Turkish J. Eng. Env. Sci.
29 (2005) , 105 – 111.
© TÜBİTAK

The Longest Run Statistic Associated with Exchangeable Binary Variables

Serkan ERYILMAZ
İzmir University of Economics, Department of Mathematics,
İzmir, TURKEY
e-mail: serkan.eryilmaz@ieu.edu.tr

Received 22.07.2004

Abstract

Since the hydrologic risk is defined as a risk of exceedance of a critical level during a period, the number of the longest successive exceedances which is referred to as the longest run statistic shows the length of the longest hydrologic risk period. This paper is concerned with the longest run statistic associated with exceedances. In the derivation of a binary sequence, an independent sequence of random variables and a random threshold are considered. The elements of a binary sequence are assigned with respect to a random threshold. In such a consideration, the elements of the corresponding binary sequence are not independent, but are symmetrically dependent, i.e. exchangeable. The distribution of the longest run statistic is derived for the sequence of exchangeable binary variables and its hydrological use is discussed.

Key words: Longest run, Exchangeable binary variables, Exceedances, Hydrologic risk.

Introduction

There has been considerable research on the longest run statistic in the literature on runs. The history starts with the investigation of the distribution of the longest run in a sequence of independent and identically distributed (i.i.d.) Bernoulli trials. The exact expression for the distribution of the longest run in i.i.d. Bernoulli trials was first presented in Burr and Cane (1961). Many authors have contributed to the longest run: see Philippou and Makri (1985), Philippou (1986), and Philippou and Makri (1986) among others. A Markov chain imbedding technique developed by Fu and Koutras (1994) has been successfully applied to obtain the distribution of the longest run. In a sequence of Markov dependent trials, the longest run is discussed in studies by Lou (1996), Vaggelatou (2003), and Eryılmaz (2005). For a lucid review of the longest run we refer to Balakrishnan and Koutras (2002).

The longest run statistic is representative of critical droughts and floods. See, Millan and Yevjevich (1971), Salazar and Yevjevich (1975), Şen (1976), and Şen (1991). In the use of the longest run statistic in hydrology, binary sequences are derived by considering the i.i.d. random variables that represent droughts or floods and a fixed threshold. More explicitly, let $X_1, X_2, ..., X_n$ be a sequence of i.i.d. random variables that has hydrological information in a certain locality. By using the fixed threshold c, a binary sequence with 2 distinct and independent elements can be obtained. In this case, the corresponding binary sequence consists of i.i.d. Bernoulli trials with success probability $p = P\{X_i > c\}$ and failure probability $q = 1 - p = P\{X_i \leq c\}$. In Şen (1991) the longest run is studied under this type of truncation. In the present work, another type of truncation which can be viewed as a random truncation is considered by using 2 independent sequences of random variables such that one of them contains past observations and the other contains future observations. This type of modelling was used by Thomas (1948), Chow (1951), and Chow (1953) in making predictions about the recurrences of floods and droughts

in the future on the basis of what is known from past data.

Since a random truncation model includes both past and future observations, the findings of the present paper provide the opportunity of making predictions about the recurrences of floods and droughts in the future. The results may also be useful for the computation of risk probabilities required for planning, design and evaluation in a management process. In the following section, we provide an explicit description of the model.

Model Description

In hydrologic frequency analysis, exceedance probabilities are used for floods, being flows larger than given in what we designed, and non-exceedance probabilities are used for droughts, which are flows less than those given by the prepared design . According to Gumbel's equation, the non-exceedance probability is computed by the following equation (see, for example Stedinger et al., (1992)):

$$P_0 = \frac{r}{n+1},$$

where r shows the rank of a drought, n is the number of drought events and P_0 is the non-exceedance probability. The rank of an observation is the number that denotes its position among all observations.

The non-exceedance probability P_0 is the probability that a drought less than the tabulated flow will occur. The statistical description can be given as follows:

Let $X_1, X_2, ..., X_n$ be a sequence of i.i.d. random variables that represent floods or droughts for the past n years in a certain locality. Suppose that $X_{n+1}, X_{n+2}, ..., X_{n+m}$ includes the corresponding variables for future m years in the same locality. It is assumed that $X_1, X_2, ..., X_n$ and $X_{n+1}, X_{n+2}, ..., X_{n+m}$ are independent and have the same continuous distribution function, F. We determine the threshold from the sequence that belongs to the past n years. The rth smallest element, i.e. the rth order statistics $X_{r:n}$ of $X_1, X_2, ..., X_n$ is chosen as a threshold. This threshold is necessary to determine the extreme values $X_{1:n}$ and $X_{n:n}$, which are of special importance in hydrological inferences. In this case, we use the random threshold $X_{r:n}$, the rth smallest flood or drought for the truncation of $X_{n+1}, X_{n+2}, ..., X_{n+m}$. Define the following indicator random variables:

$$\xi_i = \begin{cases} 1 & , X_{n+i} > X_{r:n} \\ 0 & \text{otherwise} \end{cases} \quad , \quad i = 1, 2, ..., m. \tag{1}$$

In order to facilitate the understanding of the variables included in the model, we provide the following illustrative figure:

$$\underbrace{\left.\begin{matrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{matrix}\right\}}_{\text{past}} \xrightarrow{X_{r:n}} \underbrace{\left\{\begin{matrix} X_{n+1} & > X_{r:n} & \Rightarrow \xi_1 = 1 \\ X_{n+2} & < X_{r:n} & \Rightarrow \xi_2 = 0 \\ \vdots & \vdots & \vdots \\ X_{n+m} & > X_{r:n} & \Rightarrow \xi_m = 1 \end{matrix}\right.}_{\text{future}}$$

Figure 1. Illustrative figure of the variables included in the model.

The exceedance and non-exceedance probabilities associated with $\xi_i s$ are given, respectively, by

$$P\{X_{n+i} > X_{r:n}\} = 1 - \frac{r}{n+1} \tag{2}$$

and

$$P\{X_{n+i} \leq X_{r:n}\} = \frac{r}{n+1}. \tag{3}$$

It is clear that the non-exceedance probability P_0 coincides with the probability $P\{X_{n+i} \leq X_{r:n}\}$, i.e. the probability that a drought less than the tabulated flow will occur. Similarly, the exceedance probability given in (2) is the probability that a flood greater than the tabulated flow will occur if $X_i s$ represent floods.

In an experiment involving 2 possible outcomes, 1 "success" and 0"failure", a run of "1" type of element is an uninterrupted sequence of such elements bordered at each end by the other type of element "0". For example, in the binary sequence given in (4), we have a run of two 1s, a run of one 1 and a run of three 1s. In the present work, our main interest is the number of the longest successive exceedances, which is referred to as the **longest run statistic**.

$$110001011100 \tag{4}$$

For the sequence given above the longest run statistic takes the value 3. Actually, the elements of the sequence given in (4) are $\xi_1, \xi_2, ..., \xi_m$. Since we have

used a random threshold instead of a certain threshold, the elements of the binary sequence $\xi_1, \xi_2, ..., \xi_m$ are not i.i.d. Bernoulli trials, but are exchangeable, i.e. the joint distribution of $\xi_1, \xi_2, ..., \xi_m$ is invariant under permutations of the subscripts. Hence, we do not have classical Bernoulli trials here.

The sum of the random variables $\xi_1, \xi_2, ..., \xi_m$ is called an exceedance statistic and is used in many fields such as reliability, hydrology, and statistical quality control. The statistic $S_m = \#\{i \leq m : X_{n+i} > X_{r:n}\}$ and different types of exceedance statistics are discussed in studies by Epstein (1954), Sarkadi (1957), Bairamov (1997), Wesolowski and Ahsanullah (1998), Bairamov and Eryılmaz (2000), and Eryılmaz (2003a, 2003b). Under the assumption of independence of $X_1, X_2, ..., X_n$ and $X_{n+1}, X_{n+2}, ..., X_{n+m}$ and having the same continuous distribution function, the probability distribution of the random variable S_m is

$$P\{S_m = k\} = \frac{\binom{m-k+r-1}{m-k}\binom{n-r+k}{k}}{\binom{m+n}{n}}, \quad k = 0, 1, ..., m. \tag{5}$$

A general formula for the distribution of the sum of exchangeable binary random variables has been derived by George and Bowman (1995). Let $\xi_1, \xi_2, ..., \xi_m$ be a set of exchangeable binary random variables and let

$$\lambda_k = P\{\xi_1 = \xi_2 = ... = \xi_k = 1\}.$$

Using an inclusion and exclusion principle, George and Bowman (1995) obtained

$$P\{S = s\} = \binom{m}{s}\sum_{j=0}^{m-s}(-1)^j\binom{m-s}{j}\lambda_{s+j} \tag{6}$$

where $S = \sum_{i=1}^{m}\xi_i$. For the variables given in (1), $\lambda_k = \frac{\binom{n-r+k}{k}}{\binom{n+k}{n}}$ and λ_1 represents the exceedance probability given in (2).

Our aim is to find the distribution of the longest run statistic for the sequence of binary random variables given in (1). Since the random variables defined in (1) are exchangeable (or symmetrically dependent), we have to derive the distribution of the longest run statistic for the sequence of exchangeable binary random variables.

The Distribution of the Longest Run Length in Exchangeable Binary Trials

In this section we provide the distribution of the longest run L_m for the sequence of exchangeable binary trials.

Let us consider the conditional distribution of the length of the longest run in m trials, given the number of $S_m = l$ $(0 \leq l \leq m)$ successes. Clearly,

$$P\{L_m < k \mid S_m = l\} = \frac{P\{L_m < k, S_m = l\}}{P\{S_m = l\}}.$$

Since $\xi_1, \xi_2, ..., \xi_m$ are exchangeable

$P\{L_m < k, S_m = l\} = N(m, k, l)P\{l$ of the ξ s are 1 and $m - l$ of the ξs are $0\}$

where the coefficient $N(m, k, l)$ shows the number of ways corresponding with the event $\{L_m < k, S_m = l\}$. We already know that

$P\{S_m = l\} = \binom{m}{l}P\{l$ of the ξs are 1 and $m - l$ of the ξs are $0\}$

hence,

$$P\{L_m < k \mid S_m = l\} = \frac{N(m, k, l)}{\binom{m}{l}}. \tag{7}$$

The equality given in (7) also holds for i.i.d. trials and it is known that for i.i.d. trials

$$P\{L_m < k \mid S_m = l\} = \binom{m}{l}^{-1}\sum_{i=0}^{\left[\frac{l}{k}\right]}(-1)^i\binom{m-l+1}{i}\binom{m-ik}{m-l} \tag{8}$$

where $[x]$ denotes the integer part of x (see Riordan (1958)). We observe that the conditional distribution of L_m given the number of successes remains valid for the sequence of exchangeable binary trials. Using the total probability law, we may for all $k = 1, 2, ..., m$ obtain

$$P\{L_m < k\} = \sum_{l=0}^{m}P\{L_m < k \mid S_m = l\}P\{S_m = l\} \tag{9}$$

and by using (6) and (8) in (9)

$$P\{L_m < k\} = \sum_{l=0}^{m} \sum_{i=0}^{\min(\left[\frac{l}{k}\right], m-l+1)} \sum_{j=0}^{m-l} (-1)^i (-1)^j \binom{m-l+1}{i} \binom{m-ik}{m-l} \binom{m-l}{j} \lambda_{l+j} \tag{10}$$

Equation (10) gives the cumulative distribution of the longest run length for the sequence of exchangeable binary trials.

Now, let us consider the random variables given in (1). In this case, since $\lambda_{l+j} = \frac{\binom{n-r+l+j}{l+j}}{\binom{n+l+j}{n}}$, for all $k = 1, 2, ..., m$ we have

$$\begin{aligned} P\{L_m < k\} &= \binom{m+n}{n}^{-1} \sum_{l=0}^{m} \sum_{i=0}^{\min(\left[\frac{l}{k}\right], m-l+1)} \left[(-1)^i \binom{m}{l}^{-1} \binom{m-l+1}{i} \binom{m-ik}{m-l} \right. \\ &\quad \left. \times \binom{m-l+r-1}{m-l} \binom{n-r+l}{l} \right] \end{aligned} \tag{11}$$

In Figures 2-4, graphs of the cumulative distribution given in (11) for various exceedance probabilities (or various values of n and r) are illustrated. In each figure for sample sizes $m = 5, 10, 20$, and 50 from the inner towards the outer one, respectively, cumulative distributions are graphed. According to Figures 2-4, increasing the exceedance probability leads to a decrease in the longest run length occurrence.

In Table 1, numerical values for the $E(L_m)$, expectation of the longest run are presented for various values of n, r and m.

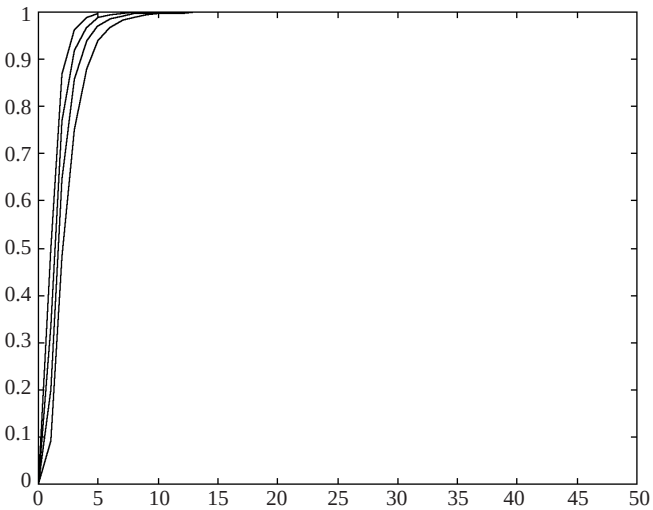

Figure 2. Cumulative distribution for $n = 5, r = 5$ (exceedance probability = 0.1667).

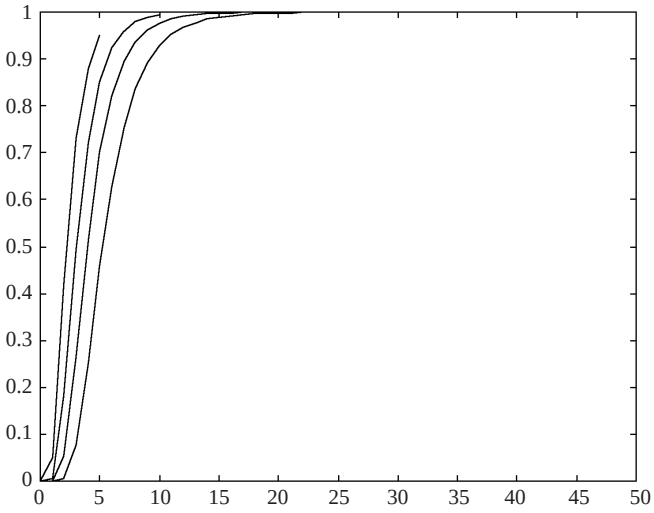

Figure 3. Cumulative distribution for $n = 15, r = 8$ (exceedance probability $= 0.5000$).

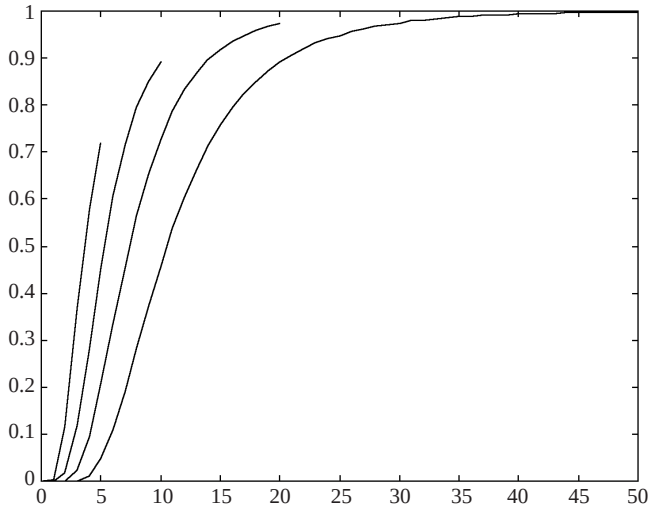

Figure 4. Cumulative distribution for $n = 15, r = 4$ (exceedance probability $= 0.7500$).

Table 1. Expectation of the longest run.

n	r	m	$E(L_m)$	n	r	m	$E(L_m)$
5	5	5	0.6865	15	4	5	3.2201
5	5	10	1.0355	15	4	10	5.2742
5	5	20	1.4182	15	4	20	7.8553
5	5	50	1.9409	15	4	50	11.713
9	5	5	1.9930	15	8	5	1.9737
9	5	10	2.9658	15	8	10	2.9043
9	5	20	4.0420	15	8	20	3.9209
9	5	50	5.5365	15	8	50	5.3233

The foregoing findings enable us to give an answer to the following type of question:

Example. What is the probability that the largest flood during the past 10 years will be exceeded a maximum of twice consecutively during the next 10 years?

Since $n = r = m = 10$, the corresponding probability is

$$P\{L_{10} = 2\} = P\{L_{10} < 3\} - P\{L_{10} < 2\}$$

$$= 0.9796 - 0.9021 = 0.078$$

Conclusions

The distribution function of the longest run length associated with exchangeable binary trials in a random truncation model was derived by the conditioning method. In a random truncation model,

the threshold is chosen randomly from n past years and r and $X_{r:n}$ denote the rank and the corresponding observation of the chosen year among $X_1, X_2, ..., X_n$, respectively. The truncation of a sequence $X_{n+1}, X_{n+2}, ..., X_{n+m}$ (future observations) at a random level $X_{r:n}$ creates a dependent sequence $\xi_1, \xi_2, ..., \xi_m$ of binary variables. We observe that the longest run length based on a sequence $\xi_1, \xi_2, ..., \xi_m$ depends on n, r and m.

Although the different values of n and r give the same exceedance probability, the dependence structure of a binary sequence $\xi_1, \xi_2, ..., \xi_m$ affects the distribution of L_m, and hence its expectation. As seen from Table 1, the same exceedance probabilities for $n = 15, r = 8$ and $n = 9$,and $r = 5$ give different expected values.

In Şen (1991), expectations of the longest run length are tabulated for a certain truncation level, c, with exceedance probability $p = P\{X_i > c\} = 0.5$. According to his computations, for $m = 10, 20$,and 50, $E(L_m)$ takes the values $2.799, 3.729$,and 5.007 respectively. One observes from Table 1, for the same exceedance probability, 0.5, (taking $n = 15, r = 8$ or $n = 9, r = 5$) that expectations depart from the foregoing values. This demonstrates that the longest run length distribution is sensitive to both the variation of the values of n,and r and to the dependence structure of $\xi_1, \xi_2, ..., \xi_m$. This sensitivity constitutes the difference between certain and random truncation models.

Acknowledgment

The author would like to thank the referees for their helpful comments and suggestions, which led to an improvement of the presentation of the paper.

References

Bairamov, I.G., "Some Distribution Free Properties of Statistics Based on Record Values and Characterizations of the Distributions through a Record", J. Appl. Statist. Sci., 5, 17-25, 1997.

Bairamov, I.G. and Eryılmaz, S., "Distributional Properties of Statistics Based on Minimal Spacing and Record Exceedance Statistics", Journal of Statistical Planning and Inference, 90, 21-33, 2000.

Balakrishnan, N. and Koutras, M.V. , Runs and Scans with Applications, Wiley Series in Probability and Statistics, 2002.

Burr, E.J. and Cane, G., "Longest Run of Consecutive Observations Having a Specified Attribute", Biometrika, 48, 461-465, 1961.

Chow, V.T., "A General Formula for Hydrologic Frequency Analysis", Trans. Amer. Geophys. Union, 32, 231-282, 1951.

Chow, V.T., "Frequency Analysis of Hydrologic Data with Special Application to Rainfall Intensities", University of Illinois Bulletin, 50, 1953.

Epstein, B., "Tables for the Distribution of the Number of Exceedances", Ann. Math. Statist., 25, 762-768, 1954.

Eryılmaz, S., "On the Distribution and Expectation of Success Runs in Nonhomogeneous Markov Dependent Trials", Statistical Papers, 46, 117-128, 2005.

Eryılmaz, S., "Random Threshold Models Based on Multivariate Observations", Journal of Statistical Planning and Inference, 113, 557-568, 2003a.

Eryılmaz, S., "Records and Exceedances When Underlying Distribution Contains Atoms", Pakistan Journal of Statistics, 19, 25-39, 2003b.

Fu, J. and Koutras, M., "Distribution Theory of Runs: Markov Chain Approach", J. Amer. Stat. Assoc., 89, 1050-1058, 1994.

George, E.O. and Bowman, D., "A Full Likelihood Procedure for Analyzing Exchangeable Binary Data", Biometrics, 51, 512-523, 1995.

Lou, W.Y.W., "On Runs and Longest Run Tests: A Method of Finite Markov Chain Imbedding", J. Amer. Statist. Assoc., 91, 1595-1601, 1996.

Millan, J. and Yevjevich, V. , "Probabilities of Observed Droughts", Hydrol. Pap., Colorado State University, Fort Collins, 51 pp., 1971.

Philippou, A.N. and Makri, F.S. , "Longest Success Runs and Fibonacci Type Polynomials", The Fibonacci Quarterly, 23, 338-346, 1985.

Philippou, A.N. and Makri, F.S., "Successes, Runs and Longest Runs", Statist. Probab. Lett., 4, 101-105, 1986.

Philippou, A.N. , "Distributions and Fibonacci Polynomials of Order k, Longest Runs, and Reliability of Consecutive $k-out-of-n$: F Systems", Fibonacci Numbers and Their Applications (edited by A.N. Philippou, G.E. Bergum and A.F. Horadam), Reidel, Dordrecht, 203-227, 1986

Riordan, J., An Introduction to Combinatorial Analysis, John Wiley & Sons, New York, 1958.

Salazar, P.G. and Yevjevich, V. , "Analysis of Drought Characteristics by the Theory of Runs", Hydrol. Pap. No. 80, Colorado State University, Fort Collins, 1975.

Sarkadi, K., "On the Distribution of the Number of Exceedances", Ann. Math. Stat., 28, 1021-1022, 1957.

Stedinger, J.R., Vogel, R.M. and Georgiou, E.F., Frequency Analysis of Extreme Events. Handbook of Hydrology. D.S. Maidment. New York, McGraw-Hill, 1992.

Şen, Z., "Wet and Dry Periods of Annual Flow Series", J. Hydraul. Eng., ASCE, Prof. Pap. 12457, 102 (HY10), 1503-1514, 1976.

Şen, Z., "On the Probability of the Longest Run Length in an Independent Series", Journal of Hydrology, 125, 37-46, 1991.

Thomas, H.A., "Frequency of Minor Floods", J. Boston Soc. Civil Engineers, 35, 425-442, 1948.

Vaggelatou, E., "On the Length of the Longest Run in a Multi State Markov Chain", Statist. Probab. Lett., 62, 211-221, 2003.

Wesolowski, J. and Ahsanullah, M., "Distributional Properties of Exceedance Statistics", Ann. Inst. Statist. Math., 50, 543-565, 1998.