RUNS BASED ON DISCRETE ORDER STATISTICS

A. STEPANOV[1] §

Abstract. In the present paper, we study runs based on discrete order statistics. Limit results for the total number of such runs and the length of the longest run are derived.

Keywords: Runs, order statistics, limit laws.

AMS Subject Classification: 60F05, 60F15

1. Introduction

Runs are commonly used in different applications including reliability, quality control and molecular biology. They are intensively discussed in the statistical literature. For a thorough review on runs we refer to Balakrishnan and Koutras (2002), Fu and Lou (2003), and the references therein.

At the first stage of the research on runs, most studies were conducted for runs obtained from the sequences of integer-valued random variables. Attention has been turned later towards runs based on the sequences of continuous random variables. Among some recent publications in that direction we mention the papers of Eryilmaz (2005, 2007, 2008), Eryilmaz and Fu (2008), Fan et al. (2008), Eryilmaz and Stepanov (2008), and Stepanov (2011a, 2011b), where runs based on spacings and ratios of order statistics of continuous random variables have been investigated.

In the present work, we study runs based on order statistics obtained from discrete distributions. The problem discussed in our paper is new.

Let in the following, $X_1, \ldots, X_n$ be independent and identically distributed random variables taking the positive integers and having a distribution $F(n) = P\{X \leq n\}$. Let

$$-\infty = X_{0,n} < X_{1,n} \leq \ldots \leq X_{n,n} < X_{n+1,n} = \infty$$

be the order statistics obtained from this sample. We say that the sequence of order statistics $X_{i,n}, \ldots, X_{j,n}$ forms a run of the length $j - i - 1$ $(j \geq i+3,\ i \geq 0,\ j \leq n+1)$ if

$$X_{i,n} < X_{i+1,n} = X_{i+2,n} = \ldots = X_{j-1,n} < X_{j,n}.$$

If $j = i+1$, i.e. we have two order statistics $X_{i,n} < X_{i+1,n}$, we say that the order statistics $X_{i,n}, X_{i+1,n}$ form a run of the length 1; if $j = i+2$ and $X_{i,n} < X_{i+1,n} < X_{i+2,n}$, we say that $X_{i,n}, X_{i+1,n}, X_{i+2,n}$ form two runs, where each of these runs has the length 1.

Let us define auxiliary variables $\xi_{i,n}$ by

$$\xi_{1,n} = 1, \qquad \xi_{n+1,n} = 1,$$

[1] Department of Mathematics, Izmir University of Economics, 35330, Balcova, Izmir, Turkey, e-mail: alexei.stepanov@ieu.edu.tr

§ Manuscript received 16 August 2011.

$$\xi_{i,n} = \begin{cases} 1, & \text{if} \quad X_{i-1,n} < X_{i,n} \quad i = 2, \dots, n, \\ 0, & \text{otherwise.} \end{cases}$$

In terms of these variables, if

$$\xi_{i+1,n} = 1, \xi_{i+2,n} = \dots = \xi_{j-1,n} = 0, \xi_{j,n} = 1,$$

then we have a run of the length $j - i - 1$. We also define a variable $R_n = 1, 2, \dots, n$ by

$$R_n = \sum_{l=1}^{n} \xi_{l,n}.$$

We refer to R_n as to the total number of runs based on the order statistics $X_{1,n}, \dots, X_{n,n}$. In the case of a single run, when $\xi_i = 0$ $(2 \leq i \leq n)$, we have $R_n = 1$. Observe that the total number of runs indicates how many 'strong' (different) order statistics are in a discrete sample. For the length of the longest run we will use the following designation L_n $(1 \leq L_n \leq n)$.

The above notions are illustrated by the following example.

Example 1.1. *Let us consider a sequence of integer order statistics:*

$$\underbrace{2,2,2,}\underbrace{\;4\;,}\underbrace{\;5\;,}\underbrace{8,8,8,8,}\underbrace{9,9,9,}\underbrace{10,10,}\underbrace{14,14,14,14,14,14}.$$

We have seven run groups, and $R_{20} = 7,\ L_{20} = 6$.

The rest of this paper is organized as follows. In Section 2, we present distributional and limit results for R_n and L_n. In Subsection 2.1, we discuss these distributional and limit results in the finite case. In Subsection 2.2, we analyze the asymptotic behavior of runs based on discrete order statistics when the support is unbounded. In particular, we will show that for any unbounded support

$$R_n \to \infty \text{ a.s.}, \qquad \frac{R_n}{n} \to_p 0,$$

$$\frac{R_n}{\sum_{i=1}^{\infty}(1-(1-p_i)^n)} \to_p 1 \qquad \text{and} \qquad \frac{L_n}{n} \to_p p_M,$$

where $p_M = max\{p_1, p_2, \dots\}$.

2. Results

Let $p_i = P\{X = i\}$ and N $(1 \leq N \leq \infty)$ be the smallest number such that $p_i = 0$ $(i > N)$.

2.1. **Finite case.** In this subsection we assume that $N < \infty$. The probability mass function of $\xi_{i,n}$ and the expected value of R_n can be written as

$$P\{\xi_{i,n} = 1\} = \sum_{j=1}^{N-1} P\{j = X_{i-1,n} < X_{i,n}\} =$$

$$\frac{n!}{(n-i+1)!} \sum_{j=1}^{N-1} (1 - F(j))^{n-i+1} \sum_{l=1}^{i-1} \frac{p_j^l F^{i-l-1}(j-1)}{l!(i-l-1)!} =$$

$$\frac{n!}{(n-i+1)!(i-1)!} \sum_{j=1}^{N-1} (1 - F(j))^{n-i+1} [F^{i-1}(j) - F^{i-1}(j-1)]$$

and

$$ER_n = N - F^n(N-1) - \sum_{j=1}^{N-1}(1-p_j)^n. \qquad (1)$$

It follows from (1) that $ER_n \to N$ as $n \to \infty$. Generally speaking, in the finite case we have a Bernoulli scheme, where $R_n \to N$ *a.s.* and $L_n/n \to p_M$ *a.s.* for $p_M = \max\{p_1, \ldots, p_N\}$.

2.2. **Infinite case.** Let $N = \infty$. The probability mass function of $\xi_{i,n}$ and the joint mass function of $\xi_{i,n}, \xi_{k,n}$ are given here by

$$P\{\xi_{i,n} = 1\} = \frac{n!}{(n-i+1)!(i-1)}\sum_{j=1}^{\infty}(1-F(j))^{n-i+1}[F^{i-1}(j) - F^{i-1}(j-1)] \qquad (2)$$

and

$$P\{\xi_{i,n} = 1, \xi_{k,n} = 1\} = \qquad (2 \le i < k \le n)$$

$$\sum_{j=1}^{\infty}\sum_{t=j+1}^{\infty} P\{X_{i-1,n} = j, X_{k-1,n} = t\} = \frac{n!}{(n-k+1)!}\sum_{j=1}^{\infty}\sum_{l=1}^{i-1}\frac{p_j^l F^{i-l-1}(j-1)}{l!(i-l-1)!}$$

$$\sum_{t=j+1}^{\infty}(1-F(j))^{n-k+1}\sum_{m=1}^{k-i}\frac{p_t^m[F(t-1)-F(j)]^{k-i-m}}{m!(k-i-m)!} =$$

$$\frac{n!}{(n-k+1)!}\sum_{j=1}^{\infty}\frac{F^{i-1}(j) - F^{i-1}(j-1)}{(i-1)!}$$

$$\sum_{t=j+1}^{\infty}(1-F(j))^{n-k+1}\frac{[F(t)-F(j)]^{k-i} - [F(t-1)-F(j)]^{k-i}}{(k-i)!}. \qquad (3)$$

The conditional mass function of R_{n+1} given $\{R_n = k\}$ can be written as:

$$P\{R_{n+1} = k+1 | R_n = k\} = P\{X_{n+1} \neq X_i\} = \qquad (i = 1, \ldots, n)$$

$$\sum_{j=1}^{\infty} p_j(1-p_j)^n.$$

Then

$$E\{R_{n+1}|R_n = k\} = k + \sum_{j=1}^{\infty} p_j(1-p_j)^n.$$

From the properties of the conditional expectation, one can obtain

$$ER_n = \sum_{j=1}^{\infty}[1-(1-p_j)^n] = E_n. \qquad (4)$$

It should be mentioned that (4) can be derived directly from (2). It follows from (4) that $ER_n \to \infty$.

Let us consider other properties of R_n. It is easily seen that $R_{n+1} \ge R_n$ *a.s.* This observation leads us to the following limit result.

Theorem 2.1. *For any discrete F such that $N = \infty$, we have*

$$R_n \to \infty \quad a.s.$$

Proof. Since R_n is monotone, it converges almost surely either to a finite limit or to infinity. Thus, it is sufficient to show that R_n converges in probability to infinity.

The event $\{R_n = k\}$ means that the variables $X_1, \ldots, X_n$ take exactly k different positive integer values, i.e.

$$P\{R_n \le m\} = \sum_{k=1}^{m} P\{R_n = k\} =$$

$$\sum_{k=1}^{m} \frac{n!}{(n-k+1)!} \sum_{i_1=1}^{\infty} p_{i_1} \sum_{i_2=i_1+1}^{\infty} p_{i_2} \cdots \sum_{i_k=i_{k-1}+1}^{\infty} p_{i_k} (p_{i_1}+p_{i_2}+\ldots+p_{i_k})^{n-k} \le \sum_{k=1}^{m} \frac{n!}{(n-k+1)!} p_{sup,k}^{n-k} \to 0, \tag{5}$$

where $p_{sup,k} = \sup_{i_1,i_2,\ldots,i_k}(p_{i_1}+p_{i_2}+\ldots+p_{i_k}) < 1$. □

In light of the previous result it is interesting to trace the limit behavior of the ratio $\frac{R_n}{n}$. It is easily seen that

$$\frac{ER_n}{n} > \frac{ER_{n+1}}{n+1} \qquad (n \ge 1).$$

The following asymptotic result holds.

Theorem 2.2. *For any discrete F such that $N = \infty$, we have*

$$\frac{ER_n}{n} \to 0 \qquad \textit{and} \qquad \frac{R_n}{n} \to_p 0.$$

Proof. Indeed,

$$\frac{ER_n}{n} = o_n(1) + \frac{1}{n} \sum_{j=J}^{\infty} [1 - (1 - p_j)^n].$$

By the inequality

$$n(a-b)b^{n-1} < a^n - b^n < n(a-b)a^{n-1} \qquad (a > b > 0), \tag{6}$$

we get

$$\frac{ER_n}{n} < o_n(1) + 1 - F(J-1).$$

Choosing J as large one can make the right-hand side of the last inequality smaller than any $\varepsilon > 0$. The first statement of Theorem 2.2 readily follows.

The second statement of Theorem 2.2 follows from Chebyshev's inequality and the first statement. □

We see that the sequence R_n converges to infinity, and R_n/n converges to zero. Properly normalized R_n can tend to one.

Theorem 2.3. *For any discrete F such that $N = \infty$, we have*

$$\frac{R_n}{E_n} \to_p 1$$

Proof. By Chebyshev's inequality,

$$P\{|R_n - ER_n| > \varepsilon ER_n\} \le \frac{VarR_n}{\varepsilon^2 E_n^2}.$$

Let us estimate $VarR_n$. We have

$$VarR_n = 2 \sum_{i=2}^{n-1} \sum_{k=i+1}^{n} E\xi_{i,n}\xi_{k,n} + 3E_n - E_n^2.$$

It follows from (3) that

$$\sum_{i=2}^{n-1}\sum_{k=i+1}^{n} E\xi_{i,n}\xi_{k,n} = 1 - E_n + \sum_{j=1}^{\infty}\sum_{l=j+1}^{\infty}(1-(1-p_j)^n-(1-p_l)^n+(1-p_j-p_l)^n)$$

and

$$VarR_n = 2 + E_n - E_n^2 + 2\sum_{j=1}^{\infty}\sum_{l=j+1}^{\infty}(1-(1-p_j)^n-(1-p_l)^n+(1-p_j-p_l)^n) =$$

$$2+\sum_{j=1}^{\infty}((1-p_j)^n-(1-2p_j)^n)+\sum_{j=1}^{\infty}\sum_{l=1}^{\infty}((1-p_j-p_l)^n-(1-p_j-p_l+p_jp_l)^n) \leq$$

$$2+\sum_{j=1}^{\infty}((1-p_j)^n-(1-2p_j)^n).$$

Then

$$\frac{VarR_n}{E_n^2} = \frac{\sum_{j=1}^{\infty}((1-p_j)^n-(1-2p_j)^n)}{\sum_{j=1}^{\infty}(1-(1-p_j)^n)E_n}.$$

By (5),

$$\frac{VarR_n}{E_n^2} < \frac{n\sum_{j=1}p_j(1-p_j)^{n-1}}{\sum_{j=1}^{\infty}(1-(1-p_j)^n)E_n} < \frac{1}{E_n} \to 0.$$

□

In the above theorem we found that $\frac{R_n}{E_n} \to_p 1$. What one can say about the limiting behavior of E_n? Obviously, $E_n \leq n$. In the following result we find the lower bound for E_n.

Proposition 2.1. *For any $n \geq 1$*

$$E_n \geq \sum_{j=1}^{n}\frac{(1-F(1))^{j-1}}{j}.$$

Proof. Observe that

$$E_n = \sum_{i=1}^{\infty}p_i + \sum_{i=1}^{\infty}p_i(1-p_i) + \ldots + \sum_{i=1}^{\infty}p_i(1-p_i)^{n-1}.$$

However,

$$\sum_{i=1}^{\infty}p_i(1-p_i)^j \geq \sum_{i=1}^{\infty}(F(i)-F(i-1))(1-F(i))^j \geq \int_{F(1)}^{1}(1-x)^j dx = \frac{(1-F(1))^j}{j+1}.$$

□

It is reasonable to assume that for some distributions the sum E_n can tend to finite limit as $n \to \infty$ and for other distributions it can tend to infinity. We focus now on these matters.

In Eisenberg *et al.* (1993) and Brands *et al.* (1994) the asymptotic behavior of the number of maxima in the discrete case is discussed. The following limit

$$\lim_{i\to\infty}\frac{p_{i+1}}{p_i} = p \in [0,1] \tag{7}$$

is used for distribution tail classification.

Theorem 2.4. *1) Let the limit in (7) exist and $p \in [0,1)$, then $E_n \to E < \infty$.*
1) Let the limit in (7) exist and $p = 1$, then $E_n \to \infty$.

Proof. Observe first that if the limit in (7) exist and $p \in [0,1]$, then for any $n \geq 1$

$$\lim_{i\to\infty} \frac{1-(1-p_{i+1})^n}{1-(1-p_i)^n} = p \in [0,1].$$

2) Let us choose $\varepsilon > 0$ and I such that

$$\frac{1-(1-p_{i+1})^n}{1-(1-p_i)^n} < p+\varepsilon < 1 \quad (i > I).$$

Then

$$\sum_{i=I+1}^{\infty} [1-(1-p_i)^n] < \frac{1-(1-p_{I+1})^n}{1-p-\varepsilon}.$$

Obviously, E_n is bounded and increasing, which implies the first statement of Theorem 2.4.

2) Choose $\varepsilon > 0$ and I such that

$$\frac{1-(1-p_{i+1})^n}{1-(1-p_i)^n} > 1-\varepsilon \quad (i > I).$$

Then

$$\sum_{i=I+1}^{\infty} [1-(1-p_i)^n] > \frac{1-(1-p_{I+1})^n}{\varepsilon}.$$

Taking ε as small we can do the right-hand side of the last inequality as great. This implies the truth of the second statement of Theorem 2.4. □

The last result can be commented in the following way. When $p \in [0,1)$, i.e. the distribution tail is not heavy, the order statistics are concentrated in the "beginning" of the tail, each positive integer is occupied by many sample observations and, consequently, we have a finite number of runs. When the tail is heavy, i.e. $p = 1$, the sample disperse is great. We have many "strong" (isolated) order statistics, which give us the infinite number of runs.

In the conclusion, we present a limit result for the longest run.

Proposition 2.2. *For any discrete F such that $N = \infty$, we have*

$$\frac{L_n}{n} \to_p p_M,$$

where $p_M = \max\{p_1, p_2 \ldots\}$.

Proof. We apply the same argument that was already used in the end of Subsection 2.1. Choose K rather large such that $K > J \geq 1$, where $p_M = p_J$, and $p_M > \widetilde{p}_K = p_K + p_{K+1} + \ldots$ Now we can suppose that we have a finite case, where $X = 1, \ldots, K-1, K$ with $p_1, \ldots, p_{K-1}, \widetilde{p}_K$. The argument that was once used for a Bernoulli scheme can be applied now again. □

References

[1] Balakrishnan, N. and Koutras, M.V., (2002), Runs and scans with applications, Wiley Series in Probability and Statistics.

[2] Brands, J.J.A.M., Steutel, F.W. and Wilms, R.J.G., (1994), On the number of maxima in a discrete sample, Statistics & Probability Letters, 20, 209–217.

[3] Eisenberg, B., Stengle, G. and Strang, G., (1993), The asymptotic probability of a tie for first place, Annals of Applied Probability, 3, 731–745.

[4] Eryilmaz, S., (2005), On the distribution and expectation of success runs in nonhomogeneous Markov dependent trials, Statistical Papers, 46(1), 117–128.

[5] Eryilmaz, S., (2007), Extension of runs to the continuous valued sequences, Statistics & Probability Letters, 77(4), 383–388.

[6] Eryilmaz, S., (2008), Distribution of runs in a sequence of exchangeable multi-state trials, Statistics & Probability Letters, 78, 1505–1513.

[7] Eryilmaz, S. and Fu, J., (2008), Runs in continuous-valued sequences, Statistics & Probability Letters, 78, 759-765.

[8] Eryilmaz, S. and Stepanov, A., (2008), Runs in an ordered sequence of random variables, Metrika, 67, 299–313.

[9] Fan, A. H., Wang, Z. Z. and Ding, F. Q., (2008), Some limit theorems of runs to the continuousvalued sequence, Kybernetes, 37, 1279-1286.

[10] Fu, J.C. and Lou W.Y.W., (2003), Distribution theory of runs and patterns and its applications, World Scientific Publishing, Singapore.

[11] Stepanov, A., (2011a), Limit theorems for runs based on 'small' spacings, Statistics & Probability Letters, 81, 54–61.

[12] Stepanov, A., (2011b), Runs based on the ratios of consecutive order statistics, Communications in Statistics -Theory and Methods, 40(18), 3252–3268.

Alexei Stepanov is presently an Associate Professor at the Department of Mathematics of Izmir University of Economics, Turkey. He graduated from the Faculty of Mathematics and Mechanics of St.Petersburg (Leningrad) State University, USSR, in 1985 and received his PhD degree in probability and statistics there in 1989. In the period of 1990-2008 he worked at Kaliningrad State Technical University, Russia. He has been working at Izmir University of Economics since 2008. His research interests are probability theory, statistics, stochastic processes and simulation with focusing in these topics on the theories of records and order statistics, runs, limit theorems and multivariate distributions.