Elsevier Editorial System(tm) for Statistics and Probability Letters
Manuscript Draft

Manuscript Number: STAPRO-D-16-01227R2

Title: On the length of the longest head run

Article Type: Letter

Keywords: Longest head run, extremes in samples of random size.

Corresponding Author: Dr. S Y Novak,

Corresponding Author's Institution:

First Author: S Y Novak

Order of Authors: S Y Novak

Abstract: We evaluate the accuracy of approximation to the distribution of the length of the longest head run in a Markov chain with a discrete state space. An estimate of the accuracy of approximation in terms of the total variation distance is established for the first time.

Detailed Response to Reviewers

Statistics and Probability Letters No: STAPRO-D-16-01227R1

Title: On the length of the longest head run

Response to the reviewer comments

1. I've replaced "length of the longest head run" with "longest run" as suggested by the reviewer.
2. I've added "{\xi_{k}=1} represents failure of the *k*th component" as suggested by the reviewer.
3. The text on page 2 states "N_n(k) of head runs with lengths ≥*k*".
4. Statistics & Probability Letters is among journals cited in the manuscript.

20.06.2017

On the length of the longest head run

S.Y.Novak

MDX University, School of Science & Technology,

The Burroughs, London NW44BT, UK

Submitted: December 24, 2016. Revised: June 20, 2017.

Abstract

We evaluate the accuracy of approximation to the distribution of the length of the longest head run in a Markov chain with a discrete state space. An estimate of the accuracy of approximation in terms of the total variation distance is established for the first time.

Key words: longest head run, extremes in samples of random size.
AMS Subject Classification: 60E15, 60G70.

1 Introduction

Let $\{\xi_i, i \geq 1\}$ be a sequence of random 0's and 1's (i.e., "tails" and "heads"). Then

$$L_n = \max\{k : \xi_{i+1} = ... = \xi_{i+k} = 1 \quad (\exists i \leq n-k)\} \tag{1}$$

is the length of the longest head run (LLHR) among $\xi_1, ..., \xi_n$.

Statistic L_n has applications in biology, reliability theory, finance, and nonparametric statistics (see, e.g., [1, 2, 3, 17]). In particular, the reliability of a consecutive k-out-of-n system with n components can be expressed via $\mathbb{P}(L_n < k)$, where the event $\{\xi_k = 1\}$ represents failure of the kth component: the system fails if and only if k consecutive components fail [1, 4, 6, 13].

The study of the distribution of LLHR has a long history. Apparently, the task was first formulated by de Moivre [10], Problem LXXIV. Renewed interest to the topic arose in connection with the Erdös–Rényi strong law of large numbers [5].

A limit theorem for LLHR in the case of independent Bernoulli $\mathbf{B}(p)$ trials was established by Goncharov [8]. The limiting distribution of LLHR was found in more general situations as well, see [1, 12, 14, 19] and references therein. In particular, a limit theorem for LLHR in a Markov chain with a finite state space $\mathcal{X}$ where hitting a subset of $\mathcal{X}$ is considered a "success" is given in [12]. An estimate of the rate of convergence and asymptotic expansions in the limit theorem for LLHR in a two-state Markov chain have been established in [13]. Concerning LIL-type results, see [16] and references therein.

An exact formula for $\mathbb{P}(L_n < k)$ in terms of combinatorial coefficients in the case of independent Bernoulli trials was found by Uspensky [18]. In the case of a two-state Markov chain Fe et al. [6] present an exact formula for $\mathbb{P}(L_n < k)$ in terms of a specially constructed matrix of transition probabilities, and establish the asymptotics of $\ln \mathbb{P}(L_n < k)$ as $n \to \infty$ if k is fixed (see also Theorem 2 in [13]).

Note that L_n can be represented as a sample maximum in a sample of random size ν_n, where ν_n is a certain renewal process (cf. [13, 14]). References concerning extremes in samples of random size can be found, e.g., in [7, 9, 16].

It is known that the accuracy of approximation to the distribution of LLHR in terms of the uniform distance is $n^{-1}\ln n$ [13]. The result has been generalised to the case of a Markov chain with a finite state space [14] as well as to the case of m-dependent random variables [15]. Asymptotic expansions in the limit theorem for LLHR in a two-state Markov chain [13] confirm that the rate $n^{-1}\ln n$ cannot be improved.

There is a simple relation between LLHR and the number $N_n(k)$ of head runs with lengths $\geq k$:

$$\{L_n < k\} = \{N_n(k) = 0\}.$$

Note that the estimates of the accuracy of approximation to the distribution of $N_n(k)$ have been established in terms of the total variation distance (see [1, 2, 16] and references therein). However, the problem of evaluating the accuracy of approximation to the distribution of LLHR in terms of the total variation distance remained open for a long while.

In this paper we derive an estimate of order $n^{-1}\ln n$ to the total variation distance between $\mathcal{L}(L_n)$ and the approximating distribution.

2 Results

Let $\{X_i, i \geq 1\}$ be a homogeneous Markov chain with a finite state space $\mathcal{X}$ and transition probabilities $\|p_{ij}\|_{i,j\in\mathcal{X}}$. We denote by

$$\bar{\pi} = \|\pi_i\|_{i\in\mathcal{X}}$$

the stationary distribution of the chain.

Given a subset $A \subset \mathcal{X}$, let LLHR be defined by (1), where

$$\xi_i = \mathbb{1}\{X_i \in A\}$$

(hitting A is considered a "success"). We set

$$U \ = \ \|p_{ij}\|_{i,j\in A}\,, \quad \bar{\pi}_A = \|\pi_i\|_{i\in A}\,,$$

and let

$$q(k) \ = \ \bar{\pi}_A U^{k-1}(E-U)\bar{1} \quad (k \geq 1),$$

where $\bar{1}$ is a vector of 1's and E is a unit diagonal matrix.

Let ζ_n, Z_n be random variables (r.v.s) with distribution functions (d.f.s)

$$\mathbb{P}(\zeta_n < k) = (1-q(k))^{n-k}\,,\ \mathbb{P}(Z_n < k) = \exp(-nq(k)) \qquad (k \geq 1).$$

Recall the definition of the total variation distance between the distributions of r.v.s X and Y:

$$d_{TV}(X;Y) \equiv d_{TV}(\mathcal{L}(X);\mathcal{L}(Y)) = \sup_{A\in\mathcal{A}} |\mathbb{P}(X\in A) - \mathbb{P}(Y\in A)|\,,$$

where $\mathcal{A}$ is a Borel σ-field.

The distribution of LLHR L_n can be well approximated by $\mathcal{L}(\zeta_n)$ or $\mathcal{L}(Z_n)$; the accuracy of such approximation in terms of the uniform distance is known to be of order $n^{-1}\ln n$. In Theorem 1 below we show that the result holds in terms of the stronger total variation distance.

Theorem 1 *Assume that*

(P0) there is only one class C of essential states that consists of periodic subclasses $C_1, ..., C_d$;

(P1) $A \cap C_i \neq \emptyset$ $(1 \leq i \leq d)$;

(P2) $0<\lambda<1$, *where* λ *is the largest eigenvalue of matrix* U*;*
(P3) if $i\in C_\ell$ *for some* $\ell\in\{1,...,d\}$*, then*

$$|p_{ij}(m)-d\pi_j|\le u_m\,,\ \ H:=\sum_{m\ge 1}u_m<\infty \tag{2}$$

if $j\in A\cap C_k$ *and* $k-m=\ell\ (\mathrm{mod}\,d)$*; if* $i\notin C_1\cup...\cup C_d$*, then* (2) *holds for all* $j\in A$*;*
(P4) $z_i>0$ *(*$\forall i\in A$*), where* $\bar z=\|z_i\|_{i\in A}$ *is the corresponding to* λ *right eigenvector of matrix* U*.*

Then there exists a positive constant $C=C(\lambda,\bar z,\bar\pi_A)$ *such that*

$$d_{TV}(L_n;Z_n)\le Cn^{-1}\ln n \qquad (n\ge C). \tag{3}$$

The result holds if Z_n *in (3) is replaced with* ζ_n*.*

3 Proofs

Proof of Theorem 1 makes use of Theorem 2 from [14], which is presented below (note that the argument of Theorem 2 in [14] is valid for any fixed $d\in\mathbb{N}$). In the particular case of a stationary Markov chain the result of Theorem 2 is given by Theorem 2.1 in [12].

Theorem 2 *Let* $\{X_i,i\ge 1\}$ *be a homogeneous Markov chain with a discrete state space* $\mathcal{X}$*, transition probabilities* $\|p_{ij}\|_{i,j\in\mathcal{X}}$ *and stationary distribution* $\bar\pi$*. Assume conditions P(0) – P(4). Then there exists a positive constant* $c_\star=c_\star(\lambda,\bar z,\bar\pi_A)$ *such that as* $n>2k\ge c_\star$*,*

$$\begin{aligned}&|\mathbb{P}(L_n<k)-\mathbb{P}(Z_n<k)|\\ &\le c_\star\lambda^k+c_\star k\lambda^k\exp\Big(-nq(k)(1-c_\star k\lambda^k)\Big)\,.\end{aligned} \tag{4}$$

Taking into account the obvious inequality

$$|e^x-e^y|\le|x-y|e^{\max\{x;y\}} \qquad (x,y\in\mathbb{R}), \tag{5}$$

we notice that (4) holds true if Z_n is replaced with ζ_n.

In the case of independent observations inequalities of this kind with explicit constants are presented in [15, 11]. In the case of a two-state Markov chain with

$\alpha := p_{11} \in (0;1)$, $\beta := p_{00} < 1$, a sharp bound of this kind is given in [13], Theorem 2: there exist constants $q \in (0;1)$, $C < \infty$ such that

$$\sup_{k>C} \left| \mathbb{P}(L_n < k) - A(t_0)/t_0^{n+1} \right| \leq C q^n \tag{6}$$

for particular t_0 and function $A(t)$ obeying $|A(t_0) - 1| \leq C_1 \gamma k \alpha^k$, $|t_0 - 1 - \gamma\alpha^k| \leq C_1 k (\gamma\alpha^k)^2$ for some $C_1 < \infty$, where $\gamma = (1-\alpha)(1-\beta)/\alpha(2-\alpha-\beta)$. In the case of independent Bernoulli $\mathbf{B}(\alpha)$ trials (6) holds with $q = \alpha$, $C = (2+\alpha-\alpha^2)/(1-\alpha)(1-\alpha^2)$.

By a well-known property of the total variation distance,

$$2d_{_{TV}}(L_n; Z_n) = \sum_{k\geq 0} |\mathbb{P}(L_n = k) - \mathbb{P}(Z_n = k)|. \tag{7}$$

The idea of the prof is to split the sum in (7) into appropriate fragments and show that the desired estimate holds for each fragment.

Recall that $\bar\pi_A = \|\pi_i\|_{i\in A}$, and set

$$\begin{aligned} c_* &= <\bar\pi_A; \bar z> (1-\lambda)/\lambda z^*, \quad c^* = <\bar\pi_A; \bar z> (1-\lambda)/\lambda z_*, \\ \bar z_* &= \inf\{z_j : j \in A\}, \qquad \bar z^* = \sup\{z_j : j \in A\}. \end{aligned}$$

Note that

$$0 < c_* \leq c^* < \infty.$$

It is easy to see that

$$c_* \lambda^k \leq q(k) \leq c^* \lambda^k \tag{8}$$

(cf. (8) in [14]). Let

$$k(n) = \log n - \log \ln n + \log(c_*/2).$$

Hereinafter $\log$ is to the base $1/\lambda$, symbol c (with or without indexes) denotes positive constants.

Using (4) and (8), we check that

$$\begin{aligned} &\sum_{k\leq k(n)} |\mathbb{P}(L_n = k) - \mathbb{P}(Z_n = k)| \qquad (9) \\ &\leq \mathbb{P}(L_n \leq k(n)) + \mathbb{P}(Z_n \leq k(n)) \leq c_1 n^{-1} \ln n. \end{aligned}$$

It remains to evaluate

$$\sum_{k>k(n)} |\mathbb{P}(L_n{=}k) - \mathbb{P}(Z_n{=}k)|.$$

According to (4) and (8), there exists a positive constant c_2 such that

$$|\mathbb{P}(L_n{=}k) - \mathbb{P}(Z_n{=}k)| \leq c_2\lambda^k + c_2 k\lambda^k e^{-n\lambda^k c_*/2} \tag{10}$$

as $n{>}2k{\geq}c_2$. Evidently,

$$\sum_{k>k(n)} \lambda^k \leq \lambda^{k(n)}/(1{-}\lambda) = 2n^{-1}(\ln n)/(1{-}\lambda)c_*. \tag{11}$$

Thus, it remains to evaluate $\sum_{k>k(n)} k\lambda^k e^{-n\lambda^k c_*/2}$.

Note that function $f(x) = xe^{-x}$ decreases in $[1;\infty)$. Clearly, $n\lambda^k c_*/2 \in [1; \ln n]$ as $k(n){<}k{<}\log(nc_*/2)$. Therefore,

$$\begin{aligned} \sum_{k(n)<k<\log(nc_*/2)} k\lambda^k e^{-n\lambda^k c_*/2} &\leq n^{-1}\log(nc_*/2) \sum_{k(n)<k<\log(nc_*/2)} n\lambda^k e^{-n\lambda^k c_*/2} \\ &\leq n^{-1}\log(nc_*/2) \int_{k(n)}^{\log(nc_*/2)} n\lambda^x e^{-n\lambda^x c_*/2} dx \\ &\leq 2n^{-1}\log(nc_*/2)/\ln(1/\lambda)c_*. \end{aligned} \tag{12}$$

Since

$$\sum_{k\geq m} k\lambda^k \leq m\lambda^m/(1{-}\lambda)^2 \qquad (m{\geq}1),$$

we have

$$\sum_{k\geq\log(nc_*/2)} k\lambda^k e^{-n\lambda^k c_*/2} \leq \sum_{k\geq\log(nc_*/2)} k\lambda^k \leq 2\lceil\log(nc_*/2)\rceil/nc_*(1{-}\lambda)^2\,, \tag{13}$$

where $\lceil x\rceil$ denotes the smallest integer greater than or equal to x.

Combining estimates (9) – (13), we derive (3). The proof is complete. □

Acknowledgments. The author is grateful to the Associate Editor and the reviewers for helpful comments.

References

[1] Balakrishnan N., Koutras M.V. (2001) *Runs and scans with applications.* New York: Wiley.

[2] Barbour A.D., Holst L. and Janson S. (1992) *Poisson approximation.* Oxford: Clarendon Press.

[3] Bateman G.I. (1948) On the power function of the longest run as a test for randomness in a sequence of alternatives. — Biometrika, v. 35, 97–112.

[4] Chryssaphinou O., Papastavridis S.G. (1990) Limit distribution for a consecutive-k-out-of-n:F system. — Adv. Appl. Prob., v. 22, 491–493.

[5] Erdös P., Rényi A. (1970) On a new law of large numbers. — J. Anal. Math., v. 22, 103–111.

[6] Fu J.C., Wang L., 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. Probab., v. 40, No 2, 346–360.

[7] Galambos J. (1987) *The asymptotic theory of extreme order statistics.* — Melbourne: R.E. Krieger Publishing Co.

[8] Goncharov V.L. (1944) On the field of combinatory analysis. — Amer. Math. Soc. Transl., v. 19, No 2, 1–46.

[9] Lebedev A.V. (2015) *Non-classical problems in extreme value theory.* — DSc thesis. Moscow: Moscow State University.

[10] de Moivre A. (1738) *The doctrine of chances.* — London: H. Woodfall.

[11] Muselli M. (2000) New improved bounds for reliability of consecutive-k-out-of-n:F systems. — J. Appl. Prob., v. 37, 1164–1170.

[12] Novak S.Y. (1988) Time intervals of constant sojourn of a homogeneous Markov chain in a fixed subset of states. — Siberian Math. J., v. 29, No 1, 100–109.

[13] Novak S.Y. (1989) Asymptotic expansions in the problem of the longest head–run for a Markov chain with two states. — Trudy Inst. Math. (Novosibirsk), 1989, v. 13, 136–147 (in Russian).

[14] Novak S.Y. (1991) Rate of convergence in the limit theorem for the length of the longest head run. — Siberian Math. J., v. 32, No 3, 444–448.

[15] Novak S.Y. (1992) Longest runs in a sequence of m–dependent random variables. — Probab. Theory Rel. Fields, v. 91, 269–281.

[16] Novak S.Y. (2011) *Extreme value methods with applications to finance.* — London: Chapman & Hall/CRC Press. ISBN 9781439835746

[17] Schwager S.J. (1983) Run probabilities in sequences of Markov-dependent trials. — J. Amer Stat. Assoc., v. 78, No 1, 168–175.

[18] Uspensky J.V. (1937) *Introduction to Mathematical Probability.* — McGraw-Hill Book Company.

[19] Vaggelatou E. (2003) On the length of the longest run in a multi-state Markov chain. — Statistics Probab. Letters, v. 62, 211–221.