SORT 38 (2) July-December 2014, 251-270

Exact prediction intervals for future current records and record range from any continuous distribution

H. M. Barakat[1], E. M. Nigm[1] and R. A. Aldallal[2]

Abstract

In this paper, a general method for predicting future lower and upper current records and record range from any arbitrary continuous distribution is proposed. Two pivotal statistics with the same explicit distribution for lower and upper current records are developed to construct prediction intervals for future current records. In addition, prediction intervals for future observations of the record range are constructed. A simulation study is applied on normal and Weibull distributions to investigate the efficiency of the suggested method. Finally, an example for real lifetime data with unknown distribution is analysed.

MSC: 62G30, 62G32, 62M20, 62F25.

Keywords: Current record values, record range, pivotal quantity, prediction interval, coverage probability.

1. Introduction

Let $\{X_i;\ i \geq 1\}$ be a sequence of iid continuous random variables each distributed according to cumulative distribution function (cdf) $F_X(x) = P(X \leq x)$ and probability density function (pdf) $f_X(x)$. An observation X_j will be called an upper record value if its value exceeds that of all previous observations. Thus, X_j is an upper record if $X_j > X_i$ for every $i < j$. An analogous definition, with the inequality being reversed, deals with lower record values. The times at which the records occur are called record times.

[1] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt
[2] Department of Mathematics and Statistics, MSA University, Cairo, Egypt
Received: September 2013
Accepted: July 2014

There are some situations wherein upper and lower records are observed together, such as the case of weather data. In these cases, It is quite conceivable to consider lower and upper records jointly, when a new record of either kind (upper or lower) occurs, and these records are called current records. In this paper, we denote them by U_n^c and L_n^c, respectively, and call the nth upper current record and the nth lower current record of the sequence $\{X_n\}$ when the nth record of any kind (either an upper or lower) is observed. It can be noticed that $U_{n+1}^c = U_n^c$ if $L_{n+1}^c < L_n^c$ and that $L_{n+1}^c = L_n^c$ if $U_{n+1}^c > U_n^c$. That is, the upper current record value is the largest observation seen to date at the time when the nth record (of either kind) is observed. According to the definition, $L_0^c = U_0^c = X_1$. For $n \geq 1$, the interval (L_n^c, U_n^c) is then referred to as the record coverage. The record range is then defined by $R_n^c = U_n^c - L_n^c$. The record range may also be defined as the nth record range in the sequence of the usual sample range $R_n = \max(X_1, X_2, \ldots, X_n) - \min(X_1, X_2, \ldots, X_n)$, where by definition $R_0^c = 0$ and $R_1^c = R_2$. Notice that a new record range is attained once a new upper or lower record is observed (see, Basak, 2000). Both current record values and record range can be detected in several real-life situations. For example, the consistency of the production process is required to meet a product's specifications. If the record range is large, then it is likely that large number of products will lie outside the specifications of the product. Predictions of future upper and lower current records, as well as record range, are of natural interest in this context. Prediction of future events is a problem of great interest and plays an important role in many applications, such as meteorology, hydrology, industrial stress testing and athletic events. Several authors have considered prediction problems involving record values. For example, Ahmadi and Balakrishnan (2004) derived distribution-free confidence intervals to estimate the fixed quantiles of an arbitrary unknown distribution, based on current records of an iid sequence from that distribution. Raqab and Balakrishnan (2008) obtained distribution-free prediction intervals for records from the Y-sequence based on record values from the X-sequence of iid random variables from the same distribution. Raqab (2009) obtained prediction intervals for the current records from a future iid sequence based on observed current records from an independent iid sequence of the same distribution. Ahmadi and Balakrishnan (2011) discussed the prediction of future order statistics based on the current record values. In this paper, we consider two pivotal quantities for the lower and upper current records based on an arbitrary cdf F_X with the same explicit distribution-free (not depending on the cdf F_X). By using these pivotal quantities, prediction intervals of future observations of lower-upper current records and record range are explicitly derived. Moreover, simulation study is applied on normal and Weibull distributions to investigate the efficiency of the suggested method. Finally, an example of real lifetime data is analysed, where it is assumed that the distribution of the data is unknown.

2. Auxiliary results

Houchens (1984) used an inductive argument to derive the pdf of U_n^c, L_n^c and R_n^c, based on an arbitrary cdf F_X, (in the sequel we write $U_n^c \parallel X$, $L_n^c \parallel X$ and $R_n^c \parallel X$ to indicate that these statistics are based on the cdf F_X), respectively by

$$f_{U_n^c \parallel X}(x) = 2^n f_X(x)\Big[1 - \bar{F}_X(x)\sum_{k=0}^{n-1}\frac{[-\log \bar{F}_X(x)]^k}{k!}\Big], \tag{2.1}$$

$$f_{L_n^c \parallel X}(x) = 2^n f_X(x)\Big[1 - F_X(x)\sum_{k=0}^{n-1}\frac{[-\log F_X(x)]^k}{k!}\Big]$$

and

$$f_{R_n^c \parallel X}(r) = \frac{2^n}{(n-1)!}\int_{-\infty}^{\infty} f_X(r+x)f_X(x)$$

$$\Big[-\log\left(1 - F_X(r+x) + F_X(x)\right)\Big]^{n-1} dx,\ 0 < r < \infty,$$

where $\bar{F}_X(x) = 1 - F_X(x)$.

Houchens (1984) deduced a useful representation for $U_n^c \parallel Y$, when Y has a negative exponential with parameter 2, i.e., $Y \sim \mathrm{EX}(2)$. Namely,

$$U_n^c \parallel Y \stackrel{d}{=} Y_0 + Y_1 + \ldots + Y_n, \tag{2.2}$$

where "$\stackrel{d}{=}$"means identical in distribution and Y_i's are independent random variables such that $Y_0 \sim \mathrm{EX}(2)$ and the remaining $Y_i \sim \mathrm{EX}(1)$. An analogous representation for the lower current record can be easily obtained by noting that

$$f_{-U_n^c \parallel X}(x) = f_{U_n^c \parallel X}(-x) = 2^n f_X(-x)\Big[1 - \bar{F}_X(-x)\sum_{k=0}^{n-1}\frac{(-\log \bar{F}_X(-x))^k}{k!}\Big]$$

$$= 2^n f_{-X}(x)\Big[1 - \bar{F}_{-X}(x)\sum_{k=0}^{n-1}\frac{(-\log \bar{F}_{-X}(x))^k}{k!}\Big],$$

which yields

$$-U_n^c \parallel X \stackrel{d}{=} L_n^c \parallel -X. \tag{2.3}$$

Applying (2.3), we get $-U_n^c \parallel Y \stackrel{d}{=} -Y_0 - Y_1 - \cdots - Y_n \stackrel{d}{=} Z_0 + Z_1 + \cdots + Z_n$, where $Z_0 \sim \mathrm{EX}^+(2)$, $Z_i \sim \mathrm{EX}^+(1), i = 1,2,\ldots,n$, and $\mathrm{EX}^+(\beta)$ is the positive exponential cdf

with parameter β. Thus, by applying again (2.3) and noting that $Y \sim \mathrm{EX}(\beta) \Rightarrow Z = -Y \sim \mathrm{EX}^+(\beta)$, we get

$$L_n^c \parallel Z \stackrel{d}{=} Z_0 + Z_1 + \ldots + Z_n,$$

where $Z \sim \mathrm{EX}^+(2)$, $Z_0 \sim \mathrm{EX}^+(2)$ and $Z_i \sim \mathrm{EX}^+(1), i = 1,2,\ldots,n$.

3. Main results

The following theorem is the main result of this article. In what follows we assume that F_X is a continuous cdf with the generalized inverse function $F_X^{-1}(y) = \inf\{x : F_X(x) \geq y\}$.

Theorem 3.1. *Let $U_n^c = U_n^c \parallel X$, $L_n^c = L_n^c \parallel X$ and $R_n^c = R_n^c \parallel X$ be the upper current record, the lower current record and the record range based on the cdf F_X, respectively. Furthermore, let $0 < \alpha, \beta < 1$ and $m = 1,2,\ldots$ Then,*

1. $\left(U_n^c, F_X^{-1}\left(1 - \bar{F}_X^{1+t_{m:\alpha}}(U_n^c)\right)\right)$ *is $(1-\alpha)\%$ confidence interval for U_{n+m}^c.*
2. $\left(F_X^{-1}(F_X^{1+t_{m:\beta}}(L_n^c)), L_n^c\right)$ *is $(1-\beta)\%$ confidence interval for L_{n+m}^c,*
3. $\left(R_n^c = U_n^c - L_n^c, F_X^{-1}\left(1 - \bar{F}_X^{1+t_{m:\alpha}}(U_n^c)\right) - F_X^{-1}\left(F_X^{1+t_{m:\beta}}(L_n^c)\right)\right)$ *is $\gamma\%$ confidence interval for R_{n+m}^c, where $\gamma \geq \max(1-\alpha-\beta, 0)$ (e.g., $\gamma \geq 0.98$ if $\alpha = \beta = 0.01$).*

Theorem 3.1 will follow from the following lemma, which is proved in the Appendix and individually expresses an interesting fact.

Lemma 3.1. *Let $U_n^\star = U_n^c \parallel Y$ and $L_n^\star = L_n^c \parallel Z$, where $Y \sim EX(2)$ and $Z \sim EX^+(2)$. Then, for every $m = 1,2,\ldots$, the two pivotal statistics $\bar{T}_m = \frac{U_{n+m}^\star - U_n^\star}{U_n^\star}$ and $T_m = \frac{L_{n+m}^\star - L_n^\star}{L_n^\star}$ have the same pdf $f(t)$, where*

$$f(t) = \frac{2^{n-1} m t^{m-1}}{(t+\frac{1}{2})^{m+1}} - \sum_{k=0}^{n-1} \binom{k+m}{k} \frac{2^{n-k-1} m t^{m-1}}{(t+1)^{k+m+1}}. \tag{3.1}$$

Remark 3.1. One can easily check that $\int_0^\infty f(t)dt = 1$, by using the two formulas

$$\int_0^\infty \frac{t^N}{(t+a)^M} dt = a^{N-M+1} \sum_{i=0}^{N} \binom{N}{i} \frac{(-1)^{i+1}}{N-i-M+1}, \; a > 0,$$

and

$$\sum_{i=0}^{N}\frac{(-1)^i}{M+i}\binom{N}{i}=\frac{N!(M-1)!}{(M+N)!},$$

for any two positive integers N and M, for which $N<M-1$.

Proof of Theorem 3.1. On applying Lemma 3.1, we get $P\Big(0\le \bar{T}_m \le t_{m:\alpha}\Big)=1-\alpha$, and $P\Big(0\le T_m \le t_{m:\beta}\Big)=1-\beta$. Therefore, we get

$$P\Big(0\le \frac{U^*_{n+m}-U^*_n}{U^*_n}\le t_{m:\alpha}\Big)=P\Big(U^*_n\le U^*_{n+m}\le U^*_n(1+t_{m:\alpha})\Big)=1-\alpha \tag{3.2}$$

and

$$P\Big(0\le \frac{L^*_{n+m}-L^*_n}{L^*_n}\le t_{m:\beta}\Big)=P\Big(0\ge L^*_{n+m}-L^*_n\ge L^*_n t_{m:\beta}\Big)=1-\beta \tag{3.3}$$

(note that $L^*_n\le 0$). Thus, the first two relations of Theorem 3.1 (1. and 2.) follow immediately by applying the transformations $U^\star_n=-2\log(\bar{F}_X(U^c_n))$ and $L^\star_n=2\log(F_X(L^c_n))$, respectively, on the relations (3.2) and (3.3).

In order to find the confidence interval for the record range we use the two well-known relations

$$P(C_1C_2)\ge \max(P(C_1)+P(C_2)-1,0),$$

for any two events C_1 and C_2, and

$$\{a+\bar{a}\le X+Y\le b+\bar{b}\}\subset\{\bar{a}<X<\bar{b},a<Y<b\},$$

for any two random variables X and Y, to get

$$P\Big(R^c_n=U^c_n-L^c_n\le R^c_{n+m}\le F_X^{-1}\Big(1-\bar{F}_X^{1+t_{m:\alpha}}(U^c_n)\Big)-F_X^{-1}\Big(F_X^{1+t_{m:\beta}}(L^c_n)\Big)\Big)$$

$$\ge P\Big(U^c_n\le U^c_{n+m}\le F_X^{-1}\Big(1-\bar{F}_X^{1+t_{m:\alpha}}(U^c_n)\Big),-L^c_n\le -L^c_{n+m}\le -F_X^{-1}\Big(F_X^{1+t_{m:\beta}}(L^c_n)\Big)\Big)$$

$$=\gamma\ge\max(1-\alpha-\beta,0).$$

This completes the proof. □

By using an argument similar to the one applied in Lemma 3.1, the proofs of the following two results are in the appendix.

Lemma 3.2. *The joint pdf's of* $U_1^\star, U_2^\star, \ldots, U_n^\star$ *and* $L_1^\star, L_2^\star, \ldots, L_n^\star$ *are given respectively by*

$$f_{U_n^\star, U_{n-1}^\star, \ldots, U_1^\star}(y_n, y_{n-1}, \ldots, y_1) = e^{-y_n}[e^{y_1/2} - 1], 0 < y_1 < y_2 < \cdots < y_n,$$

and

$$f_{L_n^\star, L_{n-1}^\star, \ldots, L_1^\star}(z_n, z_{n-1}, \ldots, z_1) = e^{z_n}[e^{-z_1/2} - 1], \;\; z_n < z_{n-1} < \cdots < z_1 < 0.$$

Lemma 3.2 opens the way for interesting inferential study based on the current records. Actually, by noting that $U_n^\star = -2\log(\bar{F}_X(U_n^c \parallel X))$ and $L_n^\star = 2\log(F_X(L_n^c \parallel X))$, we can obtained the likelihood functions based on the upper and lower current records, respectively, as

$$f_{U_n^c \parallel X, \ldots, U_1^c \parallel X}(x_n, \ldots, x_1) = \frac{\bar{F}_X^2(x_n) F_X(x_1)}{\bar{F}_X(x_1)} \left(\prod_{j=1}^{n} \frac{2 f_X(x_j)}{\bar{F}_X(x_j)} \right), \; x_1 < x_2 < \cdots < x_n$$

and

$$f_{L_n^c \parallel X, \ldots, L_1^c \parallel X}(x_n, \ldots, x_1) = \frac{\bar{F}_X^2(x_n) \bar{F}_X(x_1)}{F_X(x_1)} \left(\prod_{j=1}^{n} \frac{2 f_X(x_j)}{F_X(x_j)} \right), x_n < x_{n-1} < \cdots < x_1.$$

The above likelihood functions can be used to obtain the point estimators of any unknown parameters of the cdf F_X, especially if the available data are the current record values.

Lemma 3.3. *Each of the sequence* $\{U_n^c \parallel X\}$ *and* $\{L_n^c \parallel X\}$ *forms a Markov chain.*

Tables 1, 2 and 3 give the values of $t_{m:\theta}$, where $\int_0^{t_{m:\theta}} f(t)dt = 1 - \theta$, for the values of $n = 2, 3, \ldots, 20, m = 1, 2, \ldots, 5$ and $\theta = 0.1, 0.05, 0.01$. The calculations in these tables are carried out by Mathematica 8.

Table 1: $P(\bar{T}_m \le t_{m:0.1}) = P(T_m \le t_{m:0.1}) = 0.9$.

n	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
2	0.893932	1.64789	2.12161	3.09928	3.81681
3	0.637903	1.15382	1.64826	2.13481	2.61746
4	0.496616	0.887298	1.25887	1.62313	1.98369
5	0.406947	0.720864	1.01764	1.30767	1.59422
6	0.34491	0.607108	0.853803	1.09426	1.33144
7	0.299402	0.524443	0.735349	0.940467	1.14249
8	0.264573	0.461651	0.645749	0.824454	1.00024
9	0.237047	0.41233	0.575618	0.733861	0.88935
10	0.214737	0.37256	0.519236	0.661177	0.80051
11	0.196285	0.339808	0.472924	0.601579	0.727758
12	0.180767	0.312366	0.434205	0.551829	0.667099
13	0.167533	0.289036	0.401353	0.509676	0.615756
14	0.156111	0.268957	0.373128	0.473505	0.571739
15	0.146152	0.251494	0.348617	0.442127	0.533588
16	0.137392	0.236166	0.327132	0.41465	0.500204
17	0.129626	0.222602	0.308145	0.390389	0.470749
18	0.122693	0.210515	0.291243	0.368811	0.444567
19	0.116466	0.199676	0.276101	0.349494	0.421142
20	0.110841	0.1899	0.262458	0.332101	0.400062

Table 2: $P(\bar{T}_m \le t_{m:0.05}) = P(T_m \le t_{m:0.05}) = 0.95$.

n	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
2	1.33466	2.36039	3.35038	4.32794	5.29962
3	0.917775	1.58465	2.22095	2.84604	3.46562
4	0.699294	1.18883	1.65183	2.10471	2.55247
5	0.565044	0.950237	1.31202	1.66461	2.01244
6	0.474206	0.791113	1.08708	1.37465	1.6578
7	0.408643	0.677553	0.927536	1.16979	1.4079
8	0.359082	0.592484	0.808619	1.01759	1.2227
9	0.320292	0.526398	0.716627	0.900193	1.08012
10	0.2891	0.473584	0.643377	0.806939	0.967069
11	0.263469	0.430412	0.583685	0.731109	0.875286
12	0.242029	0.394463	0.534115	0.668256	0.799318
13	0.223828	0.364064	0.492298	0.615323	0.735419
14	0.208182	0.338022	0.45655	0.570139	0.680937
15	0.194588	0.315462	0.42564	0.531123	0.633941
16	0.182666	0.29573	0.39865	0.497096	0.592992
17	0.172124	0.278324	0.374878	0.46716	0.556997
18	0.162736	0.262857	0.353783	0.440621	0.525111
19	0.154321	0.249021	0.334935	0.416931	0.49667
20	0.146736	0.23657	0.317995	0.395657	0.471145

Table 3: $P(\bar{T}_m \leq t_{m:0.01}) = P(T_m \leq t_{m:0.01}) = 0.99$.

n	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
2	2.85847	4.79726	6.66544	8.50916	10.3413
3	1.79354	2.91229	3.97659	5.02093	6.05546
4	1.29678	2.06118	2.78104	3.4839	4.17816
5	1.01294	1.58618	2.12161	2.64218	3.15505
6	0.830235	1.28585	1.70856	2.11803	2.52051
7	0.703094	1.07977	1.42729	1.76285	2.09202
8	0.609623	0.929976	1.22414	1.50739	1.78474
9	0.538056	0.816357	1.07087	1.31536	1.55434
10	0.481519	0.727304	0.951294	1.166	1.37557
11	0.435735	0.655671	0.855492	1.04666	1.23301
12	0.397907	0.596827	0.777067	0.949212	1.11681
13	0.366128	0.547641	0.711716	0.868181	1.02035
14	0.339055	0.505923	0.656439	0.799775	0.939036
15	0.315716	0.470098	0.609086	0.741278	0.869594
16	0.295387	0.439003	0.568075	0.690695	0.80962
17	0.277521	0.411761	0.532217	0.646532	0.757316
18	0.261696	0.387699	0.500602	0.607646	0.711309
19	0.247582	0.366292	0.472522	0.573149	0.670533
20	0.234914	0.347123	0.447416	0.542341	0.63415

4. Simulation study

In order to check the efficiency of the presented method in Theorem 3.1, a simulation study is conducted for two important lifetime distributions: $Weibull[1,2]$, with scale and shape parameters 1 and 2, respectively, and $Normal[0,1]$. For each of these distributions, we generate a random sample of size 100. Moreover, for each of these random samples, the lower and upper current record values are picked up and then the corresponding record ranges are computed. By accident, we got the same number, 12, of current records (lower and upper) for the two random samples (i.e., for the two distributions). Table 4 gives these 12 observed values of $U_n^c \parallel X$ and $L_n^c \parallel X$, as well as $R_n^c \parallel X$, where $X \sim Weibull[1,2]$, or $X \sim Normal[0,1]$. Now, we assume that we have only observed the first 9 values of current records (lower and upper) (i.e., 75% of the observed values of the current records) and we want to predict the three next ones (i.e., 25% of the observed values of the current records). Theorem 3.1 enables us to get predictive confidence intervals for these three next values. Tables 5 and 6 give these predictive confidence intervals for $U_{9+m}^c \parallel X$, $L_{9+m}^c \parallel X$ and $R_{9+m}^c \parallel X$, where $m = 1,2,3$, for the cdf's $X \sim Weibull[1,2]$ and $X \sim Normal[0,1]$, respectively.

Algorithm

Step 1: select the cdf F_X from which the data will come,
Step 2: choose the values of N,

Step 3: generate a random sample of size N from F_X,
Step 4: pick up the lower and upper current record values from the observed data and compute the corresponding record range values. Let the number of the observed lower and upper current record values be n. Choose the value of M, which is about 25% of n,
Step 5: choose a significant coefficient θ and numerically solve the equation

$$\int_0^{t_{m:\theta}} f(t)dt = 1-\theta,\ m = 1,2,\ldots,M,$$

using (3.1) (after replacing n in (3.1) by $n-M$) and Mathematica 8,
Step 6: determine the lower and upper bounds of the predictive confidence intervals for $U^c_{n-M+m} \parallel X$, $L^c_{n-M+m} \parallel X$ and $R^c_{n-M+m} \parallel X$, $m = 1,2,..,M$, by using Theorem 3.1 and the step 5.

The presented results in Tables 5 and 6 show that all the true values of $U^c_{9+m} \parallel X$, $L^c_{9+m} \parallel X$ and $R^c_{9+m} \parallel X$, where $m = 1,2$, are included in their predictive confidence intervals for the two cdf's $X \sim Weibull[1,2]$ and $X \sim Normal[0,1]$. Moreover, almost, the true values of these statistics are also included in their predictive confidence intervals for the two cdf's, for $m = 3$. Nevertheless, the length of the predictive confidence interval increases (i.e., we get less accuracy) with increasing the value of m, i.e. the number of the unobserved data is increased. Therefore, we advise predicting no more than one fourth of the data that we have.

Table 4: *Current records and record range from* $Weibull[1,2]$ *and* $Normal[0,1]$.

$Weibull[1,2]$				$Normal[0,1]$			
n	U^c_n	L^c_n	R^c_n	n	U^c_n	L^c_n	R^c_n
1	3.84915	3.84915	0	1	−0.187968	−0.187968	0
2	3.84915	0.446312	3.402838	2	−0.187968	−0.35455	0.166582
3	5.64291	0.446312	5.196598	3	0.1652	−0.35455	0.51975
4	5.64291	0.375142	5.267768	4	0.1652	−1.21013	1.37533
5	5.64291	0.192999	5.449911	5	1.40996	−1.21013	2.62009
6	6.1647	0.192999	5.971701	6	1.40996	−1.37108	2.78104
7	10.2282	0.192999	10.035201	7	1.40996	−1.66077	3.07073
8	10.2282	0.108285	10.119915	8	2.07656	−1.66077	3.73733
9	10.2282	0.0235643	10.2046357	9	2.07656	−1.90336	3.97992
10	10.5855	0.0235643	10.5619357	10	2.10684	−1.90336	4.0102
11	12.9219	0.0235643	12.8983357	11	2.10684	−2.15466	4.2615
12	12.9219	0.0202959	12.9016041	12	2.96574	−2.15466	5.1204

Table 5: *Predictive confidence intervals for the next three observations of current records and record range from Weibull*$[1,2]$*, with different significance levels (SL's)* 90%, 95% *and* 99%.

for $m=1$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U_{10}^{c}	(10.2282,12.6528)	(10.2282,13.5042)	(10.2282,15.7315)
L_{10}^{c}	(0.00818032,0.0235643)	(0.00564575,0.0235643)	(0.00214174,0.0235643)
R_{10}^{c}	(10.2046357,12.6446)	(10.2046357,13.4986)	(10.2046357,15.7294)
for $m=2$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U_{11}^{c}	(10.2282,14.4456)	(10.2282,15.6123)	(10.2282,18.5781)
L_{11}^{c}	(0.00374765,0.0235643)	(0.00225577,0.0235643)	(0.000621026,0.0235643)
R_{11}^{c}	(10.2046357,14.4418)	(10.2046357,15.61)	(10.2046357,18.5774)
for $m=3$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U_{12}^{c}	(10.2282,16.1157)	(10.2282,17.558)	(10.2282,21.1813)
L_{12}^{c}	(0.00181212,0.0235643)	(0.000967742,0.0235643)	(0.000200224,0.0235643)
R_{12}^{c}	(10.2046357,16.1139)	(10.2046357,17.577)	(10.2046357,21.1811)

Table 6: *Predictive confidence intervals for the next three observations of current records and record range from Normal*$[0,1]$*, with different SL's* 90%, 95% *and* 99%.

for $m=1$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U_{10}^{c}	(2.07656,2.43784)	(2.07656,2.55498)	(2.07656,2.84252)
L_{10}^{c}	$(-2.24886,-1.90336)$	$(-2.36081,-1.90336)$	$(-2.63544,-1.90336)$
R_{10}^{c}	(3.97992,4.6867)	(3.97992,4.91579)	(3.97992,5.47796)
for $m=2$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U_{11}^{c}	(2.07656,2.67961)	(2.07656,2.82774)	(2.07656,3.17785)
L_{11}^{c}	$(-2.47987,-1.90336)$	$(-2.62134,-1.90336)$	$(-2.9555,-1.90336)$
R_{11}^{c}	(3.97992,5.15948)	(3.97992,5.44908)	(3.97992,6.13335)
for $m=3$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U_{12}^{c}	(2.07656,2.88969)	(2.07656,3.06129)	(2.07656,3.45983)
L_{12}^{c}	$(-2.68049,-1.90336)$	$(-2.84427,-1.90336)$	$(-3.22445,-1.90336)$
R_{12}^{c}	(3.97992,5.57018)	(3.97992,5.90556)	(3.97992,6.68428)

5. The case when the cdf F is unknown and real data example

Undoubtedly the lack of knowledge of the distribution of the resulted data in any statistical experiment is the most frequent case. In fact the assumption that the distribution F is known is unreal. However, we can overcome this problem by using the observed data that we have (i.e., $X_1, X_2, \ldots, X_N$) to select a statistical distribution that best fits this data set. Actually, we cannot "just guess" and use any other particular distribution without testing several alternative models as this can result in analysis errors. In most cases, we need to fit two or more distributions, compare the results, and select the most valid model (see Example 5.1). Naturally, the "candidate" distributions we fit should be chosen depending on the nature of our observed data. For example, in the case of a life testing experiment we should fit non-negative distributions such as Gamma or Weibull. Obviously when this procedure is applied, all we need, is that the size N of the observed data to be large enough to carry the necessary identification methods (e.g., build a histogram) and goodness-of-fit tests (e.g., the Kolmogorov-Smirnov test) based on the empirical cdf of $X_1, \ldots, X_N$. In Example 5.1, we consider $N = 130$ realistic observations (cf. Arnold, et al. 1998, Page 49) with unknown distribution. These data yield 14 current records (lower-upper). The first 11 of them resulted from the first 48 observations. Thus, we look for the best distribution F that fits these data (the 48 observations). After that we predict the last three current records and their corresponding record ranges by applying the results of Theorem 3.1 on the first 11 current records and their corresponding record ranges. We find almost all the predictions are accurate even when we select another fitted distribution for the data but with less goodness-of-fit to the data than the first one.

Example 5.1. The following data (read row-wise) represent the average July temperatures (in degrees centigrade) of Neuenburg, Switzerland, during the period 1864-1993 (from Klupppelberg and Schwere, 1995).

19.0 20.1 18.4 17.4 19.7 21.0 21.4 19.2 19.9 20.4 20.9 17.2 20.2 17.8 18.1
15.6 19.4 21.7 16.2 16.4 19.0 20.6 19.0 20.7 15.8 17.7 16.8 17.1 18.1 18.4
18.7 18.7 18.4 19.2 18.0 18.7 20.7 19.4 19.2 17.4 22.0 21.4 19.3 16.8 18.2
16.2 15.9 22.1 17.5 15.3 16.5 17.4 17.0 18.3 18.3 15.3 18.2 21.5 17.0 21.6
18.2 18.1 17.6 18.2 22.6 19.9 17.1 17.2 17.3 19.4 20.1 20.1 17.0 19.4 17.5
16.8 17.0 19.9 18.2 19.2 18.5 20.8 19.5 21.1 15.8 21.3 21.2 18.8 22.3 18.6
16.8 18.2 17.2 18.4 18.7 21.1 16.3 17.4 18.0 19.5 21.2 16.8 17.4 20.7 18.4
19.8 18.7 20.5 18.3 18.2 18.2 19.2 20.2 18.2 17.4 19.2 16.3 17.4 20.3 23.4
19.2 20.2 19.3 19.0 18.8 20.3 19.7 20.7 19.6 18.1

The above data yield 14 current records. These current records and their corresponding record ranges are presented in Table 7. First, we try to fit the first 48 observations, for several cdf's such as exponential, logistic, Gamma, normal, Weibull, Gumbel, Laplace

and inverse Gamma distributions. The methods of maximum likelihood and moments are used to estimate the parameters of the candidate cdf's. After that we apply the Anderson-Darling, Cramér-von Mises, and Kolmogorov-Smirnov goodness of fit tests to check the fitting of these cdf's. Among these cdf's, we found that only the Gamma, normal and logistic distributions fit these data. Moreover, the $Gamma[119.277, 0.157808]$ distribution is the best cdf that fits these data (in the average w.r.t the three applied goodness of fit tests and the two used methods of estimation) the second cdf is $Normal$ $[18.8229, 1.71722]$, while the third is logistic distribution $Logistic[18.8205, 1.01236]$, see Tables 8-10 and Figures 1-3. The predictive confidence intervals for the next three statistics U^c_{11+m}, L^c_{11+m} and $R^c_{11+m}, m = 1,2,3$, for the Gamma, normal and logistic cdf's are represented in Tables 11-13, respectively. These tables show that almost all the true values of the above three statistics are included in the predictive confidence intervals. This result shows that our suggested method is stable regardless the choice of the cdf that fits the data.

Table 7: *Current records and record ranges which are resulted from all our data.*

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U^c_n	19.0	20.1	20.1	20.1	21.0	21.4	21.4	21.4	21.7	22.0	22.1	22.1	22.6	23.4
L^c_n	19.0	19.0	18.4	17.4	17.4	17.4	17.2	15.6	15.6	15.6	15.6	15.3	15.3	15.3
R^c_n	0	1.1	1.7	2.7	3.6	4.0	4.2	5.8	6.1	6.4	6.5	6.8	7.3	8.1

Table 8: *Fitting the first* 48 *observations for gamma cdf.*

Distribution/Test-Method	$Gamma[\alpha, \beta]$	
Maximum Likelihood	$\hat{\alpha}_{ML} = 119.277$ $\hat{\beta}_{ML} = 0.157808$	
	P-Value	*Statistic*
Kolmogorov-Smirnov	0.995234	0.0569809
Anderson-Darling	0.977713	0.235002
Cramér-Von-Mises	0.983675	0.0274912
Moments	$\hat{\alpha}_M = 120.149$ $\hat{\beta}_M = 0.156663$	
	P-Value	*Statistic*
Kolmogorov-Smirnov	0.994289	0.0578043
Anderson-Darling	0.974785	0.241202
Cramér-Von-Mises	0.981763	0.0281783

Table 9: *Fitting the first* 48 *observations for normal cdf.*

Distribution/Test-Method	$Normal[\mu,\sigma]$	
Maximum Likelihood	$\hat{\mu}_{ML} = 18.8229$ $\hat{\sigma}_{ML} = 1.71722$	
	P-Value	*Statistic*
Kolmogorov-Smirnov	0.994086	0.0579686
Anderson-Darling	0.982812	0.222963
Cramér-Von-Mises	0.987088	0.0261305
Moments	$\hat{\mu}_{M} = 18.8229$ $\hat{\sigma}_{M} = 1.71722$	
	P-Value	*Statistic*
Kolmogorov-Smirnov	0.994086	0.0579686
Anderson-Darling	0.982812	0.222963
Cramér-Von-Mises	0.987088	0.0261305

Table 10: *Fitting the first* 48 *observations for logistic cdf.*

Distribution/Test-Method	$Logistic[\mu,\beta]$	
Maximum Likelihood	$\hat{\mu}_{ML} = 18.8205$ $\hat{\beta}_{ML} = 1.01236$	
	P-Value	*Statistic*
Kolmogorov-Smirnov	0.98876	0.061264
Anderson-Darling	0.964482	0.260431
Cramér-Von-Mises	0.979247	0.02903
Moments	$\hat{\mu}_{M} = 18.8229$ $\hat{\beta}_{M} = 0.946754$	
	P-Value	*Statistic*
Kolmogorov-Smirnov	0.927317	0.0756047
Anderson-Darling	0.838543	0.409448
Cramér-Von-Mises	0.882778	0.0489246

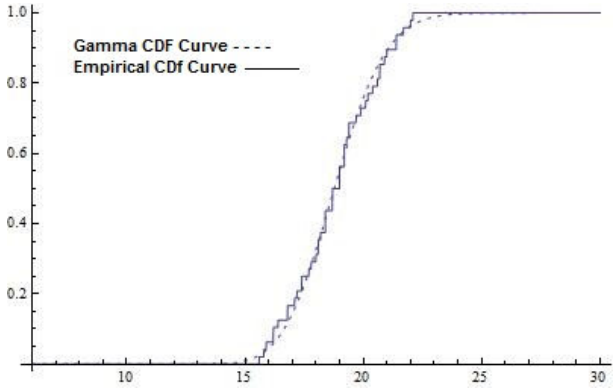

Figure 1: *Plot showing the goodness-of-fit for gamma cdf.*

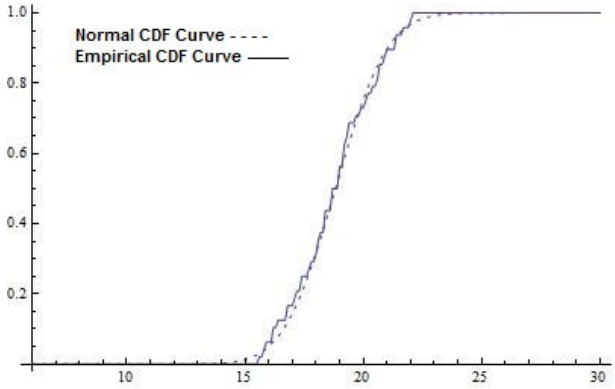

Figure 2: *Plot showing the goodness-of-fit for normal cdf.*

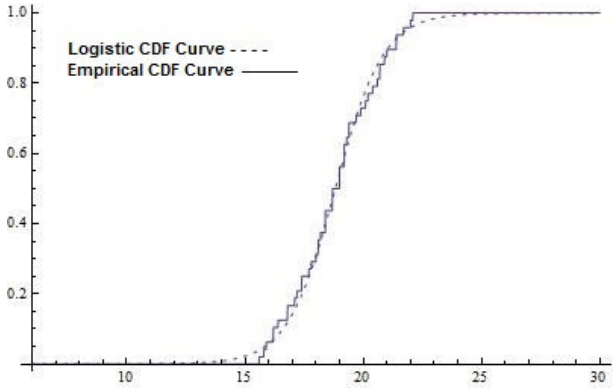

Figure 3: *Plot showing the goodness-of-fit for logistic cdf.*

Table 11: *Predictive confidence intervals for* U^c_{11+m}, L^c_{11+m} *and* R^c_{11+m}, $m = 1,2,3$, *from* $Gamma[119.277, 0.157808]$.

for $m=1$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{12}	(22.1,22.6482)	(22.1,22.8253)	(22.1,23.2593)
L^c_{12}	(15.1588,15.6)	(15.0195,15.6)	(14.6846,15.6)
R^c_{12}	(6.5,7.4894)	(6.5,7.8058)	(6.5,8.5747)
for $m=2$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{13}	(22.1,23.021)	(22.1,23.2463)	(22.1,23.7782)
L^c_{13}	(14.8674,15.6)	(14.6945,15.6)	(14.2959,15.6)
R^c_{13}	(6.5,8.1536)	(6.5,8.5516)	(6.5,9.4823)
for $m=3$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{14}	(22.1,23.3496)	(22.1,23.6122)	(22.1,24.222)
L^c_{14}	(14.6161,15.6)	(14.4189,15.6)	(13.9732,15.6)
R^c_{14}	(6.5,8.7335)	(6.5,9.1933)	(6.5,10.2488)

Table 12: *Predictive confidence intervals for* U^c_{11+m}, L^c_{11+m} *and* R^c_{11+m}, $m = 1,2,3$, *from* $Normal[18.8229, 1.71722]$.

for $m=1$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{12}	(22.1,22.5884)	(22.1,22.756)	(22.1,23.1421)
L^c_{12}	(15.107,15.6)	(14.9494,15.6)	(14.5665,15.6)
R^c_{12}	(6.5,7.4814)	(6.5,7.8066)	(6.5,8.5756)
for $m=2$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{13}	(22.1,22.9307)	(22.1,23.1306)	(22.1,23.5979)
L^c_{13}	(14.7762,15.6)	(14.578,15.6)	(14.1147,15.6)
R^c_{13}	(6.5,8.1545)	(6.5,8.5526)	(6.5,9.4832)
for $m=3$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{14}	(22.1,23.2219)	(22.1,23.4528)	(22.1,23.9826)
L^c_{14}	(14.4875,15.6)	(14.2586,15.6)	(13.7333,15.6)
R^c_{14}	(6.5,8.7344)	(6.5,9.1942)	(6.5,10.2493)

Table 13: *Predictive confidence intervals for U^c_{11+m}, L^c_{11+m} and R^c_{11+m}, $m = 1,2,3$, from $Logistic[18.8205, 1.01236]$.*

for $m=1$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{12}	(22.1,22.77)	(22.1,22.997)	(22.1,23.5757)
L^c_{12}	(14.9403,15.6)	(14.7169,15.6)	(14.1475,15.6)
R^c_{12}	(6.5,7.8297)	(6.5,8.2801)	(6.5,9.4282)
for $m=2$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{13}	(22.1,23.2539)	(22.1,23.5578)	(22.1,24.3102)
L^c_{13}	(14.4641,15.6)	(14.1651,15.6)	(13.4251,15.6)
R^c_{13}	(6.5,8.7898)	(6.5,9.3927)	(6.5,10.8851)
for $m=3$	$SL=90\%$	$SL=95\%$	$SL=99\%$
U^c_{14}	(22.1,23.7001)	(22.1,24.0702)	(22.1,24.9755)
L^c_{14}	(14.0251,15.6)	(13.6612,15.6)	(12.771,15.6)
R^c_{14}	(6.5,9.675)	(6.5,10.409)	(6.5,12.2045)

6. Conclusion

In this paper we focused on the prediction of upper and lower records. The obtained results are useful when people are interested in knowing extreme values on different periods, areas, etc. and their range of variation. Theorem 3.1 suggests a new method to estimate confidence intervals for upper, lower and range records. This new method depends on constructing two pivotal statistics with the same distribution for lower and upper current records. The real data Example 5.1, shows that when the cdf of the data is unknown, this method is applicable with acceptable degree of accuracy, even if we fail to assign the type of the distribution of the data with a high accuracy. It is worth mentioning that the result and the method of the proofs of this paper are quite different from the known results concerning the prediction problems of record values. For example, Ahmadi and Balakrishnan (2004) used only the current records to estimate the fixed quantiles of the given cdf (unknown cdf), while Raqab and Balakrishnan (2008) obtained distribution-free prediction intervals for the usual records (not the current records). Finally Raqab (2009) predicted the current records, by using the two-sample prediction plan, where the variable to be predicted comes from an independent future sample. In this paper, we consider the one-sample prediction plan, where the variable to be predicted comes from the same sample so that it may be correlated with the observed data.

Acknowledgement

The authors would like to thank the anonymous referees for constructive suggestions and comments that improved the representation substantially.

Appendix

Proof of Lemma 3.1. By using (2.2), we get

$$P(U^{\star}_{n+m} \le x | U^{\star}_{n} = y) = P(Y_0 + Y_1 + \cdots + Y_n + \cdots + Y_{n+m} \le x | Y_0 + Y_1 + \ldots + Y_n = y)$$

$$= P(Y_{n+1} + \cdots + Y_{n+m} \le x - y | Y_0 + Y_1 + \cdots + Y_n = y) = P(Y_{n+1} + \cdots + Y_{n+m} \le x - y). \quad (1)$$

On the other hand, since $Y_i \sim \mathrm{EX}(1)$, for $i = n+1, \ldots, n+m$, then

$$f_{U^{\star}_{n+m}|U^{\star}_{n}}(x|y) = f_{Y_{n+1}+\cdots+Y_{n+m}}(x-y) = \frac{(x-y)^{m-1}}{(m-1)!} e^{-(x-y)} I_{(0,\infty)}(x-y), \quad (2)$$

where $I_A(.)$ is the usual indicator function of the set A. Therefore, by combining (1) and (2) with (2.1), we get

$$f_{U^{\star}_{n+m},U^{\star}_{n}}(x,y) = f_{U^{\star}_{n+m}|U^{\star}_{n}}(x|y) f_{U^{\star}_{n}}(y)$$

$$= \frac{(x-y)^{m-1}}{(m-1)!} e^{-(x-y)} 2^n \left(\frac{1}{2} e^{-y/2}\right) \left[1 - e^{-y/2} \sum_{k=0}^{n-1} \frac{(-\log e^{-y/2})^k}{k!}\right]$$

$$= \frac{2^{n-1}(x-y)^{m-1} e^{-(x-y/2)}}{(m-1)!} \left[1 - e^{-y/2} \sum_{k=0}^{n-1} \frac{y^k}{2^k k!}\right]. \quad (3)$$

Now, by using the transformation $\bar{T}_m = \frac{U^{\star}_{n+m} - U^{\star}_{n}}{U^{\star}_{n}}$ and $W = U^{\star}_{n}$, we get

$$f_{\bar{T}_m,W}(t,w) = \frac{2^{n-1} w^m t^{m-1} e^{-w(t+\frac{1}{2})}}{(m-1)!} - \frac{2^{n-1} t^{m-1} e^{-w(t+1)}}{(m-1)!} \sum_{k=0}^{n-1} \frac{w^{k+m}}{2^k k!}.$$

Thus, we conclude that

$$f_{\bar{T}_m}(t) = \int_0^{\infty} f_{\bar{T}_m,W}(t,w) dw = \frac{2^{n-1} m t^{m-1}}{(t+\frac{1}{2})^{m+1}} - \sum_{k=0}^{n-1} \binom{k+m}{k} \frac{2^{n-k-1} m t^{m-1}}{(t+1)^{k+m+1}}.$$

Similarly, we can show, for any $x \le z \le 0$, that $P(L^\star_{n+m} \le x | L^\star_n = z) = P(Z_{n+1} + \cdots + Z_{n+m} \le x - z)$. Since $Z_i \sim \mathrm{EX}^+(1)$, for $i = n+1, \ldots, n+m$, then

$$f_{L^\star_{n+m}|L^\star_n}(x|z) = f_{Z_{n+1}+\cdots+Z_{n+m}}(x-z) = \frac{(-(x-z))^{m-1}}{(m-1)!} e^{(x-z)} I_{(-\infty,0)}(x-z).$$

Thus,

$$f_{L^\star_{n+m},L^\star_n}(x,z) = f_{L^\star_{n+m}|L^\star_n}(x|z) f_{L^\star_n}(z)$$

$$= \frac{2^{n-1}(-(x-z))^{m-1} e^{(x-z/2)}}{(m-1)!} \Big[1 - e^{z/2} \sum_{k=0}^{n-1} \frac{(-z)^k}{2^k k!}\Big],\ x \le z \le 0.$$

Now, by using the transformation $T_m = \frac{L^\star_{n+m} - L^\star_n}{L^\star_n}$ and $V = L^\star_n$, we get

$$f_{T_m,V}(t,v) = \frac{2^{n-1}(-v)^m t^{m-1} e^{v(t+\frac{1}{2})}}{(m-1)!} - \frac{2^{n-1} t^{m-1} e^{v(t+1)}}{(m-1)!} \sum_{k=0}^{n-1} \frac{(-v)^{k+m}}{2^k k!}, v \le 0, t \ge 0.$$

Then, we conclude that

$$f_{T_m}(t) = \int_{-\infty}^{0} f_{T_m,V}(t,v)dv = \frac{2^{n-1} m t^{m-1}}{(t+\frac{1}{2})^{m+1}} - \sum_{k=0}^{n-1} \binom{k+m}{k} \frac{2^{n-k-1} m t^{m-1}}{(t+1)^{k+m+1}}.$$

This completes the proof. □

Proof of Lemma 3.2. Clearly, (3) yields

$$f_{U^\star_n, U^\star_{n-1}}(y_n, y_{n-1}) = 2^{n-2} e^{-(y_n - y_{n-1}/2)} \Big[1 - e^{-y_{n-1}/2} \sum_{k=0}^{n-2} \frac{\left(y_{n-1}/2\right)^k}{k!}\Big].$$

On the other hand, by applying the same argument as in Lemma 3.1, we can show that

$$P(U^\star_n \le y_n, U^\star_{n-1} \le y_{n-1} | U^\star_{n-2} = y_{n-2})$$
$$= P(Y_{n-1} + Y_n \le y_n - y_{n-2}, Y_{n-1} \le y_{n-1} - y_{n-2} | Y_0 + Y_1 + \cdots + Y_{n-2} = y_{n-2})$$
$$= P(Y_{n-1} + Y_n \le y_n - y_{n-2}, Y_{n-1} \le y_{n-1} - y_{n-2}).$$

Since, $f_{Y_{n-1},Y_n}(y_{n-1}, y_n) = e^{-y_{n-1} - y_n}$, we get

$$f_{Y_{n-1},Y_{n-1}+Y_n}(y_{n-1} - y_{n-2}, y_n - y_{n-2}) = e^{-(y_n - y_{n-2})},\ y_{n-2} < y_{n-1} < y_n.$$

Therefore, $f_{U_n^\star,U_{n-1}^\star|U_{n-2}^\star}(y_n,y_{n-1}|y_{n-2}) = e^{-(y_n-y_{n-2})}$, which by using (2.1) implies

$$f_{U_n^\star,U_{n-1}^\star,U_{n-2}^\star}(y_n,y_{n-1},y_{n-2}) = e^{-(y_n-y_{n-2})} f_{U_{n-2}^\star}(y_{n-2})$$

$$= 2^{n-3} e^{-(y_n-y_{n-2}/2)} \Big[1 - e^{-y_{n-2}/2} \sum_{k=0}^{n-3} \frac{(y_{n-2}/2)^k}{k!}\Big].$$

Therefore, by induction we get the claimed result for the upper current records and the result for the lower current records can be proved by applying the same argument. □

Proof of Lemma 3.3. Since the proof of the lemma for the two sequences $\{U_n^c \parallel X\}$ and $\{L_n^c \parallel X\}$ are very similar, we only prove the lemma for the 1st sequence. For any two positive integers $t < s$, we can easily, by applying the same argument in the proof of Lemmas 3.1, 3.2, to show that

$$P(U_s^c \parallel X \le x_s | U_1^c \parallel X = x_1, \ldots, U_t^c \parallel X = x_t)$$

$$= P(U_s^\star \le x_s^\star | U_1^\star = x_1^\star, \ldots, U_t^\star = x_t^\star) = P(Y_{t+1} + \cdots + Y_s \le x_s^\star - x_t^\star),$$

where $x_i^\star = -2\log[\bar{F}_X(x_i)]$, $i = t, s$. Therefore,

$$f_{U_s^c \parallel X | U_1^c \parallel, \ldots, U_t^c \parallel X}(x_s | x_1, \ldots, x_t) = \frac{(x_s^\star - x_t^\star)^{m-1}}{(m-1)!} e^{-(x_s^\star - x_t^\star)} I_{(0,\infty)}(x_s^\star - x_t^\star).$$

This completes the proof. □

References

Ahmadi, J. and Balakrishnan, N. (2011). Distribution-free prediction intervals for order statistics based on record coverage. *Korean Statistical Society*, 40, 181–192.

Ahmadi, J. and Balakrishnan, N. (2008). Prediction intervals for future records. *Statistics & Probability Letters*, 78, 395–405.

Ahmadi, J. and Balakrishnan, N. (2004). Confidence intervals for quantiles in terms of record range. *Statistics & Probability Letters*, 68, 1955–1963.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records*. Wiley, New York.

Basak, P. (2000). An application of record range and some characterization results. In: Balakrishnan, N. (ed.) *Advances on Methodological and Applied Aspects of Probability and Statistics*. Gordon and Breach Science Publishers, New york: 83–95.

Houchens, R. L. (1984). *Record Value, Theory and Inference*, Ph. D. Dissertation, University of California, Riverside, CA.

Raqab, M. R. (2009). Distribution-free prediction intervals for the future current record statistics. *Statistical Papers*, 50, 429–439.