Mathematical Theory of Records

Alexei Stepanov[1]

[1]Immanuel Kant Baltic Federal University, Kaliningrad, Russia

Ostrava 2019

Outline

Records are commonly used in different areas such as sport, finance, reliability, hydrology and others.

The first paper of Chandler (1952) attracted the attention of many researchers and inspired many new publications.

The mathematical theory of records is amply discussed in the books of Arnold *et al.* (1998), Nevzorov (2001) and Ahsanullah and Nevzorov (2015); see also the references therein.

Some examples after definitions.

Let $X_1, X_2, \ldots$ be a sequence of random variables (rv's). The sequences of (upper) record times $L(n)$ $(n \geq 1)$ and record values $X(n)$ $(n \geq 1)$ are defined as follows:

$$L(1) = 1, \quad X(1) = X_1,$$

$$L(n) = \min\{j, j > L(n-1); X_j > X_{L(n-1)}\} \quad (n = 2, 3, \ldots), \tag{1.1}$$

$$X(n) = X_{L(n)} \quad (n = 1, 2, \ldots).$$

Let also $M_n = \max\{X_1, \ldots, X_n\}$.

If in (1.1) we replace the second sign $>$ with sign $<$, then we obtain the sequences of lower record times $l(n)$ and values $x(n)$.

Let $X_1, X_2, \ldots$ be a sequence of iid rv's with continuous F. Then $-X_1, -X_2, \ldots$ is a sequence of iid rv's with continuous $G(x) = 1 - F(-x)$. If some $X_{i_1} < X_{i_2} < \ldots$ $(i_1 < i_2 < \ldots)$ are upper records in the sequence $X_1, X_2, \ldots$ with F, then $-X_{i_1} > -X_{i_2} > \ldots$ are lower records in $-X_1, -X_2, \ldots$ with G. So results for $x(n)$ follow from results for $X(n)$.

Examples:

(1) Construction of dams $M_n, X(k) \leq h$ for large n, k.

(2) Insurance. Near maximum, near record observations: $X_i \in [M_n - a, M_n],\ X_i \in [X(n) - a, X(n)]$. Their sum $S = \sum X_i$. Sums of 10 percent of large claims can cause 90 percent of insurance payments.

(3) Low records, mimima in survival analysis.

(4) Sport records.

Let $X_1, X_2, \ldots$ be iid rv's with continuous F. Let us introduce record indicators ξ_n ($n \geq 1$):

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

Lemma 2.1 (Rényi 1976) *The variables* $\xi_1, \xi_2, \ldots$ *are independent and*

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

The distribution of L(2) can be found as

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

The sequence $L(n)$ $(n \geq 1)$ forms a Markov chain and

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

It follows that

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

One can show that

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

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

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

When $n = 2$ we have

$$P(L(2) = k) = \frac{1}{k(k-1)} \ (k \geq 2).$$

Relations between $L(n)$ and ξ_n.

$$P(L(n) > m) = P(\xi_1 + \xi_2 + \ldots + \xi_m < n)$$

Let us denote $N(n) = \xi_1 + \xi_2 + \ldots + \xi_n$. Then $N(n)$ is the number of records in $X_1, X_2, \ldots, X_n$. We have

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

At average, in a sample $X_1, X_2, \ldots, X_{100}$, we have

$$\log 100 \approx 4.6$$

and in a sample $X_1, X_2, \ldots, X_{1000}$, we have

$$\log 1000 \approx 6.9$$

record values.

Let $X_1, X_2, \ldots$ be iid rv's with $F(x) = 1 - e^{-x}$ $(x > 0)$.

Theorem 2.1 (Tata 1969) *The variables*

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

are iid with $F(x) = 1 - e^{-x}$ $(x > 0)$.
So $X(n) \stackrel{d}{=} Y_1 + \ldots + Y_n \sim Gamma(n)$ and

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

Let $X_1, X_2, \ldots$ be iid rv's with arbitrary continuous $F(x)$. Then

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

are iid rv's with $1 - e^{-x}$. 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

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

Let $X_1, X_2, \ldots$ be iid rv's with absolutely continuous $F(x)$ and $f(x)$. Then

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

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

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

Let $X_1, X_2, \ldots$ be iid rv's with arbitrary continuous $F(x)$. We know that

$$P(L(n) > m) = P(N(m) < n),$$

where $N(n) = \xi_1 + \xi_2 + \ldots + \xi_n$ and $P(\xi_n = 1) = 1/n$. Then

$$\frac{N(n)}{\log n} \xrightarrow{p} 1 \quad \text{and} \quad \frac{N(n) - \log n}{\sqrt{\log n}} \xrightarrow{d} Z,$$

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

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

Let $X_1, X_2, \ldots$ be iid rv's with $F(x) = 1 - e^{-x}$ $(x > 0)$. By Tata's representation,

$$X(n) \stackrel{d}{=} Y_1 + \ldots + Y_n.$$

where Y_i are iid standard exponential rv's. Let now $X_1, X_2, \ldots$ be iid rv's with arbitrary continuous F. Then

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

Assume that $X, X_1, X_2, \ldots$ are iid rv's with support on non-negative integers and $F(n) = P(X \leq n) < 1$ for all $n \geq 0$. Let $p_n = P(X = n)$ and $q_n = P(X \geq n)$ be the distribution tail.

The joint pmf of the first n discrete record values is

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

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

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

$$\begin{aligned} P(X(n+m) = k_{n+m} \quad | \quad X(n) = k_n) \\ = \quad \frac{p_{k_{n+m}}}{q_{k_n+1}} \sum_{l_1=k_n+1}^{k_{n+m}-m+1} \frac{p_{l_1}}{q_{l_1+1}} \cdots \sum_{l_{m-1}=l_{m-2}+1}^{k_{n+m}-1} \frac{p_{l_{m-1}}}{q_{l_{m-1}+1}}, \end{aligned}$$

where $n-1 \leq k_n \leq k_{n+m} - m,\ m \geq 1,\ \prod_{i=n+1}^{n} \frac{p_{k_i}}{q_{k_i+1}} = 1$ and the sum

$$\sum_{l_1=k_n+1}^{k_{n+m}-m+1} \frac{p_{l_1}}{q_{l_1+1}} \cdots \sum_{l_{m-1}=l_{m-2}+1}^{k_{n+m}-1} \frac{p_{l_{m-1}}}{q_{l_{m-1}+1}}$$

is equal to 1 when $m = 1$ and to $\sum_{l_1=k_n+1}^{k_{n+2}-1} \frac{p_{l_1}}{q_{l_1+1}}$ when $m = 2$.

Define random indicators ξ_i $(= 0, 1;\ i = 0, 1, \ldots)$: $\xi_i = 1$ if there is a record value $X(n)$ such that $X(n) = i$.

Lemma 4.1 (Shorrock 1972) *The rv's ξ_i $(i = 0, 1, \ldots)$ are independent and*

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

Representation 4.1 *Under the conditions of Lemma 4.1,*

$$P(X(n) > m) = P(\xi_0 + \xi_1 + \cdots + \xi_m < n) \quad (n \geq 1).$$

Theorem 4.1 *Assume that $X_1, X_2, \ldots$ are iid rv's with support on non-negative integers and $F(n) = P(X \leq n) < 1$ for all $n \geq 0$. Then*

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

Theorem 4.2 *Assume that $X_1, X_2, \ldots$ are iid rv's with support on non-negative integers and $F(n) = P(X \leq n) < 1$ for all $n \geq 0$ and $\lim_{n\to\infty} \frac{p_n}{q_n} = a < 1$. Then*

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

The concept of weak records is introduced in Vervaat (1973). Weak records were discussed in Stepanov (1992, 1993) and others.

Let $X_1, X_2, \ldots$ be a sequence of iid rv's with support $\{1, 2, \ldots, N\}$, $N \leq \infty$. The sequences of weak record times $L^w(n)$ and weak record values $X^w(n)$ are defined as follows:

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

$$X^w(n) = X_{L^w(n)}, \quad n \geq 1.$$

The joint probability mass function

$$P(X^w(1) = k_1, \ldots, X^w(n) = k_n) = p_{k_n} \prod_{i=1}^{n-1} \frac{p_{k_i}}{q_{k_i}},$$

for any $1 \leq k_1 \leq \ldots \leq k_n \leq N$ (if $N = \infty$ the last inequality is, obviously, sharp). Define weak record indicators ξ_i^w, $i = 1, 2, \ldots$ as follows

$\xi_i \quad = \quad k$

if there are exactly k weak record values that are equal to i.

Lemma 4.3 (Stepanov 1992) *The rv's ξ_i^w, $i = 1, 2, \ldots$ are independent and*

$$P(\xi_i = k) = \frac{q_{i+1}}{q_i}\left(1 - \frac{q_{i+1}}{q_i}\right)^k, \qquad k = 0, 1, \ldots, i = 1, 2, \ldots N-1,$$

where $P(\xi_N = \infty) = 1$ *if* $N < \infty$ *and* $P(\xi_{N+j} = 0) = 1, \quad j \geq 1$.

Representation 4.2 *Under the conditions of Lemma 4.3,*

$$P(X^w(n) > m) = P(\xi_0^w + \xi_1^w + \ldots + \xi_m^w < n) \quad (n \geq 1).$$

Let $X_1, X_2, \ldots$ be a sequence of iid rv's with $F(n) < 1$ for any $n \geq 0$.

Theorem 4.3 *Let*

$$a = \sup_{n \geq 0} \beta_n < 1.$$

Then

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{X^w(n)} \frac{p_i}{q_{i+1}}}{n} \overset{a.s.}{=} 1.$$

Theorem 4.4 *Let* $a = \sup_{n\geq 0} \beta_n < 1$ *and* $\frac{\sum_{i=0}^{n} \frac{p_i}{q_{i+1}}}{\sum_{i=0}^{n} \frac{p_i q_i}{q_i^2}} \to \varepsilon \in [1-a, a]$. *Then*

$$\sqrt{\varepsilon}\frac{\sum_{i=0}^{X^w(n)} \frac{p_i}{q_{i+1}} - n}{\sqrt{n}} \xrightarrow{d} Z.$$

If $F^{-1}(x)$ can be found analytically,
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, we 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 – the rejection method.

Rejection Method Suppose we can generate a rv $\tilde{X}$ having density function q 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, we should find a constant $c > 1$ such that $c = \sup_x \frac{h(x)}{q(x)}$.

Algorithm

Step 1: Generate $\tilde{X} = \tilde{x}$ (with density function q) and a random number $U = u$;

Step 2: If $u \leq \frac{h(x)}{cq(x)}$, set $X = \tilde{x}$. Otherwise, return 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 rv with mean c.

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 obtained. One then generates variables X_i till one of them, say X_j, is greater than $X(n)$. Then $X(n+1) = X_j$ becomes the next record value, which is also kept.*

Sequences of records form Markov chains and

$$P(X(n+1) \leq x_{n+1} \mid X(n) = x_n) = \frac{F(x_{n+1}) - F(x_n)}{1 - F(x_n)} \quad (x_{n+1} > x_n).$$

Let Z_i $(i \geq 1)$ be iid with standard normal distribution Φ and $Z(n)$ $(n \geq 1)$ be the corresponding records. The conditional density of $Z(n+1)$ given $Z(n) = Z_n$

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

Let $\beta_n^* = \frac{z_n + \sqrt{z_n^2 + 4}}{2}$.

Algorithm (Pakhteev, Stepanov 2019) *The sequence $Z(n)$ $(n \geq 1)$ can be generated as follows.*
STEP 1: Generate $Z(1) = Z_1, Z(2), \ldots, Z(i)$ $(i \geq 1)$ by the direct algorithm of record generation till $Z(i) > 0$.
For $n \geq i$, apply the rejection method and the recursive approach. Assume that $Z(n) = z_n$ is already obtained.
STEP 2: Generate random numbers $U_1 = u_1, U_2 = u_2$.
STEP 3: If

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

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

We have to generate negative normal records by the direct algorithm. We compare $f_{Z(n+1)|Z(n)}(z_{n+1} \mid z_n)$ with $g(z_{n+1} \mid z_n, \beta_n) = \beta_n e^{-\beta_n(z_{n+1}-z_n)} \quad (z_{n+1} > z_n)$, 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 rv such that $Z_1 < 0, \ldots, Z_{\tau-1} < 0$ and $Z_\tau > 0$. Observe that τ is a geometric rv and $E\tau = 2$. In a simulation experiment the number of first negative normal records is insufficient.

In Algorithm 4.1 $c^*(z_n) = \sup_{z_{n+1} > z_n} \frac{f_{Z(n+1)|Z(n)}(z_{n+1}|z_n)}{g(z_{n+1}|z_n, \beta_n)}$. One can prove that $c^*(z_n) \to 1$ as $z_n \to \infty$.

It is known that $Z(n) \overset{\text{a.s.}}{\to} \infty$. Algorithm 4.1, which is based on the rejection method, eventually works as an algorithm based on the inverse-transform method. With time almost every generation in a generation experiment is accepted and becomes a new record.

If one generates directly standard normal rv one cannot obtain (with nowadays 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:

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

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 4.1 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.

Using records for testing some statistical hypotheses: tests for randomness, for homoscedasticity, for trend against natural alternatives.
Foster and Stuart (1954), Foster and Teichroew (1954) and others.
Test for trend Let

$$S(n) = N_1(n) - N_2(n),$$

be the difference between the number of upper and lower records in the sample $X_1, \dots, X_n$. Let

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

where Y_k are iid rv's and δ is a nonstochastic constant.

If $\delta > 0$, then the number of upper records is stochastically larger and 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

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

and $\frac{S(n)}{\sqrt{2 \log n}}$ is asymptotically normal.

$H_0 : \ \delta = 0$ against $H_1 : \ \delta \neq 0$.

reject if

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

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

Let $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ two iid samples with F_X and F_Y, respectively. A problem of 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 interesting 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 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 $Y_1, \cdots, Y_{n_2}, X_1, \cdots, X_{n_1}$. The null hypothesis H_0 is rejected in favor of H_1 if a large value of W_{n_1,n_2} is observed.

Let

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

where $Y(0) = -\infty$ and $X(i), Y(i)$ $i = 1, 2, \ldots$

Theorem 6.1 Shorrock (1972) *Let* $\mu^X_{(a,b]} = \#\{j \in \{1, 2, \ldots\} : X(j) \in (a, b]\}$*. Then random variables* μ*, taken for different non-overlapping intervals, are independent and*

$$P(\mu(x, y) = i) = \frac{e^{-\lambda_{x,y}} \lambda_{x,y}^i}{i!} \qquad (i \geq 0),$$

where $\lambda_{x,y} = -\log\left(\frac{1-F(y)}{1-F(x)}\right)$.

Theorem 4.2 Balakrishnan, Dembinska, Stepanov, (2008)
Under $H_0 : F_X = F_Y$, the rv's $R_1, R_2, \ldots$ are iid 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 $Rank(Y(2)) = 5$.

Then

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

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

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

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

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

Let $Z = (X, Y), Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \dots$ be iid random vectors with a continuous $F(x, y) = P(X \leq x, Y \leq y)$, survival function $\bar{F}(x, y) = P(X > x, Y > y)$, marginal distributions $H(x) = P(X \leq x),\ G(y) = P(Y \leq y)$, marginal survival functions $\bar{H}(x) = P(X > x),\ \bar{G}(y) = P(Y > y)$ and densities $f(x, y), h(x)$ and $g(y)$.

There are many definitions of bivariate records; on page 266 of Arnold *et al.* (1998), four different definitions of bivariate records have been introduced.

The third definition of bivariate records states "A new bivariate record occurs at time i if X_i exceeds the current X record and Y_i exceeds the current Y record." We call such records north-east (ne) bivariate records. Let us first set $L^{ne}(1) = 1$ and

$$Z_1^{ne} = (X^{ne}(1), Y^{ne}(1)) = (X(1), Y(1)) = (X_1, Y_1).$$

Next, we set

$$L^{ne}(n+1) = \min\left\{j > L^{ne}(n) : \ X_j > X_{L^{ne}(n)} \ \& \ Y_j > Y_{L^{ne}(n)}\right\} \quad (n \geq 1),$$

$$Z_n^{ne} = (X^{ne}(n), Y^{ne}(n)) = (X_{L^{ne}(n)}, Y_{L^{ne}(n)}) \quad (n \geq 2).$$

Let us consider $S = \{Z_1, \ldots, Z_{L^{ne}(n)}\}$. There is no S observation located in the quarter-plane

$$QP = (X^{ne}(n), \infty) \times (Y^{ne}(n), \infty).$$

If we consider now the sample $T = \{Z_{L^{ne}(n)+1}, Z_{L^{ne}(n)+2}, \ldots\}$, then the first T observation, that falls in QP, becomes the next north-east bivariate record $Z^{ne}_{n+1} = (X^{ne}(n+1), Y^{ne}(n+1))$.

The probability mass function of $L^{ne}(2)$

$$P(L^{ne}(2) = k) = \int_{\mathbb{R}^2} (1 - \bar{F}(x,y))^{k-2} \bar{F}(x,y)\, dF(x,y) \quad (k \geq 2).$$

It follows that

$$P(L^{ne}(2) \geq k) \geq \int_{\mathbb{R}^2} (1 - \bar{H}(x))^{k-2}\, dF(x,y) = \frac{1}{k-1} \quad (k \geq 2)$$

It is easily seen that

$$E(L^{ne}(2) - 1) = \int_{\mathbb{R}^2} \frac{dF(x,y)}{\bar{F}(x,y)} \geq \int_{\mathbb{R}^2} \frac{dF(x,y)}{\bar{H}(x)} = \int_{\mathbb{R}} \frac{dH(x)}{\bar{H}(x)} = \infty..$$

For univariate records that if $H(x)$ is continuous, then $P(L(2,x) \geq k) = \frac{1}{k-1}$ $(k \geq 2)$ and $EL(2,x) = \infty$.

The joint density of $Z_1^{ne}, \ldots, Z_{n-1}^{ne}, Z_n^{ne}$ is

$$f_{Z_1^{ne},\ldots,Z_{n-1}^{ne},Z_n^{ne}}(x_1, y_1, \quad \ldots \quad, x_{n-1}, y_{n-1}, x_n, y_n) = \frac{f(x_1, y_1)}{\bar{F}(x_1, y_1)} \cdots \frac{f(x_{n-1}, y_{n-1})}{\bar{F}(x_{n-1}, y_{n-1})} f(x_n, y_n).$$

Sequences of north-east bivariate record times and record vectors form Markov chains, and

$$f_{Z_{n+1}^{ne}|Z_n^{ne}}(x_{n+1}, y_{n+1} \mid x_n, y_n) = \frac{f(x_{n+1}, y_{n+1})}{\bar{F}(x_n, y_n)} \quad (x_{n+1} > x_n,\ y_{n+1} > y_n),$$

$$P(L^{ne}(n+1) = k \quad \mid L^{ne}(n) = i) = \int_{\mathbb{R}^2} \quad (1 - \bar{F}(x, y))^{k-i-1} \bar{F}(x, y) \, dF_{Z_n^{ne}}(x, y).$$

Example

Consider

$$F(x,y) = 1 - e^{-x} - \frac{1}{y+1} + \frac{e^{-x(y+1)}}{y+1} \quad (x, y \geq 0)$$

with the survival function $\bar{F}(x,y) = \frac{e^{-x(y+1)}}{y+1}$, the marginal distributions

$$H(x) = 1 - e^{-x} \ (x > 0), \quad G(y) = 1 - \frac{1}{y+1} \ (y > 0),$$

and the conditional distributions $Y \mid X = x$ $G_{Y|X}(y \mid x) = 1 - e^{-xy} \ (x, y > 0)$ and

$$H_{X^{ne}(n+1)|Z_n^{ne}}(x_{n+1} \mid x_n, y_n) = 1 - e^{-(y_n+1)(x_{n+1}-x_n)}.$$

We can generate the consecutive north-east X record values by the inverse-transform method as

$$X^{ne}(n+1) = X^{ne}(n) + \frac{-\log u_n}{Y^{ne}(n)+1},$$

where $U_n = u_n$ is the generation of a random number. Then

$$G_{Y^{ne}(n+1)|Z_n^{ne}}(y_{n+1} \mid x_n, y_n) = 1 - \frac{y_n+1}{y_{n+1}+1} e^{-x_n(y_{n+1}-y_n)} \quad (y_{n+1} > y_n).$$

From the form of $G_{Y^{ne}(n+1)|Z_n^{ne}}(y_{n+1} \mid x_n, y_n)$, we observe that the inverse-transform method is not useful here, and so we apply the rejection method. Let us take $q(y_{n+1} \mid x_n, y_n) = x_n e^{-x_n(y_{n+1}-y_n)}$ $(x_n \in R, y_{n+1} > y_n)$ as a dominated density function for

$$\begin{aligned} & g_{Y^{ne}(n+1)|Z_n^{ne}}(y_{n+1} \mid x_n, y_n) = G'_{Y^{ne}(n+1)|Z_n^{ne}}(y_{n+1} \mid x_n, y_n) \\ = \quad & \frac{(1+x_n(y_{n+1}+1))(y_n+1)}{(y_{n+1}+1)^2} e^{-x_n(y_{n+1}-y_n)}. \end{aligned}$$

We then find that

$$1 < c(x_n, y_n) = \sup_{y_{n+1} > y_n} \frac{g_{Y^{ne}(n+1)|Z_n^{ne}}(y_{n+1} \mid x_n, y_n)}{q(y_{n+1} \mid x_n, y_n)} = \frac{(1 + x_n(y_n + 1))}{(y_n + 1)x_n}.$$

Observe that the choice of the dominated density function q is a good one here since $c(x_n, y_n) = 1 + \frac{1}{x_n(1+y_n)} \to 1$ as $x_n \to \infty$ or $y_n \to \infty$. In the algorithm, we should compare

$$\frac{g_{Y^{ne}(n+1)|Z_n^{ne}}(y_{n+1} \mid x_n, y_n)}{c(x_n, y_n) q(y_{n+1} \mid x_n, y_n)}$$

with random number U, i.e., using previously obtained $X^{ne}(n) = x_n$ and $Y^{ne}(n) = y_n$, we should compare

$$\left(\frac{y_n + 1}{y_{n+1} + 1}\right)^2 \frac{1 + x_n(y_{n+1} + 1)}{1 + x_n(y_n + 1)}$$

with $U = u$.

Algorithm 7.1 *Step 1: First, generate* $(X^{ne}(1), Y^{ne}(1)) = (x_1, y_1)$*. For this purpose, generate random numbers* $U_1 = u_1, V_1 = v_1$*, and set*

$$x_1 = -\log u_1, \quad y_1 = \frac{-\log v_1}{x_1}.$$

For $n \geq 2$*, apply the following recursive approach. Assume that* $(X^{ne}(n), Y^{ne}(n)) = (x_n, y_n)$ *is already obtained.*

. *Step 2: 2.1: Generate random number* $U_n = u_n$. *Set*

$$X^{ne}(n+1) = x_{n+1} = x_n + \frac{-\log u_n}{y_n + 1};$$

2.2: Generate random numbers $V_n = v_n, T_n = t_n$;
2.3: Set

$$\tilde{y}_n = y_n + \frac{-\log v_n}{x_n}.$$

If

$$t_n < \left(\frac{y_n + 1}{\tilde{y}_n + 1}\right)^2 \frac{1 + x_n(\tilde{y}_n + 1)}{1 + x_n(y_n + 1)},$$

set $Y^{ne}(n+1) = y_{n+1} = \tilde{y}_n$. *Otherwise, return to 2.2.*

By using Algorithm 7.1, we generated 10000 times the first ten X and Y north-east record values. Then, for every n, we found mean values of $\bar{X}^{ne}(n)$, $\bar{Y}^{ne}(n)$ $(n = 1, \ldots, 10)$.

Table 7.1

$n =$	1	2	3	4	5	6
$\bar{X}^{ne}(n)$	1.0051	1.4946	1.7680	1.9755	2.1511	2.2962
$\bar{Y}^{ne}(n)$	8.0807	27.2174	36.0187	43.5699	49.6565	56.0334

Ahsanullah, M., Nevzorov V. B. (2015). *Record via Probability Theory*. Atlantis Press.

Arnold B., Balakrishnan N., Nagaraja H. (1992). *A first course in order statistics*. Wiley, New York.

Balakrishnan, N., Dembinska, A. and Stepanov, A. (2008). Precedence-type tests based on record values, *Metrika*, **68**, 233–255.

Chandler K.N., 1952. The distribution and frequency of record values. J. Royal Statist. Soc. Ser. B. 14, 220–228.

Foster, F.G. and Stuart, A. (1954). Distribution free tests in time-series band on the breaking records, *J. Royal Statist. Soc.*, Ser. B, **16** (1), 1–22.

Foster, F.G. and Teichroew, D. (1955). A sampling experiment on the powers of the record tests for trend in a time series, *J. Royal Statist. Soc.*, Ser. B, **17** 115–121.

Nevzorov, V. B. (2001). *Records: Mathematical Theory*, Translation of Mathematical Monographs, Vol. **194**, American Mathematical Society, Providence, Rhode Island.

Pakhteev, A. and Stepanov, A. (2019). On simulation of normal records, *Communication in Statistics – Simulation and Computation*, **48** (8), 2413–2424.

Rényi, A. (1962). On the extreme elements of observations, In *Selected papers of Alfred Renyi*, Budapest: Akademiai Kiado, **3**, 50–65, 1976. (Translation of the paper – Edy megfigyelessorozat kiemelkedo elemeirol, *Mag. Tud. Acad. 3 Oszt. Kozl.*, 1962, **12**, 105–121.)

Shorrock, R.W. (1972). A limit theorem for inter-record times, *J. Appl. Probab.*, **9** (1), 219–223.

Stepanov, A. V. (1992). Limit theorems for weak records, *Theory of Probability and Its Applications*, **37**, 570-574 (English translation).

Stepanov, A. V. (1993). A characterization theorem for weak records, *Theory of Probability and Its Applications*, **38**, 762-764 (English translation).

Tata, M.N. (1969). On outstanding values in a sequence of random variables, *Z. Wahrscheinlichkeitheor. verw. Geb.*, **12** (1), 9–20.

Vervaat, W. (1973). Limit theorems for records from discrete distributions, *Stochastic Processes and their Applications* **1**, 317-334.