Communications in Mathematics

Alexei Stepanov
On the Mathematical Theory of Records

Communications in Mathematics, Vol. 29 (2021), No. 1, 151–162

Persistent URL: http://dml.cz/dmlcz/148996

sciendo

Communications in Mathematics 29 (2021) 151–162
DOI: 10.2478/cm-2021-0009

On the Mathematical Theory of Records

Alexei Stepanov

Abstract. In the present work, we briefly analyze the development of the mathematical theory of records. We first consider applications associated with records. We then view distributional and limit results for record values and times. We further present methods of generation of continuous records. In the end of this work, we discuss some tests based on records.

1 Introduction

Let $X_1, X_2, \ldots$ be a sequence of random variables. By this sequence of variables, let us define the sequences of upper record times $L(n)(n \geq 1)$ and record values $X(n)(n \geq 1)$ as follows:

$$L(1) = 1, \quad L(n+1) = \min\left\{j : j > L(n),\, X_j > X_{L(n)}\right\}, \tag{1}$$

$$X(n) = X_{L(n)}.$$

If in (1) we replace the second sign $>$ with the sign $<$, then instead of the sequences of upper record times $L(n)$ and values $X(n)$ we obtain the sequences of lower record times and values. Since the theory of lower records follows from the theory of upper records, we basically discuss upper records in this work. Wherein (with the exception of Section 5) the term "upper" is not used.

The first mathematical paper on records was published by Chandler [6]. The paper attracted the attention of many researchers and inspired many new publications. For references, see the books of Arnold et al. [2], Nevzorov [12] and Ahsanullah and Nevzorov [1]; see also the references therein.

It should be noted that records are commonly used in different areas such as sport, finance, reliability, hydrology, survival analysis and others. Let us consider two examples, one relating to insurance and the other to hydrology.

2020 MSC: 60G70, 62G30

Key words: Record times and values, distributional and limit results, inverse-transform and rejection methods, generation techniques, statistical tests.

Affiliation:
Institute of Physics, Mathematics and Information Technology, Immanuel Kant Baltic Federal University, A. Nevskogo 14, Kaliningrad, 236041 Russia
E-mail: alexeistep45@mail.ru

Example 1. Assume that $X_i > 0$ $(1 \le i \le L(n))$ is a set of client claims of some large insurance company. We call a claim X_i a near record claim if $X_i \in [X(n) - a, X(n)]$ for some non-stochastic constant $a > 0$. Let also

$$S_n = \sum_{i=1}^{L(n)} X_i I_{X_i \in [X(n)-a, X(n)]},$$

where I_A is the indicator-function of the event A. It is known that the sums of approximately 10 percent of large claims can cause 90 percent of the total insurance payment of a company. This fact motivates researchers to study the asymptotic behavior of sums of near record claims S_n.

The research on large claims was conducted, among others, in Hashorva and Hüsler [10], [11], Hashorva [9] and Balakrishnan et al. [5]. Let us consider another example.

Example 2. Assume now that $X_i > 0$ $(i \ge 1)$ is a sequence of water levels of some river. Suppose we are going to construct a dam with height H and, correspondingly, going to estimate H. On one hand, it should be such that $\lim_{n\to\infty} P(X(n) < H) = 1$ and, on the other hand, H should be chosen rather small in order to avoid excessive construction spending. This problem also inspires researchers to investigate asymptotic properties of $X(n)$.

The development of the mathematical theory of records is analyzed in the present paper. In Section 2, we discuss distributional results for record values and times. In Section 3, we present limit results for records. Methods of generation of continuous records are presented in Section 4. In final Section 5, we describe some statistical procedures based on records.

2 Distributional Results

2.1 Distributional Results for Record Times

Let us discuss distributional results for record times in the general continuous case. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with continuous distribution F. Let us introduce the record indicators ξ_n $(n \ge 1)$ as follows:

$$\xi_n = \begin{cases} 1, & \text{if } X_n \text{ is a record value,} \\ 0, & \text{otherwise.} \end{cases}$$

The following result was proved by the famous Hungarian mathematician Rényi [18].

Lemma 1. *The variables $\xi_1, \xi_2, \ldots$ are independent and*

$$P(\xi_n = 1) = 1/n \quad (n \ge 1).$$

Making use of Lemma 1, one can find the distribution of $L(2)$. Indeed,

$$\begin{aligned} P(L(2)=k) &= P(\xi_1=1,\xi_2=0,\ldots,\xi_{k-1}=0,\xi_k=1) \\ &= \frac{1}{(k-1)k}. \end{aligned} \tag{2}$$

It also follows from Lemma 1 that the sequence $L(n)$ $(n \geq 1)$ forms a Markov chain and

$$P(L(n)=k|L(n-1)=j) = \frac{j}{(k-1)k} \quad (n \geq 2,\ n-1 \leq j < k).$$

Identity (2) implies that

$$EL(2) = \infty,$$

and then $EL(2) \leq EL(n) = \infty$ $(n \geq 2)$. For $n \geq 2$ one can show that

$$P(L(n)=k) = \frac{|\ S_{k-1}^{n-1}\ |}{k!},$$

where S_k^n are the Stirling numbers of the first kind, defined by

$$x(x-1)\ldots(x-k+1) = \sum_{n=0}^{k} S_k^n x^n.$$

Let us denote the number of records in a sample $X_1,\ldots,X_n$ as $N(n)$. We have $N(n) = \xi_1 + \xi_2 + \ldots + \xi_n$. It should be noted that

$$P(L(n) > m) = P(\xi_1 + \xi_2 + \ldots + \xi_m < n). \tag{3}$$

Let us now estimate the expected value of the number of record values in a continuous sample $X_1,\ldots,X_n$. Obviously,

$$\begin{aligned} EN(n) &= E\xi_1 + E\xi_2 + \ldots + \xi_n \\ &= 1 + 1/2 + \ldots + 1/n \approx \log n. \end{aligned}$$

That way, an observer at average can register $\log 100 \approx 4.6$ record values in a sample $X_1,\ldots,X_{100}$ and $\log 1000 \approx 6.9$ record values in a sample of size $n = 1000$. No doubt record values appear rarely.

Unfortunately, the distribution function of $L(n)$ in the discrete case depends on the underlying distribution F. It can be found only for each particular discrete distribution.

2.2 Distributional Results for Record Values

We first discuss distributional results for continuous record values. The following result was obtained by the Czech mathematician Tata [22].

Theorem 1. *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $H(x) = 1 - e^{-x}$ $(x > 0)$. Then the variables*

$$Y_1 = X(1), Y_2 = X(2) - X(1), Y_3 = X(3) - X(2), \ldots$$

are also independent and identically distributed with H.

Tata's result allows to find the distribution of $X(n)$ when the underlying continuous distribution F is arbitrary. First, observe that in the standard exponential case

$$X(n) \stackrel{d}{=} Y_1 + \ldots + Y_n, \tag{4}$$

where $X \stackrel{d}{=} Y$ means equality of the distributions of X and Y. Then $X(n)$ has a gamma distribution with parameters $(n, 1)$. We, consequently, have

$$P(X(n) \leq x) = \frac{1}{(n-1)!} \int_0^x e^{-u} u^{n-1} du.$$

Let now $X_1, X_2, \ldots$ be independent and identically distributed random variables with arbitrary continuous F. Observe that variables

$$E_1 = -\log(1 - F(X_1)),\ E_2 = -\log(1 - F(X_2)), \ldots$$

are independent and identically distributed with H. It should also be noted that if X_j is a record value among $X_1, X_2, \ldots$, then E_j is a record value among $E_1, E_2, \ldots$ Then if F is an arbitrary continuous distribution, then

$$P(X(n) \leq x) = \frac{1}{(n-1)!} \int_0^{-\log(1-F(x))} e^{-u} u^{n-1} du. \tag{5}$$

Let us consider the form of the joint density of $X(1), \ldots, X(n)$ in the absolutely continuous case when $f(x) = F'(x)$ is the underlying density. Here, we have

$$\begin{aligned} f_{X(1),\ldots,X(n-1),X(n)} \qquad & (x_1, \ldots, x_{n-1}, x_n) \\ = \quad & \frac{f(x_1)}{1 - F(x_1)} \cdots \frac{f(x_{n-1})}{1 - F(x_{n-1})} f(x_n). \end{aligned}$$

It follows that the sequence $X(1), X(2), \ldots$ forms a Markov chain and

$$P(X(n+1) \leq y \mid X(n) = x) = \frac{F(y) - F(x)}{1 - F(x)} \quad (x < y). \tag{6}$$

Let us now discuss distributional results for record values in the discrete case. These results were obtained, in particular, in the works of Shorrock [19], Vervaat [23] and Pakhteev and Stepanov [16]. Assume that $X, X_1, X_2, \ldots$ are independent and identically distributed random variables with support on non-negative integers and for all $n \geq 0$

$$F(n) = P(X \leq n) < 1. \tag{7}$$

Condition (7) guarantees the existence of the sequence $X(n)$ $(n \geq 1)$ with probability one. Indeed, let there be a non-negative integer n_0 such that $F(n_0 - 1) < 1$ and $F(n_0) = 1$. Then $P(X(1) = n_0) = F(n_0) - F(n_0 - 1)$ and $P(X(2)$ exists$) = 1 - (F(n_0) - F(n_0 - 1)) < 1$. Let also $p_n = P(X = n)$ and $q_n = P(X \geq n)$. The joint probability mass function of the first n discrete record values has the form

$$\begin{aligned} P(X(1) = k_1, \quad \ldots \quad & , X(n) = k_n) \\ = \quad & p_{k_n} \prod_{i=1}^{n-1} \frac{p_{k_i}}{q_{k_i+1}} \quad (0 \leq k_1 < \ldots < k_n). \end{aligned}$$

It follows that the sequence $X(n)$ $(n \geq 1)$ forms a Markov chain and

$$P(X(n+m) = k_{n+m}, \quad \ldots \quad , X(n+1) = k_{n+1} | X(n) = k_n) = \frac{p_{k_{n+m}}}{q_{k_n+1}} \prod_{i=n+1}^{n+m-1} \frac{p_{k_i}}{q_{k_i+1}} \quad (m \geq 1).$$

Let us define random indicators η_i $(= 0, 1;\ i = 0, 1, \ldots)$. Let us set $\eta_i = 1$ if there is a record value $X(n)$ such that $X(n) = i$. The following lemma was proved by Shorrock [19].

Lemma 2. *The random variables η_i $(i = 0, 1, \ldots)$ are independent and*

$$P(\eta_i = 1) = \frac{p_i}{q_i}.$$

Representation 1. *Under the conditions of Lemma 2*

$$P(X(n) > m) = P(\eta_0 + \ldots + \eta_m < n) \quad (n \geq 1).$$

3 Limit Results

Continuous case Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with arbitrary continuous F. Let us recall that $N(n) = \xi_1 + \xi_2 + \ldots + \xi_n$, where the variables ξ_i $(1 \leq i \leq n)$ are independent and $P(\xi_i = 1) = 1/i$. By applying the limit theorems apparatus to independent variables ξ_i, one can obtain the following limit results for the number of records $N(n)$:

$$\frac{N(n)}{\log n} \xrightarrow{a.s.} 1 \quad \text{and} \quad \frac{N(n) - \log n}{\sqrt{\log n}} \xrightarrow{d} Z, \tag{8}$$

where by the symbols $\xrightarrow{d}$ and $\xrightarrow{a.s.}$ we denote the convergence in distribution and with probability one, respectively, and Z is a standard normal random variable, i.e.,

$$P(Z \leq x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du.$$

By applying (3) to the limit results in (8), we obtain

$$\frac{\log L(n)}{n} \xrightarrow{a.s.} 1 \quad \text{and} \quad \frac{\log L(n) - n}{\sqrt{n}} \xrightarrow{d} Z.$$

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $H(x) = 1 - e^{-x}$ $(x > 0)$. Using (4), we can easily get that

$$\frac{X(n)}{n} \xrightarrow{a.s.} 1 \quad \text{and} \quad \frac{X(n) - n}{\sqrt{n}} \xrightarrow{d} Z.$$

Let now $X_1, X_2, \ldots$ be independent and identically distributed random variables with arbitrary continuous F. Applying the argument that was used for obtaining (5), we get the limit results

$$\frac{-\log(1 - F(X(n))}{n} \xrightarrow{a.s.} 1 \quad \text{and} \quad \frac{-\log(1 - F(X(n)) - n}{\sqrt{n}} \xrightarrow{d} Z.$$

Discrete case In the end of this section, we assume that $X_1, X_2, \ldots$ are independent and identically distributed random variables with support on non-negative integers and $F(n) = P(X \le n) < 1$ for all $n \ge 0$. Observe that $\sum_{n=1}^{\infty} p_n = 1$ and then by the Dini test $\sum_{n=1}^{\infty} \frac{p_n}{q_n} = \infty$. In turn, the Abel-Dini test implies that $\sum_{n=1}^{\infty} \frac{p_n}{q_n \left(\sum_{i=1}^{n} \frac{p_i}{q_i}\right)^2} < \infty$. Then $\sum_{n=1}^{\infty} \frac{D\eta_n}{\left(\sum_{i=1}^{n} \frac{p_i}{q_i}\right)^2} < \infty$ and Kolmogorov's strong law of large numbers states that $\frac{\sum_{i=1}^{n} \eta_i}{\sum_{i=1}^{n} \frac{p_i}{q_i}} \overset{a.s.}{\to} 1$. Theorem 2 follows now from the last strong convergence and Representation 1.

Theorem 2. *The following asymptotic law*

$$\frac{\sum_{i=0}^{X(n)} \frac{p_i}{q_i}}{n} \overset{a.s.}{\to} 1 \quad (n \to \infty)$$

holds true.

One can derive a version of the central limit theorem for the sequence $X(n)$ $(n \ge 1)$.

Theorem 3. *Let* $\lim_{n\to\infty} \frac{p_n}{q_n} = a < 1$. *Then*

$$\frac{\sum_{i=0}^{X(n)} \frac{p_i}{q_i} - n}{\sqrt{(1-a)n}} \overset{d}{\to} Z \quad (n \to \infty).$$

4 Generation of Continuous Records

In this section, we assume that $X, X_1, X_2, \ldots$ are independent and identically distributed random variables with absolutely continuous distribution F and density f. In this section we first briefly describe some basic methods of generation of random variables. When the inverse continuous distribution function F^{-1} can be found analytically, one can apply the inverse-transform method for generating X.
Inverse-Transform Method *By this method, we can obtain* $X = x$ *as*

$$x = F^{-1}(u),$$

where $U = u$ *is the generation of a random number.*
The method works only for "simple" distributions. When the inverse F^{-1} can be found only numerically, one can use the inverse-transform method along with a numerical method for F^{-1}. An alternative method of generation in the case when F^{-1} cannot be found analytically is the rejection method.
Rejection Method Suppose we can generate a random variable $\tilde{X}$ having density function g by the inverse-transform method. Suppose X with density function h cannot be generated by the inverse-transform method and X and $\tilde{X}$ have the same support. Then, one should find a constant $c > 1$ such that $c = \sup_x \frac{h(x)}{g(x)}$.

Algorithm 1. *The rejection method.*
STEP 1: Generate $Y = y$ *with density* g.
STEP 2: Generate a random number $U = u$.
STEP 3: If $u < \frac{h(y)}{cg(y)}$, *set* $X = y$. *Otherwise, go to STEP 1.*

The choice of $\tilde{X}$ is determined by the fact that $c > 1$ should get the smallest possible value. The number of iterations in this method for obtaining X is a geometric variable with mean c.

Simulation methods of records were discussed in the works of Bairamov and Stepanov [3], Nevzorov and Stepanov [13], Balakrishnan et al. [4], Stepanov et al. [21], Pakhteev and Stepanov [14], [15], [16], [17] and Stepanov [20]; see also the references therein. Various algorithms of record generation are known. The first and most natural algorithm of record generation is the direct one.

The direct algorithm of record generation *The value $X(1) = X_1$ is generated and kept. For $n \geq 1$, one can apply the recursive approach, which assumes that $X(n)$ is already generated. One then generates variables X_i till one of them, say $X_j = x_j$, is greater than $X(n)$. Then $X(n+1) = x_j$ becomes the next record value, which is also kept.*

The direct algorithm allows to obtain sequences of record values in both discrete and continuous cases. However, if a large number of records is needed this algorithm is computationally burdensome and slow. Other (more effective) algorithms of generation of records are based on the fact that sequences of records form Markov chains. For record generation we can use the conditional distribution given by (6). Further in this section we discuss only algorithms of generation of normal records.

Let now Z_i $(i \geq 1)$ be independent and identically distributed random variables with standard normal distribution Φ and density ϕ, and let $Z(n)$ $(n \geq 1)$ be the corresponding normal records. It follows from (6) that the conditional density of $Z(n+1)$ given $Z(n) = z_n$ has the form

$$f_{Z(n+1)|Z(n)}(z_{n+1} \mid z_n) = \frac{\phi(z_{n+1})}{1 - \Phi(z_n)} \quad (z_{n+1} > z_n).$$

The following algorithm was proposed in Pakhteev and Stepanov [17].

Algorithm 2. *The sequence $X(n)$ $(n \geq 1)$ can be generated as follows.*

STEP 1: Generate $X(1) = X_1, X(2), \ldots, X(i)$ $(i \geq 1)$ by the direct algorithm of record generation till $X(i) > 0$.

For $n \geq i$, apply the rejection method and the following recursive approach. Assume that $X(n) = x_n$ is already generated.

STEP 2: Generate random numbers $U_1 = u_1, U_2 = u_2$ and set $\beta_n^* = \frac{x_n + \sqrt{x_n^2 + 4}}{2}$.

STEP 3: If

$$-2 \log u_2 > (x_n - \log u_1/\beta_n^* - \beta_n^*)^2$$

set $X(n+1) = x_n - \log u_1/\beta_n^*$. Otherwise, return to STEP 2.

We explain why do we have to generate negative normal records by the direct algorithm. We compare the conditional density $f_{Z(n+1)|Z(n)}(z_{n+1} \mid z_n)$ with density $g(z_{n+1} \mid z_n, \beta_n) = \beta_n e^{-\beta_n(z_{n+1} - z_n)}$ $(z_{n+1} > z_n)$, where $\beta_n > 0$ is such that g approximates f in the "best" way. For positive z_n the forms of the curves $f_{Z(n+1)|Z(n)}(z_{n+1} \mid z_n)$ and $g(z_{n+1} \mid z_n, \beta_n)$ are similar. The forms of g and f when z_n is negative are different and f cannot be approximated by g for any choice

of β_n. Let $\tau = 1, 2, \ldots$ be a random variable such that $Z_1 \leq 0, \ldots, Z_{\tau-1} \leq 0$ and $Z_\tau > 0$. Observe that for $k \geq 1$

$$\begin{aligned} P(\tau = k) &= P(Z_1 \leq 0, \ldots, Z_{k-1} \leq 0, Z_k > 0) \\ &= \Phi^{k-1}(0)(1 - \Phi(0)) = \frac{1}{2^k}. \end{aligned}$$

It follows that τ is a geometric random variable with parameter $1/2$. Then $E\tau = 2$. That way, in a simulation experiment the number of first negative normal records is insufficient and they can be generated by the direct algorithm.

Remember that in Algorithm 2 $c^*(z_n) = \sup_{z_{n+1} > z_n} \frac{f_{Z(n+1)|Z(n)}(z_{n+1}|z_n)}{g(z_{n+1}|z_n, \beta_n)}$. One can prove that $c^*(z_n) \to 1$ as $z_n \to \infty$. It is known that $Z(n) \overset{a.s.}{\to} \infty$. The convergence $c^*(z_n) \to 1$ tells us that Algorithm 4.2, which is based on the rejection method, eventually works as an algorithm based on the inverse-transform method. With time every generation in a generation experiment is accepted and becomes a new record value. The last argument guaranties efficiency and good performance of Algorithm 2.

If one generates directly standard normal random variables one cannot obtain (with today's best computer software) a standard normal generation which exceeds, say, value 50. We generated in MatLab (by the computer AMD FX(tm)-8350 Eight-Core Processor 4.00GHZ 16 GB.) a single sequence of normal records and obtained:

$$\begin{aligned} X(10^3) &= 43.7085 \\ X(10^4) &= 140.4020 \\ X(10^5) &= 447.2026 \\ X(10^6) &= 1414.59097 \\ X(10^7) &= 4472.6570 \\ X(10^8) &= 14142.3753 \\ X(10^9) &= 44721.3003 \\ X(2 * 10^9) &= 63251.0830. \end{aligned}$$

This shows the power of our indirect Algorithm 2. We made another simulation experiment. Making use of numerical integration, we computed in the standard normal case the means of 110 normal records. Then we generated by Algorithm 2 one million times 110 first records and found the corresponding sample means.

EX(30)	=	7.3226,	$\bar{X}$ (30)	=	7.3234,
EX(50)	=	9.6483,	$\bar{X}$(50)	=	9.6491,
EX(70)	=	11.5214,	$\bar{X}$(70)	=	11.5219,
EX(90)	=	13.1335,	$\bar{X}$(90)	=	13.1337,
EX(110)	=	14.5705,	$\bar{X}$(110)	=	14.5708.

The visual comparison shows that the differences between the means and their statistical estimates are small. This again indicates that Algorithm 2 allows to generate "long" sequences of normal records efficiently.

5 Statistical Procedures Related to Records

In the two years following the publication of the first paper on records by Chandler [6], there appeared some papers in which records were used for testing some statistical hypotheses; see Foster and Stuart [7], Foster and Teichroew [8]. These papers were followed by other statistical works on records. They treated testing for randomness, for homoscedasticity, for trend against natural alternatives. In this section we discuss two examples of using records in statistics. In the first example we present a test for trend against natural alternatives and in the second one we consider a precedence test based on records.

Test for trend Let $S(n) = N_1(n) - N_2(n)$ be the difference between the number of upper and lower records in a sample $X_1, \ldots, X_n$. Let

$$X_k = Y_k + \delta k \quad (k = 1, \ldots, n),$$

where Y_k are independent and identically distributed random variables and δ is a nonstochastic constant. It is clear that if $\delta > 0$ then the number of upper records is stochastically larger than the number of lower records. If $\delta = 0$,

$$S(n) = \nu_1 + \ldots + \nu_n,$$

where $\nu_k = 1$ if X_k is an upper record, $\nu_k = -1$ if X_k is a lower record and $\nu_k = 0$, otherwise. We have

$$ES(n) = 0, \quad Var\ S(n) = \sum_{k=1}^{n} \frac{2}{k} \sim 2 \log n.$$

Observe that $\frac{S(n)}{\sqrt{2 \log n}}$ is asymptotically normal. Let the null hypothesis H_0 be $\delta = 0$ and the alternative hypothesis H_1 be $\delta \neq 0$. By the test construction, we reject H_0 if

$$\text{either} \quad S(n) > z_{\alpha/2} \sqrt{2 \log n}, \quad \text{or} \quad S(n) < -z_{\alpha/2} \sqrt{2 \log n},$$

where $\alpha = 1 - \Phi(z_\alpha)$.

Precedence test based on records Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be two samples of independent and identically distributed random variables with distributions F_X and F_Y, respectively. A commonly encountered problem in practice is the comparison of F_X and F_Y. It appears, for example, when we wish to test whether a new manufacturing process or a new medical treatment is better than the existing one. Thus we are interested in testing the null hypothesis

$$H_0 : F_X = F_Y$$

against

$$H_1 : F_X > F_Y$$

or

$$H_1' : F_X < F_Y.$$

A known procedure for testing H_0 in terms of order statistics is the Wilcoxon rank-sum test with the test statistic

$$W_{n_1,n_2} = \sum_{i=1}^{n_2} Rank(Y_i),$$

where $Rank(Y_i)$ is the rank of Y_i in the ordered sample consisting of $Y_1, \ldots, Y_{n_2}$,X_1,..., X_{n_1}. The null hypothesis H_0 is rejected in favor of H_1 if a large value of W_{n_1,n_2} is observed.

A similar precedence test can be proposed in terms of records; see Balakrishnan *et. al.* (2008). Let

$$R_i = \#\{j \in \{1, 2, \ldots\} : Y(i-1) < X(j) \leq Y(i)\},$$

where $Y(0) = -\infty$ and $X(i), Y(i)$ $i = 1, 2, \ldots$ The following theorem was proved in Balakrishnan *et. al.* (2008).

Theorem 4. *Under $H_0 : F_X = F_Y$, the random variables $R_1, R_2, \ldots$ are independent and identically distributed and*

$$P(R_i = k \mid H_0) = \left(\frac{1}{2}\right)^{k+1}, \quad i = 1, 2, \ldots, \ k = 0, 1, \ldots$$

Let $Rank(Y(i)),\ i = 1, 2, \ldots$ be the rank of $Y(i)$ in an ordered sequence consisting of X- and Y-records. For example, if we have $X(1) < X(2) < Y(1) < X(3) < Y(2) < X(4)\ldots$, then $Rank(Y(1)) = 3$ and $Rank(Y(2)) = 5$. Let us define the following test statistic

$$RW_{(r)} = \sum_{i=1}^{r} Rank(Y(i)).$$

Since $Rank(Y(1)) = RM_1 + 1$ and $Rank(Y(i)) - Rank(Y(i-1)) = RM_i + 1$, $i = 2, 3, \ldots$, Theorem 4.2 enables us to establish the null distribution of $RW_{(r)}$ as

$$\begin{aligned}
P(RW_{(r)} < s | H_0 : F_X = F_Y) = & \\
& \sum_{\mathcal{A}_{(r)}(s)} P(Rank(Y(1)) = i_1, \ldots, Rank(Y(r)) = i_r \mid H_0) \\
= & \sum_{\mathcal{A}_{(r)}(s)} P(Rank(Y(1)) = i_1, Rank(Y(2)) - Rank(Y(1)) = i_2 - i_1 - 1, \ldots, \\
& \qquad Rank(Y(r)) - Rank(Y(r-1)) = i_r - i_{r-1} - 1 \mid H_0) \\
= & \sum_{\mathcal{A}_{(r)}(s)} (1/2)^{i_r},
\end{aligned}$$

where $\mathcal{A}_{(r)}(s) = \{(i_1, i_2, \ldots, i_r) : 0 < i_1 < \ldots < i_r$ and $i_1 + i_2 + \ldots + i_r < s$. Large values of $RW_{(r)}$ lead to the rejection of H_0 in favor of H_1. Therefore, for a specified value of significance α, the critical region will be $\{s_W, s_W + 1, \cdots\}$,

where the critical value s_W (corresponding to an exact level $\hat{\alpha}$ closest to α but not exceeding α) is the largest integer s satisfying

$$P(RW_{(r)} \geq s | H_0 : F_X = F_Y) = 1 - \sum_{\mathcal{A}_{(r)}(s)} (1/2)^{i_r} = \hat{\alpha} \leq \alpha.$$

With this test we finish considering statistical tests based on records and complete our discussion.

References

[1] M. Ahsanullah, V.B. Nevzorov: *Records via Probability Theory*. Springer (2015).

[2] B.C. Arnold, N. Balakrishnan, H.N. Nagaraja: *Records*. John Wiley & Sons, New York (1998).

[3] I. Bairamov, A. Stepanov: Numbers of near bivariate record-concomitant observations. *J. Multivariate Anal.* 102 (5) (2011) 908–917.

[4] N. Balakrishnan, H.Y. So, X.J. Zhu: On Box–Muller transformation and simulation of normal record data. *Comm. Statist. Simulation Comput.* 45 (10) (2016) 3670–3682.

[5] N. Balakrishnan, A.G. Pakes, A. Stepanov: On the number and sum of near-record observations. *Adv. in Appl. Probab.* 37 (3) (2005) 765–780.

[6] K.N. Chandler: The distribution and frequency of record values. *J. Roy. Statist. Soc. Ser. B* 14 (2) (1952) 220–228.

[7] F.G. Foster, A. Stuart: Distribution-free tests in time-series based on the breaking of records. *J. Roy. Statist. Soc. Ser. B* 16 (1) (1954) 1–22.

[8] F.G. Foster, D. Teichroew: A sampling experiment on the powers of the records tests for trend in a time series. *J. Roy. Statist. Soc. Ser. B* 17 (1) (1955) 115–121.

[9] E. Hashorva: On the number of near-maximum insurance claim under dependence. *Insurance Math. Econom.* 32 (1) (2003) 37–49.

[10] E. Hashorva, J. Hüsler: The neighbourhood of the bivariate maxima: with application to insurance. *In: IV International Conference in Stochastic Geometry, Convex Bodies, Empirical Measures & Applications to Engineering Science, Rend. Circ. Mat. Palermo (2) Suppl. part I* 70 (2002) 361–376.

[11] E. Hashorva, J. Hüsler: Estimation of tails and related quantities using the number of near-extremes. *Comm. Statist. Theory Methods* 34 (2) (2005) 337–349.

[12] V.B. Nevzorov: *Records: Mathematical Theory*. American Mathematical Society, Providence (2001).

[13] V.B. Nevzorov, A. Stepanov: Records with confirmation. *Statist. Probab. Lett.* 95 (2014) 39–47.

[14] A. Pakhteev, A. Stepanov: Simulation of gamma records. *Statist. Probab. Lett.* 119 (2016) 204–212.

[15] A.I. Pakhteev, A.V. Stepanov: Generation of large sequences of normal record values and maxima. *Vestnik St. Petersburg Univ. Math.* 51 (3) (2018) 260–266.

[16] A. Pakhteev, A. Stepanov: Discrete records: Limit theorems for their spacings and generation methods. *Statist. Probab. Lett.* 148 (2019) 134–142.

[17] A. Pakhteev, A. Stepanov: On simulation of normal records. *Comm. Statist. Simulation Comput.* 48 (8) (2019) 2413–2424.

[18] A. Rényi: On the extreme elements of observations. *In: Selected papers of Alfred Rényi* 3 (1976) 50–65. Translation of On the outliers of a series of observations (Hungarian), Magyar Tud. Akad. Mat. Fiz. Oszt. Kozl. 12 (1962) 105–121.

[19] R.W. Shorrock: On record values and record times. *J. Appl. Probability* 9 (2) (1972) 316–326.

[20] A. Stepanov: On simulation of weak records. *Comm. Statist. Simulation Comput.* 48 (3) (2019) 797–806.

[21] A. Stepanov, A. Berred, V.B. Nevzorov: Concomitants of records: limit results, generation techniques, correlation. *Statist. Probab. Lett.* 109 (2016) 184–188.

[22] M.N. Tata: On outstanding values in a sequence of random variables. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 12 (1969) 9–20.

[23] W. Vervaat: Limit theorems for records from discrete distributions. *Stochastic Process. Appl.* 1 (4) (1973) 317–334.

Received: 21 November 2019
Accepted for publication: 21 December 2019
Communicated by: Pasha Zusmanovich