INSTITUTE OF MATHEMATICS
OF THE POLISH ACADEMY OF SCIENCES

Sniadeckich 8, 00-956 Warsaw, P.O. Box 21, Poland, www.impan.pl, e-mail: im@impan.pl
tel. +48 22522 81 00, fax: +48 22629 39 97

Prof. dr habil. Tomasz Rychlik
Department of Mathematical Statistics

Report
of the member of the Dissertational Council on the thesis
Simulation of records and extreme values
by Artem Igorevich Pakhteev
submitted for the granting procedure
of the Candidate degree in the Mathematical and Physical Sciences
in speciality 01.01.05: Probability Theory and Mathematical Statistics

The dissertation is devoted to analysis of extreme random phenomena: record values and times, and maximal values in the sequences of independent identically distributed random variables. The main topic of the thesis is an effective simulation of maximal and record values, and record times. The problem of asymptotic behavior of differences of record values in discrete populations is also studied there.

The main body of the thesis has 80 pages (I focus on the English version only), and it is completed by 18 pages long appendix with program codes of the algorithms presented in the thesis. The reference list contains 32 papers and books cited in the manuscript, and 4 papers of the author that the thesis is based on. It is worth pointing out that these papers were published in good international mathematical journals: two in *Statistics and Probability Letters*, one in *Communications in Statistics*, and one in *Vestnik of St. Petersburg State University*. This means that the value and correctness of the results presented in the dissertation have been already positively verified by international experts refereeing the papers. It should be noted that all these papers were published jointly with the supervisor Professor Alexei Stepanov.

The main achievement of the thesis was construction of effective algorithms of generating records from distributions whose density functions f have analytic forms, but the distribution functions F do not have (in particular, the gamma and normal distributions). Two basic tools were used here. One was coincidence of the conditional distribution of the next record given the previous one with the left-truncated parent distribution at the value of the preceding record. The other one was the rejection method of simulating random variables. The idea consists in generating (e.g., by the classic inverse-transform method) another random variable Y with a density g which with a multiplicative constant $c > 1$ majorizes f. The value of

09/2-02-483 от 18.09.2020

Y is accepted if $\frac{f(Y)}{cg(Y)} > U$ for an independent standard uniform random variable U. Since the procedure gives the acceptance probability $\frac{1}{c}$, it is useful to find the smallest possible c.

My first objection to the author is that he did not properly describe the rejection method. He made much effort to determine minimal $c = \sup \frac{f(x)}{g(x)}$, but he did not explain the purpose of these efforts. It looks as if $c = \sup \frac{f(x)}{g(x)}$ is the only c for which the method works. For the gamma and Gaussian models studied in the thesis $g(x)$ was chosen as the two-parameter exponential distribution with varying parameters at each step of simulating the consecutive record values. The location parameter was selected naturally as the value of the previous record. The scale was chosen so to minimize the constant c. I have checked the calculations, and I can confirm that the scale parameters for generating the gamma and normal distributions were chosen optimally.

The algorithms presented in the paper are efficient and useful by two reasons. One is that their numerical calculations are very fast. In particular, for large n optimal constants tend to 1 (which was formally proven) and so the probability of accepting record in each step tends to 1. Secondly, the algorithms allow us to generate very many values of records with is not available by other methods. This was confirmed by numerical comparisons with various simulation methods presented in the literature. What is lacking for me is the precise description of the way of generating the initial observation (and record) in the gamma model with the shape parameter $\alpha < 1$. The respective density function is unbounded in the neighborhood of 0, and I wonder which density function was used for the rejection procedure then.

I appreciate a clever method of simulating many values of maxima in normal populations. It consists in generating record values and record times, and taking the preceding record values in the inter-record times. It should be noted that the method works for general continuous and unnecessarily normal samples.

Both the inverse-transform and rejection methods were adapted to discrete populations in Chapter 5, and several interesting examples were presented there. I may only regret that no example of discrete record sequence simulation with use of rejection method was presented.

Finally, I discuss the limit theorems for the spacings (adjacent as well non-adjacent ones) of records in discrete populations. The candidate noticed that the limit behavior of them depends on the parameter

$$\beta = \lim_{n\to\infty} \frac{q_{n+1}}{q_n} = \lim_{n\to\infty} \frac{\mathbb{P}(X \geq n+1)}{\mathbb{P}(X \geq n)}.$$

In the cases $\beta = 0$, $\beta = 1$, and $0 < \beta < 1$, the differences of records $X(n+m) - X(n)$, $n \to \infty$, tend in distribution to the minimal and maximal constants m and $+\infty$, and to the negative binomial distribution with parameters m and $1-\beta$, respectively. Strong limits for the differences of adjacent records were presented as well.

Conclusion. The results presented in the Ph. D. thesis of A. I. Pakhteev constitute new, interesting and valuable contributions to the analysis of ordered random variables. They are practically applicable as well. The thesis written by Artem Igorevich Pakhteev on the subject *Simulation of records and extreme values* fulfils essential requirements established in the decree *On the procedure of granting academic degrees in the St. Petersburg State University*. The applicant Artem Igorevich Pakhteev deserves granting with the Candidate degree in the Mathematical and Physical Sciences in speciality 01.01.05: Probability Theory and Mathematical Statistics. Point 11 of the decree mentioned above has not been violated.

Member of the Dissertational Council

Prof. dr habil. Tomasz Rychlik
Department of Mathematical Statistics
Institute of Mathematics
Polish Academy of Sciences

Tomasz Rychlik

Warsaw, August 24, 2020

Below I list a number of minor mistakes I have found while reading the manuscript. They can be easily corrected, and by no means they do not diminish the merits of research achievements of Mr. Pakhteev.

p. 111, l. 7 and l. 5 bottom: notation q_n and p_n were not introduced yet.

p. 114, l. 3: the formula holds for *all* parent distribution functions F, discrete as well.

p. 115, l. 9: continuity and identical distribution of $X_1, \ldots, X_n$ should be explicitly assumed.

p. 116, l. 5 bottom: the formula is valid for $x_1 < \ldots < x_n$ only. Moreover, absolute continuity of F should be assumed here.

p. 118, Representation 2.2.1: In fact, Lemma 2.2.1 does not have assumptions.

p. 126, l. 7: $\left(x_{n+1}^{\alpha-1} e^{-x_{n+1}\left(1-\frac{1}{\mu}\right)}\right)' = x_{n+1}^{\alpha-2} e^{-x_{n+1}\left(1-\frac{1}{\mu}\right)} \left[\alpha - 1 - \left(1 - \frac{1}{\mu}\right) x_{n+1}\right] < 0.$

p. 126, l. 1 bottom: $c^*(x_n) \leq 1 + \frac{1-\alpha}{x_n} - \varepsilon$, because $1 + \frac{1-\alpha}{x_n}$ is the limit.

p. 133, l. 3: The last factor is $\left(-\log\left(\frac{1-F(x_n|\alpha)}{1-F(x_m|\alpha)}\right)\right)^{n-m-1}$. Moreover, $x_m < x_n$ should be assumed here.

p. 137, lines 6–1 bottom and p. 138, l. 1: the numbers look suspicious: $X(2 \cdot 10^i) \approx 2 \cdot 10^i$.

p. 139, lines 6–8: for $\alpha = 0.05$ and $k = 7$, the null hypothesis should be rejected as well, because $30.236 > 30.144$.

p. 143, Algorithm 3.2.2: instead of repeating several operations at each step, it suffices to perform Algorithm 3.2.1 and take $\sigma X(n) + a$.

p. 159, l. 3: the exponent should have the from $\ell_{n+1} - \ell_n - 1$.

p. 163, lines 6–5 bottom: the argument is based on the Borel-Cantelli lemma which appears later in p. 166.

p. 167, lines 5–3 bottom: replace ξ_0 by ξ_1.

p. 168, l. 2 bottom: $\sum_{n=1}^{\infty} \sum_{i=n-1}^{\infty} \frac{q_i + 3p_{i+1}}{q_{i+1} q_{i+2}} \mathbb{P}(X(n) = i)$.

p. 168, l. 1 bottom: $\sum_{i=0}^{\infty} \sum_{n=1}^{i+1} \frac{q_i + 3p_{i+1}}{q_{i+1} q_{i+2}} \mathbb{P}(X(n) = i) \leq \sum_{i=0}^{\infty} \sum_{n=1}^{\infty} \frac{q_i + 3p_{i+1}}{q_{i+1} q_{i+2}} \mathbb{P}(X(n) = i)$.

p. 170, Algorithm 5.1.1: write rather "$F_{X(n+1)|X(n)}$ can be found explicitly".