\epsilon_j = 0$. It is easily checked that the latter probability equals

$$\prod_{h=i}^{i+l} P(s_h \ \Delta_h \ s_{h-1}) = \Big(\frac{1}{k}\Big)^{j} \Big(\frac{k-1}{k}\Big)^{l-j},$$

where $j = \sum_{h=i}^{i+l} \epsilon_h$ is the number of 1's among $\epsilon_i, \ldots, \epsilon_{i+l}$. But as already noted,

$$P(\hat{s}_h = \epsilon_h) = \begin{cases} \frac{1}{k} & \text{if} \quad \epsilon_h = 1 \\ \frac{k-1}{k} & \text{if} \quad \epsilon_h = 0 \end{cases},$$

so

$$P(\hat{s}_i = \epsilon_i, \ldots, \hat{s}_{i=l} = \epsilon_{i+l}) = \prod_{h=i}^{i+l} P(\hat{s}_h = \epsilon_h),$$

establishing Lemma 2.1. □

Remark: The lemma is not true if the letters a_i occur with different probabilities.

Now let $\hat{R}_{k,n-1}$ denote the longest run of 1's in $\{\hat{s}_j\}$. Note that

$$L_{k,n}(\{s_j\}) \;=\; 1 \;+\hat{R}_{k,n-1}(\{\hat{s}_j\}) \;\text{ and }\; \hat{R}_{k,n-1} \stackrel{D}{=} R_{k,n-1}.$$

The probability of any sequence $\{\hat{s}_j\}$ is $\frac{(k-1)^{N_0}}{k^{n-1}}$, where N_0 is the number of zeros in the sequence, not counting $\hat{s}_1$. The number of sequences $\{s_j\}$ that map into a given $\{\hat{s}_j\}$ under T is just $k(k-1)^{N_0}$, because every time there is a change of letter in $\{s_j\}$, there are $k-1$ possibilities for the next letter; and there are k possible choices of initial letter. Hence, if s denotes a generic sequence $\{s_j\}$ from the alphabet $\mathcal{A}$,

$$E(L_{k,n}) \;=\; \sum_s L_{k,n}(s)k^{-n} \;=\sum_{\hat{s}}[\hat{R}_{k,n-1}(\hat{s}) \;+\; 1]\frac{k(k-1)^{N_0}}{k^n}$$

$$=\; E[\hat{R}_{k,n-1} \;+\; 1] = E(R_{k,n-1}) \;+\; 1.\square$$

The argument given above extends easily to show

Corollary 2.1. $Var(L_{k,n}) = Var(R_{k,n-1})$.

This variance is essentially independent of n (Waterman, p. 277).

Thus the expected length of the longest run of heads or tails in a sequence of (fair) coin tosses grows as $log_2(n)$, the same rate as the expected longest head run. If $\mathcal{A} = \{a, c, g, t\}$, the DNA alphabet, and we assume independence and equal probabilities for the bases, the longest run of any of the four bases grows in length as $log_4(n)$, the same rate as the longest run of a given base (compare the statement on p. 162 of the noteworthy book of Ewens and Grant, 2001, where the authors have apparently misinterpreted a result of Karlin et al., 1983, personal communication).

The result of Proposition 2.1 may seem a bit counterintuitive, despite the heuristic justification we began with. However, taking for illustration the case of k=2, it is still true for large n that $P(L_{2,n} \geq j)$ is nearly twice as large as $P(R_{2,n} \geq j)$ when j becomes moderately large; this is a simple consequence of the near-independence of the events {Head run of length at least j} and {Tail run of length at least j} for large n. Taking $n = 2,000$, where the expected length of the longest head run is about 10, Poisson approximation (cf. Arratia et

al., 1990, for a clear explanation of Poisson approximation in this context) gives, for example,

$$P(R_{2,n} \geq 13) = .11432 \pm 2.000 \text{ x } 10^{-4},$$

$$P(R_{2,n} \geq 14) = .058922 \pm 5.369 \text{ x } 10^{-5}.$$

It is simple to modify the approach of Arratia et al. (1990) to get tail probabilities for $L_{2,n}$ (but note that b_2 in Theorem 1 of Arratia et al. will no longer be zero in this case). We get

$$P(L_{2,n} \geq 13) = .21556 \pm 4.594 \text{ x } 10^{-4},$$

$$P(L_{2,n} \geq 14) = .11432 \pm 1.185 \text{ x } 10^{-4}.$$

If the events {Head run of length at least j} and {Tail run of length at least j} were independent, the calculations for $R_{2,n}$ would give

$$P(L_{2,n} \geq 13) = .21557,$$

$$P(L_{2,n} \geq 14) = .11437,$$

in very good agreement with the previous results.

3 The Markov Chain Case

The results of the preceding section for sequences of independent coin tosses can be generalized to a special case of sequences generated by Markov chains. Assume again an alphabet $\mathcal{A}$ of letters $a_1, \ldots, a_k$, generated by a symmetric Markov chain with transition probabilities c on the diagonal and $(1-c)/(k-1)$ on the off-diagonals. The stationary distribution for such a Markov chain is uniform on $\{a_1, \ldots, a_k\}$ and we assume the chain is started with this initial distribution. Define $\mathcal{T}$ as before, mapping sequences $\{s_j\}$ from $\mathcal{A}$ into dyadic (0,1) sequences $\{\hat{s}_j\}$.

Lemma 3.1. *$\{\hat{s}_j\}$ is a Bernoulli sequence with $p = c$.*

Proof. It is straightforward to show that $\{\hat{s}_j\}$ is a Markov chain, with transition probabilities $1 - c$ in the first column and c in the second. □

Let $L^c_{k,n}$ denote the longest run of any letter in the Markov chain $\{s_j\}$ and $\hat{R}^c_{k,n-1}$ the longest run of 1's in $\{\hat{s}_j\}$. As before, $\hat{R}^c_{k,n-1} \stackrel{D}{=}$

$R^c_{k,n-1}$, the longest head run in a Bernoulli(c) sequence, and from our construction,

$$L^c_{k,n} \ = 1 + \hat{R}^c_{k,n-1}.$$

Proposition 3.1. $E(L^c_{k,n}) \ = E(R^c_{k,n-1}) \ + \ 1.$

Proof. A sequence $\{s_j\}$ with N_0 changes of state has probability

$$\left(\frac{1-c}{k-1}\right)^{N_0} c^{n-1-N_0},$$

and there are $(k-1)^{N_0}$ sequences of this type. The result follows as in Proposition 2.1. □

4 Unequal Probabilities

. Returning now to the independent case, consider a more general model, again with the alphabet $\mathcal{A}$, but where the probabilities for the letters $\{a_i\}_{i=1}^k$ are given by $\{p_i\}_{i=1}^k$. Again let $L_{k,n}$ denote the longest run of any letter. Let $p*$ denote the maximum of $p_1, p_2, \ldots, p_k$ and suppose first that there is only one letter $a*$ appearing with probability $p*$.

For clarity we begin with the case $k = 2$, and suppose that $p > 1/2$. Apply again the heuristic of Section 1. With $q = 1 - p$, we now have

$$E[\text{number of runs of length } m] \ \approx \ n(p^m + q^m) = np^m[1 \ + \ (q/p)^m],$$

which leads to

$$L_{2,n} \ \approx \ log_{1/p}(n) + log_{(1/p)}[1 \ + \ (q/p)^m] \ \approx log_{(1/p)}(n).$$

Similar reasoning in the case of $k > 2$ would give

$$L_{k,n} \ \approx \ log_{(1/p*)}(n).$$

This reasoning is confirmed by the a.s. result for this case; it is straightforward to modify the classical proof for longest head runs (Erdös and Rényi, 1970) to show that

$$P(\lim_{n\to\infty} \frac{L_{k,n}}{log_{1/p*}(n)} \ = \ 1) \ = \ 1. \qquad (1)$$

As for $E(L_{k,n})$, (1) suggests that its limit behavior is the same as that of $E(R^{a*}_{k,n})$, where $R^{a*}_{k,n}$ now denotes the longest run of the letter $a*$ appearing with probability $p*$. Indeed, we expect that in this case as n increases, the longest run of any letter will, with ever higher probability, be the longest run of $a*$. We have

$$E(L_{k,n}) \ = E(L_{k,n}; L_{k,n} = R^{a*}_{k,n}) \ + \ E(L_{k,n}; L_{k,n} \neq R^{a*}_{k,n}) \qquad (2)$$

Using Hölder's inequality on the second term on the right-hand side of (2),

$$E(L_{k,n}; L_{k,n} \neq R^{a*}_{k,n}) \ \leq [E(L_{k,n})^2 P(L_{k,n} \neq R^{a*}_{k,n})]^{1/2}.$$

Because the variance of $L_{k,n}$ does not grow with n, the term $E({L_{k,n}}^2)$ is dominated by $log^2_{1/p*}(n)$. Consider again the case $k = 2$, with $p > 1/2$. The longest head run $R^H_{2,n}$ grows as $log_{1/p}(n)$, and the longest tail run $R^T_{2,n}$ as $log_{1/q}(n)$. Using again Poisson approximation to estimate $P(R^T_{2,n} < t)$ and $P(R^H_{2,n} < t)$ for $log_{1/q}(n) \ < t \ < log_{1/p}(n)$, we have

$$P(R^T_{2,n} \geq t) \approx \ nq^t, \qquad P(R^H_{2,n} < t) \approx \ exp\{-np^t\}.$$

Thus if t is chosen to make

$$log^2_{1/p}(n)[nq^t + exp\{-np^t\}] \ \longrightarrow \ 0 \text{ as } \ n \to \infty,$$

the second term on the right-hand side of (2) converges to zero and

$$E(L_{2,n}) - E(R^H_{2,n}) \ \longrightarrow 0 \ \text{ as } \ n \to \infty.$$

Choosing the closest integer to the value of t satisfying

$$t = log_{1/q}(n) + (2 + \epsilon) log_{1/q} log_{1/p}(n)$$

for $\epsilon > 0$ will accomplish this. A similar approach will work for $k > 2$ as long as there is a unique $a*$ corresponding to the maximal probability $p*$. Thus in the non-equiprobable case with a unique $a*$, the value of $E(L_{k,n})$ is essentially the same, for large n, as that of $E(R^{a*}_{k,n})$.

Table 1 below (based on 5,000 simulation runs) confirms our intuition that as p gets further from .5, the difference between $E(L_{2,n})$ and $E(R^H_{2,n})$ diminishes quickly.

Table 1. Comparison of $E(L_{2,n})$ and $E(R^H_{2,n})$ for $p > .5$

p	$E(L_{2,500})$	$E(R^H_{2,500})$	$E(L_{2,1000})$	$E(R^H_{2,1000})$	$E(L_{2,2000})$	$E(R^H_{2,2000})$
.51	9.2966	8.5228	10.3092	9.5336	11.3852	10.6202
.55	9.7952	9.5572	10.8918	10.6856	11.9500	11.7834
.67	13.5642	13.5636	15.2722	15.2720	17.0014	17.0012

We turn finally to the non-equiprobable case when the letter a* corresponding to the maximal probability $p*$ is not unique. Suppose without loss of generality that $a_1, a_2, \ldots, a_r$ all appear with the maximal probability $p*$, where $2 \leq r \leq k$. Then our heuristic gives

$$L_{k,n} \approx log_{1/p*}(nr) + log_{1/p*}[1 + o(1)]$$

and we might expect

$$E(L_{k,n}) \approx E(R^{a_1}_{k,n}) + log_{(1/p*)}(r). \tag{3}$$

Some suppport for (3) appears to be given in a theorem of Karlin et al. (1985), p. 36. (No proof of the theorem in Karlin et al. is given, and I have not found a published proof elsewhere.) A special case of the much more general theorem of Karlin et al. considers s independently generated sequences of length n, on the same k-letter alphabet, but with possibly different probabilities in the s sequences. Denote by $K_{2,s}$ the longest word present in at least two of the s sequences. Let

$$\lambda := \max_{u,v} \Big(\sum_{i=1}^{k} p_i^{(u)} p_i^{(v)} \Big),$$

where $p_i^{(l)}$ denotes the probabilities for the l^{th} sequence, $l = 1, 2, \ldots, s$ and the maximum is taken over all pairs $1 \leq u < v \leq s$.

Theorem 4.1. *(Karlin et al., 1985). As* $n \to \infty$, $E(K_{2,s})$ *has precise growth order*

$$[log\Big(\binom{s}{2} n^2\Big) + log\lambda(1-\lambda) + 0.577]/(-log\lambda).$$

In our case, we take $s = r+1$. For sequence 1 we take probabilities $\{p_i\} = \{p_i^{(1)}\}$, where $p_1 = p_2 = \ldots = p_r$, and $2 \leq r \leq k$. Sequence j, $2 \leq j \leq r+1$, will consist entirely of the letter a_{j-1}. It is easy to check that in this case $\lambda = p*$, and the longest word common to at least two of the $r+1$ sequences will be the longest run of any one of the letters $a_i, 1 \leq i \leq r$.

At first glance, the theorem appears to give the wrong result, owing to the term $log\left(\binom{r+1}{2}n^2\right)$. However, the theorem refers to the case of "shifts allowed"; i.e., a word of a given length starting in position i in one sequence could be matched by the same word starting at position j in a different sequence. For any pair of sequences, this makes roughly n^2 potential sites for a match, and there are $\binom{r+1}{2}$ pairs of sequences to examine. For our situation, matches can only occur between sequence 1 and one of the other sequences, so there are only r pairs to consider; and "shifts" are irrelevant, as sequences 2 through (r+1) consist of identical letters. Thus the relevant number of comparisons is rn, rather than $\binom{r+1}{2}n^2$.

This would give the growth of $E(L_{k,n})$ in the non-equiprobable case as $[log(n) + log(r)]/log(1/p*)$. For the letters $a_1, a_2, \ldots, a_r$, the rate of growth of $E(R^{a_i}_{k,n})$ is, by results quoted earlier, $log(n)/log(1/p*)$. Thus allowing runs of any letter increases the expected length of a run by an amount $log(r)/log(1/p*)$, where r is the number of letters appearing with the highest probability $p*$, in agreement with the heuristic. The maximum increase of 1 is achieved when $r = k$, i.e., in the equiprobable case, an intuitively reasonable result.

One thousand simulation runs with n = 10,000, using $\mathbf{p} = (.4, .4, .2)$ and $\mathbf{p} = (.3, .3, .3, 1)$, appear to provide support for the result of (3).

Table 2. Comparison of predicted and calculated difference of $E(L_{r,10000})$ and $E(R^{p*}_{r,10000})$

$\{p_i\}$	$p*$	r	$log_{1/p*}(r)$	$E(L_{r,10000}) - E(R^{p*}_{r,10000})$
(.4, .4, .2)	.4	2	.756	.741
(.3, .3, .3, .1)	.3	3	.912	.918

References

Arratia, R., Goldstein, L. and Gordon, L. (1990), Poisson approximation and the Chen-Stein method. Statistical Science, **5**, 403-434.

Ewens, W. J. and Grant, G. R. (2001), Statistical Methods in Bioinformatics. New York: Springer-Verlag.

Erdös, P. and Rényi, A. (1970), On a new law of large numbers. Journal of Anal. Math, **22**, 103-111.

Gordon, L., Schilling, M. and Waterman, M. S. (1986), An extreme value theory for long head runs. Probab. Theory Rel. Fields, **72**, 279-287.

Karlin, S., Ghandour, G., Ost, F., Tavare, S. and Korn, L. (1983), New approaches for computer analysis of nucleic acid sequences. Proc. Natl. Acad. Sci., USA, **80**, 5660-5664.

Karlin, S., Ghandour, G., and Foulser, D. (1985), DNA sequence comparisons of the human, mouse, and rabbit immunoglobulin kappa gene. Mol. Biol. Evol., **2**, 35-52.

Waterman, M. S. (1995), Introduction to Computational Biology. London: Chapman and Hall.